Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2010) 1–10

Procedia Computer Science

Maximum a posteriori estimation for Markov chains based on Gaussian Markov random fields

H. Wu, F. Noé*

Free University of Berlin, Arnimallee 6, 14195 Berlin, Germany

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.

Keywords:
Markov chain, Gaussian Markov random field, maximum a posteriori, cross validation

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) $\boldsymbol{T} = \left[T_{ij}\right] \in \mathbb{R}^{n\times n}$ with T_{ij} the transition probability from the i-th state to the j-th state. 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 [1–3], for MC models

*Corresponding author
Email addresses: hwu@zedat.fu-berlin.de (H. Wu), frank.noe@fu-berlin.de (F. Noé)

can be transformed to linear stochastic systems with zero-mean noise. However, the conventional LS estimators may violate the nonnegative constraints on TMs. Thus, some restricted LS methods [4–6] based on constrained quadratic programming algorithms were developed to avoid this problem. Some researchers [2, 5] 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 [7], for it is consistent and asymptotically normally distributed as the sample size increases [8], and can be efficiently calculated by counting transition pairs. Some experiments show ML estimator is superior to the LS estimators [9]. Moreover, the ML method can be applied to revsersible TM estimation for some physical and chemical processes [10].

Recently, the Bayesian approach [11, 12] 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 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 $p(\boldsymbol{T}|\boldsymbol{\Theta}) \propto \prod_{i,j} T_{ij}^{\theta_{ij}-1}$. It is a conjugate prior 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 $\boldsymbol{\Theta} = \mathbf{1}$ [13], and $\boldsymbol{\Theta} = \mathbf{0}$ makes the posterior mean of the TM identical to the ML estimate [14]. The matrix $\boldsymbol{\Theta}$ can also be optimized by using the empirical Bayes approach [15]. The matrix beta prior distribution based Bayesian estimation of revsersible TM was investigated in [13]. 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 [16, 17] 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 [18–21] 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. 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 [19, 22], 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, and $d(\cdot,\cdot)$ denotes a distance measure between vertices. It is clear that the neighboring data points are desired to have the similar values.

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) $\boldsymbol{T} = \left[T_{ij}\right] \in \mathbb{R}^{n\times n}$ whose entries are given by

$$T_{ij} = p\left(x_{t+1} = s_j | x_t = s_i\right) \tag{2}$$

where

$$\sum_j T_{ij} = 1, \quad T_{ij} > 0 \tag{3}$$

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

$$p(x_{0:m}|\boldsymbol{T}) = \prod_{i,j} T_{ij}^{C_{ij}} \tag{4}$$

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

$$C_{ij} = \left|\left\{(x_t, x_{t+1})|x_t = s_i, x_{t+1} = s_j, 0 \leq t \leq m-1\right\}\right| \tag{5}$$

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 to provide a GMRF model based prior distribution for the TM $\boldsymbol{T} = \left[T_{ij}\right]$. Assuming a neighborhood structure on the state space, we construct a neighborhood relation between the transition pairs as

$$\left(s_i, s_j\right) \sim (s_k, s_l) \Leftrightarrow \left(s_i, s_j\right) \in (\partial s_k \cup \{s_k\}) \times (\partial s_l \cup \{s_l\}) \setminus \{(s_k, s_l)\} \tag{6}$$

Then the unknown matrix $\boldsymbol{T}$ can be modeled by GMRF with distribution

$$p_{GMRF}(\boldsymbol{T}|\sigma) \propto \exp\left(-u(\boldsymbol{T}, \sigma)\right) \tag{7}$$

where

$$u(\boldsymbol{T}, \sigma) = \frac{1}{2\sigma^2}\sum_{(s_i,s_j)\sim(s_k,s_l)}\left(\frac{T_{ij}-T_{kl}}{d_{ijkl}^2}\right)^2 \tag{8}$$

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

$$d_{ijsk} = \sqrt{d^2\left(s_i, s_k\right) + d^2\left(s_j, s_l\right)} \tag{9}$$

However, the realization of distribution (7) does not satisfy (3) in the general case. Therefore we modify the prior distribution as

$$p_{GMC}\left(\boldsymbol{T}|\sigma\right) = p_{GMRF}\left(\boldsymbol{T}|\sigma, \boldsymbol{T} \in \Omega_n\right) = \begin{cases} \frac{1}{z(\sigma)} \exp\left(-u\left(\boldsymbol{T}, \sigma\right)\right), & \boldsymbol{T} \in \Omega_n \\ 0, & \boldsymbol{T} \notin \Omega_n \end{cases} \tag{10}$$

where

$$z\left(\sigma\right) = \int_{\Omega_n} \exp\left(-u\left(\boldsymbol{T}, \sigma\right)\right) \mathrm{d}\boldsymbol{T} \tag{11}$$

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 $\boldsymbol{C} = \left[C_{ij}\right]$ is given by

$$\hat{\boldsymbol{T}} = \arg\max_{\boldsymbol{T}} \left\{\log p\left(\boldsymbol{C}|\boldsymbol{T}\right) + \log p\left(\boldsymbol{T}\right)\right\} \tag{12}$$

Using the proposed GMRF prior model and assuming the parameter σ is known, (12) is equivalent to the following optimization problem

$$\hat{\boldsymbol{T}}\left(\sigma\right) = \arg\min_{\boldsymbol{T} \in \Omega_n} \left\{ -\sum_{i,j} C_{ij} \log T_{ij} + u\left(\boldsymbol{T}, \sigma\right) \right\} \tag{13}$$

It is a convex problem and can be solved without any spurious local minima. In this paper, we perform the optimization by the diagonalized Newton (DN) method (see [23] for details).

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. But it is intractable to compute the joint prior distribution $p(\boldsymbol{T}, \sigma) = p(\sigma) p_{GMC}(\boldsymbol{T}|\sigma)$ for $z\left(\sigma\right)$ has no closed form.

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 [24]. 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 train 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

$$\hat{\boldsymbol{T}}_k\left(\sigma\right) = \arg\max_{\boldsymbol{T} \in \Omega^n} \left\{\log p\left(\boldsymbol{C}_k^{train}|\boldsymbol{T}\right) - u\left(\boldsymbol{T}, \sigma\right)\right\} \tag{14}$$

and the predictive log-likelihood

$$CV_k\left(\sigma\right) = \log p\left(\boldsymbol{C}_k^{test}|\hat{\boldsymbol{T}}_k\left(\sigma\right)\right) \tag{15}$$

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

$$\sigma^* = \arg\max_{\sigma} CV(\sigma) \tag{16}$$

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

It can be seen from (15) 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

$$\log\left(T_{ij}\right) \approx \mathrm{PL}_\eta\left(T_{ij}\right) = \frac{1}{\eta}\left(T_{ij}^\eta - 1\right) \tag{17}$$

when calculating $CV_k(\sigma)$, where $\eta \in (0, 1)$ is a small number ($\eta = 0.1$ in this paper). It is easy to prove that $\lim_{\eta\to 0} \mathrm{PL}_\eta(x) = \log(x)$ for $x > 0$.

4. Estimation with Stochastic Observations

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

$$o_t | x_t \sim p(o_t | x_t) \tag{18}$$

for $t = 0, \ldots, m$ are available. In this case, the MAP estimator of the TM with prior parameter σ can be expressed by

$$\hat{\boldsymbol{T}}(\sigma) = \arg\max_{\boldsymbol{T}\in\Omega_n} \{\log p(O|\boldsymbol{T}) - u(\boldsymbol{T}, \sigma)\} \tag{19}$$

where $O = \{o_0, \ldots, o_m\}$, and computed with the expectation maximization (EM) algorithm [25] 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

$$\begin{aligned} Q\left(\boldsymbol{T}|\boldsymbol{T}^{(k)}\right) &= \mathbb{E}\left[\log\left(\boldsymbol{C}(X)|\boldsymbol{T}\right) - u(\boldsymbol{T}, \sigma)|\boldsymbol{T}^{(k)}, O\right] \\ &= \sum_{i,j} \bar{C}_{ij} \log T_{ij} - u(\boldsymbol{T}, \sigma) \end{aligned} \tag{20}$$

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

$$\bar{\boldsymbol{C}} = \left[\bar{C}_{ij}\right] = \mathbb{E}\left[\boldsymbol{C}(X)|\boldsymbol{T}^{(k)}, O\right] \tag{21}$$

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

$$\boldsymbol{T}^{(k+1)} = \arg\min_{\boldsymbol{T}\in\Omega_n} \left\{ -\sum_{i,j} \bar{C}_{ij} \log T_{ij} + u(\boldsymbol{T}, \sigma) \right\} \tag{22}$$

Step 4. Terminate if

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

is small enough.

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

Note that (22) has the same form as (13) with $\bar{C}_{ij} \geq 0$ for any i, j, so (22) is a convex optimization problem and can be solved by the DN algorithm too.

Further, in a similar manner to Section 3.3, the value of σ can be designed through the MCCV algorithm. Due to space limitations, we omit details here.

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

$$\mathrm{d}r = -f(r)\,\mathrm{d}t + \rho \mathrm{d}W \tag{23}$$

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

$$V(r) = \begin{cases} -111.01r^3 + 178.63r^2 - 82.27r + 10.55, & r < 0.75 \\ 182.8915r^3 - 482.64r^2 + 413.69r - 113.44, & 0.75 \leq r < 1 \\ -153.36r^3 + 526.11r^2 - 595.06r + 222.81 & 1 \leq r < 1.25 \\ 84.94r^3 - 367.53r^2 + 521.98r - 242.62, & 1.25 < r \end{cases} \tag{24}$$

Discretizing the motion equation (23) with time step $\Delta t = 10^{-3}$ and decomposing the state space $\{r|0 \leq r \leq 2\}$ into $n = 100$ "cells" $S = \{s_1, \ldots, s_n\}$ with $s_i = \frac{2i-1}{n}$, we can get the grid based approximate MC model

$$p\left(x_{k+1} = s_j | x_k = s_i\right) \propto \exp\left(-\frac{\left(s_j - s_i + \Delta t f(s_i)\right)^2}{2\rho^2 \Delta t}\right) \tag{25}$$

The corresponding TM $\boldsymbol{T} = \left[T_{ij}\right]$ is shown in Fig. 2. Furthermore, the neighborhood structure on S is here defined by $\partial s_i = \{s_{i-1}, s_{i+1}\} \cap S$ with distance measure $d\left(s_i, s_j\right) = |i - j|$.

5.2. TM estimation

Here, we will use the MAP method presented in Section 3 to estimate the TM $\boldsymbol{T}$ based on a realization $\{r(t)|0 \leq t \leq 3\}$ of (23) (see Fig. 3), and compare it with the ML method [11]. 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 metric [26] defined as

$$KLR\left(\hat{\boldsymbol{T}} || \boldsymbol{T}\right) = \sum_{ij} \hat{\pi}_i \hat{T}_{ij} \log \frac{\hat{T}_{ij}}{T_{ij}} \tag{26}$$

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

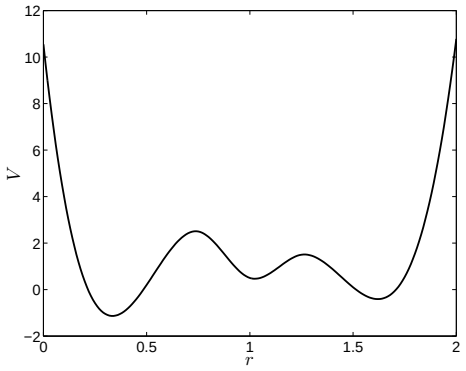


Figure 1: Potential Function

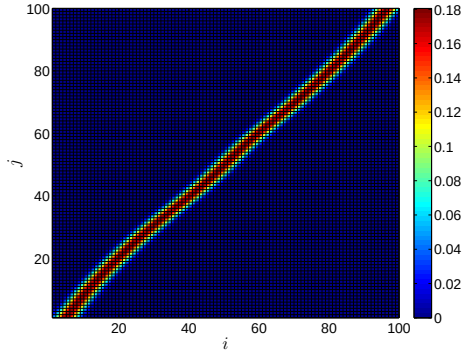


Figure 2: $\boldsymbol{T}$

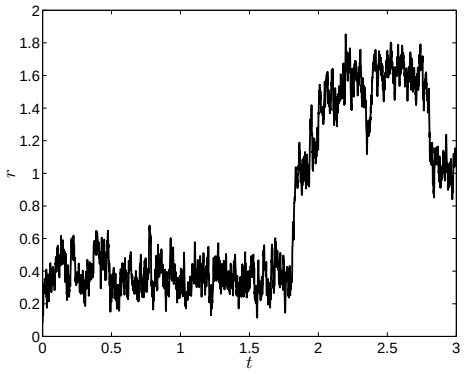


Figure 3: $r(t)$

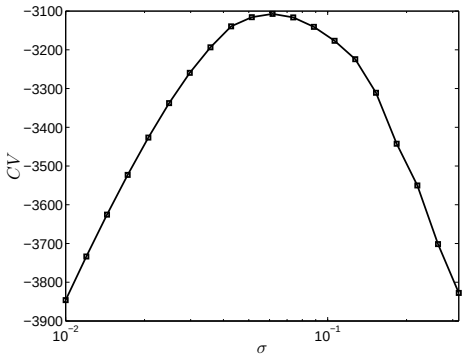


Figure 4: $CV(\sigma)$

Fig. 5 shows the estimation results of the proposed MAP method with different σ and ML method. 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.

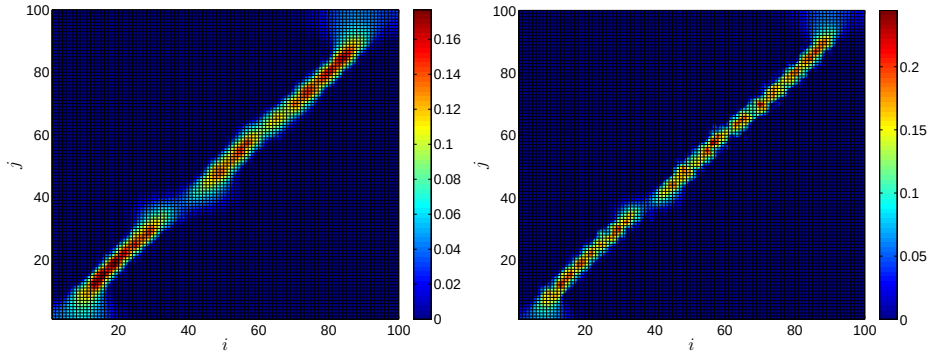
5.3. TM estimation with noisy data

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

$$o(t)|r(t) \sim \mathcal{N}\left(r(t), v^2\right) \tag{27}$$

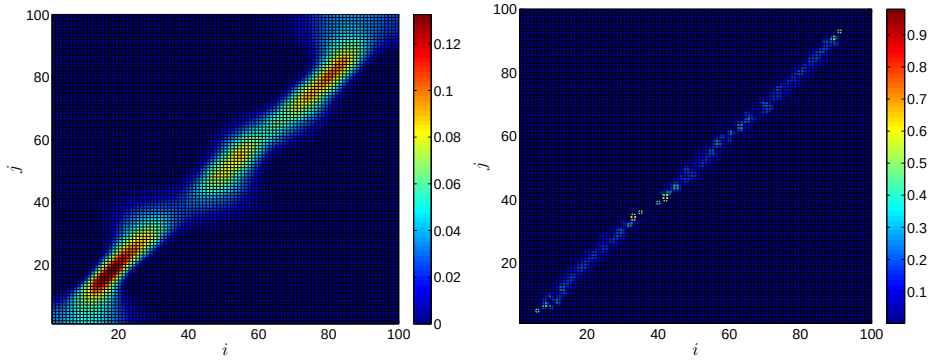
with $v = 0.1$.

The MAP estimator with GMRF prior in Section 4 will now be compared to the ML estimator implemented using Baum-Welch algorithm [27]. The MCCV results are shown in Fig. 6 and the optimal $\sigma^* = 0.1947$.



(a) MAP ($\sigma = \sigma^*$), $KLR\left(\hat{\boldsymbol{T}}||\boldsymbol{T}\right) = 0.0872$

(b) MAP ($\sigma = 3\sigma^*$), $KLR\left(\hat{\boldsymbol{T}}||\boldsymbol{T}\right) = 0.1232$



(c) MAP ($\sigma = \sigma^*/3$), $KLR\left(\hat{\boldsymbol{T}}||\boldsymbol{T}\right) = 0.7290$

(d) ML, $KLR\left(\hat{\boldsymbol{T}}||\boldsymbol{T}\right) = 0.2316$

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

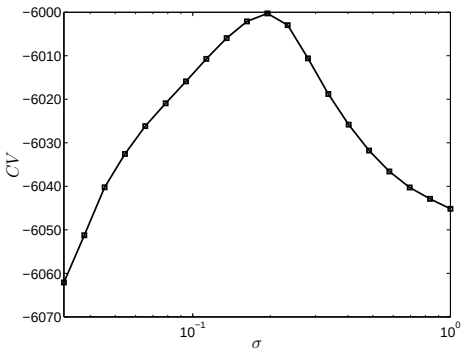
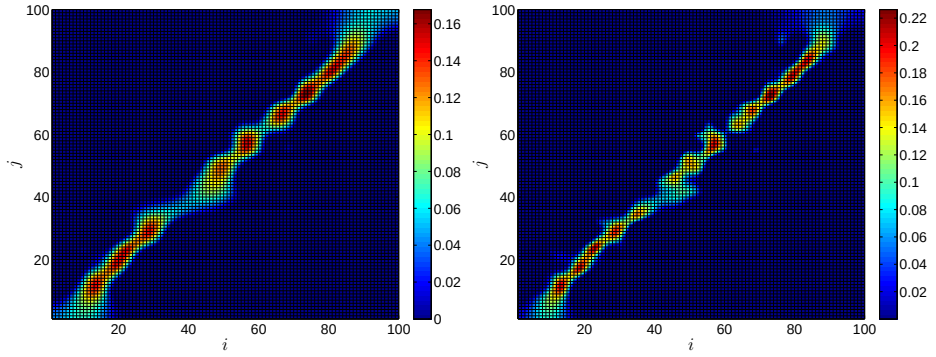
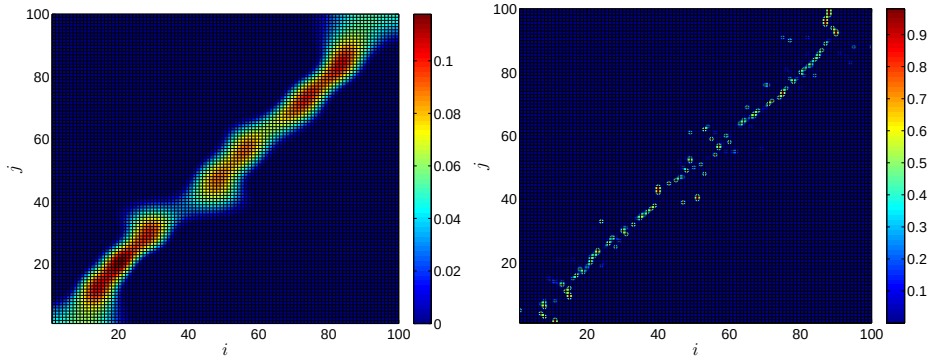


Figure 6: $CV\,(\sigma)$



(a) MAP ($\sigma = \sigma^*$), $KLR\left(\hat{\boldsymbol{T}}||\boldsymbol{T}\right) = 0.1700$

(b) MAP ($\sigma = 3\sigma^*$), $KLR\left(\hat{\boldsymbol{T}}||\boldsymbol{T}\right) = 0.2406$



(c) MAP ($\sigma = 0.06159$), $KLR\left(\hat{\boldsymbol{T}}||\boldsymbol{T}\right) = 0.4902$

(d) ML, $KLR\left(\hat{\boldsymbol{T}}||\boldsymbol{T}\right) = 2.2819$

Figure 7: $\hat{\boldsymbol{T}}$

In Fig. 7, 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. As can be seen from Fig. 7d, the ML estimator exhibits strong overfitting. With comparison to ML method, the proposed MAP estimator avoids overfitting by the regularization term $u(\boldsymbol{T}, \sigma)$, which penalizes excessively large value of T_{ij}.

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 7c, 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.

6. Conclusions

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.

References

[1] G. Miller, Finite Markov processes in psychology, Psychometrika 17 (2) (1952) 149–167.
[2] A. Madansky, Least squares estimation in finite Markov processes, Psychometrika 24 (2) (1959) 137–144.
[3] L. Telser, Least-squares estimates of transition probabilities, Measurement in Economics (1963) 270–292.
[4] T. Lee, G. Judge, T. Takayama, On estimating the transition probabilities of a Markov process, Journal of Farm Economics 47 (3) (1965) 742–762.
[5] H. Theil, G. Rey, A quadratic programming approach to the estimation of transition probabilities, Management Science 12 (9) (1966) 714–721.
[6] G. Judge, T. Takayama, Inequality restrictions in regression analysis, Journal of the American Statistical Association 61 (313) (1966) 166–181.
[7] T. Anderson, L. Goodman, Statistical inference about Markov chains, The Annals of Mathematical Statistics 28 (1) (1957) 89–110.
[8] M. Kendall, A. Stuart, The advanced theory of statistics, Vol. 2, Charles Griffin, London, 1961.
[9] T. Lee, G. Judge, A. Zellner, Estimating the parameters of the Markov probability model from aggregate time series data, North-Holland, 1970.
[10] G. Bowman, K. Beauchamp, G. Boxer, V. Pande, Progress and challenges in the automated construction of Markov state models for full protein systems, The Journal of Chemical Physics 131 (2009) 124101.
[11] T. Lee, G. Judge, A. Zellner, Maximum likelihood and Bayesian estimation of transition probabilities, Journal of the American Statistical Association 63 (324) (1968) 1162–1179.
[12] J. Martin, Bayesian decision problems and Markov chains, Wiley New York, 1967.
[13] F. Noé, Probability distributions of molecular observables computed from Markov models, The Journal of Chemical Physics 128 (2008) 244103.
[14] C. Fuh, T. Fan, A Bayesian bootstrap for finite state Markov chains, Statistica Sinica 7 (1997) 1005–1020.
[15] M. Meshkani, L. Billard, Empirical Bayes estimators for a finite Markov chain, Biometrika 79 (1) (1992) 185–193.
[16] S. Assoudou, B. Essebbar, A Bayesian Model for Markov Chains via Jeffrey's Prior, Communications in Statistics 32 (11).
[17] S. Assoudou, B. Essebbar, A Bayesian model for binary Markov chains, International Journal of Mathematics and Mathematical Sciences 2004 (8) (2004) 421–429.
[18] J. Besag, Spatial interaction and the statistical analysis of lattice systems, Journal of the Royal Statistical Society. Series B (Methodological) 36 (2) (1974) 192–236.
[19] J. Besag, C. Kooperberg, On conditional and intrinsic autoregressions, Biometrika 82 (4) (1995) 733–746.
[20] H. Rue, L. Held, Gaussian Markov random fields: theory and applications, Chapman & Hall, 2005.
[21] S. Li, Markov random field modeling in image analysis, Springer, 2009.
[22] S. Saquib, C. Bouman, K. Sauer, ML parameter estimation for Markov random fields with applications to Bayesian tomography, IEEE Transactions on Image Processing 7 (7) (1998) 1029–1044.
[23] T. Larsson, M. Patriksson, 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, no. 702, Department of Mathematics, Linköping University, 1998.
[24] J. Shao, Linear model selection by cross-validation, Journal of the American Statistical Association (1993) 486–494.
[25] 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) 39 (1) (1977) 1–38.
[26] Z. Rached, F. Alajaji, L. Campbell, The Kullback-Leibler divergence rate between Markov sources, IEEE Transactions on Information Theory 50 (5) (2004) 917–921.
[27] L. Welch, Hidden Markov models and the Baum-Welch algorithm, IEEE Information Theory Society Newsletter 53 (4) (2003) 1–10.