Estimation of Transition Matrices from Data

June 17, 2016

Consider that we have observed a trajectory $[s_1, ..., s_T]$ amongst n discrete states labeled 1 through n. The number of transitions between any pair of states i and j are denoted $\mathbf{C} = [c_{ij}^{\text{obs}}]$, i.e.:

$$c_{ij}^{\text{obs}} = |\{s_t = i \wedge s_{t+1} = j \mid t = 1, ..., T-1\}| \tag{0.1}$$

and the total number of transitions out of state i is c_i^{obs}:

$$c_i^{\text{obs}} = \sum_{j=1}^{n} c_{ij}^{\text{obs}}. \tag{0.2}$$

The likelihood of a Markov chain with transition matrix $\mathbf{P} = [p_{ij}] \in \mathbb{R}^{n\times n}$ is given by:

$$p(\mathbf{C}^{\text{obs}}|\mathbf{T}) = \prod_{i,j=1}^{n} p_{ij}^{c_{ij}^{\text{obs}}} \tag{0.3}$$

Our aim is to estimate the transition matrix that maximizes (0.3). In order to generalize our approach we also consider the maximization of a posterior probability which uses a conjugate prior, i.e. a prior with the same functional form as the likelihood:

$$p(\mathbf{T}|\mathbf{C}^{\text{obs}}) \propto \prod_{i,j=1}^{n} p_{ij}^{c_{ij}^{\text{prior}} + c_{ij}^{\text{obs}}} = \prod_{i,j=1}^{n} p_{ij}^{c_{ij}}, \tag{0.4}$$

Where c_{ij}^{prior} are prior counts. We now seek the matrix that maximizes (0.4). For the choice $c_{ij}^{\text{prior}} \equiv 0$ (uniform prior), this is identical to the maximum likelihood estimator. The probability (0.4) is difficult to work with due to the product. For optimization purposes it is therefore a common "trick" to instead work with the logarithm of the likelihood (log-likelihood):

$$Q = \log p(\mathbf{P} \mid \mathbf{C}) = \sum_{i,j} c_{ij} \log p_{ij}. \tag{0.5}$$

This is useful since the logarithm is a monotonic function: as a result, the maximum of $\log f$ is also the maximum of f. However, this function is not bounded from above, since for $p_{ij} \to \infty$, $Q \to \infty$. Of course, we somehow need to restrict ourselves to sets of variables which actually form transition matrices, i.e., they satisfy the constraint:

$$\sum_j p_{ij} = 1. \tag{0.6}$$

When optimizing with equality constraints, one uses Langrangian multipliers. The Lagrangian for Q is given by:

$$F = Q + \lambda_1(\sum_j p_{1j} - 1) + ... + \lambda_m(\sum_j p_{mj} - 1). \tag{0.7}$$

This function is maximized by the maximum likelihood transition matrix. It turns out that F only has a single stationary point, which can be easily found by setting the partial derivatives,

$$\frac{\partial \log F}{\partial p_{ij}} = \frac{c_{ij}}{p_{ij}} + \lambda_i, \tag{0.8}$$

to zero:

$$\frac{c_{ij}}{\hat{p}_{ij}} + \lambda_i = 0 \tag{0.9}$$

$$\lambda_i \hat{p}_{ij} = -c_{ij}. \tag{0.10}$$

We now sum over j on both sides and make use of the transition matrix property:

$$\lambda_i \sum_{j=1}^{m} \hat{T}_{ij} = \lambda_i = -\sum_{j=1}^{m} c_{ij} = -c_i. \tag{0.11}$$

Now we inserte the Langrange multipliers, $\lambda_i = -c_i$, in Eq. (0.9) and thus:

$$\frac{c_{ij}}{\hat{p}_{ij}} - c_i = 0 \tag{0.12}$$

$$\hat{p}_{ij} = \frac{c_{ij}}{c_i}. \tag{0.13}$$

It turns out that $\hat{\mathbf{P}}(\tau)$, as provided by Eq. (0.13), is the maximum of $p(\mathbf{P}|\mathbf{C}^{\text{obs}})$ and thus also of $p(\mathbf{C}^{\text{obs}}|\mathbf{P})$ when transition matrices are assumed to be uniformly distributed *a priori*.

In the limit of infinite sampling, i.e., trajectory length $T \to \infty$, $p(\mathbf{P}|\mathbf{C}^{\text{obs}})$ converges towards a Dirac delta distribution with its peak at $\hat{\mathbf{P}}(\tau)$. Given that π_i is the equilibrium probability of state i, the expected number of transitions out of state i is

$$\mathbb{E}[c_{ij}] = N\pi_i p_{ij} \tag{0.14}$$

and thus

$$\mathbb{E}[c_i] = N\pi_i. \tag{0.15}$$

In this case the prior contribution vanishes:

$$\lim_{N\to\infty}\hat{p}_{ij}=\lim_{N\to\infty}\frac{c_{ij}}{c_i}=\lim_{N\to\infty}\frac{c_{ij}^{\mathrm{prior}}+N\pi_i p_{ij}}{c_{ij}^{\mathrm{prior}}+\pi_i p_{ij}}=p_{ij}, \tag{0.16}$$

i.e., the estimator is "asymptotically unbiased".

Note that the nonreversible maximum likelihood estimator $\hat{p}_{ij}=c_{ij}/\sum_{k=1}^{n}c_{ik}$ does usually not fulfill detailed balance ($\pi_i\hat{p}_{ij}\neq\pi_j\hat{p}_{ji}$), even if the underlying dynamics is in equilibrium and thus $\pi_i p_{ij}=\pi_j p_{ji}$ holds for the exact transition matrix $\mathbf{P}(\tau)$. In many cases it is desirable and advantageous to estimate a transition matrix that does fulfill detailed balance. If the underlying microscopic dynamics are reversible, detailed balance of the discrete transition matrix $\mathbf{P}$ is a consequence and therefore it is meaningful to impose this structure onto $\hat{\mathbf{P}}$. Two practical implications are: (1) The additional constraints reduce the number of free parameters in $\hat{\mathbf{P}}$ which can result in a significant reduction of statistical uncertainty [2]. (2) The eigenvalues and eigenvectors of a reversible transition matrix are real.

There is no known closed form solution for the maximum probability estimator with the detailed balance constraint. We present two iterative methods subsequently.

Let $x_{ij}=\pi_i p_{ij}$ be the unconditional transition probability to observe a transition $i\to j$. These variables fulfill the constraint $\sum_{i,j}x_{ij}=1$, and the detailed balance condition is given by $x_{ij}=x_{ji}$. It is hence sufficient to store the x_{ij} with $i\leq j$ in order to construct a reversible transition matrix. A reversible transition matrix $\mathbf{P}=[p_{ij}]$ can be expressed as

$$p_{ij}=\frac{x_{ij}}{x_i}=\frac{x_{ij}}{\pi_i} \tag{0.17}$$

where $\mathbf{X}=[x_{ij}]$ is a nonnegative symmetric matrix and $x_i=\sum_j x_{ij}=\pi_i$. Then the log-likelihood of $\mathbf{X}$ with count matrix $\mathbf{C}=[c_{ij}]$ is

$$Q=\log p\left(\mathbf{C}|\mathbf{X}\right)=\sum_i c_{ii}\log\frac{x_{ii}}{x_i}+\sum_{i<j}\left(c_{ij}\log\frac{x_{ij}}{x_i}+c_{ji}\log\frac{x_{ji}}{x_j}\right) \tag{0.18}$$

We now consider solving

$$\hat{\mathbf{X}}=\arg\max_{\mathbf{X}}Q \tag{0.19}$$

We first derive the optimality conditions as in [1]. The partial derivatives of Q are given by:

$$\begin{aligned}\frac{\partial Q}{\partial x_{ij}} &= \frac{c_{ij}}{x_{ij}}+\frac{c_{ji}}{x_{ji}}-\sum_{j'=1}^{n}\frac{c_{ij'}}{\sum_{k=1}^{n}x_{ik}}-\sum_{j'=1}^{n}\frac{c_{jj'}}{\sum_{k=1}^{n}x_{jk}}\\ &= \frac{c_{ji}+c_{ij}}{x_{ij}}-\frac{c_i}{x_i}-\frac{c_j}{x_j}\end{aligned}$$

where we have used $c_i = \sum_{k=1}^{n} c_{ik}$. Q is maximized at $\frac{\partial Q}{\partial x_{ij}} = 0$, yielding the optimality conditions:

$$\frac{c_{ji} + c_{ij}}{x_{ij}} - \frac{c_i}{x_i} - \frac{c_j}{x_j} = 0 \qquad \forall i, j. \tag{0.20}$$

Direct fixed-point iteration: In [1] the following approach to search $\hat{\mathbf{X}}$ has been suggested. Solving 0.20 for x_{ij}, we can write the optimality conditions as the following fixed-point iteration:

$$x_{ij}^{(k+1)} = \frac{c_{ij} + c_{ji}}{\frac{c_i}{x_i^{(k)}} + \frac{c_j}{x_j^{(k)}}}, \tag{0.21}$$

where k is the iteration number. This approach iteratively updates all elements of $\mathbf{X}$ until the log-likelihood converges. The transition matrix $\hat{\mathbf{P}}$ can then be calculated from Eq. (0.17).

We can get a more memory-efficient algorithm by summing both sides over j and using $\sum x_{ij} = x_i$ and $x_i = \pi_i$ [5]:

$$\pi_i^{(k+1)} = \sum_j \frac{c_{ij} + c_{ji}}{\frac{c_i}{\pi_i^{(k)}} + \frac{c_j}{\pi_j^{(k)}}}. \tag{0.22}$$

In this approach, only the n variables of the $\boldsymbol{\pi}$-vector are varied and need to be stored. After convergence, the reversible transition matrix estimate is obtained in one step from the final estimates $\pi_i = \pi_i^{(k)}$:

$$\hat{p}_{ij} = \frac{c_{ij} + c_{ji}}{c_i + \frac{\pi_i}{\pi_j} c_j}. \tag{0.23}$$

Acknowledgements: Part of this article has been published in [4].

Citing maximum likelihood transition matrix estimation: This is textbook knowledge. Cite any suitable book on Markov chains, such as [3].

Citing maximum likelihood reversible transition matrix estimation: The fixed-point reversible estimator was first introduced in [1].

References

[1] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.*, 131:124101, 2009.

[2] F. Noé. Probability Distributions of Molecular Observables computed from Markov Models. *J. Chem. Phys.*, 128:244103, 2008.

[3] J. R. Norris. *Markov Chains.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

[4] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134:174105, 2011.

[5] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. Estimation and uncertainty of reversible markov models. *in preparation.*