A closer look at Markov chain transition matrix estimation when the model cycle length is shorter than the data observation interval

Duncan Ermini Leaf

University of Southern California
Leonard D. Schaeffer Center for Health Policy & Economics

dleaf@healthpolicy.usc.edu

December 2, 2021

Problem

- Data collected at long observation intervals
- Want to simulate a time-homogeneous Markov chain at shorter cycles (time steps) than the observation interval
- Example: monthly simulation from biennial data
- How to get the maximum likelihood estimate of the short-cycle transition matrix?

Notation

$$\mathbf{Q} = \mathbf{P}^k$$

- $\mathbf{Q}$ = transition matrix over data observation interval
- $\mathbf{P}$ = transition matrix over model cycle length
- k = number of cycles per observation interval
- Example: monthly simulation from biennial data...

$$\mathbf{Q} = \mathbf{P}^{24}$$

Craig & Sendi (2002)[1] diagonalization

1. Get maximum likelihood estimate $\hat{\mathbf{Q}}$ using the transitions rates observed in the data
2. Use eigendecomposition of $\hat{\mathbf{Q}}$ to obtain
   - $\mathbf{A}$, the matrix of eigenvectors, and
   - $\mathbf{D}$, the diagonal matrix of eigenvalues,

   such that,

$$\hat{\mathbf{Q}} = \mathbf{A}\mathbf{D}\mathbf{A}^{-1}$$

3. The maximum likelihood estimate of $\mathbf{P}$ is:

$$\hat{\mathbf{P}} = \mathbf{A}\mathbf{D}^{1/k}\mathbf{A}^{-1}$$

[1] Craig, B. A., & Sendi, P. P. (2002). Estimation of the transition matrix of a discrete-time Markov chain. *Health Economics*, *11*(1), 33–42.

Diagonalization doesn't always work...

- $\hat{\mathbf{P}} = \mathbf{A}\mathbf{D}^{1/k}\mathbf{A}^{-1}$ requires real-valued, non-negative eigenvalues and linearly independent eigenvectors
- When diagonalization fails, Craig & Sendi recommend iterative optimization of the likelihood of $\mathbf{P}$ in order to find $\hat{\mathbf{P}}$ directly, but they warn...

"*Convergence to the MLE is not guaranteed (may converge to local maximum) so several initial transition matrices are recommended.*"

Other indirect methods

- See Chhatwal et al. (2016)[2], Jahn et al. (2019)[3], and references therein.
- "Indirect" because they start with $\hat{\mathbf{Q}}$ and try to find $\hat{\mathbf{P}}$
- Failure modes:
  - $\hat{\mathbf{P}}^k = \hat{\mathbf{Q}}$, but $\hat{\mathbf{P}}$ is not a transition matrix (negative probabilties)
  - $\hat{\mathbf{P}}$ is a transition matrix, but $\hat{\mathbf{P}}^k \neq \hat{\mathbf{Q}}$
    - not a maximum likelihood estimate
- Can also result in $\hat{\mathbf{P}}^k = \hat{\mathbf{Q}}$ for multiple $\hat{\mathbf{P}}$

---

[2] Chhatwal, J., Jayasuriya, S., Elbasha, E. H. (2016). Changing cycle lengths in state-transition models: challenges and solutions. *Medical Decision Making, 36*(8), 952–964.

[3] Jahn, B., Kurzthaler, C., Chhatwal, J., Elbasha, E. H., Conrads-Frank, A., Rochau, U., ... Siebert, U. (2019). Alternative conversion methods for transition probabilities in state-transition models: validity and impact on comparative effectiveness and cost-effectiveness. *Medical Decision Making, 39*(5), 509–522.

Direct approach

- Maximize the likelihood of $\mathbf{P}$ instead of likelihood of $\mathbf{Q}$
- Constrain $\mathbf{P}$ to be a transition matrix
- This is avoids the failures of Craig & Sendi's diagonalization and the other indirect methods

Log-likelihood function

- $s =$ number of states in in the model
- $n_{ij} =$ number of transitions from state i to state j observed in the data
- Log-likelihood:

$$l(\mathbf{P}) = \sum_{i=1}^{s} \sum_{j=1}^{s} n_{ij} \ln q_{ij}(\mathbf{P})$$

where $q_{ij}(\mathbf{P})$ is the (i, j) element of $\mathbf{P}^k$

Optimization setup

- Constraints on $\mathbf{P}$:
  - $p_{ij} \geq 0$
  - $\sum_{j=1}^{s-1} p_{ij} \leq 1$
  - $p_{is} = 1 - \sum_{j=1}^{s-1} p_{ij}$
- Using constrOptim function in R, an interior-point (barrier) method
  - Adds boundary function to log-likelihood that pushes objective toward $-\infty$ when any p_{ij} gets very close to zero
  - Inner iteration uses BFGS optimization
  - See Chapter 16 of Lange (2010)[4]

[4] Lange, K. (2010). *Numerical Analysis for Statisticians* (2nd ed.). New York: Springer.

Grid search setup

- Optimization needs a initial value for $\mathbf{P}$
- Convergence depends on this inital value
- For each p_{ij} $(j \neq s)$, select R values evenly spaced within the $[0, 1]$ interval
  - For $R = 20$: 0.0476, 0.0952, 0.1429, ..., 0.9524
- Use the cross-product to create a set of initial values for $\mathbf{P}$: $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_M$
- Run optimization starting at each of these initial values and see where it converges

Counting convergence points

- How many *meaningfully* different convergence points are there?
- In publication, $\hat{\mathbf{P}}$ would be rounded to two or three decimal places
- Count the number of unique convergence points after rounding

Counting convergence clusters

- Initial values of $\mathbf{P}$ are equally spaced neighbors
- Convergence points should be clustered around local maxima
- Count the number of clusters such that
  1. points within clusters are closer than the initial distance between neighbors, and
  2. distance between clusters is greater that the initial distance between neighbors

- Any convergence points that differ by at most 0.005 are practically the same
- Count the number of clusters separated by at least 0.005 distance

Study setup

- $s = 3$ states, $M = 6,859,000$ initial values for $\mathbf{P}$
- $k = 2, 24, 100$ cycles per observation interval
- Study 1: $\hat{\mathbf{Q}}$ diagonalization fails due to one negative eigenvalue

$$\mathbf{N}_1 = \begin{bmatrix} 200 & 650 & 400 \\ 350 & 350 & 100 \\ 250 & 300 & 50 \end{bmatrix}$$

- Study 2: $\hat{\mathbf{Q}}$ diagonalization fails due to two negative eigenvalues

$$\mathbf{N}_2 = \begin{bmatrix} 100 & 200 & 650 \\ 300 & 350 & 100 \\ 250 & 300 & 50 \end{bmatrix}$$

Results

	Cycles per observation interval (k)		
	2	24	100
Study 1			
Non-converging	5	2	0
Unique to 2 decimal places	29,275	5,333,775	5,920,163
Unique to 3 decimal places	472,017	6,597,787	6,858,483
Clusters at 0.0476 distance	864	33	3
Clusters at 0.0050 distance	32,510	5,935,634	6,473,178
Study 2			
Non-converging	13	56	30
Unique to 2 decimal places	87,065	5,105,778	6,085,119
Unique to 3 decimal places	495,642	6,154,129	6,849,961
Clusters at 0.0476 distance	1,465	58	6
Clusters at 0.0050 distance	112,572	5,537,140	6,548,330

Concluding remark 1/3

For larger k the log-likelihood flattens, requiring tighter convergence criteria, but tighter convergence criteria can push the optimzation into the constraint boundary where it fails

- Alternative optimization approach of Craig & Sendi *might* avoid this, but needs to be investigated

Concluding remark 2/3

This is a lot of computational work. Are these worth the effort?

- likelihood maximization
- discrete time
  - Consider survival / time-to-event models
- time-homogeniety

Concluding remark 3/3

Further work:

- complex eigenvalues
- $s > 3$ states
- Any suggestions?

Acknowldegment

The author acknowledges the Center for Advanced Research Computing (CARC) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication. URL: `https://carc.usc.edu`.