Least square estimation of a hidden Markov chain parameters

Junko Murakami [1] and Thomas Taylor [2]

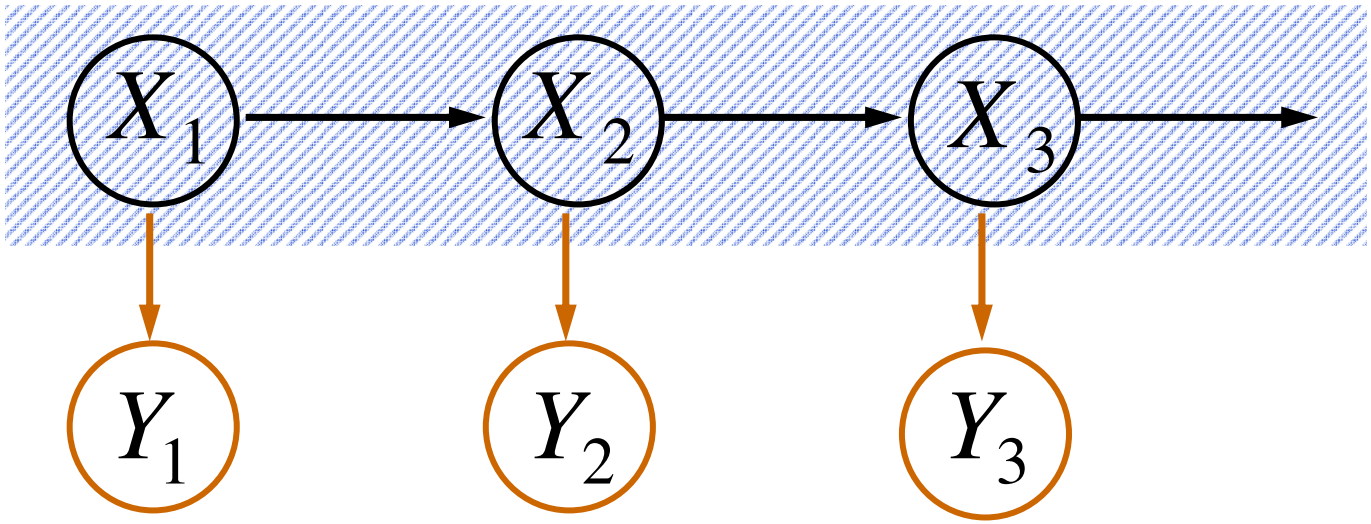
1. Victoria University of Wellington
2. Arizona State University

Outline

- Description of ‘simple’ hidden Markov models.
- Description of least square error estimate (LSE) (or Bayes estimate). – *mean*
- Comparison of the *mode*, the maximum likelihood estimate obtained by the Baum-Welch (B-W) algorithm, and the *mean.*

What do I mean by ‘simple’ HMMs?

‘simplest’ HMM (1)



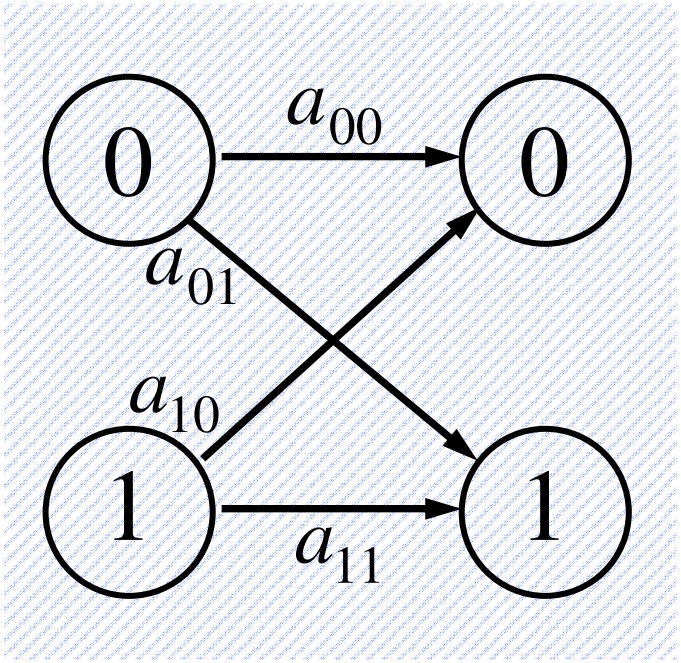
State sequence (Markov chain) $X^{1,n} = (X_1, X_2, \ldots, X_n)$

Observation sequence $Y^{1,n} = (Y_1, Y_2, \ldots, Y_n)$

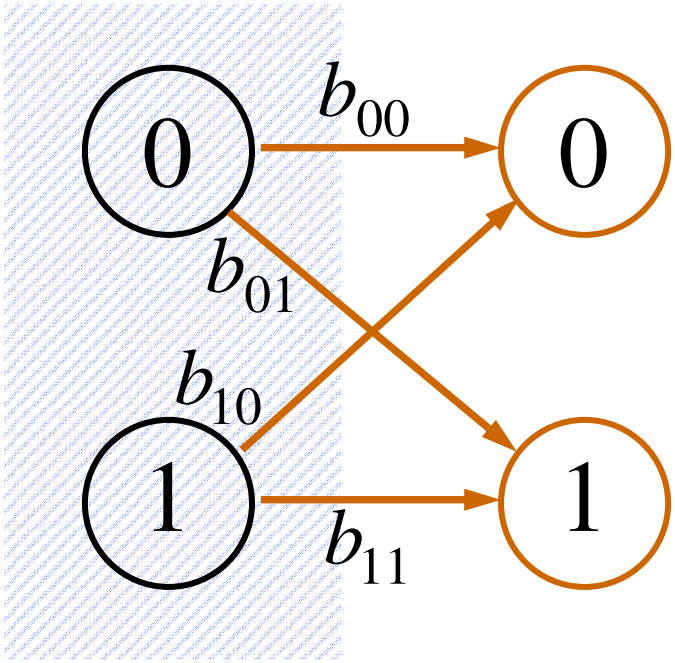
$X^{1,n} \in \{0,1\}$ and $Y^{1,n} \in \{0,1\}$

‘simplest’ HMM (2)

Conditional Probabilities



X_t X_{t+1}



X_t Y_t

Also, let $r_0 = P(X_1 = 0)$ and $r_1 = P(X_1 = 1)$.

Note : $a_{i1} = 1 - a_{i0}$, $b_{i1} = 1 - b_{i0}$, and $r_1 = 1 - r_0$ for $i = 0, 1$.

Least Square Error (LSE) Estimate (or Bayes Estimate)

LSE (Bayes) Estimate (1)

Finds the expected value of the parameter set given an observation sequence; i.e.,

$$\hat{\theta} = E[\theta] = \int \theta \, P\left(\theta \mid Y^{1,n}\right) d\theta.$$

Assuming the uniform distribution of θ (i.e., letting $P(\theta) = 1$), and using Bayes' theorem, we have

$$E[\theta] = \frac{\sum_{X^{1,n} \in \Omega_n} \int \theta \, P\left(Y^{1,n}, X^{1,n} \mid \theta\right) d\theta}{\sum_{X^{1,n} \in \Omega_n} \int P\left(Y^{1,n}, X^{1,n} \mid \theta\right) d\theta}$$

LSE (Bayes) Estimate (2)

NOTE

As it is, the summation is over

$$\Omega_n = \{\text{all the possible values of } X^{1,n}\},$$

which has the size 2^n.

LSE (Bayes) Estimate (3)

First, we let

$$k_{ij} = \#\left(X_t = i \text{ and } X_{t+1} = j\right) \text{ and}$$
$$l_{iu} = \#\left(X_t = i \text{ and } Y_t = u\right)$$

for $i, j, u \in \{0, 1\}$, where $\#\left(\text{event}\right)$ means the total number of events over $t \in \{1, 2, \ldots, n\}$.

Let $K = \{k_{ij}\}$ and $L = \{l_{iu}\}$.

LSE (Bayes) Estimate (4)

If we let $r_i = 1/2$ for simplicity, $P\left(Y^{1,n}, X^{1,n} \mid \theta\right)$ is in the form

$$\frac{1}{2} a_{00}^{k_{00}} \left(1 - a_{00}\right)^{k_{01}} a_{11}^{k_{11}} \left(1 - a_{11}\right)^{k_{10}} b_{00}^{l_{00}} \left(1 - b_{00}\right)^{l_{01}} b_{11}^{l_{11}} \left(1 - b_{11}\right)^{l_{10}},$$

and so both $\int_\theta \theta\, P\left(Y^{1,n}, X^{1,n} \mid \theta\right) d\theta$ and $\int_\theta P\left(Y^{1,n}, X^{1,n} \mid \theta\right) d\theta$ are functions of K and L, where $\theta = \{r_0, a_{00}, a_{11}, b_{00}, b_{11}\}$.

NOTE: Because of the symmetry in the probability distribution, the integration should be under some restriction; e.g., $a_{00} \geq a_{11}$.

LSE (Bayes) Estimate (5)

Fact

1. The integrals are functions of $\{K, L\}$.

2. $\{K, L\}$ can be expressed as a function of $\omega_n = \{k_1, k_{11}, l_{11}, X_1, X_n\}$ instead,

where k_1 is the number of 1's in $X^{1,n}$.

The integrals $\int_\theta \theta\, P(Y^{1,n}, X^{1,n} \mid \theta)\, d\theta$ and $\int_\theta P(Y^{1,n}, X^{1,n} \mid \theta)\, d\theta$ are functions of ω_n.

LSE (Bayes) Estimate (6)

Fact

Given a particular $Y^{1,n}$, some different state sequences $X^{1,n}$ produce the same value of $\omega_n = \{k_1, k_{11}, l_{11}, X_1, X_n\}$.

Let $h_n(\omega_n)$ be the number of distinct $X^{1,n}$ values that corresponds to the same input ω_n given $Y^{1,n}$.

LSE (Bayes) Estimate (7)

Then the summations can be over $\widehat{\Omega}_n = \{\omega_n \mid h_n(\omega_n) > 0\}$ of sizes that are polynomial of n

instead of over all the possible values of $X^{1,n} \in \Omega_n$ of size 2^n.

Note: The algorithm describe in this presentation can be extended to **any** state sizes of the Markov chain and observation sequences.

LSE (Bayes) Estimate (8)

Let $h_1(0, 0, 0, 0, 0) = 1$

for t from 1 to $n - 1$

with all $\omega_t = (k_1, k_{11}, l_{11}, 0, X_t)$ such that $h_t(\omega_t) > 0$

increment $h_{t+1}(k_1, k_{11}, l_{11}, 0, 0)$ and

$h_{t+1}(k_1 + 1, k_{11} + X_t, l_{11} + X_{t+1}, 0, 1)$

by the value $h_t(\omega_t)$

end for

(Because of the symmetry, we can find $h_n(\omega_n)$ for $X_1 = 1$ once the ones for $X_1 = 0$ is obtained.)

LSE (Bayes) Estimate (9)

Computational complexity reduction

n	size of Ω_n	size of $\widehat{\Omega}_n$		
	2^n	median	minimum	maximum
50	1.1×10^{15}	14647.5	2398	17283
100	1.3×10^{30}	117825	4951	143817
150	1.4×10^{45}	399459.5	43785	483721

LSE (Bayes) Estimate (10)

Advantages

- On average, closer than MLE (e.g., Baum-Welch estimates) to the true parameters when the data size is small.
- Online computation is possible.
- Finds the exact expected values (unbiased).
- One-time computation.

LSE (Bayes) Estimate (11)

Disadvantage

- Computational complexity grows still exponentially in the state space size.

Simulation: B-W and LSE estimates with a small data set (1)

Outline:

Generate 200 θ - values, randomly with respect to (i) the determinant of $A = \{a_{ij}\}$ and to (ii) the difference $b_{00} - b_{10}$.

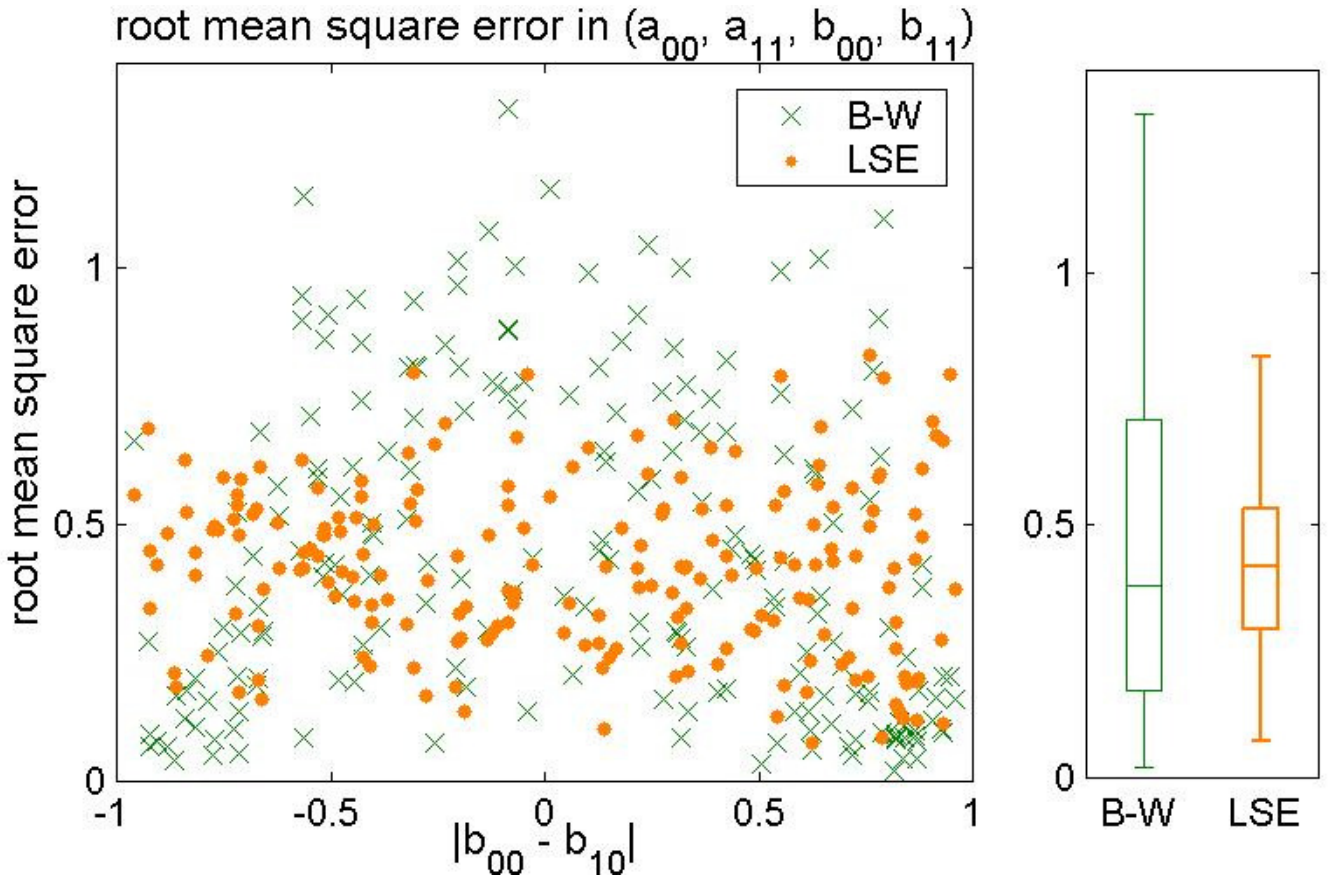
For each θ, generate a set of $\{X^{1,n}, Y^{1,n}\}$, $n = 150$, and obtain the estimates.

As for the Baum-Welch algorithm:

Find 15 estimates using randomly picked initial estimates.

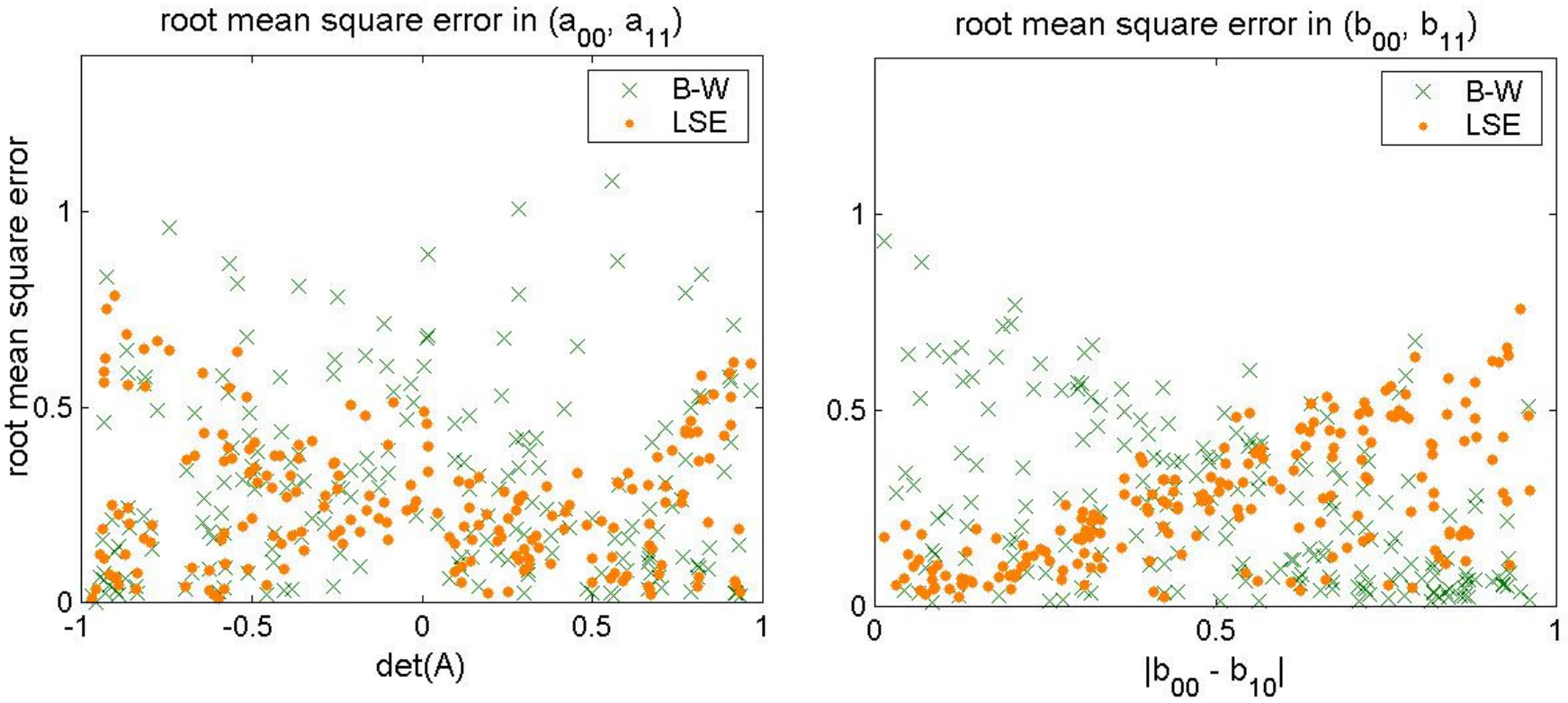
Pick the one with the largest basin.

Simulation: B-W and LSE estimates with a small data set (2)



On average, the B-W estimates (green dots) were farther away from the true parameters and less stable than LSE estimates (orange dots).

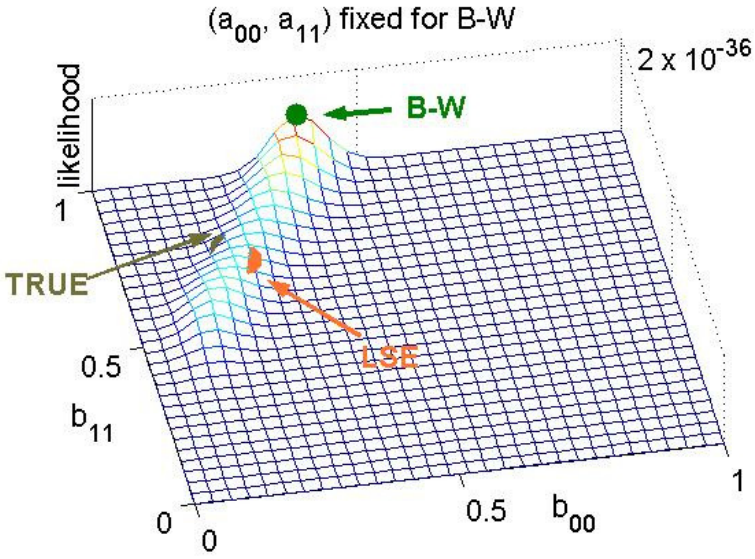
Simulation: B-W and LSE estimates with a small data set (3)

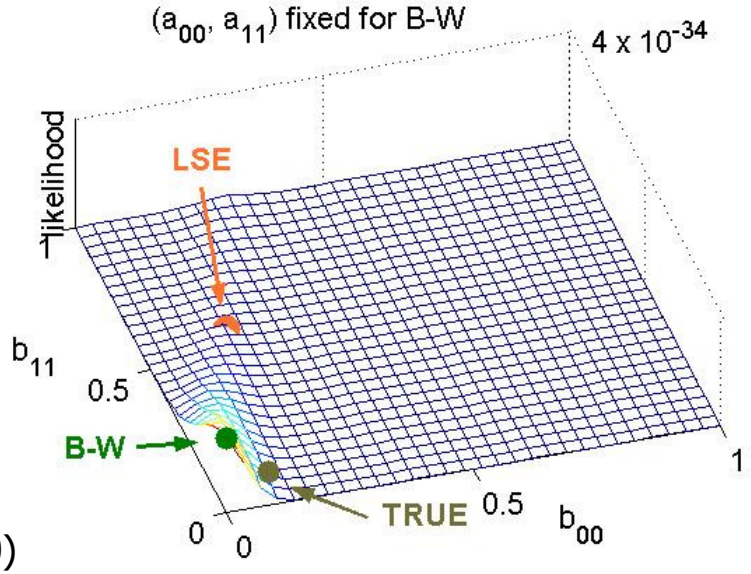


The estimation error in the matrices for the Markov chain and observation sequence are plotted separately.

From the figure on the right, we see that the less the current state matters to what we observe, the better the LSE becomes than the B-W.

Simulation: B-W and LSE estimates with a small data set (4)





(n = 150)

$|b_{00} - b_{10}| = 0.05$

	a_{00}	a_{11}
TRUE	0.90	0.40
LSE	0.65	0.39
B-W	0.43	0.00

$|b_{00} - b_{10}| = 0.95$

	a_{00}	a_{11}
TRUE	0.90	0.40
LSE	0.77	0.51
B-W	0.82	0.38

- On the left, the effect of the current state to the probability distribution for the observation is small. The mode is unstable, while the mean is stable.
- On the right, the effect is large. The mode tend to be closer to the true parameter value, while the mean often stays in some distance from the mode.

Conclusions

- Depending on the application, the LSE estimates might be more suitable than the popular B-W estimates.

- Extension of the LSE algorithm to various HMMs coming.

- Further comparison of the mode and mean of the likelihood surface for various types of HMMs could be interesting.

Referenes

- J. A. Bilms, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," International Computer Science Institute, Tech. Rep. ICSI-TR-97-021, April 1998.

- L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1-8, 1972.

- J. Murakami, "Parameter estimate of a hidden Markov chain," Unpublished Ph.D. Dissertation, Arizona State University, Tempe, AZ, USA, May 2005.