Hidden Markov Models with Applications in Cell Adhesion Experiments

Ying Hung
Department of Statistics and Biostatistics
Rutgers University
Outline

• Introduction to cell adhesion experiments
• Challenges
• Hidden Markov models
• Order selection in HMMs
• Simulations
• Three real examples
• Summary
References


An Example of Cell Biology Experiment

• Inspired by the statistical analysis of biomechanical experiments.

• Cell adhesion
  ➢ The binding of a cell to another cell or surface.
  ➢ Mediated by specific interaction between cell adhesion proteins (receptors) and the molecules that they bind to (ligands).

• Biologists describe the receptor-ligand binding as a key-to-lock type relation.

• What makes cells sticky? When, how, and to what cells adhere?

• Why important?
  ➢ Cell adhesion plays an important role in many physiological and pathological processes.
  ➢ Plays a key role in tumor metastasis in cancer study.
Thermal fluctuation experiments

- Thermal fluctuation experiment is a specific type of cell adhesion experiment, which uses the reduced thermal fluctuations to indicate the presence of receptor-ligand bonds.
- Objective: Identify association and dissociation points for receptor-ligand bonds.
- Accurate estimation of these points is essential because
  - they are required for precise measurement of bond lifetimes and waiting times,
  - they form the basis for subsequent estimation of the kinetic parameters.
Experimental settings

• In thermal fluctuation assay, red blood cell (RBC) is used as an adhesion sensor.
• Receptor surface (target bead on the right of the Figure) and ligand surface (probe bead linked to a RBC on the left of the Figure) are brought into contact, allowing receptor-ligand bonds to form by thermal fluctuation of RBC.
• When a bond forms, thermal fluctuations are reduced.
• Decrease/resumption of thermal fluctuations of a biomembrane force probe (left bead linked to RBC) pinpoints association/dissociation of receptor-ligand bonds.

![Diagram of experimental setup](image)

**Ligand surface**
VWF-A1 is one type of ligand

**Receptor surface**
GC is the receptor
Data

- The position of the probe bead is tracked by image analysis software to produce the data.
- Horizontal position of the edge of the probe bead is plotted versus time.
- Bond formation is equivalent to adding a molecular spring in parallel to the force transducer spring to stiffen the system (Marshall et al. 2006). Therefore, the fluctuation decreases when a receptor-ligand bond forms and resumes when the bond dissociates.
- Bond lifetime: from bond formation to dissociation
- Bond waiting time: from bond dissociation to formation
Challenges

- Challenges in identifying the bond association/dissociation points:
  ✓ Points are not directly observable.
Challenges

• Challenges in identifying the bond association/dissociation points:
  ✓ Points are not directly observable. Can only be detected by variance changes.
Challenges

- Challenges in identifying the bond association/dissociation points:
  - Points are not directly observable. Can only be detected by variance changes.
  - Observations are not independent.
Challenges

- Challenges in identifying the bond association/dissociation points:
  - Points are not directly observable. **Can only be detected by variance changes.**
  - Observations are not independent. The thermal fluctuations are independently distributed given their binding status (e.g., binding or not), but transition from one status to another can be dependent due to *cell memory* (Hung et al. 2008).
Challenges

• Challenges in identifying the bond association/dissociation points:
  ✓ Points are not directly observable. Can only be detected by variance changes.
  ✓ Observations are not independent. The thermal fluctuations are independently distributed given their binding status (e.g., binding or not), but transition from one status to another can be dependent due to cell memory (Hung et al. 2008).
  ✓ In practice, data contains an unknown number of bond types and each bond associated with different fluctuation decreases, depending on the stiffness of the molecules.
Challenges

• Challenges in identifying the bond association/dissociation points:
  ✓ Points are not directly observable. Can only be detected by variance changes.
  ✓ Observations are not independent. The thermal fluctuations are independently distributed given their binding status (e.g., binding or not), but transition from one status to another can be dependent due to cell memory (Hung et al. 2008).
  ✓ In practice, data contains an unknown number of bond types and each bond associated with different fluctuation decreases, depending on the stiffness of the molecules. The following figure shows two types of bonds.
An naive approach

- Moving standard deviation plot based on the thermal fluctuation data.
- A plot with standard deviations calculated by 15 consecutive observations.
- Some periods (marked by arrows) in which the standard deviations decrease significantly indicate the presence of bonds.
Limitations

- Standard deviation plots are intuitive and easy to implement but have limitations:
  - the analysis relies heavily on subjective judgments and is time-consuming.
  - the accuracy of identifying the association and dissociation points is susceptible to the number of consecutive points used in the calculation.
  - Inconsistent identification of association/dissociation points.
  - has no clear decision rule and theoretical justification, especially when the observations are not independent.
  - This issue becomes more serious when there is more than one type of bonds.
A framework based on Hidden Markov models

- The probe fluctuations with different variations that correspond to different underlying binding states.
- Use hidden states in HMM to represent the binding status.
- These unobservable states, including no bond and a number of distinct types of bonds, are not assumed to be independent, but rather to have a Markovian structure so that the cell memory effects can be captured.
- Given the hidden states, probe locations are assumed to be independent and normally distributed with some unknown parameters that capture the variations associated with different bond types.
Hidden Markov Model with Two States

\[ Y_0 \sim N(\mu_0, \sigma_0^2) \quad Y_1 \sim N(\mu_1, \sigma_1^2) \]

\( Y \): possible observation, random variables

\( 0 \)-Non-Bond
\( 1 \)-Bond

\( 2\sigma_1 \)
\( 2\sigma_0 \)
Hidden Markov Model with two states

\[ Y_0 \sim N(\mu_0, \sigma_0^2) \quad Y_1 \sim N(\mu_1, \sigma_1^2) \]

- \( a_{00} \): state transition probabilities
- \( Y \): possible observation, random variables
- 0-Non-Bond
- 1-Bond

\[ 2\sigma_1 \quad 2\sigma_0 \]
Hidden Markov Model with Two States

Y: possible observation, random variables
a: state transition probabilities

\[ Y_0 \sim N(\mu_0, \sigma_0^2) \]
\[ Y_1 \sim N(\mu_1, \sigma_1^2) \]

\( a_{01} \)
Definitions for the HMM

• Suppose $y_s$ represents the probe location at time $s$. There is an unobservable binding state, denoted by $x_s$, associated with $y_s$.
• The change of states can be captured by a stationary Markov chain on $K$ states with transition probability $P_{ij} = P(x_{s+1} = j | x_s = i)$ and stationary probability $\pi_i$, where $i, j = 1, \ldots, K$.
• Different order of cell memory effect can be captured and assessed by the use of transition matrix.
• Conditional on the underlying binding states, the observed probe locations are assumed to be mutually independent and normally distributed with density
  \[ f(y_s; \sigma_{x(s)}, \phi_{x(s)}) \], where $\phi_{x(s)}$ and $\sigma_{x(s)}^2$ are the mean and variance.
• Hidden states are defined only according to the variance in the study.
• The mean functions are allowed to be different with respect to the states because the probe can be pulled/pushed by a small force due to the presence of a bond.
Hidden Markov Models

• The standard thermal fluctuation experiment is usually conducted with several independent replicates.

• Notation:

Assume \( Y_i = (y_{i1}, \ldots, y_{it}) \). Let \( X_i = (x_{i1}, \ldots, x_{it}) \), \( \Phi_i = (\phi_{i1}, \ldots, \phi_{it}) \), and \( \Sigma_i = (\sigma_{x(i1)}, \ldots, \sigma_{x(it)}) \) be the hidden states, mean, and variance for the \( i \)th sequence. Then, the density for \( Y_i \) can be written as

\[
F(Y_i; \Sigma_i, \Phi_i) = \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} \prod_{j=1}^{t} f(y_{ij}; \sigma_{x(ij)}, \phi_{x(ij)}) \pi_{x(i1)}P_{x(i1)x(i2)} \cdots P_{x(i,t-1)x(it)}.
\]

• Number of the hidden states \( \rightarrow \) number of bond types
• Starting and ending points of each state \( \rightarrow \) association and dissociation points of the corresponding bond.
Problems solved?

- Challenges in identifying the bond association/dissociation points:
  - Points are not directly observable. Hidden states in HMM.
  - Observations are not independent. Cell memory can be handled by transition matrix.

In practice, data contains an unknown number of bond types and each bond associated with different fluctuation decreases, depending on the stiffness of the molecules.

**Objective:** Estimate the order of the HMM.
Existing work in HMM

• HMMs have proven to be useful in many areas (Rabiner 1989) and theoretical properties of HMMs, given the number of state is known, have been extensively studied (Leroux 1992, Bickel et al. 1998).

• The unknown number of bond types in the current experiment requires the study of a new problem, which is important but has not yet been satisfactorily resolved.

• In practice, AIC or BIC type of methods are commonly used. However, they have not been theoretically justified in the context of HMMs (MacDonald and Zucchini 1997).
A double penalized log-likelihood function is defined as

\[ \tilde{l}_n(\Sigma, \Phi) = l_n(\Sigma, \Phi) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_n(\eta_k), \]

where \( \eta_k = \sigma_{k+1} - \sigma_k \), for \( k = 1, 2, \ldots, K - 1 \), and \( \sigma_1 \leq \sigma_2 \cdots \leq \sigma_K \).

- \( K \) is an upper bound for the order of the states.
- The true order, denoted by \( K_0 \), is unknown.
- The first penalty is used to prevent small values of stationary probabilities.
- The second penalty is a nonnegative function that shrinks small \( \eta_k \). Thus, it prevents overfitting by different normal distributions with variances close to each other.
- \( p_n \) is assumed to be the SCAD penalty (Fan and Li 2001). Other penalties can also be used.
- **Theorem:** Under some identifiability and regularity conditions, \( \hat{K}_0 \) tends to \( K_0 \) with probability tending to one.
Applications beyond cell adhesion experiments

• Although the proposed order selection approach in HMMs is motivated by a study of cell adhesion experiment, it has many areas of application, including signal processing, environmental science, and bioinformatics.

• The number of underlying states is often unknown in practice. Efficient estimation of order not only improves the prediction but also provides valuable scientific information.

• Examples:
  ✓ MacKay (1994) proposes an HMM to model lesions experienced on the brain stem given an unknown number of disease states in the study of multiple sclerosis.
  ✓ Hughes and Guttorp (1994) models the rainfall process given unobserved and unknown number of weather states.
  ✓ In the study of heart rate variability in sleeping neonates (Celeus and Clairambault 1992), the proposed method can be used to characterize the unknown number of periods in the neonate sleep.
## Simulation settings

<table>
<thead>
<tr>
<th>case</th>
<th>$K_0$</th>
<th>Mean</th>
<th>Variance</th>
<th>Transition Matrix</th>
<th>$n$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>$(0,0)'$</td>
<td>$(0.5,4)'$</td>
<td>$P_{ij} = 0.5$</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$(0,0)'$</td>
<td>$(0.5,4)'$</td>
<td>$P_{ij} = 0.5$</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>$(0,0)'$</td>
<td>$(0.5,4)'$</td>
<td>$P_{ij} = 0.5$</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>$(0,0)'$</td>
<td>$(0.5,4)'$</td>
<td>$P_4 = \begin{pmatrix} 0.9 &amp; 0.1 \ 0.3 &amp; 0.7 \end{pmatrix}$</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>$(0,0)'$</td>
<td>$(0.5,4)'$</td>
<td>$P_4$</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>$(0,0)'$</td>
<td>$(0.5,4)'$</td>
<td>$P_4$</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>$(9,20,1,9)'$</td>
<td>$(0.3,0.5,1,2)'$</td>
<td>$P_8 = \begin{pmatrix} 0.75 &amp; 0.1 &amp; 0.1 &amp; 0.05 \ 0.2 &amp; 0.7 &amp; 0.05 &amp; 0.05 \ 0.2 &amp; 0.1 &amp; 0.6 &amp; 0.1 \ 0.3 &amp; 0.1 &amp; 0.1 &amp; 0.5 \end{pmatrix}$</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>$(9,20,1,9)'$</td>
<td>$(0.3,0.5,1,2)'$</td>
<td>$P_8$</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>$(3,10,7,1)'$</td>
<td>$(0.3,0.5,0.8,1.1)$</td>
<td>$P_{ij} = 0.25$</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>$(3,10,7,1)'$</td>
<td>$(0.3,0.5,0.8,1.1)$</td>
<td>$P_{ij} = 0.25$</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>$(3,10,7,1)'$</td>
<td>$(0.3,0.5,0.8,1.1)$</td>
<td>$P_{ij} = 0.25$</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>$(20,1,5,9,17)'$</td>
<td>$(0.6,0.8,1.5,1.7,2)$</td>
<td>$P_{ij} = 0.2$</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>$(20,1,5,9,17)'$</td>
<td>$(0.6,0.8,1.5,1.7,2)$</td>
<td>$P_{ij} = 0.2$</td>
<td>10</td>
<td>500</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>$(20,1,5,9,17)'$</td>
<td>$(0.6,0.8,1.5,1.7,2)$</td>
<td>$P_{ij} = 0.2$</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>$(0,10,-16,20,15,0,2,0.5,0.8,0.9,1.2,1,3,1.4,2,2.1)'$</td>
<td>$P_{ij} = 1/9$</td>
<td>10</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>$(0,10,-16,20,15,0,2,0.5,0.8,0.9,1.2,1,3,1.4,2,2.1)'$</td>
<td>$P_{ij} = 1/9$</td>
<td>50</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

Note: The table represents various simulation settings with different parameters for the mean and variance, along with the transition matrix for each case. Each case indicates the number of replications ($n$) and the number of iterations ($t$).
Simulation results

• Comparison with AIC, BIC, and minimal-distance (MS) criterion (MacKay 2002).

Table 2: Simulation results

<table>
<thead>
<tr>
<th>case</th>
<th>order</th>
<th>AIC</th>
<th>BIC</th>
<th>MS</th>
<th>DP</th>
<th>case</th>
<th>order</th>
<th>AIC</th>
<th>BIC</th>
<th>MS</th>
<th>DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.22</td>
<td>0.7</td>
<td>0.23</td>
<td>0.09</td>
<td>1</td>
<td>0.18</td>
<td>0.7</td>
<td>0.16</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.78</td>
<td>0.3</td>
<td>0.74</td>
<td>0.84</td>
<td>2</td>
<td>0.82</td>
<td>0.29</td>
<td>0.83</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.06</td>
<td>2</td>
<td>0.81</td>
<td>0.59</td>
<td>0.78</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.17</td>
<td>0.66</td>
<td>0.17</td>
<td>0.07</td>
<td>4</td>
<td>0.01</td>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.16</td>
<td>0.37</td>
<td>0.11</td>
<td>0.05</td>
<td>6</td>
<td>3</td>
<td>0.01</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.84</td>
<td>0.63</td>
<td>0.85</td>
<td>0.88</td>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Simulation results

- Comparison with AIC, BIC, and minimal-distance (MS) criterion (MacKay 2002).

The estimated order equals to a value between 1 and 11. The value with the highest frequency is indicated by a boldface. When the true order is 2 (i.e., cases 1 to 6), the double penalized approach (DP) is consistently the best and has more than 84% success rate in identifying the true order. AIC also works reasonably well in these settings, while BIC tends to underestimate in some cases. When sample size increases, the selection accuracy of DP is improved which is consistent with the asymptotic results. This result is observed throughout the simulations, i.e., cases 7-8, cases 9-11, cases 12-14, and cases 15-16. For $K_0 = 4$, both MS and DP outperform the other methods in the unbalanced cases (of 7 and 8). It appears to be more difficult to identify the correct order in the balanced cases (cases 9 to 14). In these cases, DP consistently identifies the correct order with the highest frequency while most of the other methods underestimate the order even for larger sample size. In cases 15 and 16 with $K_0 = 9$, DP outperforms the other methods.

Table 2: Simulation results

<table>
<thead>
<tr>
<th>case</th>
<th>order</th>
<th>AIC</th>
<th>BIC</th>
<th>MS</th>
<th>DP</th>
<th>case</th>
<th>order</th>
<th>AIC</th>
<th>BIC</th>
<th>MS</th>
<th>DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.22</td>
<td>0.7</td>
<td>0.23</td>
<td>0.09</td>
<td>1</td>
<td>0.18</td>
<td>0.7</td>
<td>0.16</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.78</td>
<td>0.3</td>
<td>0.74</td>
<td>0.84</td>
<td>2</td>
<td>0.82</td>
<td>0.29</td>
<td>0.83</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.06</td>
<td>3</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.17</td>
<td>0.66</td>
<td>0.17</td>
<td>0.07</td>
<td>1</td>
<td>0.18</td>
<td>0.41</td>
<td>0.14</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.79</td>
<td>0.34</td>
<td>0.81</td>
<td>0.89</td>
<td>2</td>
<td>0.81</td>
<td>0.59</td>
<td>0.78</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.04</td>
<td>0</td>
<td>0.02</td>
<td>0.03</td>
<td>3</td>
<td>0.01</td>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.16</td>
<td>0.37</td>
<td>0.11</td>
<td>0.05</td>
<td>1</td>
<td>0.19</td>
<td>0.39</td>
<td>0.14</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.84</td>
<td>0.63</td>
<td>0.85</td>
<td>0.88</td>
<td>2</td>
<td>0.8</td>
<td>0.59</td>
<td>0.84</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.07</td>
<td>3</td>
<td>0.01</td>
<td>0.02</td>
<td>0</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Simulation results

Table 3: Simulation results

<table>
<thead>
<tr>
<th>case</th>
<th>order</th>
<th>AIC</th>
<th>BIC</th>
<th>MS</th>
<th>DP</th>
<th>case</th>
<th>order</th>
<th>AIC</th>
<th>BIC</th>
<th>MS</th>
<th>DP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td><strong>0.72</strong></td>
<td>0.02</td>
<td>0.08</td>
<td><strong>0.72</strong></td>
<td>3</td>
<td>0.39</td>
<td><strong>0.65</strong></td>
<td>0</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.42</td>
<td>0.28</td>
<td><strong>0.86</strong></td>
<td><strong>0.73</strong></td>
<td></td>
<td>5</td>
<td>0.18</td>
<td>0</td>
<td>0.12</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0.42</td>
<td>0.28</td>
<td><strong>0.86</strong></td>
<td><strong>0.73</strong></td>
<td></td>
<td>8</td>
<td>0.44</td>
<td>0.32</td>
<td><strong>0.88</strong></td>
<td><strong>0.90</strong></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.39</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.02</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>11</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.42</td>
<td>0.28</td>
<td><strong>0.86</strong></td>
<td><strong>0.73</strong></td>
<td></td>
<td>12</td>
<td>0.45</td>
<td>0.04</td>
<td>0.34</td>
<td><strong>0.75</strong></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.82</td>
<td>0.29</td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>0.01</td>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td></td>
<td>14</td>
<td>0.58</td>
<td>0.75</td>
<td><strong>0.57</strong></td>
<td><strong>0.38</strong></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.18</td>
<td>0.12</td>
<td>0.19</td>
<td></td>
<td></td>
<td>15</td>
<td>0.59</td>
<td><strong>0.74</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.28</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>0.61</td>
<td><strong>0.8</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.52</td>
<td><strong>0.72</strong></td>
<td>0.38</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.39</td>
<td>0.28</td>
<td><strong>0.59</strong></td>
<td><strong>0.74</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.03</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.01</td>
<td>0</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0</td>
<td>0.07</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.01</td>
<td>0.02</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.07</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The estimated order equals to a value between 1 and 11. The value with the highest frequency is indicated by a boldface. When the true order is 2 (i.e., cases 1 to 6), the double penalized approach (DP) is consistently the best and has more than 84% success rate in identifying the true order. AIC also works reasonably well in these settings, while BIC tends to underestimate in some cases. When sample size increases, the selection accuracy of DP is improved which is consistent with the asymptotic results. This result is observed throughout the simulations, i.e., cases 7-8, cases 9-11, cases 12-14, and cases 15-16. For \( K_0 = 4 \), both MS and DP outperform the other methods in the unbalanced cases (of 7 and 8). It appears to be more difficult to identify the correct order in the balanced cases (cases 9 to 14). In these cases, DP consistently identifies the correct order with the highest frequency while most of the other methods underestimate the order even for larger sample size. In cases 15 and 16 with \( K_0 = 9 \), DP outperforms the other methods.
Simulation results

• The true order is indicated by the number with boldface in the “order” column.
• For each method, we report the percentage of times out of 100 replications that the estimated order equals to a value between 1 and 11.
• For lower orders ($K_0=2,3$):
  – the double penalized (DP) approach is consistently the best and has more than 80% success rate in correctly selecting the number of states.
  – AIC also works reasonably well.
  – MD and BIC tend to underestimate.
• For higher orders:
  – DP consistently identifies the correct order with highest frequency.
Application in thermal fluctuation experiments

• The HMM approach is applied to real data to assess the accuracy in identifying the number of bond types and specifying their association/dissociation points.

• Three data sets.
  ✓ The first data set has one type of bond and is used to validate the level of thermal fluctuation reduction for the L-selectin-PSGL-1 bond.
  ✓ The second data set has a mixture of two different types of bonds and is used to test the proposed model and see if it can separate these two bonds.
  ✓ The third data set is used to illustrate the performance of HMM in estimating kinetic parameters. The results are compared with a carefully performed descriptive method (the only existing method for analyzing thermal fluctuation experiments).
Example 1: Two states

- Study the interaction between L-selectin and PSGL-1.
- Low densities of selectins and PSGL-1 are used to ensure that interactions formed are most likely single bonds (i.e., either no bond or a single L-selectin-PSGL-1 bond for each interaction).
- 18 independent replicates and each of them has over 300 probe positions recorded in 5 seconds.
Example 1: Two states

- $K=4$ is applied and $K_0=2$ is estimated using the double penalized approach.
- Transition matrix is
  \[
  \hat{P} = \begin{pmatrix}
  0.9924 & 0.0076 \\
  0.0242 & 0.9758
  \end{pmatrix}
  \]
  and the stationary probabilities are
  \[
  \hat{\pi} = (0.7645, 0.2355).
  \]
Example 2: Three states

• Two types of receptor-ligand bonds formed due to interactions of L-selectin and P-selectin with the PSGL-1 ligand.
• 48 independent mixture sequences.
• $K=5$ is applied and $K_0=3$ is estimated
• Transition matrix:

$$
\hat{P} = \begin{pmatrix}
0.9499 & 0.0498 & 0 \\
0.0018 & 0.8953 & 0.1029 \\
0.0449 & 0.0636 & 0.8915
\end{pmatrix},
$$

stationary distribution:

$$
\hat{\Pi} = (0.3404, 0.4264, 0.2332).
$$
Example 2: Three states

- The estimated variances for the P-selectin-PSGL-1 bond and the L-selectin-PSGL-1 bond are 16.2104 and 12.4027.
- They indicate that the formation of the L-selectin-PSGL-1 bond reduces the thermal fluctuations more than the P-selectin-PSGL-1 bond.
- This can be explained biologically because L-selectin has a higher stiffness than P-selectin (Chen et al. 2008).
Example 3: Performance on kinetic parameter estimation

- Compare the proposed HMM with the descriptive method based on on-rate and off-rate estimates.
- The descriptive method requires significantly longer training time compared to HMM method and the time to finish one analysis is much longer.
- Two states: No bond or one VWF-A1—GC bond
- The on- and off-rates ($k_{on}$ and $k_{off}$) represent statistical characteristics underlying the probabilistic kinetic processes of binding and unbinding.
- The on- and off-rates are calculated through bond life time and waiting time.

![Image of micropipette and RBC](image)

![Diagram of VWF-A1—GC bond](image)
Example 3: Performance on kinetic parameter estimation

- On-rate is calculated by the negative slope of the following plot.
- VWF-A1-GC bond on-rate calculated by Descriptive and HMM are $1.302 \pm 0.079 \text{ s}^{-1}$ and $1.395 \pm 0.046 \text{ s}^{-1}$ respectively.
- The two confidence intervals overlap, indicating that the parameter estimates are statistically close to each other.
Example 3: Performance on kinetic parameter estimation

- off-rate is calculated by the negative slope of the following plot.
- GC with VWF-A1 off-rate by the descriptive statistical algorithm and HMM are $26.58 \pm 0.92 \text{ s}^{-1}$ and $26.46 \pm 0.18 \text{ s}^{-1}$, respectively.
- The two confidence intervals overlap, indicating that the parameter estimates are statistically close to each other.
More comparisons of Descriptive and HMM

- A and B show the 95% confidence intervals of the estimated on-rate (figure A) and off-rate (figure B).
  - HMM has much narrower width of CI for three more molecular interactions tested. Thus HMM is more accurate because it reduces the errors brought by subjective judgment of the experimenter.
- C and D show the numbers of waiting times (figure C) and bond lifetime (figure D) that the two methods are respectively capable of measuring from the same set of data. Three molecular interactions are tested (WT, G1324S, and R1450E).
  - Many of the waiting times and bond lifetimes gone undetected by the descriptive method can be resolved by HMM.
  - HMM captures more events than the descriptive statistical method from the same set of data, e.g. 112 vs 40 for waiting times and 169 vs 46 for bond lifetimes for the WT case.
Summary

• A framework based on Hidden Markov models is proposed
• A double penalized order selection approach for HMM
• Applications in cell adhesion experiments
• The proposed HMM method allows us to speed up the analysis and improve the quality of estimates of receptor-ligand binding kinetics.
Thank you!
HMM for thermal Fluctuation Data

HMM (30s)

<table>
<thead>
<tr>
<th>Time (s)</th>
<th>Probe position (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>-75</td>
</tr>
<tr>
<td>1.5</td>
<td>-70</td>
</tr>
<tr>
<td>2.0</td>
<td>-65</td>
</tr>
<tr>
<td>2.5</td>
<td>-10</td>
</tr>
<tr>
<td>3.0</td>
<td>0</td>
</tr>
</tbody>
</table>

Approach | Hold (unbound) | Contact | Hold (bound) | Retract

Probe position Std. (nm) vs Time (s)

$Z_t \sim \text{Markov Chain}$

$P_{00} P_{01} P_{11} P_{10}$

$x_t | z_t=0 \sim N(\mu_0, \sigma_0^2)$

$x_t | z_t=1 \sim N(\mu_1, \sigma_1^2)$
Tuning Parameter Stability

• The probability of observing a data point in the unbound state $P_0$ is the only tuning parameter in the algorithm.

• We plot the relative error against different $P_0$. In Fig. A, the $P_0$ that gives the lowest ranges from 0.85 to 0.96 from which we choose the value in our algorithm. In Fig. B, different choices of $P_0$ do not render much inconsistency between the results from descriptive statistical method and HMM, as the relative error is smaller than 0.025. This shows the robustness in the prediction performance and the reliability of the HMM-tuning parameter.
Learning Curve and Running Time

A. Learning curves for new students

- Descriptive
- HMM

B. Time spent under B

- Descriptive
- HMM

Student 1

Student 2