Overview

- Types of hidden Markov model
- Informal and graphical approaches
- General goodness-of-fit testing
- Targeted model extensions
- Conclusion
Types of hidden Markov model

- Hidden Markov models can take many different forms
  - Applicable to time series analysis
  - Applicable to longitudinal data analysis
  - Applicable to event history analysis (multi-state modelling)
- Particular approaches to assessing goodness-of-fit may not be applicable in all cases
Types of hidden Markov model

- Discrete time versus continuous time
  - Is the underlying Markov chain defined by a transition probability matrix or a set of transition intensities?
- Ergodic chain or presence of absorbing states
  - Can the chain be assumed in equilibrium?
- Continuous response, discrete or categorical responses
- Single (long) time series of observations or multiple independent chains for different subjects
Example: Hospital infections

Monthly counts of cases of MRSA (hospital infections) in a hospital (Cooper & Lipsitch, 2004)

\[ Y = 0 \]

\[ Y | X \sim Po(\lambda X) \]

- Discrete count responses (e.g. conditionally Poisson)
- Discrete time series
  - Uses a continuous time model for the underlying Markov process
- Single time series of observations
- Assume chain begins in equilibrium
Example: Lung function in transplant patients

- Continuous response - % of baseline $FEV_1$
- Continuous time model because subjects observed at unique irregularly spaced time points and time of death known exactly
- Short chains of observations for $\sim 400$ patients

$Y \sim N(\mu_1, \sigma_1^2)$  \hspace{1cm}  $Y \sim N(\mu_2, \sigma_2^2)$
Cardiac allograft vasculopathy (CAV) in post-heart-transplantation patients

- HMM used here as a direct extension to **multi-state models**
- Categorical/ordinal response: Healthy, Mild CAV, Severe CAV, Death
  - Diagnosis subject to misclassification $\Rightarrow$ HMM
- Continuous time model because subjects observed at unique irregularly spaced time points
- Short chains of observations for 596 patients
- Patients begin reasonably healthy and progress to death so equilibrium cannot be assumed
Assumptions to be assessed

- Markov property
- Time homogeneity (stationarity)
- Conditional independence of the responses
- Conditional distribution of the responses
- Effect of covariates
**Equilibrium response distribution**

- **Mackay Altman (2004):** If the underlying Markov chain is ergodic and the HMM is stationary:
  - $\pi$ is the stationary distribution of the Markov chain
  - $F_k(y; \theta)$ is the response distribution given underlying state $k$
  - The marginal distribution of responses $Y_t$ is given by
    \[
    F(y; \theta) = \sum_k \pi_k F_k(y; \theta)
    \]

- Idea of the approach is then to compare the empirical distribution of responses $\hat{F}_n(y)$ with the parametric HMM estimate $F(y; \hat{\theta})$
Equilibrium response distribution

- If the model is correctly specified then a plot of $F(y; \hat{\theta})$ against $\hat{F}_n(y)$ should be close to a straightline.
- Main sensitivity is to deviation from assumed response distribution.
  - Can look at bivariate CDFs between consecutive observations, i.e. compare $F^2(y_1, y_2; \hat{\theta})$ with $\hat{F}^2_n(y_1, y_2)$ at a range of points $(y_1, y_2)$.
- Method relies on the assumption that the process is in equilibrium.
- Requires ordered responses taking a reasonably large number of values (e.g. continuous or count data).
Example: Simulated HMM with Poisson responses
Pseudo-residuals

- MacDonald & Zucchini (2009) advocate the use of *pseudo-residuals* to assess model fit.
- Two possible types of pseudo-residual:
  - **Ordinary pseudo-residual**: probability of seeing a less extreme response than observed given all observations except that at time $t$
    \[
    u_t = P(Y_t \leq y_t | y_{1:(t-1)}, y_{(t+1):n})
    \]
  - **Forecast pseudo-residual**: probability of seeing a less extreme response than observed given observations up to time $t$
    \[
    u_t = P(Y_t \leq y_t | y_{1:(t-1)})
    \]
Pseudo-residuals

- For well fitting models, the pseudo-residuals should be approximately $U[0, 1]$ distributed (though not independent).
- Can alternatively take $r_t = \Phi^{-1}(u_t)$ to get equivalent pseudo-residuals which should be approximately $N(0, 1)$.
- Need responses which are Normally distributed or counts with reasonably high means.
- Available for discrete-time models fitted in the `HiddenMarkov` package in R.
- Don’t need to assume chain(s) start in equilibrium.
- Can be computed for continuous time models.
Example: Papilloma counts on mice exposed to a carcinogen - Ordinary pseudo-residuals

Ridall & Pettitt (2005): Partially hidden Markov model with underlying states corresponding to i) Increase in count of papilloma ii) Decrease in count of papilloma iii) Stable count in papilloma

Homogeneous

Inhomogeneous
Example: Papilloma counts on mice exposed to a carcinogen - Forecast pseudo-residuals

Ridall & Pettitt (2005): Partially hidden Markov model with underlying states corresponding to i) Increase in count of papilloma ii) Decrease in count of papilloma iii) Stable count in papilloma
Methods based on marginal behaviour

- Suppose HMM has initial probability vector $\pi$ at time 0
- $E(Y_t) = \sum E(Y_t|X_t)P(X_t)$
- In the equilibrium case $P(X_t) = \pi$ for all $t$
- But more generally $P(X_t) = \pi^T P^t$
  where $P$ is the transition probability matrix
- Could therefore compare “observed” and expected mean observations at each time point
- Is applicable for continuous or discrete time models
  - In continuous time case may need to employ smoothing methods to estimate $\hat{E}(Y_t)$ because subjects are observed at different time points
Example: FEV₁ data

- Responses assumed to be Normally distributed.

- Can only assess $E(Y_t|X_t \neq 3)$ i.e. the mean response given the subject is still alive.
Methods based on marginal behaviour

- If the responses are categorical using the expected observation is not appropriate
- Instead, if $Y_t$ takes values on $1, \ldots, R$ consider

$$P(Y_t = r) = \sum_{X_t} P(Y_t = r | X_t) P(X_t), r = 1, \ldots, R$$

- These are effectively “prevalence counts”
- Straightforward to compute for discrete-time models where all subjects are observed at same times
- More difficult in the continuous time case, with irregular sampling and/or censoring/drop out
Methods based on marginal behaviour

Estimation of empirical $P(Y_t = r)$

- If discrete time and all subjects observed at same time:
  \[
  \hat{P}(Y_t = r) = \frac{1}{N} \sum_i I(Y_{it} = r)
  \]

- If in continuous time, observations may be at unique time points
  - Obtain $I(Y_i(t_{ij}) = r)$
  - Then use some form of non-parametric smoothing to get a non-parametric estimate of $P(Y(t) = r)$
  - Depending on the model, it may be reasonable to assume $P(Y(t) \leq r)$ is a decreasing function of time - then isotonic regression methods are applicable
Example: CAV data

- Compare “expected” $P(Y_t = r)$ with non-parametric estimate
- Exploits fact that no backward transitions are allowed in $X_t$ and misclassification probabilities are assumed fixed in time
- Hence $P(Y_t \leq r)$ are decreasing functions in $t$ for $r = 1, 2, 3$
Example: CAV data

- Compare “expected” \( P(Y_t = r) \) with non-parametric estimate.
- Exploits fact that no backward transitions are allowed in \( X_t \) and misclassification probabilities are assumed fixed in time.
Example: CAV data

Non-parametric estimate for state 4 coincides with Kaplan-Meier estimate.
Survival comparison

- Many hidden Markov models used in medical or health applications have an absorbing state
  - e.g. death
  - Usually observed without misclassification - often exact time of death (absorption) also known

- Compare fitted model for absorption times with empirical estimates
  - Kaplan-Meier if times of absorption known up to right-censoring or NPMLE for interval-censored data
  - If covariates, stratify into covariate groups and compare KM with average fitted survival
In a HMM, if we know that $X_{k-1} = X_k = X_{k+1}$ then the corresponding $Y_{k-1}, Y_k, Y_{k+1}$ should be independent and identically distributed.

This property can be exploited to provide simple graphical assessments of fit.

- Provided the underlying Markov chain makes few jumps and we have reasonably long sequences of observations.
Local correlation: FEV\textsubscript{1} data

Patient 88

Viterbi algorithm estimate of the state change timepoint
Local correlation: FEV$_1$ data

Simulated trajectory given model and change point
Autocorrelation between adjacent points much higher than expected by the HMM
Pearson-type $\chi^2$ tests

- For standard (unhidden) Markov chain models, can compare observed and expected transition counts

\[ O_{ijk} = \sum I(X_k = j, X_{k-1} = i) \]

\[ E_{ijk} = \sum I(X_{k-1} = i)P(X_k = j|X_{k-1} = i) \]

- Then

\[ X^2 = \sum \sum \sum \frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}} \sim \chi^2 \]

- Exact asymptotic distribution if discrete-time (Anderson & Goodman, 1957) or in continuous time but with balanced observation times (Kalbfleisch & Lawless, 1985)
Pearson-type $\chi^2$ tests

- When observation times are subject specific or there are continuous covariates, $P(X_k = j | X_{k-1} = i)$ is different for each subject.
- Idea is to group observations with “similar” covariates and/or times between observations together (Aguirre-Hernandez and Farewell (2002)).

$$O_{ijkg} = \sum_{s \in g} I(X_{ks} = j, X_{k-1,s} = i)$$

$$E_{ijkg} = \sum_{s \in g} I(X_{k-1,s} = i) P(X_{ks} = j | X_{k-1,s} = i)$$

$$X^2 = \sum_g \sum_k \sum_i \sum_j \frac{(O_{ijkg} - E_{ijkg})^2}{E_{ijkg}} \approx \chi^2$$
Pearson-type $\chi^2$ tests

- For hidden Markov models, we know that $P(Y_k = j)$ depends on all past observations of the process.
- Modified chi-squared test works based on considering $P(Y_k = j|Y_{k-1} = i, Y_{1:(k-2)})$ (Titman and Sharples, 2008).
- As in the covariate case, we can group together similar observations, for instance based on $Y_{k-2}$.

\[
O_{ijkg} = \sum_{s \in g} I(Y_{ks} = j, Y_{k-1,s} = i)
\]

\[
E_{ijkg} = \sum_{s \in g} I(Y_{k-1,s} = i)P(Y_{ks} = j|Y_{k-1,s} = i, Y_{1:(k-2),s})
\]

- Effectively this is looking at how well the model predicts the next observation given all previous observations.
Pearson-type $\chi^2$ tests

- In each case the null distribution is not $\chi^2$ (even asymptotically) due to the grouping of non-identically distributed transition intervals.
- Crude estimate is to assume true distribution lies between $\chi^2_{n-|\theta|}$ and $\chi^2_n$ where $n$ is number of independent cells and $|\theta|$ the number of parameters fitted.
- More accurate asymptotic approximation can be calculated (based on a weighted sum of independent $\chi^2_1$ random variables) (Titman, 2009).
- Alternatively can use a parametric bootstrap.
  - Can be slow because requires refitting of the model.
- Modification required to the test if time of death is known exactly (Titman & Sharples (2008)).
Example: CAV data

\[ T = 63.4, \, df = 31, \, p = 0.001 \]

Lack of fit due primarily to excess \(1 \rightarrow 2\) and \(2 \rightarrow 3\) transitions in certain periods

<table>
<thead>
<tr>
<th>CG</th>
<th>TQ</th>
<th>1→1</th>
<th>1→2</th>
<th>1→3</th>
<th>1→4</th>
<th>1→C</th>
<th>2→1</th>
<th>2→2</th>
<th>2→3</th>
<th>...</th>
<th>3→C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Obs</td>
<td>160</td>
<td>35</td>
<td>1</td>
<td>7.2</td>
<td>23</td>
<td>19</td>
<td>54</td>
<td>20</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exp</td>
<td>168.8</td>
<td>24.3</td>
<td>4.8</td>
<td>5.1</td>
<td>23.2</td>
<td>19.5</td>
<td>50.8</td>
<td>19.9</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Obs</td>
<td>217</td>
<td>41</td>
<td>6</td>
<td>16.1</td>
<td>18</td>
<td>9</td>
<td>24</td>
<td>11</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exp</td>
<td>216.6</td>
<td>39.8</td>
<td>10.8</td>
<td>12.7</td>
<td>18.2</td>
<td>8.5</td>
<td>24.0</td>
<td>10.8</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Obs</td>
<td>267</td>
<td>57</td>
<td>14</td>
<td>50.7</td>
<td>101</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exp</td>
<td>259.9</td>
<td>59.1</td>
<td>20.8</td>
<td>48.4</td>
<td>101.5</td>
<td>2.3</td>
<td>6.8</td>
<td>5.2</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Obs</td>
<td>215</td>
<td>25</td>
<td>8</td>
<td>3.8</td>
<td>25</td>
<td>7</td>
<td>28</td>
<td>16</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exp</td>
<td>222.9</td>
<td>19.9</td>
<td>3.4</td>
<td>5.8</td>
<td>24.8</td>
<td>14.4</td>
<td>27.1</td>
<td>9.7</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Obs</td>
<td>251</td>
<td>20</td>
<td>6</td>
<td>7.0</td>
<td>15</td>
<td>9</td>
<td>11</td>
<td>4</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exp</td>
<td>241.3</td>
<td>25.9</td>
<td>6.2</td>
<td>10.8</td>
<td>14.8</td>
<td>6.7</td>
<td>12.2</td>
<td>5.4</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Obs</td>
<td>256</td>
<td>26</td>
<td>9</td>
<td>37.2</td>
<td>94</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exp</td>
<td>249.7</td>
<td>31.6</td>
<td>9.7</td>
<td>37.1</td>
<td>94.0</td>
<td>2.3</td>
<td>3.4</td>
<td>2.7</td>
<td>...</td>
</tr>
</tbody>
</table>
Comments on Pearson-type tests

- Requires an arbitrary grouping of the observations
  - Different choices in terms of the nature and number of groups can affect the overall conclusions
- Necessity to bootstrap to get an accurate $p$-value makes it time consuming
- Low power to detect mild or moderate lack of fit for realistic sample size
- Will tend to give some idea of where the lack of fit is occurring
- Implemented within the R package $msm$ - including the modified test for exact death times
Other formal tests

- Our hidden Markov models typically have the assumption that
  \[ Y_t | X_t \perp \perp Y_{t-1} \]

- A simple, if slightly crude, way of testing this assumption is to adopt a regression model where \( Y_t | X_t \) is allowed to depend on \( Y_{t-1} \) directly. (Titman & Sharples, 2009)

- For instance, in a model with normally distributed observations,
  \[ E(Y_t | X_t = r) = \mu_r + \beta Y_{t-1} \]
  or for a misclassification model
  \[ \logit(P(Y_t = s | X_t = r)) = \alpha_{rs} + \beta I(Y_{t-1} = s) \]

- Can then use a likelihood-ratio test for \( \beta = 0 \) as a goodness-of-fit test
An alternative way of testing a specific assumption is to fit a model where the assumption has been relaxed.

- **Stationary/Time homogeneity** \(\Rightarrow\) Inhomogeneous alternative
- **Markov assumption** \(\Rightarrow\) Semi-Markov model
- **Fixed number of states** \(\Rightarrow\) Alternative number of states
- **Fixed response distribution** \(\Rightarrow\) More general responses
- **Fixed effect models** \(\Rightarrow\) Random effect models
Model Extensions: Hidden Semi-Markov models

- Relax the assumption that the sojourn times in each state are geometric (for discrete time models) or exponential (for continuous time models)
- In each case easiest computationally to use an “aggregated Markov model”
- Discrete time: Langrock, Zucchini (2011)
- Continuous time: Constrain the states to have “phase-type” distributions (Titman & Sharples 2010)
- Equivalent to adding extra latent states
- Problem is related to the issue of selecting the number of underlying states of an HMM
- Non-standard asymptotics when comparing models via likelihood ratios
Conclusions

- No single approach to assessing goodness-of-fit for HMMs
  - Some methods only appropriate for certain types of HMM
  - Depends which assumptions are most important for the application
- Worth considering a range of approaches
- Fitting a more complicated model is usually most powerful way of testing a particular assumption
  - Not always computationally feasible
References

- Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. JASA. (1985).
References (continued)