Computation of the asymptotic null distribution of goodness-of-fit tests for multi-state models

A. C. Titman †

Abstract

An approach for testing goodness-of-fit for parametric continuous-time panel observed multi-state Markov models is to use Pearson-type χ^2 tests. When observations are balanced and there are no continuous covariates, a test with an asymptotic χ^2 null distribution can be found. Aguirre-Hernandez and Farewell (2002) constructed a Pearson-type test by grouping observations with similar time intervals and covariate values, which is appropriate in more general cases. Titman and Sharples (2008) proposed an extension to accommodate models where the observed states can be subject to misclassification. While in some cases the chi-squared distribution can give a good approximation to the null distribution of these statistics, more generally, and particularly if there are continuous covariates in the model, the null distribution of the statistic has a higher mean than the χ^2 approximation. Use of the parametric bootstrap can yield a good approximation of the null distribution, however this method requires multiple refitting of the model. For many models this may be prohibitively time consuming, limiting the practical applicability of the tests. In this paper, a better approximation to the asymptotic distribution of the tests, as a weighted sum of independent χ^2_1 random variables, is given.

1 Introduction

Parametric continuous-time multi-state models are a widely used method to model the progression of a categorical response variable over time. In medical applications, the response may refer to a disease state, and this state may only be observed at discrete, irregular intervals. Such an observation scheme is referred to as panel observation. The multi-state process is often assumed to be Markov. This greatly simplifies the computation of the likelihood. For data of the form of a series of observations $x_{i0}, \ldots, x_{in_i}$ at times $t_{i0}, \ldots, x_{in_i}$, for patients $i = 1, \ldots, N$, with covariate vectors z_i, the log-likelihood can be expressed as

$$l(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log \left(p_{x_{i(j-1)} x_{ij}}(t_{i(j-1)}, t_{ij}; z_i, \theta) \right),$$

where

$$p_{rs}(t_0, t_1; z, \theta) = \mathbb{P}(X(t_1) = s | X(t_0) = r; z, \theta)$$

†Address for correspondence: Department of Mathematics and Statistics, Lancaster University, UK. Email: a.titman@lancaster.ac.uk

is the (r, s) entry of the transition probability matrix which can be found by solving the Kolmogorov Forward equations (Cox and Miller (1965))

$$\frac{dP(t_0, t)}{dt} = P(t_0, t)Q(t), \tag{1.1}$$

subject to initial condition $P(t_0, t_0) = I$. When the process is time homogeneous, the solution is given by a matrix exponential. In this situation, a closed-form solution to (1.1) can be found

$$P(t_0, t_1) = P(s) = \exp(sQ_0) = \sum_{n=0}^{\infty} \frac{s^n}{n!} Q_0^n,$$

where $s = t_1 - t_0$. The matrix exponential in this equation can be calculated using the eigen-decomposition of Q_0. Let D be a diagonal matrix of the eigenvalues and U the matrix with the corresponding eigenvectors as columns. Provided the eigenvalues are distinct, U is invertible and $Q_0 = UDU^{-1}$. Then

$$\exp(sQ) = U \exp(sD)U^{-1}.$$

Kalbfleisch and Lawless (1985) gave a numerical Fisher scoring algorithm for computing the maximum likelihood estimate.

In some observational studies, the diagnosed state at a particular time may be subject to misclassification. A method of attempting to account for this is to make the assumption that the observed states, $o_{i0}, \ldots, o_{in_i}$ are independent conditional on the values of $x_{i0}, \ldots, x_{in_i}$. The observable process is then a continuous time hidden Markov model (HMM). The observed states relate to the true underlying states through classification probabilities,

$$\pi_{rs} = \mathbb{P}(O(t) = s | X(t) = r).$$

These probabilities may be assumed unknown and to be estimated from the data.

Approaches to computation of the maximum likelihood estimates for misclassification HMMs are either to apply a continuous-time generalization of the Forward-Backward algorithm for discrete-time HMMs (Bureau et al (2003)). Alternatively the likelihood can be computed directly and maximized using derivative free optimization algorithms (Jackson and Sharples (2002); Satten and Longini (1996)).

2 Goodness-of-fit tests

In order to assess the appropriateness of the Markov assumption, as well as possible stationarity and subject homogeneity assumptions, goodness-of-fit procedures are required. When the observation times are balanced, so that all subjects are observed at the same times, $t_{1j} = \ldots = t_{Nj}$, and there are no covariates or the covariates have a small number of unique values, then a Pearson chi-squared test or asymptotically equivalent likelihood ratio test can be applied (Kalbfleisch and Lawless (1985); de Stavola (1988)). The likelihood ratio test has statistic

$$\Lambda = 2 \sum_j \sum_c \sum_{r=1}^{R} \sum_{s=1}^{R} o_{hcrs} \log\left(\frac{o_{jcrs}}{e_{jcrs}}\right)$$

while the Pearson χ^2 test has

$$X^2 = \sum_j \sum_c \sum_{r=1}^{R} \sum_{s=1}^{R} \frac{(o_{jcrs} - e_{jcrs})^2}{e_{jcrs}}$$

where

$$o_{jcrs} = \sum \mathbb{1}[X_i(t_j) = s, X_i(t_{j-1}) = r] \tag{2.1}$$

$$e_{jcrs} = p_{rs}(t_{j-1}, t_j; \hat{\theta}) n_{jcr} \tag{2.2}$$

where n_{jcr} is the number of individuals with covariate value c observed in state r at time t_{j-1}, who have an observation at time t_j.

Each statistic has an asymptotic null distribution which is χ^2 with degrees of freedom given by $C - |\theta|$, where C is the number of independent cells from the resultant contingency table and $|\theta|$ is the number of unknown parameters fitted from the data.

Such tests cannot be applied when subjects are observed at irregularly spaced intervals or when the number of unique covariate values is high. Aguirre-Hernandez and Farewell (2002) (AH/F) proposed a Pearson-type goodness-of-fit test which accommodates irregular observation times and continuous covariates. Observations are categorized by observation number into observation categories, h, and, within each observation category, by time interval category, l_h. Additionally, observations are categorized by covariate category, c, according to quantiles of the estimated transition intensity q_{rs}. Then, for each transition type, $r \to s$ for a patient with observations at times t_j, $j = 1, \ldots, n$, we calculate:

$$o_{hl_hrsc} = \sum \mathbb{1}[(X(t_{j+1}) = s, X(t_j) = r)] \tag{2.3}$$

$$e_{hl_hrsc} = \sum \mathbb{P}(X(t_{j+1}) = s | X(t_j) = r) \mathbb{1}[X(t_{j-1}) = r] \tag{2.4}$$

where the summation is over the set of observations:

$$\forall \text{patients}, j : t_{j+1} - t_j \in l_h, q(\mathbf{z}) \in c \tag{2.5}$$

where $\mathbf{z}$ is the vector of covariates for a patient.

The statistic is then given by

$$T = \sum_h \sum_{l_h} \sum_c \sum_r \sum_s \frac{(o_{hl_hrsc} - e_{hl_hrsc})^2}{e_{hl_hrsc}}. \tag{2.6}$$

Titman and Sharples (2008) (T/S) proposed an extension to this test to accommodate misclassification hidden Markov models. The same grouping method as the AH/F test applies. However, the observed and expected counts are based on the observed states, but the observed state does not have the Markov property. Therefore, it is necessary to consider all observations up to the current time in order to estimate the next observed state. o and e are then given by

$$o_{hl_hrsc} = \sum \mathbb{1}[(O(t_{j+1}) = s, O(t_j) = r)] \tag{2.7}$$

$$e_{hl_hrsc} = \sum \mathbb{P}(O(t_{j+1}) = s | O(t_j) = r, O(t_1), \ldots, O(t_{j-1})) \mathbb{1}[O(t_j) = r]. \tag{2.8}$$

In addition, Titman and Sharples showed that a modified test was required when the data consisted of a mixture of discrete observations and exact death times. However, in this paper we only consider the case of panel observation of all transitions.

3 The null distribution of the AH/F statistic

Both the AH/F test and the T/S test have null distributions which can be naively approximated by χ^2 with degrees of freedom χ^2 with degrees of freedom $C - |\theta|$ as defined above. However, even asymptotically, this approximation is usually a lower bound. In general the statistic will have an inflated mean compared to the naive degrees of freedom. It will tend to have a lower variance compared to the mean than a χ^2 distribution. The null distribution is also dependent on the choice of categorization for observations, even if the number of independent cells C remains the same.

The exact asymptotic null distribution is intractable. This is because it depends both on the unknown true value of the parameter vector θ and the observation times. Aguirre-Hernandez and Farewell proposed a parametric bootstrap to obtain an approximation of the null distribution. Obtaining one bootstrap sample involves generating realizations of the Markov process using the maximum likelihood estimate for the original data at the sampling times from the original data, refitting the Markov model to the new data and applying the goodness-of-fit test to the data. Such a method does give a test of the required size. However, Markov models and particularly hidden Markov models can be time consuming to fit, especially if there are a large number of observations or many unknown parameters. Ideally, at least 1000 bootstrap samples are required to get a reasonable approximation of the 95% or 99% point of the null distribution. Bootstrapping becomes undesirable if the model takes more than a few seconds to fit.

The non-χ^2 null distribution is due to the lack of efficiently of the maximum likelihood estimate to minimize the statistics, and the cell counts being the sum of non-identical multinomials rather than multinomial.

It is well known been that a chi-squared test based on grouped data will result in a distribution which lies between $\chi^2_{d-|\theta|}$ and χ^2_d where d is the number of independent cells in the contingency table and $|\theta|$ the number of unknown parameters fitted from the data (Chernoff (1954); Kendall and Stuart (1961)).

By considering the joint distribution of the observed counts in the contingency table and the score function of the log-likelihood, better asymptotic approximations to the null distributions of the AH/F and T/S statistics may be derived.

4 Derivation of the approximation

Suppose we have panel observed data from a Markov or hidden Markov process. For notational convenience, we suppose each observation is arbitrarily categorized into category $c = 1, \ldots, C$, where c encompasses observation quantile, time quantile, covariate and categorization by previous observed state.

Each observation from a panel observed Markov model can be considered as a multinomial random variable. An observation, $x_{c,i}$, $i = 1, \ldots, n_c$ within category c is non-identical multinomial, such that

$$x_{c,i} \sim \text{Multinomial}(1, (p_1(z_{c,i}, \theta), \ldots, p_R(z_{c,i}, \theta))) \tag{4.1}$$

where $z_{c,i}$ is the covariate vector corresponding to that observation (which we allow to include both the last observed state and the time between observations) and θ the parameter vector of length M. Similarly, for

hidden Markov models, since the likelihood for an individual can be factorized as

$$L_i = P(O(t_1))P(O(t_2)|O(t_1))\dots P(O(t_n)|O(t_1),\dots,O(t_{n-1}))$$

the observed states can be again considered Multinomial as in (4.1), provided $z_{c,i}$ also includes all the observed states up to that time.

We can therefore write either the AH/F or T/S statistic as

$$T(\mathbf{x},\hat{\theta}) = \sum_{c=1}^{C}\sum_{r}^{R}\frac{(o_{rc}-e_{rc}(\hat{\theta}))^2}{e_{rc}(\hat{\theta})} \tag{4.2}$$

where $o_{r,c} = \sum_i \mathbb{1}\{x_{c,i} = \delta_r\}$, $e_{r,c}(\hat{\theta}) = \sum_i p_r(z_{c,i},\hat{\theta})$ and δ_r is a vector of length R with rth entry 1 and all other entries zero.

For a standard chi-squared test on multinomial data, $\mathbf{O}$, the vector of length RC with entries o_{rc}, is a sufficient statistic for θ and so $\hat{\theta}$ is a deterministic function of $\mathbf{O}$. However, this isn't the case when $\mathbf{O}$ are not multinomial. The statistic is instead a function of both $\mathbf{O}$ and $\hat{\theta}$.

To proceed we first define $\upsilon(\theta)$ to be a vector of length RC with entries

$$\upsilon_{rc}(\theta) = \frac{(o_{rc}-e_{rc}(\theta))}{e_{rc}(\theta)^{\frac{1}{2}}}.$$

As

$$T(\mathbf{x},\hat{\theta}) = \upsilon(\hat{\theta})^{\mathbf{T}}\upsilon(\hat{\theta}),$$

the aim is then to determine the distribution of $\upsilon(\hat{\theta})$.

We firstly assume that the maximum likelihood estimate gives a consistent estimate such that $\hat{\theta} \xrightarrow{p} \theta$, and that we may therefore Taylor expand $\upsilon(\hat{\theta})$ about θ. This gives

$$\upsilon(\hat{\theta}) = \upsilon(\theta) + B(\hat{\theta}-\theta) + o_p(1), \tag{4.3}$$

where B is a $RC \times M$ matrix with entries $\frac{\partial e_{rc}(\theta)}{\partial \theta_m}\frac{1}{e_{rc}(\theta)^{\frac{1}{2}}}$. Also, from standard asymptotic results, $(\hat{\theta}-\theta) = (\mathbb{E}I(\theta))^{-1}U(\theta) + o_p(1)$, where $U(\theta) = \frac{\partial l(\theta)}{\partial \theta}$ is the score function and $\mathbb{E}I(\theta) = -\mathbb{E}(\frac{\partial^2 l(\theta)}{\partial \theta^T \partial \theta})$. Hence, asymptotically, $\upsilon(\hat{\theta})$ can be viewed as a linear function of $\omega = (U(\theta), v(\theta))$, a vector of length $M + RC$. Hence

$$\upsilon(\hat{\theta}) \xrightarrow{d} A\omega, \tag{4.4}$$

where

$$A = \begin{bmatrix} B\mathbb{E}(I(\theta))^{-1} & I \end{bmatrix}$$

and I is a $RC \times RC$ identity matrix.

We therefore seek the covariance matrix of ω.

Firstly note that

$$Var(U(\theta)) = \mathbb{E}I(\theta). \tag{4.5}$$

The vector $\mathbf{O}$ is made up of blocks of length R which are the sum of independent, non-identical multinomials of size 1. $\mathbf{O}$ therefore has expectation $\mathbf{e}(\theta)$ and covariance matrix which is block diagonal,

$$\Sigma = \begin{bmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_C \end{bmatrix}$$

where

$$(\Sigma_c)_{rs} = \begin{cases} \sum_i^{n_c} p_r(z_{c,i},\theta)(1-p_r(z_{c,i},\theta)) & r = s \\ \sum_i^{n_c} p_r(z_{c,i},\theta)p_s(z_{c,i},\theta) & r \neq s \end{cases}$$

for $c = 1,\ldots,C$. $\upsilon(\theta)$ therefore has mean zero and covariance matrix given by $P\Sigma P^T$, where P is a $RC \times RC$ diagonal matrix with elements $(e_{rc}(\theta))^{-\frac{1}{2}}$. Thus $Cov(\upsilon(\theta)) = P\Sigma P^T$.

We then need to find $Cov(U(\theta),\upsilon(\theta))$. Since both $U(\theta)$ and $\mathbf{O}$ are functions of the full data $\mathbf{x}$, we can condition on $\mathbf{x}$.

$$Cov(U(\theta),\mathbf{O}) = Cov[\mathbb{E}(U(\theta)|\mathbf{x}),\mathbb{E}(\mathbf{O}|\mathbf{x})] + \mathbb{E}[Cov(U(\theta),\mathbf{O}|\mathbf{x})] \tag{4.6}$$

The second term of the right hand side of equation 4.6 is zero because given $\mathbf{x}$, $U(\theta)$ and $\mathbf{O}$ are fully determined, moreover

$$Cov(U(\theta),\mathbf{O}) = Cov(U(\theta)|\mathbf{x},\mathbf{O}|\mathbf{x}).$$

Note that the mth component of $U(\theta)$ can be written as

$$U_m(\theta) = \sum_c \sum_{i=1}^{n_c} (\frac{\partial p(z_{c,i},\theta)}{\partial \theta_m}\frac{1}{p(z_{c,i},\theta)})^T x_{c,i},$$

that the (r,c) entry of $\mathbf{O}$ is just $\sum_i^{n_c} x_{(c,i)(r)}$ where $x_{(c,i)(r)}$ denotes the rth entry of the vector $x_{c,i}$, and that each of the individual observations $x_{c,i}$ are independent. Then

$$Cov(U_m(\theta),o_{rc}) = \sum_i^{n_c} Cov((\frac{\partial p(z_{c,i},\theta)}{\partial \theta_m}\frac{1}{p(z_{c,i},\theta)})^T x_{c,i},\delta_r^T x_{c,i})$$

where δ_r is a vector of length R with rth entry 1 and all other entries zero. Further

$$\begin{aligned} Cov(U_m(\theta),o_{rc}) = \sum_i^{n_c} \mathbb{E}\left[\left(\left(\frac{\partial p(z_{c,i},\theta)}{\partial \theta_m}\frac{1}{p(z_{c,i},\theta)}\right)^T x_{c,i}\right)\left(\delta_r^T x_{c,i}\right)\right] \\ -\mathbb{E}\left[(\frac{\partial p(z_{c,i},\theta)}{\partial \theta_m}\frac{1}{p(z_{c,i},\theta)})^T x_{c,i}\right]\mathbb{E}\left[\delta_r^T x_{c,i}\right]. \end{aligned} \tag{4.7}$$

The ith observation only contributes a non-zero value for both U_m and o_{rc} if the observation is in cell r with probability $p_r(z_{c,i},\theta)$. Moreover, the expected contribution to U_m from observation i is 0. Hence equation 4.7 reduces to

$$Cov(U_m(\theta),o_{rc}) = \sum_{i_c} \frac{\partial p_r(z_{c,i},\theta)}{\partial \theta_m}.$$

If we denote $\Psi = Cov(U(\theta),\mathbf{O})$, then it follows that $Cov(U(\theta),\upsilon(\theta)) = \Psi P$ and moreover that ω has covariance

$$\Omega = \begin{bmatrix} \mathbb{E}(I(\theta)) & P^T\Psi^T \\ \Psi P & P\Sigma P^T \end{bmatrix}.$$

Combining this result with (4.4), it follows that $\upsilon(\hat{\theta})$ has an asymptotic multivariate normal distribution with mean vector 0 and covariance matrix $V = A\Omega A^T$. $T(\mathbf{x}, \hat{\theta})$ can therefore be expressed as a scalar product of such a multivariate normal.

Equivalently,

$$T \xrightarrow{d} \sum_{i=1}^{M} \lambda_i X_i^2$$

where $\lambda_1, \ldots, \lambda_M$ are the eigenvalues of covariance matrix V and $X_1^2, \ldots, X_M^2$ are i.i.d. χ_1^2.

The null distribution of T therefore tends towards a distribution with characteristic function given by

$$\psi(u) = \prod_{j=1}^{R} (1 - 2\mathrm{i}\lambda_j u)^{-0.5}.$$

The p-value of a particular point can then be found numerically by applying the Gil-Pelaez formula (Gil-Pelaez (1951)). This gives

$$\mathbb{P}(T > x) = \frac{1}{2} + \int_{-\infty}^{\infty} Im(\frac{\psi(u)\exp(-\mathrm{i}ux)}{2\pi u})du.$$

Of course, $\lambda_1, \ldots, \lambda_M$, and therefore $\mathbb{P}(T > x)$, depend on the unknown true value of θ. In the same way that the parametric bootstrap approach takes $\theta = \hat{\theta}$ when simulating new datasets, in our approximation we also take $\theta = \hat{\theta}$.

In order to compute $\mathbb{P}(T > x)$ for a given θ, we need $\frac{\partial p_{rc}(z_{c,i},\theta)}{\partial \theta_m}$ and the expected Fisher information matrix $\mathbb{E}(I(\theta))$. Details of how to calculate these quantities, particularly for hidden Markov models, are given in Appendix A.1.

5 Simulation study

To validate the method on data with realistic sample sizes and observation schemes, we consider data on post-heart transplantation patients. 596 patients received heart transplants between 1979 and 2000. Patients were followed up until March 2005. Periodically patients had angiograms to assess the presence of cardiac allograpt vascopathy. Patients may be classified into one of three states. State 1, referring to absence of CAV, state 2, meaning mild CAV and state 3 referring to severe CAV. This dataset is available in the R package msm (Jackson (2008)) where data on exact death times also available. In this paper our aim is only to get a realistic observation scheme for a panel observed dataset on which to base simulations, so we remove the mortality data.

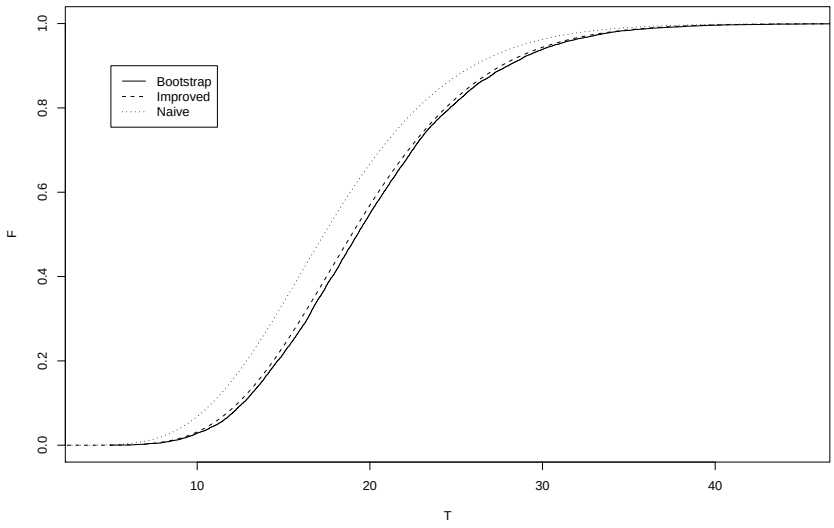
It is assumed that observation of disease free status or mild CAV is subject to classification error but that state 3 is an absorbing state. A three-state misclassification HMM was fitted to the data via maximum likelihood, giving estimate $\tilde{\theta}$. Donor age and whether the patient had ischaemic heart disease before transplantation are both significant covariates affecting CAV onset rates ($1 \rightarrow 2$ transition intensity). Donor age varied from 6 to 61 years. Ischaemic heart disease was present in 50% of subjects.

New data were simulated from a process with $\theta = \tilde{\theta}$ using the same sampling times as in the original dataset. This involves sampling both a trajectory of the Markov model and subjecting the true states to misclassification error.

To calculate the goodness-of-fit statistic we choose to stratify by time quantiles of the times between observations and also by covariate values. To avoid sparse cell counts, we limit the number of covariate groups to two. This is achieved by choosing covariate groupings based on whether $q_{12}(\mathbf{z_i}, \hat{\theta})$ for observation i, lies above or below the median for the whole sample. This means that the cell boundaries are themselves a function of the data. This contradicts an assumption made in the derivation of section 4, that the groups into which each observation is placed is fixed in advance. Such random cell boundaries are known to have some effect on the theoretical distribution of χ^2 tests (Moore (1971)). The naive χ^2 approximation has 18 degrees of freedom in this case.

10000 datasets were generated and the goodness-of-fit statistic was computed for each. Additionally, using the methods of section 4, the p-value for the particular dataset was calculated. The 95% point for the original dataset was calculated as 30.46 using the asymptotic approximation. After obtaining 10000 bootstrap samples, the 95% point was found to be at 30.78 and 5.42% of samples exceeded 30.46. In contrast the naive χ^2_{18} approximation gives the 95% point at 28.87 and has size 8.12%. Figure 1 shows a comparison of the three estimates of the null distribution.

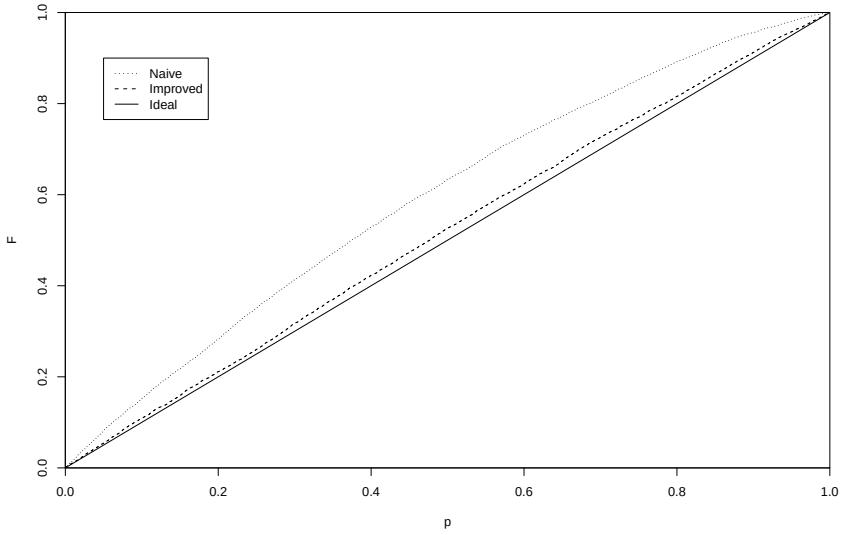
Figure 1: Comparison of empirical bootstrap null distribution with naive χ^2_{18} approximation and improved approximation.



For each bootstrap sample we also calculated the p-value of the goodness-of-fit test, based on the refitted maximum likelihood estimate and the simulated dataset, using the improved approximation. This gave similar results, with 5.44% of samples having a p-value below 0.05. Figure 2 shows the distribution of upper

p-values for the naive and improved approximations.

Figure 2: Distribution of p-values for the naive χ^2_{18} and improved approximations.



6 Discussion

An asymptotic approximation to the null distribution of the AH/F and T/S statistics has been derived. The method allows a more accurate p-value to be calculated in cases where the naive $\chi^2_{C-|\theta|}$ approximation is poor and bootstrapping is prohibitively time consuming. This significantly improves the practical applicability of the AH/F and T/S tests of goodness-of-fit.

The simulated example showed that the approximation is quite accurate for relative small sample sizes and cell frequencies. For instance, some cell categories only had 8 observations in them. The effect of small cell counts seems to be that the approximation is slightly anti-conservative. We would expect the degree of inaccuracy to increase for datasets that give tests with very sparse contingency tables. In these cases a parametric bootstrap may still be required. It is not clear to what extent cell boundaries that depend on the data have on the accuracy of the approximation, but the effect seemed minimal in the simulated example.

A goodness-of-fit suite has recently been incorporated into the R package msm (Jackson (2008)), including the AH/F and T/S goodness-of-fit tests. Code to compute more accurate p-values for models fitted using msm, based on the results of this paper, has been developed and is available from the author.

A Appendix

A.1 Calculation of first derivatives and expected Fisher information

For Markov models, $p_{rs}(z_{c,i},\theta)$ just refers to a particular entry, e.g. (r,s) of some transition probability matrix $P(t_0,t_1;\theta)$. Kalbfleisch and Lawless (1985) showed, for time homogeneous Markov models, where $P(t_0,t_1)=P(t)$,

$$\frac{\partial P(t)}{\partial \theta_m} = UV_mU^{-1}$$

where V_m is a $M\times M$ matrix with (i,j) entry

$$\begin{cases} \frac{g_{ij}^{(m)}(\exp(d_it)-\exp(d_jt))}{(d_i-d_j)} & i\neq j \\ g_{ii}^{(m)} t\exp(d_it) & i=j \end{cases}$$

where $g_{ij}^{(m)}$ is the (i,j) entry of the matrix $G_m = U^{-1}(\frac{\partial Q}{\partial\theta_m})U$ and $d_1,\ldots,d_R$ are the eigenvalues of the intensity matrix $Q(\theta)$, which are assumed to be distinct. Derivatives can also be calculated in the presence of a singular matrix U, but this is unlikely to arise at the maximum likelihood estimate in practical applications.

For hidden Markov models, $p_{rs}(z_{c,i},\theta)=\mathbb{P}(o_k=s|o_1,\ldots,o_{k-1})$. If we define a vector ξ, with rth entry $\xi_r = P(x_{k-1}=r|o_1,\ldots,o_{k-1})$, then

$$\mathbb{P}(o_k=l|o_1,\ldots,o_{k-1}) = \sum_{j=1}^{R}\sum_{s=1}^{R}\xi_j p_{js}(t_{k-1},t_k)\pi_{sl}, \tag{A.1}$$

where t_{k-1} and t_k are the $(k-1)$th and kth observation times for the subject, $p_{js}(t_{k-1},t_k)$ is the (j,s) entry of transition probability matrix $P(t_{k-1},t_k;\theta)$ and π_{sl} is the (s,l) entry in the matrix of classification probabilities. The derivative of A.1 can therefore be found using the product rule.

Calculation of $\frac{\partial\xi_j}{\partial\theta_m}$ can be done iteratively. Lystig and Hughes (2002) showed that the Forward algorithm for computing the likelihood of a HMM can be generalized to compute the first and second derivatives. Let

$$\alpha_k(i) = \mathbb{P}(X(t_k)=i,o_1,\ldots,o_k)$$

and

$$\phi_k(\theta_m,i) = \frac{\partial}{\partial\theta_m}\mathbb{P}(X(t_k)=i,o_1,\ldots,o_k).$$

Then

$$\alpha_k(j) = \sum_{j=1}^{R}\alpha_{k-1}(i)p_{ij}(t_k-t_{k-1})\pi_{jo_k}$$

and

$$\phi_k(\theta_m,i) = \sum_{i=1}^{R}p_{ij}(t_k-t_{k-1})\left(\phi_{k-1}(\theta_m,i)\pi_{jo_k}+\alpha_{k-1}(i)\frac{\partial\pi_{j,o_k}}{\partial\theta_m}\right) + \alpha_{k-1}(i)\pi_{jo_k}\frac{\partial p_{ij}(t_k-t_{k-1})}{\partial\theta_m}$$

Since $\xi_i = \frac{\alpha_k(i)}{\sum_{j=1}^{R}\alpha_k(i)}$,

$$\frac{\partial\xi_i}{\partial\theta_m} = \frac{(\sum_{j=1}^{R}\alpha_k(j))\phi_k(\theta_u,i)-\alpha_k(i)\sum_{j=1}^{T}\phi_k(\theta_m,j)}{(\sum_{j=1}^{R}\alpha_k(j))},$$

and can therefore be computed for each k by iteratively computing α and ϕ.

Finally, the expected Fisher information for both Markov and hidden Markov models is given by

$$\mathbb{E}(-\frac{\partial^2 l}{\partial\theta_u\partial\theta_v}) = \sum_{i,c}\sum_{l=1}^{R}\frac{1}{p_l(z_{c,i},\theta)}\frac{\partial p_l(z_{c,i},\theta)}{\partial\theta_u}\frac{\partial p_l(z_{c,i},\theta)}{\partial\theta_v}$$

so only requires calculation of first derivatives.

References

Aguirre-Hernandez R, Farewell V. T. A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. *Statistics in Medicine* 2002; **21**:1899-1911.

Bureau A, Shiboski S, Hughes J. P. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine* 2003; **22**: 441-462.

Chernoff H, Lehmann E.L. The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit. *The Annals of Mathematical Statistics* 1954; **25**:576-586.

Cox D.R, Miller H.D. *The theory of stochastic processes.* Chapman and Hall, London, 1965.

Fisher R. A. The conditions under which χ^2 measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society* 1924; **87**: 442-450.

Gil-Pelaez J. Note on the inversion theorem. *Biometrika* 1951; **38**: 481-482.

Jackson C.H. *msm: Multi-state Markov and hidden Markov models in continuous time. R package version 0.8.1* 2008.

Jackson C.H, Sharples L.D. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine* 2002; **21**: 113-128

Kalbfleisch J.D, Lawless J.F. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 1985; **80**:863-871.

Kendall M.G, Stuart A. *The advanced theory of statistics. Vol.2* London, 1961.

Lystig T.C, Hughes J.P. Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics* 2002; **11**: 678-689.

Moore D.S. A chi-square statistic with random cell boundaries. *The Annals of Mathematical Statistics* 1971; **42**: 147-156.

Satten G.A, Longini I.M. Markov chains with measurement error: estimating the 'true' course of a marker of the progression of Human Immunodeficiency Virus disease. *Journal of the Royal Statistical Society Series C* 1996, **45**: 265-309.

Stavola B. L de. Testing departures from time homogeneity in multistate Markov processes. *Journal of the Royal Statistical Society. Series C* 1988; **37**:242-250.

Titman A. C, Sharples L. D. A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine* 2008; **27**: 2177-2195.