Journal of Statistical Software

January 2011, Volume 38, Issue 8. http://www.jstatsoft.org/

Multi-State Models for Panel Data: The msm Package for R

Christopher H. Jackson
Medical Research Council Biostatistics Unit

Abstract

Panel data are observations of a continuous-time process at arbitrary times, for example, visits to a hospital to diagnose disease status. Multi-state models for such data are generally based on the Markov assumption. This article reviews the range of Markov models and their extensions which can be fitted to panel-observed data, and their implementation in the **msm** package for R. Transition intensities may vary between individuals, or with piecewise-constant time-dependent covariates, giving an inhomogeneous Markov model. Hidden Markov models can be used for multi-state processes which are misclassified or observed only through a noisy marker. The package is intended to be straightforward to use, flexible and comprehensively documented. Worked examples are given of the use of **msm** to model chronic disease progression and screening. Assessment of model fit, and potential future developments of the software, are also discussed.

Keywords: multi-state models, Markov models, panel data, R, **msm**.

1. Markov multi-state models for panel data

1.1. Definitions

A multi-state model describes how an individual moves between a series of states in continuous time. Suppose an individual is in state $S(t)$ at time t. The movement on the discrete state space $1, \ldots, R$ is governed by *transition intensities* $q_{rs}(t, z(t)) : r, s = 1, \ldots, R$. These may depend on time t, or, more generally, also on a set of individual-level or time-dependent explanatory variables $z(t)$. The intensity represents the instantaneous risk of moving from state r to state $s \neq r$:

$$q_{rs}(t, z(t)) = \lim_{\delta t \to 0} \mathsf{P}(S(t + \delta t) = s | S(t) = r)/\delta t.$$

The q_{rs} form a $R \times R$ matrix Q whose rows sum to zero, so that the diagonal entries are defined by $q_{rr} = -\sum_{s \neq r} q_{rs}$. An example is the general model for disease progression (Figure 1), in which individuals can advance or recover between adjacent disease states, or die from any state.

1.2. Panel data

The other articles in this issue focus on fitting multi-state models of this type to *continuously-observed* processes, where the state $S_i(t)$ of each individual $i = 1, \ldots, M$ is known at *all times* t in the study period. Survival analysis is the simplest such example, a two-state model where individuals remain alive until an observed or censored time of death.

This article focuses on multi-state models for *panel* data, in which the state $S_i(t)$ is only known at a finite series of times $t = (t_{i1}, \ldots, t_{in_i})$. Fitting multi-state models to panel data generally relies on the *Markov* assumption, that future evolution only depends on the current state. That is, $q_{rs}(t, z(t), \mathcal{F}_t)$ is independent of $\mathcal{F}_t$, the observation history $\mathcal{F}_t$ of the process up to the time preceding t. See, for example, Cox and Miller (1965) for a thorough review of continuous-time Markov chain theory. In a *time-homogeneous* Markov model, in which the q_{rs} are also independent of t, the sojourn time in each state r is exponentially-distributed with mean $-1/q_{rr}$. The probability that an individual in state r moves next to state s is $-q_{rs}/q_{rr}$.

1.3. The msm package

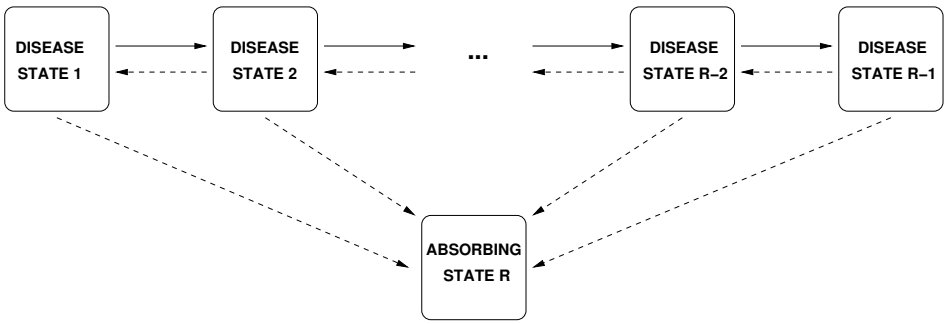
This article describes the **msm** package for R (R Development Core Team 2010), available from `http://CRAN.R-project.org/package=msm`. **msm** can be used to fit a Markov model with any number of states and any pattern of transitions to panel data, and includes several extensions such as hidden Markov models and models whose transition intensities vary with individual-specific or time-varying covariates. **msm** was motivated by studies of chronic diseases in medicine, and is frequently used in this area (Jackson *et al.* 2003; Sharples *et al.* 2003; Gani *et al.* 2007; Sweeting *et al.* 2006; Buter *et al.* 2008; Skogvoll *et al.* 2008), but it has been widely used in other fields, for example geology (Aspinall *et al.* 2006), zoology (Gautrais *et al.* 2007) and econometrics (Rummel 2009).

1.4. Likelihood for panel data

The Markov model for panel data was first described by Kalbfleisch and Lawless (1985) and Kay (1986). The likelihood for this basic model, used in **msm**, is calculated from the *transition probability matrix* $P(u, t+u)$. The (r, s) entry of $P(u, t+u)$, $p_{rs}(u, t+u)$, is the probability of being in state s at a time $t+u$, given the state at time u is r. $P(u, t+u)$ is calculated in terms of Q using the Kolmogorov differential equations (Cox and Miller 1965). If the transition intensity matrix Q is constant over the interval $(u, t+u)$, as in a time-homogeneous process, then $P(u, t+u) = P(t)$ and the equations are solved by the matrix exponential of Q scaled by the time interval,

$$P(t) = \text{Exp}(tQ).$$

The matrix exponential Exp() is notoriously difficult to calculate reliably, as discussed by Moler and van Loan (2003). It is defined by the same "power series" $\text{Exp}(X) = 1 + X^2/2! +$



$$Q = \begin{pmatrix} q_{11} & q_{12} & 0 & 0 & \dots & q_{1R} \\ q_{21} & q_{22} & q_{23} & 0 & \dots & q_{2R} \\ 0 & q_{32} & q_{33} & q_{34} & \ddots & q_{3R} \\ 0 & 0 & q_{43} & q_{44} & \ddots & q_{4R} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

Figure 1: General model for disease progression. Individuals advance between adjacent stages of disease severity, and optionally recover to an adjacent less severe state or die from any state.

$X^3/3! + ...$ as the scalar exponential, except that each term X^k in the series is defined by matrix products, not element-wise scalar multiplication. For simpler models, an analytic expression for each element of $P(t)$ can be calculated in terms of entries of Q by hand or by using symbolic algebra software. Otherwise, **msm** uses eigensystem decomposition, or, if there are repeated eigenvalues, the method of Padé approximants (Moler and van Loan 2003).

The full likelihood is then the product of probabilities of transition between observed states, over all individuals i and observation times j:

$$L(Q) = \prod_i L_i = \prod_{i,j} L_{i,j} = \prod_{i,j} p_{S(t_{ij})S(t_{i,j+1})}(t_{i,j+1} - t_{ij}). \quad (1)$$

Each component $L_{i,j}$ is the entry of the transition matrix $P(t)$ at the $S(t_{ij})$th row and $S(t_{i,j+1})$th column, evaluated at $t = t_{i,j+1} - t_{ij}$. The likelihood $L(Q)$ is maximized in terms of $\log(q_{rs})$ to compute the estimates of q_{rs}, using standard optimization algorithms, as implemented in the `optim` function in R. Standard errors are computed from the Hessian at the optimum. Some of these optimization algorithms make use of the derivatives of the likelihood, which were given by Kalbfleisch and Lawless (1985).

The likelihood (1) for this and all models in **msm** assumes that the sampling times are ignorable. That is, the fact that a particular observation is made at a certain time does not implicitly give information about the value of that observation. Sampling times are ignorable if they are fixed in advance, or otherwise chosen independently of the outcome of the process. Grüger *et al.* (1991) also showed that the sampling times are ignorable under a *"doctor's care"* sampling scheme, where the next observation time (such as a visit to a doctor) is chosen on the basis of the current state. Basing the *current* observation time on the current state would

be a non-ignorable sampling scheme. To avoid bias, non-ignorable sampling times should be modelled as part of the likelihood (Sweeting *et al.* 2010).

1.5. Likelihood under alternative observation schemes

Exact death times

In observational studies of chronic diseases, it is common that the time of death is known, but the state immediately before death is unknown. If $S(t_{i,j+1}) = D$ is such a death state, then the contribution to the likelihood at this time is summed over the unknown state m at the instant before death:

$$L_{i,j} = \sum_{m \neq D} p_{S(t_{ij}),m}(t_{i,j+1} - t_{ij}) q_{m,D}$$

Continuously-observed processes

msm allows Markov models to be fitted to processes which are continuously-observed. However, the assumption of exponential sojourn times inherent in Markov models is restrictive, and more flexible models can be fitted to such data with other software. For example, proportional hazards models with non-parametric baseline intensities can be fitted using the **mstate** package (de Wreede *et al.* 2011, 2010)

Generally, **msm** allows a dataset to be an arbitrary mixture of observations such that states are panel-observed, continuously-observed, or "exact death times".

2. Using msm for a basic Markov model

The package is illustrated with a set of data from monitoring heart transplant recipients, which is provided with **msm**. Sharples *et al.* (2003) studied the progression of coronary allograft vasculopathy (CAV), a post-transplant deterioration of the arterial walls, using these data. The dataset can be made available to the current R session using the command `data("cav")`. 30 observations from 8 individuals with missing primary diagnosis (reason for transplantation, variable `pdiag`) are dropped from the data, giving a dataset with 2816 state observations from 614 individuals.

```
R> library("msm")
R> data("cav")
R> cav <- cav[!is.na(cav$pdiag),]
```

2.1. Format of data

Approximately each year after transplant, each patient has an angiogram, at which CAV can be diagnosed. The result of the test is in the variable `state`, with possible values:

- 1, representing no CAV.

- 2, representing mild/moderate CAV.
- 3, representing severe CAV.
- 4, recorded at the date of death.

`years` gives the time of the test in years since the heart transplant.

Data are supplied to **msm** as a series of observations, grouped by patient. This should be a data frame with variables indicating the observed state of the process (`state` in the CAV data) and the time of the observation (`years` in the CAV data) If the data come from more than one individual, then a subject identification variable (`PTNUM` in the CAV data) must also be supplied. This does not need to be numeric, but observations from the same subject must be adjacent in the dataset, and observations must be ordered by time within subjects. The first eleven rows of the data `cav` give the observation series from the first two patients. Other variables are either individual-specific or time-dependent covariates (see Section 3).

```
R> cav[1:11,]
```

```
    PTNUM      age    years dage sex pdiag cumrej state firstobs
1  100002 52.49589 0.000000   21   0   IHD      0     1        1
2  100002 53.49863 1.002740   21   0   IHD      2     1        0
3  100002 54.49863 2.002740   21   0   IHD      2     2        0
4  100002 55.58904 3.093151   21   0   IHD      2     2        0
5  100002 56.49589 4.000000   21   0   IHD      3     2        0
6  100002 57.49315 4.997260   21   0   IHD      3     3        0
7  100002 58.35068 5.854795   21   0   IHD      3     4        0
8  100003 29.50685 0.000000   17   0   IHD      0     1        1
9  100003 30.69589 1.189041   17   0   IHD      1     1        0
10 100003 31.51507 2.008219   17   0   IHD      1     3        0
11 100003 32.49863 2.991781   17   0   IHD      2     4        0
```

Multi-state data can be summarized by counting, for each r and s, the number of times an observation of state r was followed by state s. This is implemented in the function `statetable.msm()`. In this example, an observation of severe CAV (state 3) was followed by a less severe state (states 1–2) on only 17 occasions.

```
R> statetable.msm(state, PTNUM, data = cav)
```

```
    to
from    1    2    3    4
   1 1348  203   44  147
   2   46  134   54   47
   3    4   13  107   55
```

2.2. Specifying the Markov model and initial values

We assume that the patient can advance or recover from consecutive states while alive, and

die from any state, as in Figure 1 with $R = 4$ states, giving a transition intensity matrix of

$$Q = \begin{pmatrix} -(q_{12} + q_{14}) & q_{12} & 0 & q_{14} \\ q_{21} & -(q_{21} + q_{23} + q_{24}) & q_{23} & q_{24} \\ 0 & q_{32} & -(q_{32} + q_{34}) & q_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

As Section 4 will explain, this model is not strictly medically realistic, but we fit it here for illustration. Note that this matrix represents transitions *in an instant* rather than the transitions *over an interval* summarized by `statetable.msm` – the 4 individuals who moved from state 3 to state 1 in successive observations are assumed to have travelled via state 2, therefore $q_{31} = 0$ but $q_{32}, q_{21} \neq 0$.

To tell **msm** what the allowed transitions of our model are, we define a matrix `twoway4.q` of the same size as Q, containing zeroes in the off-diagonal positions where the entries of Q are zero. All other off-diagonal positions contain an *initial value* for the corresponding transition intensity. Any diagonal entries q_{rr} supplied are ignored, as these are constrained to be minus the sum of all the other entries in the row. The rows and columns of `twoway4.q` are given informative names which will be used when presenting the estimates.

```
R> twoway4.q <- rbind(c(0, 0.25, 0, 0.25), c(0.166, 0, 0.166, 0.166),
+    c(0, 0.25, 0, 0.25), c(0, 0, 0, 0))
R> rownames(twoway4.q) <- colnames(twoway4.q) <- c("Well", "Mild",
+    "Severe", "Death")
```

In this example, the initial values represent a guess that the mean period in each state before moving to the next is about 2 years ($q_{rr} = -0.5$) and there is an equal probability of progression, recovery or death ($q_{rr} = -\sum_{s \neq r} q_{rs}$). Alternatively, by supplying the option `gen.inits=TRUE` to `msm()`, the initial values for non-zero entries of Q can be set to the maximum likelihood estimates under the assumption that transitions take place only at the observation times.

2.3. Running msm and interpreting results

The maximum likelihood estimate of Q is computed by the `msm()` function, as below, starting from the supplied initial values. The argument `death=4` indicates that entry times into state 4 are observed exactly but the state on the instant before is unknown (Section 1.5). The optimization in this example takes about 20 seconds on a typical current computer. Printing the object `cav.msm` returned by `msm()` displays the estimated transition intensity matrix with 95% confidence intervals. We see patients are about three times as likely to develop CAV than die without CAV (first row). After onset of mild disease, progression to severe CAV is about 50% more likely than recovery, and death from the severe disease state is rapid (mean of 1 / 0.41 = 2.4 years in state 3).

```
R> cav.msm <- msm(state ~ years, subject = PTNUM, data = cav,
+    qmatrix = twoway4.q, death = 4)
R> cav.msm
```

```
Call:
msm(formula = state ~ years, subject = PTNUM, data = cav,
    qmatrix = twoway4.q, death = 4)

Maximum likelihood estimates:
Transition intensity matrix

       Well                       Mild
Well   -0.1682 (-0.188,-0.1505) 0.1276 (0.111,0.1467)
Mild   0.2264 (0.1692,0.303)    -0.618 (-0.7195,-0.5309)
Severe 0                        0.1226 (0.07308,0.2056)
Death  0                        0
       Severe                    Death
Well   0                         0.04057 (0.03227,0.051)
Mild   0.3375 (0.2713,0.4199)    0.05405 (0.02233,0.1308)
Severe -0.4144 (-0.5245,-0.3275) 0.2919 (0.2274,0.3746)
Death  0                         0

-2 * log-likelihood:  3945.363
```

To display the fitted transition probability matrix $P(t)$ over an interval of $t = 1$ year, the function `pmatrix.msm()` is used. This suggests a 9%, 15% and 4% probability that in one year's time, an individual currently free of CAV will have mild CAV, severe CAV or be dead, respectively. The option `ci="normal"` computes a confidence interval for $P(t)$ by repeated sampling from the asymptotic normal distribution of the maximum likelihood estimates of the $\log(q_{rs})$. The output below is based on the default 1000 samples, and has converged to within 2 significant figures. Alternatively, intervals can be computed using nonparametric bootstrap resampling (`ci="boot"`). The dataset of $\sum_{i=1}^{M} n_i$ serially-correlated state observations from M individuals is rearranged as a dataset of $\sum_{i=1}^{M}(n_i - 1)$ independent transitions between pairs of states. Bootstrap datasets of transitions are drawn with replacement and the model refitted repeatedly to estimate the sampling uncertainty surrounding the estimates. This method is more accurate but much slower due to the need to refit the model for each resample.

```
R> pmatrix.msm(cav.msm, t = 1, ci = "normal")
```

```
       Well                          Mild
Well   0.8558 (0.8421,0.8691)        0.08785 (0.07671,0.09852)
Mild   0.1559 (0.1194,0.2027)        0.5602 (0.5035,0.6012)
Severe 0.009393 (0.005273,0.01624) 0.07416 (0.04487,0.1198)
Death  0                             0
       Severe                       Death
Well   0.01458 (0.01148,0.01824) 0.04175 (0.03482,0.05131)
Mild   0.2042 (0.1678,0.2445)    0.07974 (0.06067,0.1267)
Severe 0.6736 (0.6035,0.7275)    0.2429 (0.197,0.2952)
Death  0                         1 (1,1)
```

2.4. Controlling numerical optimization

The optimization may occasionally converge to a local rather than a global maximum of the likelihood surface. Therefore to ensure that the global maximum has been found, it is recommended to run `msm()` with diverse sets of initial values. However, if values too far from the optimum are chosen then the algorithm may not converge. To improve convergence, the optimization in `msm()` can be fine-tuned using all the options available to the R function `optim()`. For example, the number of iterations can be increased with `maxit`, and the log-likelihood can be rescaled during optimization (`fnscale`).

But if over-complex models are applied with insufficient data, then the parameters of the model will not be identifiable. The `fixedpars` option to `msm()` is useful for profiling likelihoods. This allows any parameters to be fixed at their initial values. The model must of course be realistic. In Markov models for panel data, it is not usually feasible to estimate a model where *instantaneous* transitions are allowed between every pair of states. For example, in chronic disease applications, transitions are generally only plausible between "adjacent" states of a disease – a patient who is observed as "well" at t_j, and "severe" at t_{j+1} must have gone through "mild" in the interval (t_j, t_{j+1}).

3. Markov models with covariates

3.1. Individual-level covariates

The effect of a vector of explanatory variables $\mathbf{z}_{ij}$ on the transition intensity for individual i at time j is modelled using proportional intensities, replacing q_{rs} with

$$q_{rs}(z_{ij}) = q_{rs}^{(0)} \exp(\boldsymbol{\beta}_{rs}^{\top} \mathbf{z}_{ij}).$$

The likelihood is then maximized over the $q_{rs}^{(0)}$ and $\boldsymbol{\beta}_{rs}$.

In the CAV example, the age of the heart transplant donor (variable `dage`) and the primary diagnosis, or reason for transplantation (variable `pdiag`), are suggested to affect the rate of onset and progression of CAV. We fit a model in which the intensities are different according to donor age and a primary diagnosis of ischaemic heart disease (IHD), after creating a binary variable `ihd` representing IHD from the categorical `pdiag`. A "formula" in standard R linear modelling syntax, `~ dage + ihd`, is supplied as the `covariates` argument to `msm()`. To facilitate convergence, the "BFGS" quasi-Newton optimization algorithm is used (see the documentation for the R function `optim()`), and the maximum number of iterations is increased to 10000. The -2× log-likelihood is also divided by 4000, since it takes values around 4000 for plausible ranges of the parameters. This ensures that optimization takes place on an approximate unit scale, to avoid numerical overflow or underflow.

```
R> ihd <- as.numeric(cav[, "pdiag"] == "IHD")
R> cav.cov.msm <- msm(state ~ years, subject = PTNUM, data = cav,
+    covariates = ~ dage + ihd, qmatrix = twoway4.q, death = 4,
+    method = "BFGS", control = list(fnscale = 4000, maxit = 10000))
```

Instead of printing the fitted model object `cav.cov.msm`, which shows the relatively uninformative baseline intensities $q_{rs}^{(0)}$ and log hazard ratios $\boldsymbol{\beta}_{rs}$, we use the function `hazard.msm()`

to display hazard ratios $\exp(\boldsymbol{\beta}_{rs})$ for each covariate on each transition with 95% confidence intervals. A primary diagnosis of IHD is associated with a 56% increase in the hazard of CAV onset, and 1 year of donor age is associated with a 2% greater risk of CAV onset and a 4% greater risk of death without CAV.

```
R> hazard.msm(cav.cov.msm)
```

```
$dage
                     HR         L        U
Well - Mild    1.0192556 1.0068692 1.031794
Well - Death   1.0381769 1.0180960 1.058654
Mild - Well    0.9981484 0.9725701 1.024399
Mild - Severe  0.9856091 0.9674640 1.004095
Mild - Death   0.9320659 0.8448829 1.028245
Severe - Mild  0.9976255 0.9476498 1.050237
Severe - Death 0.9884293 0.9648293 1.012607

$ihd
                     HR         L         U
Well - Mild    1.5647641 1.1793343  2.076160
Well - Death   1.3044011 0.8207672  2.073014
Mild - Well    0.9372774 0.5193818  1.691413
Mild - Severe  0.9578794 0.6126934  1.497540
Mild - Death   1.7858347 0.2298841 13.873100
Severe - Mild  0.7669038 0.2706515  2.173058
Severe - Death 0.7572325 0.4562969  1.256641
```

We can also use `qmatrix.msm()` to calculate the transition intensity matrix for specified covariate values as follows, in this case, a donor age of 50 years old and a primary diagnosis of IHD. Compared with the fitted intensities for the "average" person from the model without covariates (Section 2.3), we see an approximately doubled risk of CAV onset and death without CAV. (The average donor is 30 years old and about half of heart transplants are due to IHD).

```
R> qmatrix.msm(cav.cov.msm, covariates = list(dage = 50, ihd = 1))
```

```
       Well                        Mild
Well   -0.3438 (-0.4388,-0.2693) 0.2467 (0.1825,0.3335)
Mild   0.2201 (0.1153,0.4201)    -0.4811 (-0.6876,-0.3366)
Severe 0                        0.112 (0.03684,0.3404)
Death  0                        0
       Severe                   Death
Well   0                        0.09707 (0.06499,0.145)
Mild   0.2485 (0.1587,0.3891)    0.01257 (0.0007542,0.2096)
Severe -0.3233 (-0.5476,-0.1909) 0.2113 (0.1215,0.3676)
Death  0                        0
```

3.2. Model comparison

Likelihood ratio tests between nested models fitted in **msm** can be performed conveniently using the function `lrtest.msm`. Comparing a likelihood ratio statistic of 59 to a χ^2 distribution with 14 degrees of freedom shows that the model with covariates (`cav.cov.msm`) fits significantly better than the model without covariates (`cav.msm`).

```
R> lrtest.msm(cav.msm, cav.cov.msm)

            -2 log LR df            p
cav.cov.msm  58.5785 14 2.079552e-07
```

Covariate effects may be constrained to equal between different intensities, using the `constraint` argument to `msm()`. For example, in a disease progression model, the effect of a covariate on all progression rates may be equal. `constraint` is a list of vectors, one for each covariate. In the model `cav.cov2.msm` fitted below, `dage = c(1, 2, 3, 1, 2, 4, 2)` indicates that the effect of `dage` on the 1st and 4th intensities are constrained to be equal, as is the effect on the the 2nd, 5th and 7th intensities. The parameters are assumed to be ordered by reading across the rows of the transition matrix, starting at the first row: $(q_{12}, q_{14}, q_{21}, q_{23}, q_{24}, q_{32}, q_{34})$, so that in the model `cav.cov2.msm`, the effect on the CAV onset rate q_{12} equals the effect on the CAV progression rate q_{23}, and the effects on all death rates q_{14}, q_{24}, q_{34} are constrained to be equal. However, a likelihood ratio test indicates that the bigger model `cav.cov.msm` without constraints fits significantly better.

```
R> cav.cov2.msm <- msm(state ~ years, subject = PTNUM, data = cav,
+    covariates = ~ dage + ihd, constraint = list(dage =
+    c(1, 2, 3, 1, 2, 4, 2), ihd = c(1, 2, 3, 1, 2, 4, 2)),
+    qmatrix = twoway4.q, death = 4)
R> lrtest.msm(cav.cov2.msm, cav.cov.msm)

            -2 log LR df            p
cav.cov.msm  60.10682  6 4.281631e-11
```

Some intensities may not be influenced by covariates at all. In **msm**, models in which covariates affect some intensities, but not others, can be specified by fixing certain covariate effects at their default initial values of zero, by instructing the optimizer not to optimize over those parameters using the `fixedpars` argument to `msm()`. See the package help for further details.

3.3. Time-inhomogeneous models

In general, the transition probability matrix $P(u, t+u)$, hence the likelihood for panel data, cannot be calculated in closed form if Q varies over the interval $(u, t+u)$. An exception is if Q is piecewise-constant. The effect of time-dependent variables, including time itself, on the transition intensities can be modelled in **msm** under this assumption. For example, suppose a covariate varies continuously through time, but is only observed at the same times as the state of the Markov process. The approximate effect of that covariate can be estimated assuming that it is constant in between the times that it is observed, so that $P(u, t+u) = P(t)$. More

generally, time-inhomogeneous Markov models can be constructed in which piecewise-constant covariates change at times other than $(t_{i1}, \ldots, t_{in_i})$. This is accomplished by summing the likelihood over the unknown observed state at the times when the covariates change (Equation 2, Section 3.4).

msm provides a convenient facility for constructing time-inhomogeneous models in which intensities change at the same times for every individual. A vector of change points is specified in the `pci` argument to `msm()`. The following command fits an inhomogeneous model to the CAV data in which all intensities change 5 years after transplantation. This constructs a model with a single binary covariate called `timeperiod`, a *factor* in R, with levels `(-Inf, 5]` (the baseline) representing the first time period, and `[5, Inf)`, representing the second time period. A likelihood ratio test against the time-homogeneous model suggests significant time-inhomogeneity. The estimated hazard ratios from this fitted model show an increased onset rate of mild CAV in the second period, though no significant time effect on other transitions. There is weak information about the effect of time on the death rate from mild CAV.

```
R> cav.pci.msm <- msm(state ~ years, subject = PTNUM, data = cav,
+    qmatrix = twoway4.q, death = 4, pci = 5, method = "BFGS")
R> lrtest.msm(cav.msm, cav.pci.msm)

            -2 log LR df            p
cav.pci.msm  49.24128  7 2.034911e-08

R> hazard.msm(cav.pci.msm)

$`timeperiod[5,Inf)`
                        HR         L           U
Well - Mild      2.2080439 1.6418440    2.969501
Well - Death     0.6714820 0.2472622    1.823522
Mild - Well      0.6634596 0.3581871    1.228907
Mild - Severe    0.9165669 0.5747890    1.461571
Mild - Death    12.9314664 0.1392106 1201.221729
Severe - Mild    1.4253788 0.4753785    4.273867
Severe - Death   1.6828792 0.8470715    3.343381
```

Time-dependent intensities in **msm** are restricted to piecewise-constant models. More flexible alternatives are discussed in Section 6.

3.4. Censored states

In the CAV example, some patients were known to be alive but in an unknown disease state at the end of the study. We say that the disease state is *censored*, meaning that the exact value is unknown, but known to be in a certain set. Unlike in survival analysis, here it is the state, not the event time, which is censored. If the patient were alive at the end of the study but with a known state, then the standard likelihood (1) would apply.

msm allows the state observation at any time to be censored, that is, known only to be in an arbitrary subset of the state space. Suppose the $1, 2, \ldots n_i$th observations from individual i

are known only to be in the sets $C_1, C_2, \ldots, C_{n_i}$ respectively. The likelihood for this individual is a sum of the likelihoods of all possible paths through the unobserved states.

$$L_i = \sum_{s_{n_i} \in C_{n_i}} \ldots \sum_{s_2 \in C_2} \sum_{s_1 \in C_1} p_{s_1 s_2}(t_2 - t_1) p_{s_2 s_3}(t_3 - t_2) \ldots p_{s_{n_i - 1} s_{n_i}}(t_{n_i} - t_{n_i - 1}) \tag{2}$$

This likelihood is used in **msm** to fit general time-inhomogeneous models with piecewise-constant intensities, as described in Section 3.3, where the state is not observed at times when the intensities change.

Suppose the variable `state` in the data `cav` were to contain observations coded 99 on occasions where the patient is alive but in an unknown state, which could be state 1, 2 or 3. The standard Markov model could be fitted to such data using the `censor` and `censor.states` options to `msm()`, as follows.

```
R> cav.msm <- msm(state ~ years, subject = PTNUM, data = cav,
+    qmatrix = twoway4.q, death = TRUE, censor = 99,
+    censor.states = c(1, 2, 3))
```

4. Hidden Markov models

In a *hidden Markov model* (HMM), the states of the Markov chain are not observed. The observed data y_{ij} are governed by some probability distribution conditionally on the unobserved state S_{ij}. This class of model is commonly used in areas such as speech and signal processing (Juang and Rabiner 1991) and the analysis of biological sequence data (Durbin *et al.* 1998), with a discrete-time underlying Markov chain. Applications of HMMs in medicine, where continuous-time processes are usually more suitable, include Satten and Longini (1996); Bureau *et al.* (2003); Jackson and Sharples (2002); Jackson *et al.* (2003). These models can represent chronic staged diseases which can only be diagnosed by an error-prone marker.

4.1. Likelihood

The **msm** package can fit continuous-time hidden Markov models to panel-observed data with a variety of distributions for the outcome conditionally on the hidden state. HMMs are fitted in **msm** by direct maximization of the likelihood, as in Satten and Longini (1996), though Bureau *et al.* (2000) describe an alternative EM algorithm for fitting the same class of models. The contribution of individual i to the likelihood is

$$\begin{aligned} L_i &= \mathsf{P}(y_{i1}, \ldots, y_{in_i}) \\ &= \sum \mathsf{P}(y_{i1}, \ldots, y_{in_i} | S_{i1}, \ldots, S_{in_i}) \mathsf{P}(S_{i1}, \ldots, S_{in_i}) \end{aligned} \tag{3}$$

where the sum is taken over all possible paths of underlying states $S_{i1}, \ldots, S_{in_i}$. Assume that the observed states are conditionally independent given the values of the underlying states. Also assume the Markov property, $\mathsf{P}(S_{ij} | S_{i,j-1}, \ldots, S_{i1}) = \mathsf{P}(S_{ij} | S_{i,j-1})$. Then the contribution L_i can be written as a product of matrices, as follows. To derive this matrix product, decompose the overall sum in Equation 3 into sums over each underlying state. The

sum is accumulated over the unknown first state, the unknown second state, and so on until the unknown final state:

$$L_i = \sum_{S_{i1}} \mathsf{P}(y_{i1}|S_{i1})\mathsf{P}(S_{i1}) \sum_{S_{i2}} \mathsf{P}(y_{i2}|S_{i2})\mathsf{P}(S_{i2}|S_{i1}) \ldots \sum_{S_{in_i}} \mathsf{P}(y_{in_i}|S_{in_i})\mathsf{P}(S_{in_i}|S_{in_{i-1}})$$

where $\mathsf{P}(y_{ij}|S_{ij})$ is the probability density of the outcome conditional on the hidden state (also called the "emission" distribution), and $\mathsf{P}(S_{ij}|S_{i,j-1})$ is the transition probability of the hidden Markov chain, calculated as in Section 1.4.

msm allows most common distributions to be used as HMM outcome models. The modular design of **msm** allows new outcome distributions to be added easily, as described in the package documentation. These must be univariate, and **msm** is restricted to situations where only one observation is made conditionally on an underlying Markov process.

In practice, the outcome distribution may vary between individuals and through time, as well as with the hidden state. **msm** allows one *location* parameter for each class of outcome distribution to depend on covariates, for example, a linear model for the mean of a normal outcome distribution. The transition rates of the hidden Markov chain may also vary with covariates, just as for non-hidden Markov models (Section 3.1).

The distribution $\mathsf{P}(S_{i1})$ of the initial state may be estimated from the data, or fixed at plausible values. This distribution may also depend on covariates through a multinomial logistic regression.

4.2. Application of hidden Markov models: FEV_1 after lung transplants

A dataset of repeated measurements of FEV_1, forced expiratory volume in 1 second, in recipients of lung transplants (Jackson and Sharples 2002) is provided with **msm** as `data("fev")`. FEV_1 measurements are used to diagnose bronchiolitis obliterans syndrome (BOS), a chronic deterioration in lung function. FEV_1 is measured as a percentage of a baseline value for each individual, determined in the first six months after transplant, and defined to be 100% baseline at six months. Figure 2 shows a series of FEV_1 measurements from a typical patient from this dataset. BOS is modelled as a staged disease, with stages defined by

- No BOS ($\geq$ 80% baseline FEV_1).
- Mild BOS (sustained drop below $<$ 80% baseline FEV_1).
- Moderate BOS (sustained drop below $<$ 65% baseline FEV_1).
- Severe BOS (sustained drop below $<$ 50% baseline FEV_1).
- Death.

As FEV_1 is subject to high short-term variability due to acute events and natural fluctuations, the exact state at each observation time is difficult to determine, making it difficult to model the natural history of BOS as defined. Instead, we represent the BOS progression by a hidden Markov model for FEV_1, conditionally on underlying BOS states. Discrete states are considered to be an appropriate alternative to representing the underlying disease status as continuous, as the onset of BOS is often sudden.

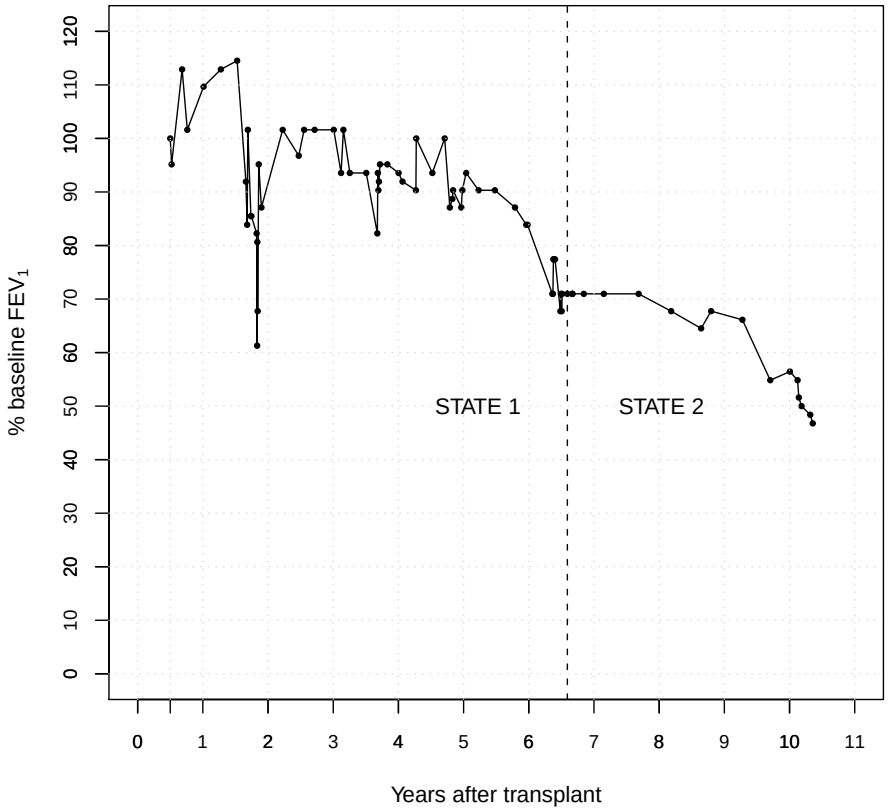


Figure 2: Measurements of lung function (FEV_1) from a lung transplant recipient and fitted BOS states from a hidden Markov model.

Here we describe a three-state "illness-death" hidden Markov model, with states representing no BOS, BOS and death, and a transition intensity matrix of

$$Q = \begin{pmatrix} -q_{12} & q_{12} & 0 \\ 0 & -q_{23} & q_{23} \\ 0 & 0 & 0 \end{pmatrix}$$

The distribution of percentage of baseline FEV_1 is Normal(μ_1, σ_1^2) in state 1 and Normal(μ_2, σ_2^2) in state 2. State 3, representing death, is observed without error and is given a label of 999 in the data. The death time is known exactly. More sophisticated four and five-state models for the FEV_1 data, using outcome distributions which separate measurement error and natural variation in the response, are described by Jackson and Sharples (2002).

4.3. Fitting hidden Markov models with msm

To fit this hidden Markov model using the `msm()` function, the argument `hmodel` is used. This is a list of objects representing the outcome distribution for each state, returned by

constructor functions. Each constructor function has arguments giving initial values for the parameters of the outcome distribution. In this example, `hmmNorm(mean = 100, sd = 16)` indicates initial values of 100 for μ_1 and 16 for σ_1. `hmmIdent(999)` represents the identity distribution, in other words, state 3 is observed without error, and is indicated by a value of 999 in the data. Initial values for the Markov transition intensities are given in an object called `three.q`, used as the `qmatrix` argument to `msm()` as before.

The FEV_1 values, conditional on the BOS state, are assumed to be affected by a time-dependent covariate indicating whether the patient suffered acute infections or rejection episodes within 14 days of the observation. To model this covariate effect we use the `hcovariates` argument to `msm()`. This takes a list of linear model formulae, which are used for the location parameter of the respective outcome distribution. In this case, the means μ_1 and μ_2 of the normal distribution have a linear model with a single binary covariate `acute`. The `hconstraint` statement (analogous to `constraint`) indicates that the effect of acute events on μ_1 and μ_2 is constrained to be the same. No covariates are assumed to affect the transition rates Q in this example, but `covariates` and `constraint` arguments could be included for this purpose just as in Section 3.1.

```
R> data("fev")
R> three.q <- rbind(c(0, exp(-6), exp(-9)), c(0, 0, exp(-6)), c(0, 0, 0))
R> hmodel1 <- list(hmmNorm(mean = 100, sd = 16), hmmNorm(mean = 54, sd = 18),
+    hmmIdent(999))
R> fev1.msm <- msm(fev ~ days, subject = ptnum, data = fev,
+    qmatrix = three.q, hmodel = hmodel1, hcovariates = list(~ acute,
+    ~ acute, NULL), hconstraint = list(acute = c(1, 1)), death = 3,
+    method = "BFGS")
R> fev1.msm
R> sojourn.msm(fev1.msm)

Call:
msm(formula = fev ~ days, subject = ptnum, data = fev, qmatrix = three.q,
    hmodel = hmodel1, hcovariates = list(~acute, ~acute, NULL),
    hconstraint = list(acute = c(1,1)), death = 3, method = "BFGS")

Maximum likelihood estimates:
Transition intensity matrix

        State 1                              State 2
State 1 -0.0007038 (-0.0008333,-0.0005945) 0.0006275 (0.0005201,0.0007572)
State 2 0                                  -0.0008011 (-0.001013,-0.0006337)
State 3 0                                  0
        State 3
State 1 7.631e-05 (3.967e-05,0.0001468)
State 2 0.0008011 (0.0006337,0.001013)
State 3 0

Hidden Markov model, 3 states
Initial state occupancy probabilities:
```

```
        Estimate LCL UCL
State 1        1  NA  NA
State 2        0  NA  NA
State 3        0  NA  NA

State 1 - normal distribution
Parameters:
       Estimate       LCL       UCL
mean  98.004361 97.34297 98.665754
sd    16.185019 15.77782 16.602730
acute -8.791807 -9.95145 -7.632163

State 2 - normal distribution
Parameters:
       Estimate       LCL       UCL
mean  51.823341 50.76293 52.883748
sd    17.676307 17.08279 18.290443
acute -8.791807 -9.95145 -7.632163

State 3 - identity distribution
Parameters:
      Estimate LCL UCL
which      999  NA  NA

-2 * log-likelihood:  51597.89

R>  sojourn.msm(fev1.msm)
        estimates       SE         L        U
State 1  1420.759 122.3921 1200.0328 1682.084
State 2  1248.389 149.3041  987.5255 1578.161
```

The estimated HMM normal outcome distributions show that in state 1, FEV_1 measurements have a mean of 98% baseline (SD 16%) and in state 2, a mean of 52% baseline (SD 18%). FEV_1 is estimated to be 9% lower within 14 days of acute illnesses. The function `sojourn.msm` presents estimates and confidence intervals for $-1/q_{rr}$, indicating the average onset and progression rates of BOS in days. BOS state 1 is estimated to begin about 3 years (estimate 1420 days) after transplantation, and state 2 a further 3 years later.

The most likely true series of states underlying the data can be estimated using the Viterbi algorithm (Viterbi 1967) through the function `viterbi.msm`. Figure 2 shows the most likely time at which the individual passed from state 1 to state 2 – the time when their decline below 80% of baseline became sustained.

4.4. Misclassification models

An important special case of HMMs is the multi-state model with misclassification, where the observed data are states, assumed to be misclassifications of the true, underlying states (Jackson *et al.* 2003). In the CAV example, it is not medically realistic for patients to recover from

a diseased state to a healthy state, as in the model of Section 2. Progression of coronary artery vasculopathy is thought to be an irreversible process. The angiography observations are actually subject to error, which leads to some false measurements of CAV states and apparent improvements in state. Thus a more realistic Markov intensity matrix Q would be as given in Figure 1, but with $q_{r+1,r} = 0$ for each r,

$$Q = \begin{pmatrix} -(q_{12} + q_{14}) & q_{12} & 0 & q_{14} \\ 0 & -(q_{23} + q_{24}) & q_{23} & q_{24} \\ 0 & 0 & -q_{34} & q_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We also assume that true state 1 (CAV-free) can be classified as state 1 or 2, state 2 (mild/moderate CAV) can be classified as state 1, 2 or 3, while state 3 (severe CAV) can be classified as state 2 or 3. Recall that state 4 represents death. Thus the matrix of misclassification probabilities is

$$E = \begin{pmatrix} 1 - e_{12} & e_{12} & 0 & 0 \\ e_{21} & 1 - e_{21} - e_{23} & e_{23} & 0 \\ 0 & e_{32} & 1 - e_{32} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

where e_{rs} is the probability of observing state s conditionally on occupying true state r.

These are hidden Markov models with a categorical outcome distribution, and as such may be fitted in **msm** using a `hmmCat()` outcome distribution for each underlying state. However **msm** provides a convenient shorthand for fitting models of this form. An `ematrix` argument to `msm()` is given a matrix of initial values for the misclassification probabilities, with zero in positions where misclassifications cannot occur. In the CAV example we initialize the four unknown misclassification parameters to 0.1, and set the initial values `oneway4.q` for Q to the approximate maximum likelihood estimates from the model without misclassification. `obstrue=firstobs` specifies that observations indicated by the binary variable `firstobs` in the data are not misclassifications, but observations of the true state. In the CAV data, these are the dates of transplantation, at which patients are known to be CAV-free, in state 1.

```
R> ematrix <- rbind(c(0, 0.1, 0, 0), c(0.1, 0, 0.1, 0),
+    c(0, 0.1, 0, 0), c(0, 0, 0, 0))
R> oneway4.q <- rbind(c(0, 0.1, 0, 0.04), c(0, 0, 0.3, 0.05),
+    c(0, 0, 0, 0.3), c(0, 0, 0, 0))
R> rownames(oneway4.q) <- colnames(oneway4.q) <- c("Well", "Mild", "Severe",
+    "Death")
R> rownames(ematrix) <- colnames(ematrix) <- c("Well", "Mild", "Severe",
+    "Death")
R> misc.msm <- msm(state ~ years, subject = PTNUM, data = cav,
+    qmatrix = oneway4.q, ematrix = ematrix, obstrue = firstobs,
+    death = TRUE, method = "BFGS")
R> misc.msm
```

```
Call:
msm(formula = state ~ years, subject = PTNUM, data = cav,
    qmatrix = oneway4.q, ematrix = ematrix, obstrue = firstobs, death = TRUE,
    method = "BFGS")

Maximum likelihood estimates:
Transition intensity matrix

       Well                       Mild
Well   -0.1317 (-0.1486,-0.1166) 0.0903 (0.07629,0.1069)
Mild   0                          -0.2917 (-0.3574,-0.238)
Severe 0                          0
Death  0                          0
       Severe                     Death
Well   0                          0.04136 (0.03318,0.05156)
Mild   0.2574 (0.1906,0.3475)     0.03429 (0.007473,0.1573)
Severe -0.3058 (-0.3878,-0.2412) 0.3058 (0.2412,0.3878)
Death  0                          0

Misclassification matrix

       Well                    Mild
Well   0.9726 (0.9539,0.9839) 0.02737 (0.01613,0.04605)
Mild   0.1751 (0.1007,0.2868) 0.7614 (0.61,0.8669)
Severe 0                      0.1143 (0.05691,0.2164)
Death  0                      0
       Severe                   Death
Well   0                        0
Mild   0.06353 (0.03667,0.1079) 0
Severe 0.8857 (0.7836,0.9431)   0
Death  0                        1 (1,1)

-2 * log-likelihood:  3910.098
```

Thus there is an estimated probability of about 0.03 that a patient truly free of CAV will be diagnosed wrongly with mild CAV, but a rather higher probability of 0.175 that underlying mild/moderate CAV will be diagnosed as CAV-free. Between the two CAV states, the mild state will be misdiagnosed as severe with a probability of 0.06, and the severe state will be misdiagnosed as mild with a probability of 0.11. The model also estimates the progression rates through underlying states. An average of 8 years (1/0.1317) is spent disease-free, an average of about 3 years is spent with mild/moderate disease, and periods of severe disease also last about 3 years on average before death.

The misclassification probabilities may also be modelled in terms of covariates, using multinomial logistic regression. This is accomplished with the `misccovariates` argument to `msm()`. For example, a disease screening test may be more sensitive for different types of individuals.

5. Model assessment

Titman and Sharples (2010a) reviewed methods for assessing the fit of Markov models to panel data. In particular, the Markov property and homogeneity of transition rates, both between individuals and through time, can be restrictive assumptions.

5.1. Diagnostic plots

One simple diagnostic compares model predictions of the entry time into a particular state with nonparametric estimates, for example Kaplan-Meier curves. If the entry time is not observed exactly, then the nonparametric estimate is an approximation. In Figure 3, the fit of four multi-state models to the exactly-observed survival times in the CAV data is assessed in this way.

```
R> par(mfrow = c(2, 2))
R> plot.survfit.msm(cav.msm, main = "cav.msm: no covariates",
+    mark.time = FALSE)
R> plot.survfit.msm(cav.cov.msm, main = "cav.cov.msm: covariates",
+    mark.time = FALSE)
R> plot.survfit.msm(cav.pci.msm, mark.time = FALSE)
R> title("cav.pci.msm: time-inhomogeneous", line = 2)
R> title("(5 year change point)", line = 1)
R> cav.pci2.msm <- msm(state ~ years, subject = PTNUM, data = cav,
+    qmatrix = twoway4.q, death = 4, pci = c(5, 10), method = "BFGS",
+    control = list(maxit = 10000))
R> plot.survfit.msm(cav.pci2.msm, mark.time = FALSE)
R> title("cav.pci2.msm: time-inhomogeneous", line = 2)
R> title("(5, 10 year change points)", line = 1)
```

Up to about 10 years, all the models predict survival reasonably accurately (within about 5%). The time-inhomogeneous model `cav.pci.msm` fits slightly better than the time-homogeneous models up to 10 years. But the first three models overestimate survival after 10 years – 106 out of 614 individuals in the data live beyond 10 years. A further time-inhomogeneous model `cav.pci2.msm` is fitted in which intensities change after 10 as well as after 5 years, which substantially improves the fit both before and after 10 years.

Another common approach to multi-state model assessment is to compare observed prevalences of states with expected prevalences under the model at a series of times. This can be done in **msm** using the functions `prevalence.msm()` and `plot.prevalence.msm()`. To compute observed prevalences precisely, all individuals should be observed at these times. If individuals are observed at different times, this relies on approximations such as assuming transitions occur only at observation times (Gentleman *et al.* 1994) or at midpoints between observation times. Figure 4 presents a plot of this type for the best-fitting model `cav.pci2.msm` for the CAV data. As time elapses, the proportions of individuals predicted to have died appear to be underestimated by the model, and the proportions alive and in states "well" and "mild" are overestimated. However, the Kaplan-Meier estimate in Figure 3 gives a more accurate estimate of the "observed" survival probability in this case. The observed prevalence of a state is simply calculated as the number of individuals known to be in that

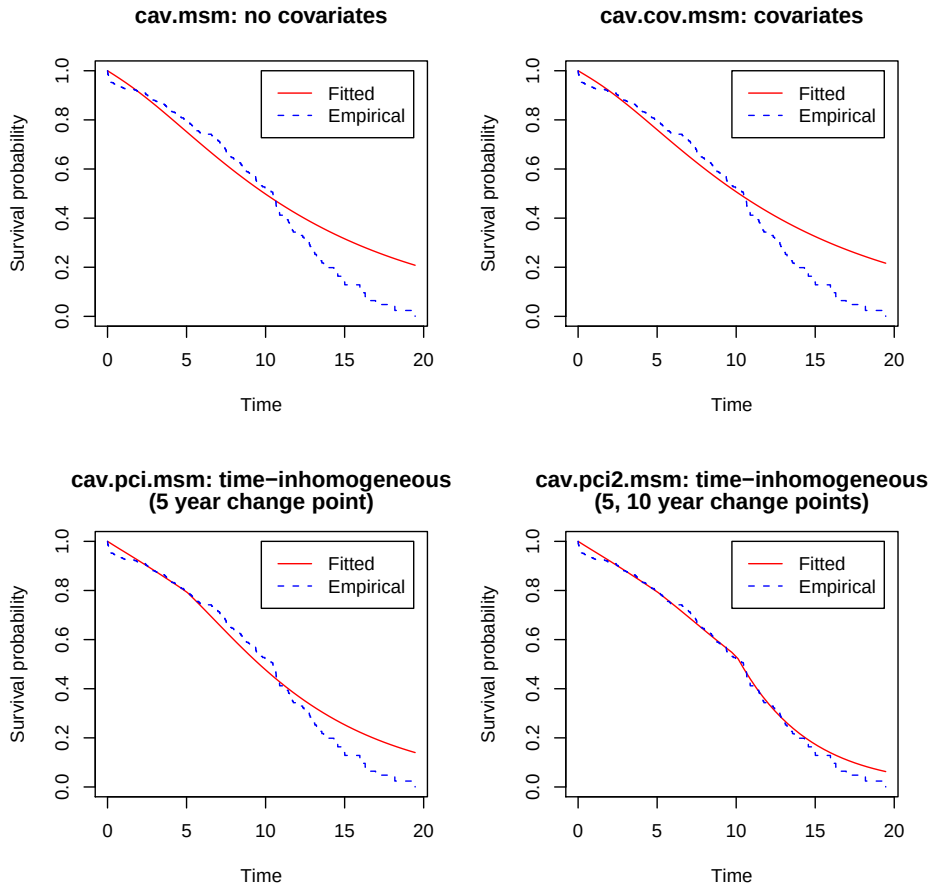


Figure 3: Comparison of observed and fitted survival for three multi-state models for the CAV data.

state, divided by the number of individuals whose state is known at that time, which ignores the information from individuals censored at earlier times.

5.2. Formal goodness-of-fit test

The previous plots are informal diagnostics to suggest potential model improvements. A formal goodness-of-fit test for the hypothesis that panel data were generated by a fitted Markov model was developed by Aguirre-Hernandez and Farewell (2002). This test was extended by Titman and Sharples (2008) to handle exactly-observed death times and misclassified states. This is implemented in **msm** as the function `pearson.msm()`. The test compares observed and expected numbers of transitions between pairs of states for a series of transition starting times, transition time intervals and covariate categories, giving a Pearson-type contingency table test statistic.

The null distribution of the statistic is not exactly χ^2, with a complex form for general panel data (Titman 2009). For simpler models without covariates, Aguirre-Hernandez and Farewell (2002) showed by simulation that the χ^2 approximation was adequate. The `pearson.msm`

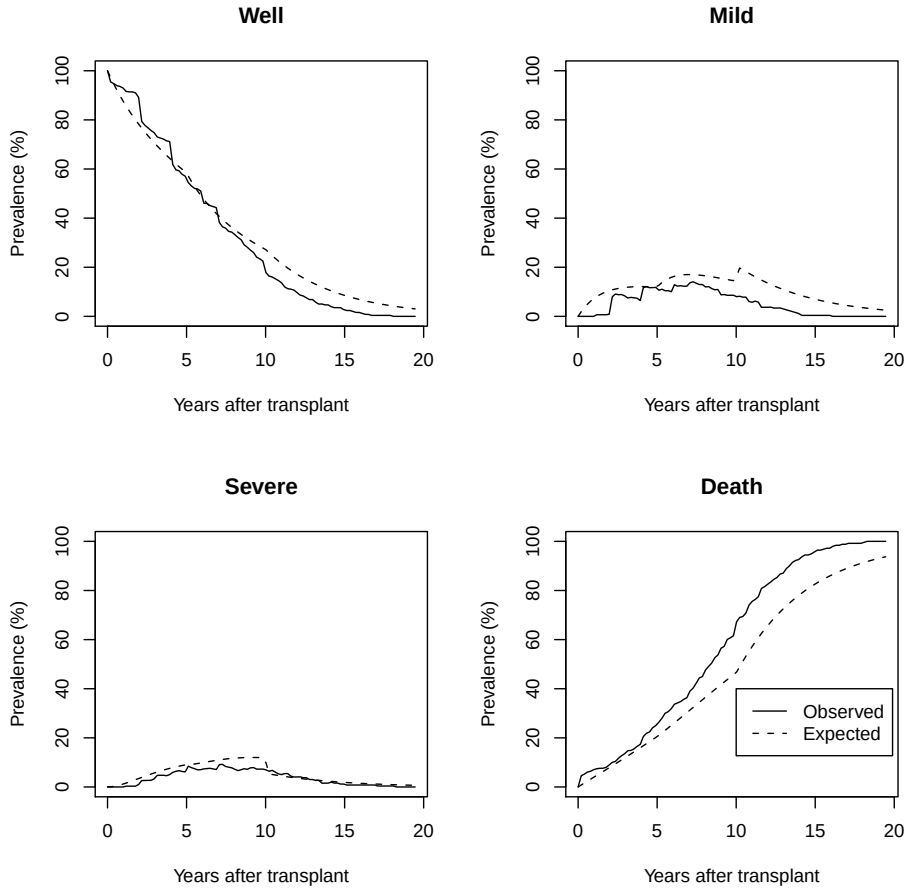


Figure 4: Comparison of observed and expected prevalence from the time-inhomogeneous model `cav.pci2.msm` for the CAV data.

function provides theoretical upper and (unless there are exact death times) lower bounds for the test p value. In general cases, the null distribution of the statistic can be estimated by the parametric bootstrap procedure of repeatedly sampling from the fitted model, refitting the model and recomputing the test statistic, resulting in an accurate p value. If the resulting contingency table is sparse, then the number of observation time, time interval or covariate categories may need to be reduced to improve the χ^2 approximation, though the power of the resulting test may be low. See the `pearson.msm` help page in the package for further details.

The Pearson-type test is performed for the four models illustrated in Figure 3. The upper p value bounds indicate that none of these models give an adequate overall fit. This suggests that even though the time-inhomogeneous model `cav.pci2.msm` fits well to survival (Figure 3), it discriminates less well between the states of CAV severity (Figure 4). A more complex pattern of time-dependence, or allowing the transition intensities to depend on covariates, would be expected to yield a better fit.

```
R>  p1 <- pearson.msm(cav.msm)
R>  p1$test
```

```
    stat df.lower p.lower df.upper      p.upper
 165.047       NA      NA       81 1.072647e-07
```

```
R>  p2 <- pearson.msm(cav.cov.msm)
R>  p2$test
```

```
     stat df.lower p.lower df.upper     p.upper
 299.9516       NA      NA      241 0.005821544
```

```
R>  p3 <- pearson.msm(cav.pci.msm)
R>  p3$test
```

```
     stat df.lower p.lower df.upper      p.upper
 136.2905       NA      NA       81 0.0001188069
```

```
R>  p4 <- pearson.msm(cav.pci2.msm)
R>  p4$test
```

```
     stat df.lower p.lower df.upper     p.upper
 125.0847       NA      NA       81 0.001216962
```

Since the method of Titman and Sharples (2008) to handle exactly-observed death times involves multiple imputation of the next scheduled observation time, these statistics and p values include some simulation error. The default 100 imputations in this example ensures the statistics have converged within 2 significant figures and the p values to within an order of magnitude.

5.3. Other issues in model assessment

The influence of each individual on the maximized likelihood can be computed and illustrated by score residuals, using the function `scoreresid.msm`. Titman and Sharples (2010a) also discussed the assessment of multi-state models with misclassification, criticising in particular the assumption of independence of the observed outcome conditionally on the underlying state.

6. Extensions of Markov models and limitations of msm

The **msm** package was designed to fit any Markov model structure to panel-observed multi-state data. Because of this aim of generality, there are limitations in handling more complex models which are only practicable for specific patterns of observations or allowed transitions.

6.1. Continuously-observed processes

For example, if the data are continuously-observed, **msm** is limited to exponential or piecewise-exponential sojourn times. More flexible models, for example, with Weibull-distributed sojourn times, are relatively easy to fit to such data. The **mstate** package (de Wreede *et al.* 2010,

2011) implements multi-state models with nonparametric baseline hazards and proportional hazards regression.

When the model is progressive, for example, a model as in Figure 1 but with all reverse transition rates $q_{r+1,r} = 0$, the number of possible pathways taken by an individual through the states is finite, so that likelihood calculations are simpler. For example, the "illness-death" model has only one disease state, and no recovery allowed from "well" to "disease". The data for such a model may only be *interval-censored*, that is, the transition to illness is known to have occurred between two observations, but at an unknown time. Flexible, non-parametric methods are possible in this case (Frydman 1995; Frydman and Szarek 2008). This is simpler than panel data, where both the type and number of transitions occurring between adjacent observations are unknown in general.

6.2. Time-inhomogeneous models

Transition intensities may vary with time, depending on either the time since the beginning of the process (a *time-inhomogeneous* model) or time since the previous transition (a *semi-Markov* model). Time-inhomogeneous models in **msm** are restricted to piecewise-constant intensities. The choice of change points is unlimited, though in practice the results may be sensitive to this choice. Continuously-changing intensities, for example with a Weibull-distributed time to the next transition, are generally more scientifically plausible and may be more parsimonious. The resulting Kolmogorov differential equations for obtaining $P(u, t+u)$, hence the likelihood for panel data, are analytically intractable, but can be solved numerically in simpler instances. For example, Chen *et al.* (2004) and Hsieh *et al.* (2002) modelled only one state with a time-varying sojourn distribution in this way. Hubbard *et al.* (2008) fitted inhomogeneous models by estimating a time transformation under which the inhomogeneous Markov model is homogeneous, assuming the ratio of transition intensities stayed constant through time.

6.3. Non-Markov models

Relaxing the Markov assumption with panel data presents more difficulties. Semi-Markov models with piecewise-constant intensities are only feasible to estimate for simpler model structures (Titman 2008). Foucher *et al.* (2010) used numerical integration to compute the likelihood for 3 or 4 state progressive semi-Markov models. Titman (2008) described an Monte Carlo EM algorithm for fitting progressive semi-Markov models to panel data. All these methods would be very difficult to implement for a general Markov model structure. In **msm**, an approximate non-Markov model might be fitted by creating artificial time-dependent covariates representing aspects of the process history, though this approach would require very frequent observations to be sufficiently accurate. A more promising approach to semi-Markov models is the *phase-type* model, in which the exponentially-distributed time spent in each state r is replaced by a series of exponential sojourns (or "phases") in hidden states $r_1, \ldots, r_k$ (Titman and Sharples 2010b). In principle, these models may be implemented as hidden Markov models in **msm**, but certain parameter constraints (currently not implemented) may be necessary for identifiability.

6.4. Random effects and Bayesian methods

Unexplained heterogeneity in transition intensities between individuals may be represented by random effects models, though these are not implemented in **msm**. Their likelihood for panel data is intractable, except for specific cases such as the "tracking" model (Satten 1999) in which the random effect acts on all intensities simultaneously, or a discrete random effects distribution (Cook *et al.* 2004).

The **msm** package is limited to maximum likelihood estimation. Multi-state models can be fitted to panel data from a Bayesian perspective using MCMC simulation (Sharples 1993), which is particularly suited to hierarchical models with random effects. Random effects Markov models with simple state structures have been implemented using the **WinBUGS** (Lunn *et al.* 2000) software for Bayesian analysis (Pan *et al.* 2007; van den Hout and Matthews 2009). Welton and Ades (2005) describe how to implement general multi-state structures using the **WBDiff** (Lunn 2004) differential equation solving interface to **WinBUGS** to calculate $P(t)$, while the **JAGS** implementation of the **BUGS** language (Plummer 2003) allows general Markov model structures to be fitted to panel data via a distribution `dmstate()`.

6.5. Discrete-time models

msm was designed for continuous-time models, but discrete-time Markov and hidden Markov models can be fitted to discrete-time data using **msm**, assuming that there is a continuous process underlying the data. The fitted transition probability matrix in one time unit, $P(1)$, is then equivalent to the transition probability matrix P of the discrete-time model. But since a discrete-time Markov model is equivalent to a series of multinomial models for each observation conditionally on the previous observation, these may be fitted more efficiently using software for multinomial logistic regression, for example, the function `multinom()` in the R package **nnet** (Venables and Ripley 2002). Currently there are several available R packages which can fit discrete-time hidden Markov models of various forms, for example **HiddenMarkov** (Harte 2010), **hsmm** (Bulla *et al.* 2008) and **mhsmm** (O'Connell and Hojsgaard 2009).

7. Further information

This article gives an overview of the **msm** package for fitting continuous-time Markov and hidden Markov models to panel data. Detailed references for all the functions for model fitting and output presentation are available as help pages in the installed package. The `doc` subdirectory of the package also contains a user guide in PDF format, which presents much of the material in this article in greater detail.

The examples in this article were run using version 1.0 of **msm**, available from `http://CRAN.R-project.org/package=msm`.

Acknowledgments

Many thanks to Linda Sharples and other colleagues at the MRC Biostatistics Unit for encouraging the development of the **msm** package, to Andrew Titman for his thorough work on model assessment and elaborated models, and to Martyn Plummer for contributing code for matrix exponentiation. Thanks also to the many users of **msm** for their comments, encour-

agement and bug reports. The author is funded by the UK Medical Research Council (grant U.1052.00.008).

References

Aguirre-Hernandez R, Farewell V (2002). "A Pearson-Type Goodness-of-Fit Test for Stationary and Time-Continuous Markov Regression Models." *Statistics in Medicine*, **21**, 1899–1911.

Aspinall W, Carniel R, Jaquet O, Woo G, Hincks T (2006). "Using Hidden Multi-State Markov Models with Multi-Parameter Volcanic Data to Provide Empirical Evidence for Alert Level Decision-Support." *Journal of Volcanology and Geothermal Research*, **153**(1-2), 112–124.

Bulla J, Bulla I, Nenadic O (2008). ***hsmm**: Hidden Semi Markov Models*. R Package version 0.3-5, URL http://CRAN.R-project.org/package=hsmm.

Bureau A, Hughes JP, Shiboski SC (2000). "An S-PLUS Implementation of Hidden Markov Models in Continuous Time." *Journal of Computational and Graphical Statistics*, **9**, 621–632.

Bureau A, Shiboski S, Hughes JP (2003). "Applications of Continuous Time Hidden Markov Models to the Study of Misclassified Disease Outcomes." *Statistics in Medicine*, **22**(3), 441–462.

Buter TC, Van den Hout A, Matthews FE, Larsen JP, Brayne C, Aarsland D (2008). "Dementia and Survival in Parkinson Disease: A 12-year Population Study." *Neurology*, **70**(13), 1017.

Chen THH, Yen M, Shiu M, Tung T, Wu H (2004). "Stochastic Model for Non-Standard Case-Cohort Design." *Statistics in Medicine*, **23**(4), 633–647.

Cook R, Yi G, Lee KA, Gladman D (2004). "A Conditional Markov Model for Clustered Progressive Multistate Processes Under Incomplete Observation." *Biometrics*, **60**(2), 436–443.

Cox DR, Miller HD (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London.

de Wreede LC, Fiocco M, Putter H (2010). "The **mstate** Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models." *Computer Methods and Programs in Biomedicine*, **99**(3), 261–274.

de Wreede LC, Fiocco M, Putter H (2011). "**mstate**: An R Package for the Analysis of Competing Risks and Multi-State Models." *Journal of Statistical Software*, **38**(7), 1–30. URL http://www.jstatsoft.org/v38/i07/.

Durbin R, Eddy S, Krogh A, Mitchison G (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Foucher Y, Giral M, Soulillou J, Daures J (2010). "A Flexible Semi-Markov Model for Interval-Censored Data and Goodness-of-Fit Testing." *Statistical Methods in Medical Research*, **19**(2), 127–145.

Frydman H (1995). "Nonparametric Estimation of a Markov 'Illness-Death' Process from Interval-Censored Observations, with Application to Diabetes Survival Data." *Biometrika*, **82**(4), 773–789.

Frydman H, Szarek M (2008). "Nonparametric Estimation in a Markov 'Illness-Death' Process from Interval Censored Observations with Missing Intermediate Transition Status." *Biometrics*, **65**(1), 143–151.

Gani R, Nixon RM, Hughes S, Jackson CH (2007). "Estimating Progression Rates in People with Highly Active Relapsing-Remitting Multiple Sclerosis." *Journal of Medical Economics*, **10**(2), 79–89.

Gautrais J, Michelena P, Sibbald A, Bon R, Deneubourg J (2007). "Allelomimetic Synchronization in Merino Sheep." *Animal Behaviour*, **74**(5), 1443–1454.

Gentleman RC, Lawless JF, Lindsey JC, Yan P (1994). "Multi-State Markov Models for Analysing Incomplete Disease History Data with Illustrations for HIV Disease." *Statistics in Medicine*, **13**(3), 805–821.

Grüger J, Kay R, Schumacher M (1991). "The Validity of Inferences Based on Incomplete Observations in Disease State Models." *Biometrics*, **47**, 595–605.

Harte D (2010). ***HiddenMarkov**: Hidden Markov Models*. R package version 1.4.2, URL http://CRAN.R-project.org/package=HiddenMarkov.

Hsieh H, Chen TH, Chang S (2002). "Assessing Chronic Disease Progression using Non-Homogeneous Exponential Regression Markov Models: An Illustration Using a Selective Breast Cancer Screening in Taiwan." *Statistics in Medicine*, **21**(22), 3369–3382.

Hubbard R, Inoue L, Fann J (2008). "Modeling Nonhomogeneous Markov Processes Via Time Transformation." *Biometrics*, **64**(3), 843–850.

Jackson CH, Sharples LD (2002). "Hidden Markov Models for the Onset and Progression of Bronchiolitis Obliterans Syndrome in Lung Transplant Recipients." *Statistics in Medicine*, **21**, 113–128.

Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E (2003). "Multistate Markov Models for Disease Progression with Classification Error." *The Statistician*, **52**(2), 1–17.

Juang BH, Rabiner LR (1991). "Hidden Markov Models for Speech Recognition." *Technometrics*, **33**, 251–272.

Kalbfleisch J, Lawless J (1985). "The Analysis of Panel Data under a Markov Assumption." *Journal of the American Statistical Association*, **80**(392), 863–871.

Kay R (1986). "A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies." *Biometrics*, **42**, 855–865.

Lunn DJ (2004). ***WinBUGS** Differential Interface (**WBDiff**)*. URL http://www.winbugs-development.org.uk/wbdiff.html.

Lunn DJ, Thomas A, Best NG, Spiegelhalter DJ (2000). "**WinBUGS** - A Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing*, **10**(4), 325–337.

Moler C, van Loan C (2003). "Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later." *SIAM Review*, **45**(1), 3–49.

O'Connell J, Hojsgaard S (2009). ***mhsmm**: Parameter Estimation and Prediction for Hidden Markov and Semi-Markov Models for Data with Multiple Observation Sequences*. R package version 0.2.3, URL http://CRAN.R-project.org/package=mhsmm.

Pan SL, Wu HM, Yen AF, Chen TH (2007). "A Markov Regression Random-Effects Model for Remission of Functional Disability in Patients Following a First Stroke: A Bayesian Approach." *Statistics in Medicine*, **26**(29), 5335–5353.

Plummer M (2003). "**JAGS**: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In K Hornik, F Leisch, A Zeileis (eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*. ISSN 1609-395X, URL http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rummel O (2009). "Distribution Dynamics and Measurement Error." In *Advances in Econometrics, Volume 24: Measurement Error: Econometrics and Practice*, volume 24. Emerald Group Publishing, Bingley, U.K.

Satten G (1999). "Estimating the Extent of Tracking in Interval-Censored Chain-of-Events Data." *Biometrics*, **55**(4), 1228–1231.

Satten GA, Longini IM (1996). "Markov Chains with Measurement Error: Estimating the 'True' Course of a Marker of the Progression of Human Immunodeficiency Virus Disease." *Applied Statistics*, **45**(3), 275–309.

Sharples LD (1993). "Use of the Gibbs Sampler to Estimate Transition Rates Between Grades of Coronary Disease Following Cardiac Transplantation." *Statistics in Medicine*, **12**, 1155–1169.

Sharples LD, Jackson CH, Parameshwar J, Wallwork J, Large SR (2003). "Diagnostic Accuracy of Coronary Angiography and Risk Factors for Post-Heart-Transplant Cardiac Allograft Vasculopathy." *Transplantation*, **76**(4), 679–682.

Skogvoll E, Eftestøl T, Gundersen K, Kvaløy J, Kramer-Johansen J, Olasveengen TM, Steen PA (2008). "Dynamics and State Transitions During Resuscitation in Out-of-Hospital Cardiac Arrests." *Resuscitation*, **78**(1), 30–37.

Sweeting M, De Angelis D, Neal K, Ramsay M, Irving W, Wright M, Brant L, Harris H (2006). "Estimated Progression Rates in Three United Kingdom Hepatitis C Cohorts Differed According to Method of Recruitment." *Journal of Clinical Epidemiology*, **59**(2), 144–152.

Sweeting MJ, Farewell V, De Angelis D (2010). "Multi-State Markov Models for Disease Progression in the Presence of Informative Examination Times: An Application to Hepatitis C." *Statistics in Medicine*, **29**(11), 1161–1174.

Titman A (2008). *Model Diagnostics in Multi-State Models of Biological Systems*. Ph.D. thesis, University of Cambridge.

Titman A (2009). "Computation of the Asymptotic Null Distribution of Goodness-of-Fit Tests for Multi-State Models." *Lifetime Data Analysis*, **15**(4), 519–533.

Titman A, Sharples LD (2008). "A General Goodness-of-Fit Test for Markov and Hidden Markov Models." *Statistics in Medicine*, **27**(12), 2177–95.

Titman A, Sharples LD (2010a). "Model diagnostics for multi-state models." *Statistical Methods in Medical Research*, **19**(6), 621–651.

Titman A, Sharples LD (2010b). "Semi-Markov Models with Phase-Type Sojourn Distributions." *Biometrics*, **66**(3), 742–752.

van den Hout A, Matthews F (2009). "Estimating Dementia-Free Life Expectancy for Parkinson's Patients Using Bayesian Inference and Microsimulation." *Biostatistics*, **10**(4), 729–743.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Viterbi J (1967). "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm." *IEEE Transactions on Information Theory*, **13**, 260–269.

Welton N, Ades A (2005). "Estimation of Markov Chain Transition Probabilities and Rates from Fully and Partially Observed Data: Uncertainty Propagation, Evidence Synthesis, and Model Calibration." *Medical Decision Making*, **25**(6), 633.

Affiliation:

Christopher Jackson
Medical Research Council Biostatistics Unit
Institute of Public Health
Forvie Site, Robinson Way
Cambridge, United Kingdom
E-mail: chris.jackson@mrc-bsu.cam.ac.uk
URL: http://www.mrc-bsu.cam.ac.uk/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 38, Issue 8 *Submitted:* 2009-07-21
January 2011 *Accepted:* 2010-08-18