# Statistical Inference and Methods

David A. Stephens

Department of Mathematics
Imperial College London

d.stephens@imperial.ac.uk
http://stats.ma.ic.ac.uk/∼das01/

31st January 2006

# Part VI

## Session 6: Filtering and Time to Event Data

## Filtering

- State-space Models
- Kalman Filter
- Non-linear/Non-Gaussian Filtering
- Particle Filters

## Time-to-Event Data

- Discrete & Continuous Time Models
- Non-parametric analysis Kaplan-Meier plots
- Proportional Hazards/Accelerated Life
- Multivariate Reliability & Competing Risks
- Multi-State Models

## State-Space Models

The state-space form of time series models has received considerable attention, as they can better represent the actual dynamics of a data generation process.

Let $y_t$ denote the observation from a (possibly multivariate) time series at time $t$, related to a vector $\alpha_t$, called the *state vector*, which is possibly unobserved and whose dimension, $m$ say, is independent of the dimension, $n$, of $y_t$.

The general form of a linear state-space model, is given by the following two equations

$$\begin{aligned} y_t &= Z_t\alpha_t + d_t + G_t\varepsilon_t, \quad t = 1, \ldots, T && \text{(1a)} \\ \alpha_{t+1} &= T_t\alpha_t + c_t + H_t\varepsilon_t. && \text{(1b)} \end{aligned}$$

Equation (1a) is the *observation* or *measurement* equation, while (1b) is *transition* equation.

In (1)

- $Z_t$ is an $(n \times m)$ matrix,
- $d_t$ is an $(n \times 1)$ vector,
- $G_t$ an $(n \times (n+m))$ matrix,
- $T_t$ is $(m \times m)$, $c_t$ is $(m \times 1)$ and
- $H_t$ is an $(m \times (n+m))$ matrix.

All of the latter matrices are referred to as the system matrices and are assumed to be non-stochastic.

Process $\varepsilon_t$ is an $((n+m)\times 1)$ vector of serially independent, identically distributed disturbances with $E(\varepsilon_t) = \mathbf{0}$ and $Var(\varepsilon_t) = \mathbf{I}$, the identity matrix.

The formulation is completed by the assumption that the initial state vector $\alpha_1$ is independent of $\varepsilon_t$ at all times $t$ and has unconditional mean and variance $a_{1|0}$ and $P_{1|0}$ respectively.

The state-space model is characterized by the properties that the system matrices, the disturbance terms $\varepsilon_t$ and initial state vector $\alpha_1$ possess. If the system matrices do not evolve with time, the state-space model is called *time-invariant* or *time-homogeneous*.

If the disturbances, $\varepsilon_t$, and initial state vector, $\alpha_1$, are assumed to have a normal distribution, then the model is termed Gaussian.

Finally, it should be noted that if $G_t H_t' = \mathbf{0}$ for all $t$, then the measurement and transition equations are uncorrelated.

Example

**AR(2) Model** Let $\{X_t\}$ be a zero-mean, Gaussian AR(2) process given by

$$X_{t+1} = \phi_1 X_t + \phi_2 X_{t-1} + u_{2t}$$

where $u_{2t} \sim NID\left(0, \sigma_2^2\right)$. Suppose also that we observe the process with noise, so that

$$Y_t = X_t + u_{1t}$$

with $u_{1t} \sim NID\left(0, \sigma_1^2\right)$ and independent of $u_{2t}$.

## Example

A state-space representation of this model can be formed as follows

$$
\begin{aligned}
Y_t &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \\
\begin{bmatrix} X_{t+1} \\ X_t \end{bmatrix} &= \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} 0 & \sigma_2^2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}
\end{aligned}
$$

where $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})'$ is $NID(\mathbf{0}, \mathbf{I})$. Clearly the latter state-space model is time-homogeneous and Gaussian and the measurement and transition equations are uncorrelated.

## Kalman Filter

The Kalman filter is used for prediction, filtering and smoothing. In particular, if we let $\Psi_t$ denote the information set up to time $t$, i.e. $\Psi_t = \{y_1, \ldots, y_t\}$, then the problem of prediction is to compute $E(\alpha_{t+1}|\Psi_t)$.

*Filtering* is concerned with calculating $E(\alpha_t|\Psi_t)$, while (fixed-interval) *smoothing* is concerned with estimating $E(\alpha_t|\Psi_T)$, for all $t < T$.

The Kalman filter consists of a set of simple recursions that compute the latter quantities.

**Linear Gaussian State-Space Model :** Assume that

$$G_t H_t' = \mathbf{0}$$

Moreover, we drop the (exogenous) terms $d_t$ and $c_t$ from the observation and transition equations for simplicity and write

$$G_t G_t' = \Sigma_t$$

and

$$H_t H_t' = \Omega_t.$$

The Kalman filter recursively computes the quantities

- $a_{t|t} = E(\alpha_t|\Psi_t)$,
- $a_{t+1|t} = E(\alpha_{t+1}|\Psi_t)$,
- $P_{t|t} = MSE(\alpha_t|\Psi_{t-1})$,
- $P_{t+1|t} = MSE(\alpha_{t+1}|\Psi_t)$.

where *MSE* is the *mean-square error* or *one-step ahead prediction variance*.

Then, starting with $a_{1|0}$ and $P_{1|0}$, $a_{t|t}$, $a_{t+1|t}$ and their MSEs are obtained by running for $t = 1, \ldots, T$, the recursions

$$
\begin{aligned}
v_t &= y_t - Z_t a_{t|t-1} \quad , \quad F_t = Z_t P_{t|t-1} Z_t' + \Sigma_t && \text{(2a)} \\
a_{t|t} &= a_{t|t-1} + P_{t|t-1} Z_t' F_t^{-1} v_t, && \text{(2b)} \\
P_{t|t} &= P_{t|t-1} - P_{t|t-1} Z_t' F_t^{-1} Z_t P_{t|t-1}, && \text{(2c)} \\
a_{t+1|t} &= T_t a_{t|t}, && \text{(2d)} \\
P_{t+1|t} &= T_t P_{t|t} T_t' + \Omega_t. && \text{(2e)}
\end{aligned}
$$

**Notes:**

- quantities $v_t$ and $F_t$ in (2a) respectively denote the one-step-ahead error in forecasting $y_t$ conditional on the information set at time $t-1$ and its MSE. This facilitates computation of the likelihood function.
- the quantities $a_{t|t}$ and $a_{t|t-1}$ are optimal estimators of $\alpha_t$ conditional on the available information, in terms of minimum mean square.
- the latter property holds under Gaussianity; if this is not assumed, $a_{t|t}$ and $a_{t|t-1}$ are optimal only within the class of linear estimators, that is, they are the minimum mean square linear estimators of $\alpha_t$ conditional on $\Psi_t$ and $\Psi_{t-1}$.

- Finally, equations (2b) and (2d ) can be combined together to yield recursions that only compute the one-step-ahead prediction estimates of $\alpha_{t+1}$ given $\Psi_t$.

  Similarly, by taking together (2c) and (2e), a single set of recursions for the MSE is obtained, which goes directly from $P_{t|t-1}$ to $P_{t+1|t}$.

- The resulting recursions, along with (2a), for $t = 1, \ldots, T-1$, are as follows:

$$
\begin{aligned}
a_{t+1|t} &= T_t a_{t|t-1} + K_t v_t && \text{(3a)} \\
K_t &= T_t P_{t|t-1} Z_t' F_t^{-1} && \text{(3b)} \\
P_{t+1|t} &= T_t P_{t+1|t} L_t' + \Omega_t && \text{(3c)} \\
L_t &= T_t - K_t Z_t. && \text{(3d)}
\end{aligned}
$$

**Parameter Estimation** Another application of the Kalman filter is the estimation of any unknown parameters $\theta$ that appear in the system matrices.

The likelihood for data $\mathbf{y} = (y_1, \ldots, y_T)$ can be constructed as

$$p(y_1, \ldots, y_T) = p(y_T|\Psi_{T-1}) \cdots p(y_2|\Psi_1)\, p(y_1) = \prod_{t=1}^{T} p(y_t|\Psi_{t-1}).$$

Assuming that the state-space model is Gaussian, by taking conditional expectations on both sides of the observation equation, with $d_t \equiv 0$ , we deduce that for $t = 1, \ldots, T$,

- $E\left(y_t|\Psi_{t-1}\right) = Z_t a_{t|t-1}$,
- $Var\left(y_t|\Psi_{t-1}\right) = F_t$.

Crucially, the one-step-ahead prediction density $p\left(y_t|\Psi_{t-1}\right)$ is the density of a multivariate normal random variable with mean $Z_t a_{t|t-1}$ and covariance matrix $F_t$.

Thus, the log-likelihood function is given by

$$\log L = -\frac{nT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(\det F_t) - \frac{1}{2}\sum_{t=1}^{T} v_t' F_t^{-1} v_t.$$

where

$$v_t = \left(y_t - Z_t a_{t|t-1}\right)$$

Typically, numerical procedures are used in order to maximize the log-likelihood to obtain the ML estimates of the parameters $\theta$ which are consistent and asymptotically normal.

**Non-Linear Non-Gaussian State-Space Models :** If the state-space model is not Gaussian, the likelihood can still be constructed in the same way using the minimum mean square linear estimators of the state vector.

However, the estimators $\widehat{\theta}$ that maximize the likelihood are the *quasi-maximum likelihood* (QML) estimators of the parameters. It can be shown that the QML estimators are also consistent and asymptotically normal.

Another algorithm can be applied to a state-space model; given a fixed set of data, estimates of the state vector are computed at each time $t$ in the sample period taking into account the full information set available.

This algorithm computes $a_{t|T} = E(\alpha_t|\Psi_T)$ along with its MSE, $P_{t|T}$ , for all $t = 1, \ldots, T-1$.

The latter quantities are computed via a set of backward recursions and use the quantities $a_{t+1|t}$, $a_{t|t}$ and the MSE matrices $P_{t|t}$, $P_{t+1|t}$, which are obtained from (2). In particular, to obtain $a_{t|T}$ and $P_{t|T}$, we start with $a_{T|T}$ and $P_{T|T}$ and we run backwards for $t = T-1, \ldots, 0$

$$\begin{aligned} a_{t|T} &= a_{t|t} + P_t^* \left( a_{t+1|T} - a_{t+1|t} \right) \\ P_{t|T} &= P_{t|t} + P_t^* \left( P_{t+1|T} - P_{t+1|t} \right) P_t^* \quad , \quad P_t^* = P_t T_t' P_{t+1|t}. \end{aligned}$$

## Markov chain Monte Carlo Applications

During the last decade, the extensive use of the MCMC, in particular the Gibbs sampler, has given rise to another smoothing algorithm called the *simulation smoother* and is also closely related to the Kalman filter.

In contrast, to the fixed interval smoother, which computes the conditional mean and variance of the state vector at each time $t$ in the sample, a simulation smoother is used for drawing samples from the density $p(\alpha_0, \ldots, \alpha_T | Y_T)$.

The first simulation smoother is based on the identity

$$p\left(\alpha_0, \ldots, \alpha_T | \Psi_T\right) = p\left(\alpha_T | \Psi_T\right) \prod_{t=0}^{T-1} p\left(\alpha_t | \Psi_t, \alpha_{t+1}\right)$$

and a draw from $p\left(\alpha_0, \ldots, \alpha_T | \Psi_T\right)$ is recursively constructed in terms of $\alpha_t$.

Starting with a draw $\widehat{\alpha}_T \sim N\left(\alpha_{T|T}, P_{T|T}\right)$, the main idea is that for a Gaussian state space model $p\left(\alpha_t|\Psi_t, \alpha_{t+1}\right)$ is a multivariate normal density and hence it is completely characterized by its first and second moments. In order to compute these moments, the usual Kalman filter recursions (2) are run, so that $\alpha_{t|t}$ is initially obtained. Then, the draw $\widehat{\alpha}_{t+1} \sim p\left(\alpha_{t+1}|\Psi_t, \alpha_{t+2}\right)$ is treated as $m$ additional observations and a second set of $m$ Kalman filter recursions is run for each element of the state vector $\widehat{\alpha}_{t+1}$.

However, the latter procedure involves the inversion of system matrices, which are not necessarily non-singular. To overcome this problem another *simulation smoother* algorithm has been devised which is computationally more efficient.

The main idea of the algorithm is to obtain a sample from the joint density of the transition equation disturbances, i.e. $p\left(H_0\varepsilon_0, \ldots, H_T\varepsilon_T|\Psi_T\right)$, by sampling

$$\widehat{\eta}_t \sim p\left(H_t\varepsilon_t|\Psi_T, H_{t+1}\varepsilon_{t+1}, \ldots, H_T\varepsilon_T\right).$$

This allows the state vector to be reconstructed, using the transition equation (1b) and substituting the simulated disturbances.

In particular, the new simulation smoother requires only the prediction equations (3) to be initially run and the quantities $v_t$, $F_t$ and $L_t$ to be stored. Then, starting with $r_T = 0$ and $U_T = 0$, the following backward recursions are run for $t = T, T-1, \ldots, 1$.

$$\begin{aligned} C_t &= \Omega_t - \Omega_t U_t \Omega_t \quad , \quad \kappa_t \sim N(0, C_t), \\ r_{t-1} &= Z_t' F_t^{-1} v_t + L_t' \left( r_t - U_t \Omega_t C_t^{-1} \kappa_t \right), \\ U_{t-1} &= Z_t' F_t^{-1} Z_t + L_t' \left( U_t + U_t \Omega_t C_t^{-1} \Omega_t U_t \right) L_t. \end{aligned}$$

Then, for $t = 1, \ldots, T$, $\eta_t = \Omega_t r_t + \kappa_t$ is a draw from

$$p\left(H_t \varepsilon_t | \Psi_T, H_{t+1}\varepsilon_{t+1}, \ldots, H_T \varepsilon_T\right).$$

For the initial disturbance term, $\eta_0 = P_{1|0} r_0 + \kappa_0$, where $\kappa_0 \sim N\left(0, P_{1|0} - P_{1|0} U_0 P_{1|0}\right)$, is a draw from $p\left(H_0 \varepsilon_0 | \Psi_T, H_1 \varepsilon_1, \ldots, H_T \varepsilon_T\right)$.

The state vector is simulated by starting with $\widehat{\alpha}_0 = 0$ and running for $t = 0, \ldots, T$ the transition equation recursion with $H_t \varepsilon_t$ substituted by $\widehat{\eta}_t$, i.e. $\widehat{\alpha}_{t+1} = T_t \widehat{\alpha}_t + \widehat{\eta}_t$.

## Inference in General State-Space Models

To recap, inference goals for state-space models are

- *filtering/tracking*: estimate state at time $t$ from all observations up to time $t$

$$p(\alpha_t|\Psi_t)$$

- *smoothing*: estimate state at time $t$ from all past and possibly some future observations

$$p(\alpha_t|\Psi_{t+s})$$

- *prediction*: estimate state at time $t$ from observations up to a previous time point $t - s$

$$p(\alpha_t|\Psi_{t-s})$$

The general (first-order Markov) state equation takes the form

$$\alpha_t = f(\alpha_{t-1}, \theta_{t-1}) + \eta_{t-1}$$

and the general observation equation takes the form

$$y_t = h(\alpha_t, \theta_t) + \epsilon_t.$$

For simplicity, we assume that the error processes $\{\eta_t\}$ and $\{\epsilon_t\}$ are independent.

The fundamental inference mechanism is Bayesian; we compute the posterior quantities of interest sequentially in the following recursive calculation:

- Prediction

$$p(\alpha_t|\Psi_{t-1}) = \int p(\alpha_t|\alpha_{t-1})p(\alpha_{t-1}|\Psi_{t-1})\, d\alpha_{t-1}$$

- Updating

$$p(\alpha_t|\Psi_t) = \frac{p(y_t|\alpha_t)p(\alpha_t|\Psi_{t-1})}{p(y_t|\Psi_{t-1})}$$

The above equations only provide an analytical solution if all densities in the state and observation equation are Gaussian, and both the state and the observation equation are linear.

- If these conditions are met, the Kalman filter provides the optimal Bayesian solution to the tracking problem.
- If these conditions are not met, we require approximations:
  - *Extended Kalman filter* (EKF): requires Gaussian densities, approximates non-linearities (using a first-order Taylor expansion)
  - *Particle filter* (PF): approximates non-Gaussian densities and non-linear equations

## Particle Filter

The particle filter uses *Monte Carlo methods*, in particular *Importance sampling*, to construct the approximations.

We approximate the integral

$$E_f[g(X)] = \int g(x)f(x)\,dx$$

by writing

$$E_f[g(X)] \equiv E_{f_0}\left[\frac{f(X)g(X)}{f_0(X)}\right] = \int \frac{g(x)f(x)}{f_0(x)} f_0(x)\,dx$$

for some importance sampling density $f_0$.

We then approximate the integral by sampling $x_1, \ldots, x_N$ from $f_0$ for large $N$, then constructing the approximation

$$\widehat{E}_{f_N} = \frac{1}{N} \sum_{i=1}^{N} \frac{g(x_i) f(x_i)}{f_0(x_i)} = \frac{1}{N} \sum_{i=1}^{N} w_i g(x_i),$$

say, where, if $f_0$ is chosen carefully

$$\widehat{E}_{f_N} \longrightarrow E_f[g(X)]$$

as $N \longrightarrow \infty$

If the function $f$ is not known exactly, but only up to proportionality, the same procedure can work if the weights $w_i$ are suitably normalized to sum to one, that is, if above

$$w_i = \frac{f(x_i)}{f_0(x_i)}$$

is replaced by

$$w_i^\star = \frac{w_i}{W} \qquad W = \sum_{i=1}^{n} w_i.$$

as $W$ approximates the integral

$$\int f(x)\, dx = \int \frac{f(x)}{f_0(x)} f_0(x)\, dx$$

Given a non-linear and non-Gaussian state space model, the goal is to compute the posterior density $p(\alpha_t|\Psi_t)$ at each time $t$ exactly.

At each time point $t$, perform the following *Sequential Importance Sampling* (SIS) procedure:

- Draw random samples ("particles") from a chosen importance density that can be sampled directly.
- Compute a normalized weight for each particle
- Approximate the true posterior density at time $t$ by the weighted sum of the particles.

Given sufficiently large numbers of particles, this IS characterization converges to the true posterior density.

In the filtering problem, at time $t$, we approximate $p(\alpha_t|\Psi_t)$ by

- particles $\mathbf{x}_t = (x_{t1}, \ldots, x_{tN})$, a sample from some importance sampling density $f_0$.
- normalized weights $\mathbf{w}_t^\star = (w_{t1}^\star, \ldots, w_{tN}^\star)$

via the discrete distribution

$$\{(w_{ti}^\star, x_{ti}), i = 1 \ldots, N\}$$

Two major questions:

- What importance density $f_0$ can or should be chosen?
- How do we compute the weights efficiently, i.e. in a recursive fashion such that for a given weight $w_{ti}^{\star}$, we just need a new observation $y_t$ to get an approximation to the posterior $p(\alpha_t|\Psi_t)$.

A simple choice for the importance sampling density at time $t$ is the *transition prior*, that is

$$f_0(\alpha_t|\mathbf{x}_{t-1}, \Psi_t) \equiv p(\alpha_t|\mathbf{x}_{t-1})$$

as then

$$w_{ti}^{\star} \propto w_{t-1,i}^{\star} p(y_t|x_{ti}^{\star}).$$

This procedure works well in general.

**Particle Degeneracy**: A problem with sequential importance sampling is that typically, after a few iterations, all but one particle have negligible weights. This phenomenon is termed *degeneracy*.

The degree of degeneracy is represented by the *effective sample size*

$$N_t = \left( \sum_{i=1}^{N} (w_{ti}^{\star})^2 \right)^{-1}$$

Possible solutions to overcome degeneracy effects:

- brute force: very large $N$
- optimized choice of importance density
- resampling

**Resampling:** Goal is to eliminate particles with small weights and replace by new particles drawn from the vicinity of particles with larger weights

- concentrate on particles with large weights
- model important parts of the posterior more precisely

Various efficient algorithms exist for resampling which operate in $O(N)$ time (e.g. *Sampling Importance Resampling (SIR)*)

## Time-to-Event Data

In many data sets the outcome of interest is the time to an event

- time between trades
- time to a credit default
- time until stock value threshold excedance
- actuarial survival
- warranty exposure to insurance risk

The distinguishing feature of such data is that at the end of the follow up period the event may not have be observed, and thus the occurrence time is **censored**.

Censoring may occur as

- truncation of study period
- loss to follow up from non-specific cause
- loss to a "competing" event unrelated to the cause of failure being studied

Typically, components are installed over a period and followed up to a fixed date beyond the end of study, and the last components installed will thus be studied for a shorter period than those installed first, and will be less likely to undergo failure.

We often can legitimately "re-align" installation times, and work in relative rather than calendar time.

**Mathematical Notation:** The principal difference between time-to-event analysis and conventional regression is that account is taken of potential **censoring** in the response variable

- we may observe some actual responses (survival, failure) times,
- censored responses where we do not observe an actual failure but rather only that the failure occurs after a **censoring time** (the end of study) – this is called **right-censoring**
- occasionally, we observe **left-censoring** or **interval-censoring**

- the response data is thus bivariate $(Y, Z)$ where $Y$ is the time at which the response is measured, and

$$Z = \begin{cases} 1 & \text{Failure is observed} \\ 0 & \text{Censored} \end{cases}$$

- The potential presence of censoring fundamentally changes how we view the modelling process - previously we have looked at probability densities and expected responses etc.
- we have previously only dealt with data $y$ for which we need to specify $P[Y = y]$; we now need to think about
  - $P[Y > y]$ for right censoring
  - $P[Y \leq y]$ for left censoring
- We now take an alternative view, and examine **reliability/survivor** and **hazard** functions.

## Survival In Discrete Time

The probability **mass function** for response variable $Y$ is $f_Y$,

$$f_Y(y) = P[Y = y] \qquad y = 0, 1, 2, ...$$

The **distribution function** $F_Y$ is

$$F_Y(y) = P[Y \leq y] = \sum_{t=0}^{y} f_Y(t) = P[Y = 0] + \ldots + P[Y = y]$$

that is a **cumulative probability function**. Note that the function $F_Y(y)$ is a **non-decreasing** function.

In conventional regression modelling, the probability contribution for data point $i$ with response $y_i$ is $f_Y(y_i)$.

For right-censored data with censoring at $y_i$, however, we only observe the event

$$Y > y_i$$

that is, death/failure has **not occurred** before $y_i$ time units. This event has probability

$$P[Y > y_i] = 1 - F_Y(y_i)$$

This motivates consideration of the **survivor (reliability) function**

$$S_Y(y) = 1 - F_Y(y)$$

Note that $S_Y(y)$ is a **non-increasing** function.

The likelihood function (via which inference and testing will be done) is thus

$$\left\{ \prod_{i:Z_i=1} f_Y(y_i) \right\} \times \left\{ \prod_{i:Z_i=0} S_Y(y_i) \right\}$$

that is

LIKELIHOOD FOR UNCENSORED DATA

$\times$

LIKELIHOOD FOR CENSORED DATA

and the role of the predictors can be introduced via the parameters of $f_Y$ and $F_Y$.

Let

$$f_Y(y) = P[Y = y] \qquad y = 0, 1, 2, ...$$

define a discrete failure distribution. Then

$$S_Y(y) = P[Y > y] = 1 - F_Y(y) = \sum_{j=y+1}^{\infty} f_Y(j)$$

Example

**Geometric Model** For some probability $\pi$)

$$f_Y(y) = (1-\pi)^y \, \pi \qquad y = 0, 1, 2, ...$$

and

$$S_Y(y) = (1-\pi)^{y+1} \qquad y = 0, 1, 2, ...$$

## The Discrete Hazard Function

As an alternative method of specification, we consider the discrete **hazard function**

$$h_Y(y) = P\left[\text{Failure at } y | \text{Survival } \geq y\right] = \frac{f_Y(y)}{S_Y(y-1)}$$

and the **integrated hazard**

$$H_Y(y) = \sum_{t=0}^{y} h_Y(t).$$

Thus

$$f_Y(y) = \left\{ \prod_{j=0}^{y-1} (1 - h_Y(j)) \right\} \times h_Y(y)$$

and

$$S_Y(y) = \prod_{j=0}^{y} (1 - h_Y(j))$$

Example

**Constant Hazard** If

$$f_Y(y) = (1-\pi)^y \pi \qquad y = 0, 1, 2, ...$$

then

$$h_Y(y) = \frac{(1-\pi)^y \pi}{(1-\pi)^{y+1}} = \frac{\pi}{1-\pi}$$

that is, a constant, independent of $y$.

## The Continuous Time Model

The probability **density function** for continuous response variable $Y$ is $f_Y$, and the expectation, likelihood function and so on that are required for regression modelling are formed from $f_Y$. The **distribution function** $F_Y$ is

$$F_Y(y) = P[Y \leq y] = \int_0^y f_Y(t)\ dt$$

In conventional regression modelling, the likelihood contribution for data point $i$ with response $y_i$ is $f_Y(y_i)$. For right-censored data with censoring at $y_i$, we have again the reliability function

$$S_Y(y) = 1 - F_Y(y)$$

## Continuous Hazards

As a further alternative method of specification, we consider the **continuous hazard function**

$$h_Y(y) = \lim_{\delta y \to 0} P\left[\text{Failure in } (y, y+\delta y) | \text{Survival } \geq y\right] = \frac{f_Y(y)}{S_Y(y)}$$

and the **integrated hazard**

$$H_Y(y) = \int_0^y h_Y(t)\ dt$$

and it can be shown that

$$S_Y(y) = \exp\left\{-H_Y(y)\right\}$$

## The Kaplan-Meier Curve

The **Kaplan-Meier curve** (or **product-limit estimate)** is a non-parametric estimate of the reliability function; it is a decreasing step-function, where the downward steps take place at the times of the failures

The estimated reliability function at the $j$th failure/censoring time as

$$\widehat{S}_j = \prod_{i=1}^{j} \left(1 - \frac{z_i}{n - i + 1}\right) \tag{4}$$

Standard errors are also available.

**Construction:** Let

- sample size $n$ comprise observed and censored failure times
- $0 < y_{(1)} < y_{(2)} < ... < y_{(m)}$, be the distinct failure times, sorted into ascending order
- $d_j$ be the number of number of failures observed at time $y_{(j)}$
  - usually $d_j = 1$
  - certainly $d_j \geq 1$ ($d_j > 1$ implies tied failure times)
- $n_j$ be the number of patients "at risk" of failure at time $t_{(j)}$, that is, the number of patients who have failure/censoring time greater than or equal to $t_{(j)}$.

Then the observed probability of surviving beyond $t_{(j)}$ (conditional on having survived that long) is

$$\widehat{p}_j = \frac{n_j - d_j}{n_j} = 1 - \widehat{q}_j$$

say, where $q_j = d_j/n_j$ is the estimated conditional probability of failure at time $t_{(j)}$. Using the chain rule for probabilities, the estimated probability of surviving at least until time $t$ is

$$\widehat{P}(t) = \prod_{j=1}^{J_t} \widehat{p}_j = \prod_{j=1}^{J_t} \left(1 - \frac{d_j}{n_j}\right) = \widehat{S}_{KM}(t) \tag{5}$$

where $J_t = \max\left\{j : t_{(j)} \leq t\right\}$.

**Standard Errors:** A number of possibilities have been suggested. Let $P_j = P\left(t_{(j)}\right)$. Then

- **Greenwood's Formula**

$$s.e.\left(\widehat{P}_j\right) = \widehat{P}_j \sqrt{\sum_{i=0}^{j-1} \frac{d_i}{n_i - d_i}}$$

- **Peto's Formula**

$$s.e.\left(\widehat{P}_j\right) = \widehat{P}_j \sqrt{\frac{1 - \widehat{P}_j}{n'_j}}$$

where $n'_j$ is an "adjusted" or "effective" sample size, the number of survivors at the beginning of the interval $\left(t_j, t_{(j+1)}\right)$.

## The Nelson-Aalen Curve

The **Nelson-Aalen estimate** is a non-parametric estimate of the **cumulative hazard** function; it takes the form

$$\widehat{H}(t) = \prod_{j=1}^{J_t} \left( \frac{d_j}{n_j} \right) \tag{6}$$

where $J_t = \max \left\{ j : t_{(j)} \leq t \right\}$. From this, we can construct another estimate of the reliability function

$$\widehat{S}_{FH}(t) = \exp \left\{ -\widehat{H}(t) \right\}$$

this is the **Fleming-Harrington estimate** of the reliability function.

**Standard Errors:** If $\widehat{H}_j = \widehat{H}(t_{(j)})$, can use

- **Greenwood**

$$s.e.\left(\widehat{H}_j\right) = \sqrt{\sum_{i=0}^{j} \frac{d_i}{n_i\left(n_i - d_i\right)}}$$

- **Tsiatis**

$$s.e.\left(\widehat{H}_j\right) = \sqrt{\sum_{i=0}^{j} \frac{d_i}{n_i^2}}$$

- **Klein**

$$s.e.\left(\widehat{H}_j\right) = \sqrt{\sum_{i=0}^{j} \frac{d_i\left(n_i - d_i\right)}{n_i^3}}$$

## The Cox Regression Model

The **Cox** (or **Proportional Hazards**) model provides a simple way of introducing exogenous variables into the survival model.

The basic components are a **baseline hazard** function, $h_0$ and a linear predictor and (positive) link function $g$. Then for observed predictor values $X_1 = x_1$, $X_2 = x_2, ..., X_K = x_K$, the hazard function takes the form

$$h_Y(y; x) = g(x^T \beta) h_0(y)$$

that is, the hazard is modified in a multiplicative fashion by the linked-linear predictor. Typically, $g$ is the exponential function.

From the previously established relationships,

$$S_Y(y;x) = \exp\left\{-\int_0^y h_Y(t)\ dt\right\} = \exp\left\{-\int_0^y g(x^T\beta)h_0(y)\ dt\right\}$$

If a coefficient $\beta_k$ is positive, the hazard is **increased**, and the expected failure time **decreased.**

The significance of a particular predictor is based on the magnitude of

$$t = \frac{\widehat{\beta}}{s.e.\left(\widehat{\beta}\right)}$$

If $|t| > 2$, then the hypothesis that $\beta = 0$ can be rejected.

**Discrete Time Proportional Modelling**: In the discrete time case, proportional hazards modelling needs to be modified to respect constraints on the hazards to be **probabilities**.

Recall that

$$h_Y(y) = P\left[\text{Failure at } y | \text{Survival } \geq y\right] = \frac{f_Y(y)}{S_Y(y-1)}$$

Then, by construction,

$$0 \leq h_Y(y) \leq 1.$$

If this hazard is a baseline hazard, and wish to recognize modification of the hazard by exogenous variables, these constraints have to be respected.

Specifically,

$$h_Y(y;x) = g(x^T\beta)h_0(y) \leq 1$$

may not be guaranteed after the multiplicative modification by $g$.

One construction that guarantees the constraints are met is to model on the transformed scale, that is

$$\left(\frac{h_Y(y;x)}{1-h_Y(y;x)}\right) = g(x^T\beta)\left(\frac{h_0(y)}{1-h_0(y)}\right).$$

This is the *proportional odds* model.

But, for general hazard probabilities $h$,

$$\frac{h(y)}{1-h(y)} = \frac{f(y)/S(y-1)}{1-f(y)/S(y-1)} = \frac{f(y)}{S(y-1)-f(y)} = \frac{f(y)}{S(y)}$$

so proportional odds modelling has a sensible interpretation.

## The Accelerated Life Model

The **Accelerated Life** model provides another way of introducing the influence of predictors into the survival model.

The basic components now are a **baseline reliability** function, $S_0$ and a linear predictor and (positive) link function $g$.. Then for an experimental unit with observed predictor values $X_1 = x_1$, $X_2 = x_2, ..., X_K = x_K$, the reliability function takes the form

$$S_Y(y; x) = S_0(g(x^T \beta)y)$$

that is, the time scale is modified in a multiplicative fashion by the linked-linear predictor; this allows direct modelling of the influence of predictors on survival.

## Frailty Modelling

The idea of frailty modelling is to introduce **random effects** terms into the linear predictor that appears in the proportional hazards and accelerated life models. For example, we extend

$$x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{iD} + ... + \beta_D x_{iD}$$

to include a random component that is specific to the individual observational unit (bond, company etc.) concerned, that is, we have

$$x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + ... + \beta_D x_{iD} + L_i$$

where $L_i$ is some (usually zero mean) random variable.

**The Log-Rank Test :** The **log-rank** test is a standard test for significant differences between two (or more) reliability functions that differ because of the influence of the different levels of a discrete predictor.

$$\begin{aligned} H_0 &: S_1 = S_2 \\ H_1 &: S_1 \neq S_2 \end{aligned}$$

It is a non-parametric test based on ranks of samples for the two or more subgroups.

Asymptotic or exact versions of the test can be carried out.

## Parametric Modelling

It is possible to fit and compare **parametric** survival models to such data. Parametric densities, reliability functions, hazards etc. can be readily used in the formation of a likelihood, potentially within the proportional hazards/accelerated life framework.

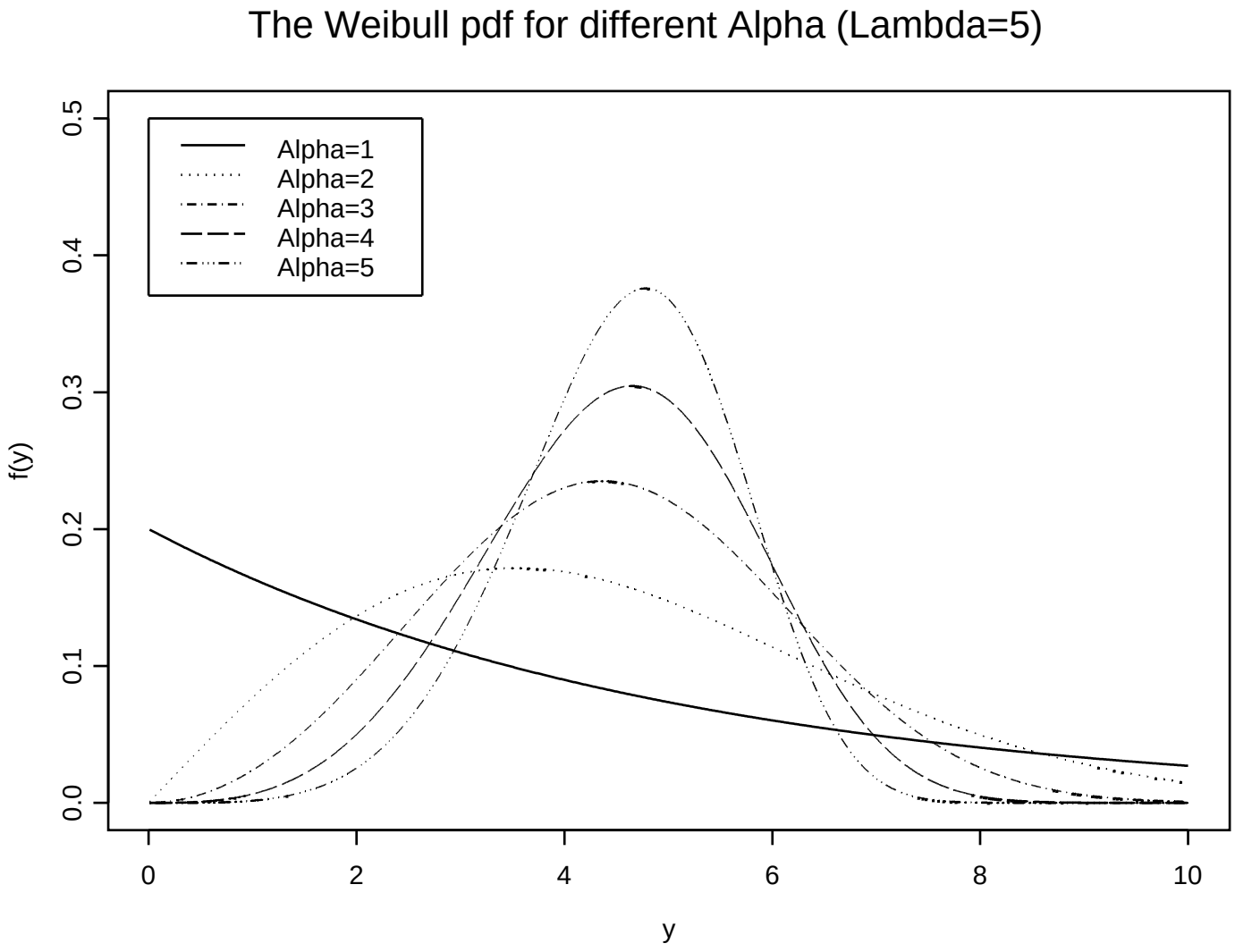
Typical models used are

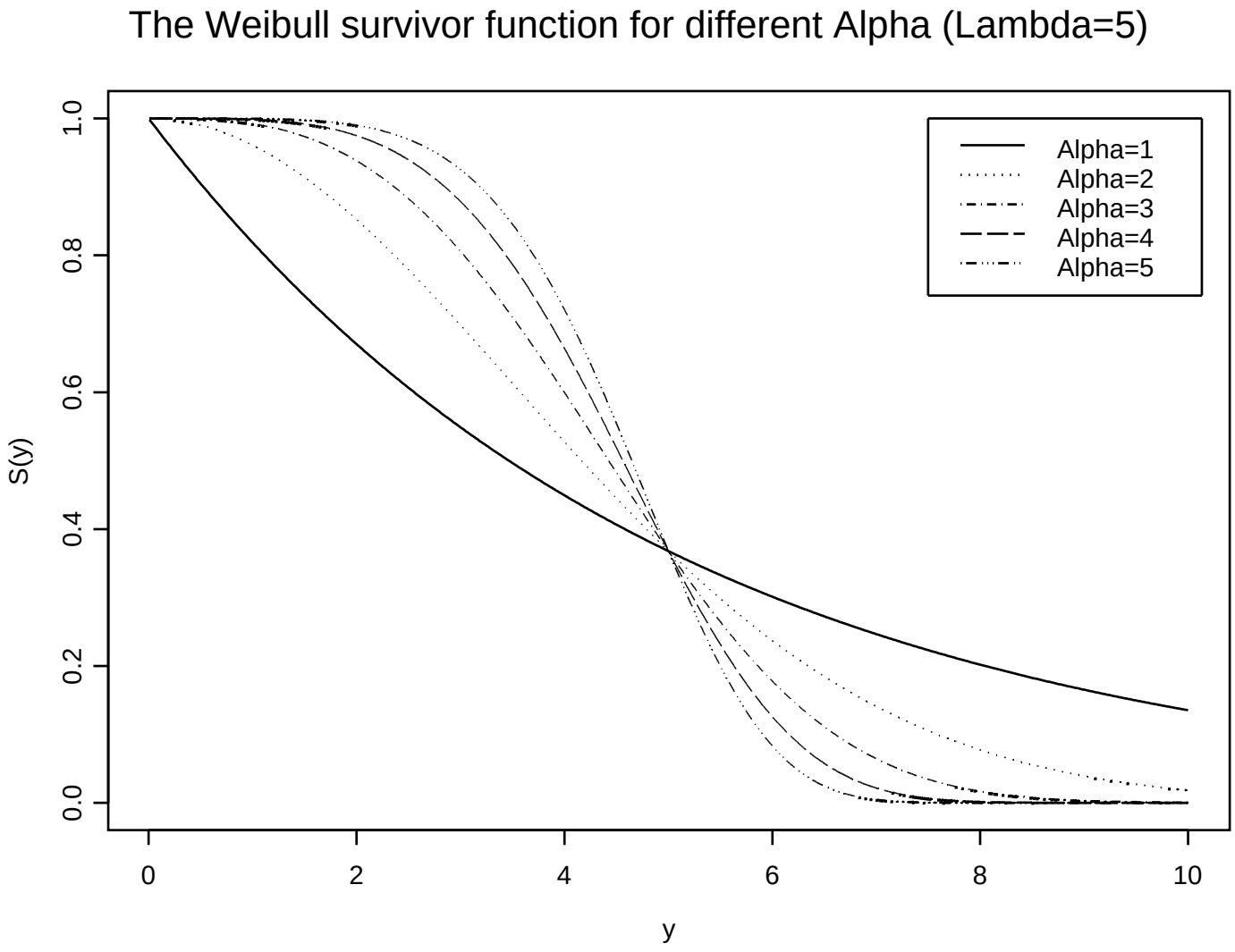
- Weibull
- Gamma
- Log-Logistic
- Log-Normal
- Pareto

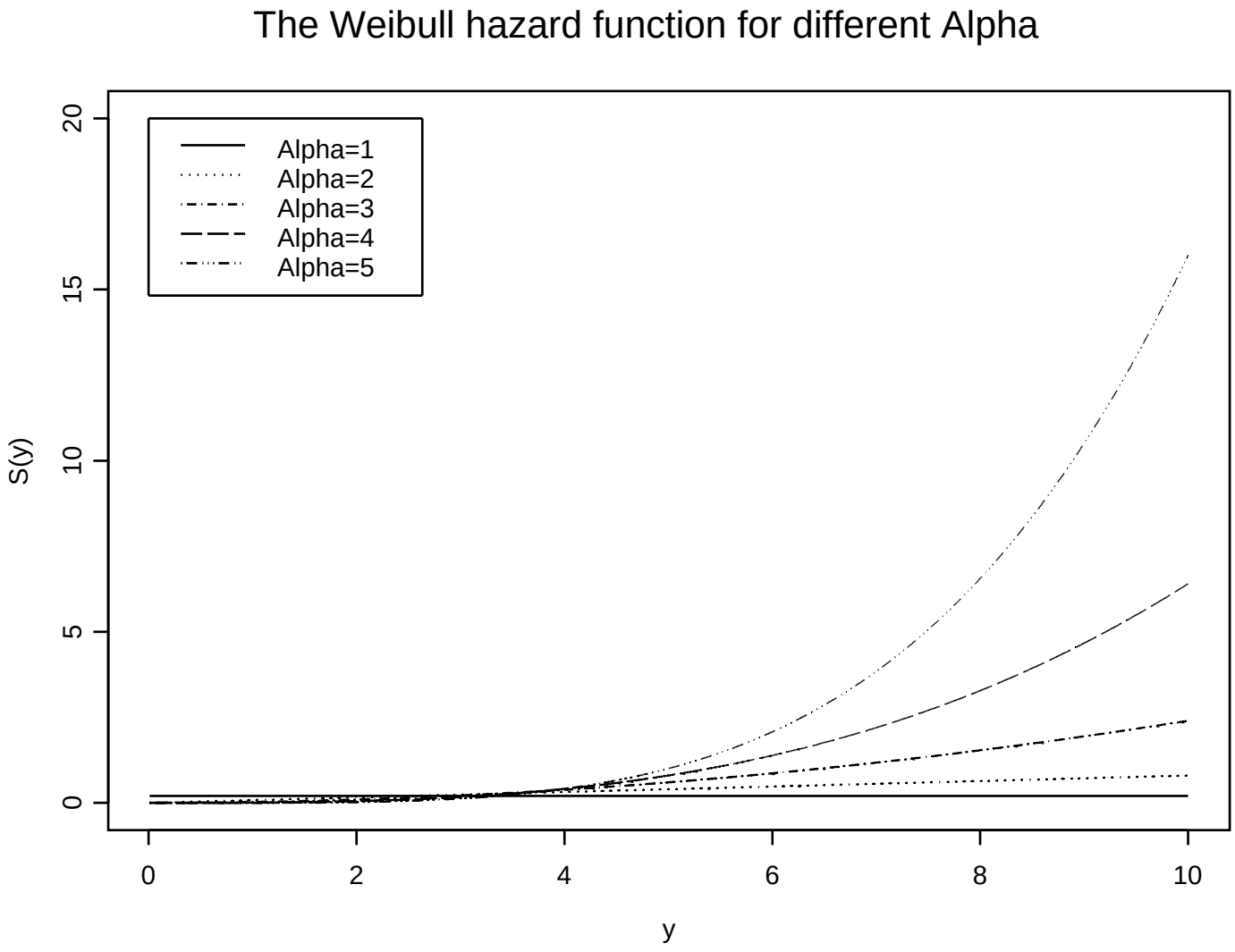
**Weibull :** for $y > 0$,

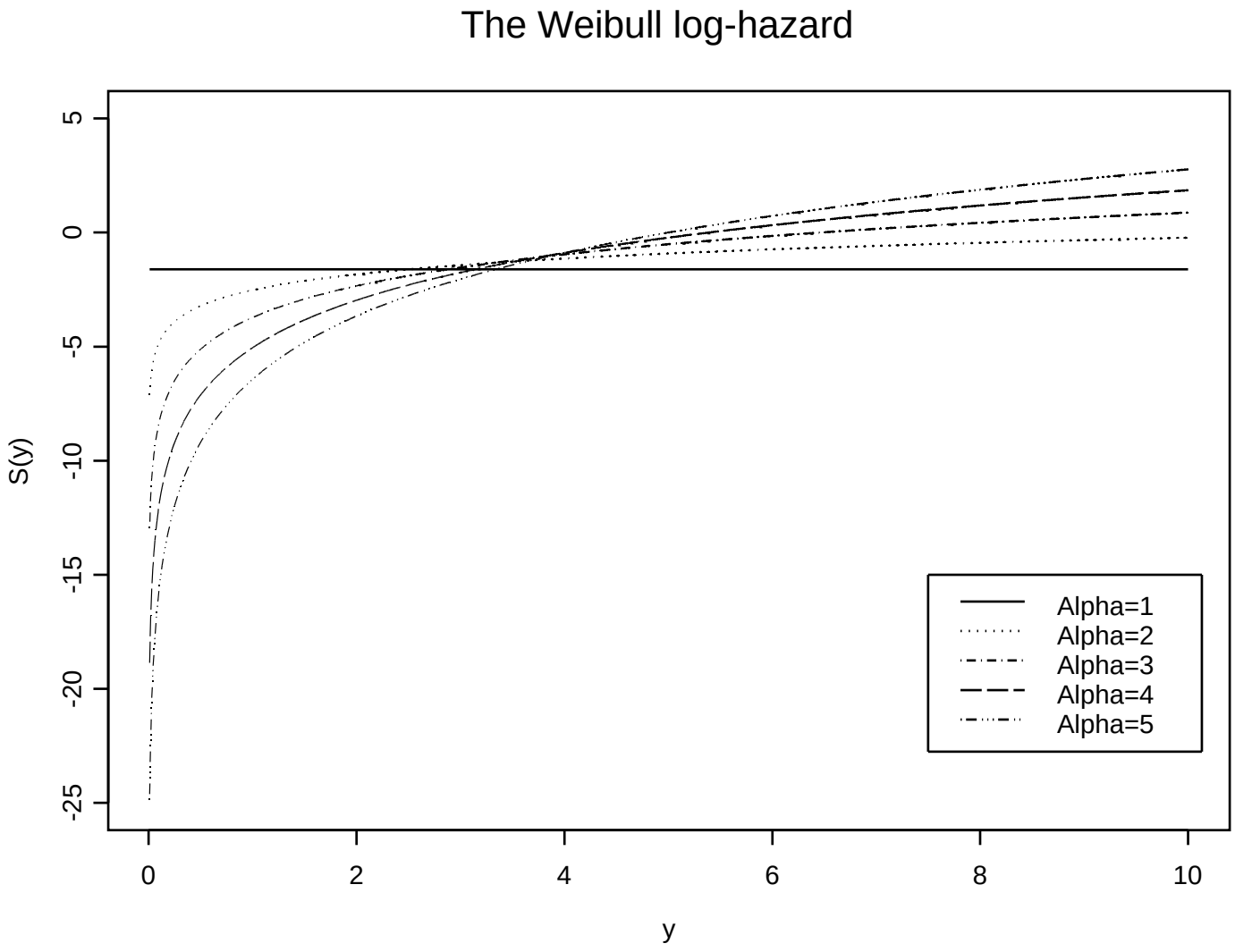
$$
\begin{aligned}
f(y) &= \frac{\alpha}{\lambda^{\alpha}} y^{\alpha-1} \exp\left\{-\left(\frac{y}{\lambda}\right)^{\alpha}\right\} \\
F(y) &= 1 - \exp\left\{-\left(\frac{y}{\lambda}\right)^{\alpha}\right\} \qquad \Longrightarrow \qquad S(y) = \exp\left\{-\left(\frac{y}{\lambda}\right)^{\alpha}\right\} \\
h(y) &= \frac{\alpha}{\lambda^{\alpha}} y^{\alpha-1} \\
H(y) &= \left(\frac{y}{\lambda}\right)^{\alpha}
\end{aligned}
$$

for parameters $\alpha, \lambda > 0$ (the *shape* and *scale* parameters respectively).


The Weibull pdf for different Alpha (Lambda=5)
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
f(y)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
y


The Weibull survivor function for different Alpha (Lambda=5)
S(y)
y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5


The Weibull hazard function for different Alpha
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5
S(y)
0
5
10
15
20
0
2
4
6
8
10
y


The Weibull log-hazard
S(y)
5
0
-5
-10
-15
-20
-25
0
2
4
6
8
10
y
Alpha=1
Alpha=2
Alpha=3
Alpha=4
Alpha=5

## Multivariate Models & Competing Risks

An important generalization extends event time $Y$ to be a **vector** quantity, that is, we have say $K$ different aspects of failure, with random variable $Y = (Y_1, ...Y_K)$ requiring a joint probability model.

A model that captures joint structure of time-to-event observations must respect marginal and conditional coherence requirements, and if possible, capture a full range of dependency structures.

Typically, such models are difficult to construct.

Some joint models exist

- Multivariate Exponential
- Multivariate Weibull
- Multivariate Lognormal
- Hougaard etc.

Some joint models can be constructed

- Copula (latent uniform)
- Latent Multivariate Normal

Typically, inference for these models is quite complex, but can be achieved using numerical procedures.

Some models use the idea of a *frailty* or factor-type structure.

- $Y_1, \ldots, Y_K$ positive random variables, potential censoring, corresponding to different items (defaultable bonds etc.)
- modelling achieved through hazard rates $h_1, \ldots, h_K$
- set

$$h_i(t) = h(t) + \eta_i(t) \qquad i = 1, \ldots, K.$$

  - $h(t)$ is some market-level "factor" determining global hazard rates, induces dependence across the individual items
  - $\eta_i(t)$ is some item specific hazard.

Another common experimental situation is one of **competing risks**; that is, there are $K$ potential causes of failure, but **at most one** is observed for each individual in the study. Then the failure time, $T$, is defined by

$$T = \min\{Y_1, ..., Y_K\}$$

If the cause of failure, $C$, is recorded as $C = k$, we observe

$$Y_1 > t, ..., Y_{k-1} > t, Y_k = t, Y_{k+1} > t, ..., Y_K > t$$

whereas if the observation is censored, we observe

$$Y_1 > t, ..., Y_{k-1} > t, Y_k > t, Y_{k+1} > t, ..., Y_K > t$$

A joint model is again often difficult to construct, and in addition there are issues to do with identifiability of the "marginal" failure processes for the components of $Y$.

i.e. without sufficient data, there are problems in estimating the models for $Y_1, ..., Y_K$ considered on their own

## Multi-State Modelling

In **multi-state** modelling, rather than just having the standard failed/not failed (dead/alive) dichotomy, with

$$Z = \begin{cases} 0 & \text{Censored} \\ 1 & \text{Failure is observed} \end{cases}$$

we have an extension to polytomy, where

$$Z(t) = \begin{cases} 0 & \text{Censored at time } t \\ 1 & \text{State 1 at time } t \\ \vdots & \vdots \\ M & \text{State } M \text{ at time } t \end{cases}$$

In such a model, we attempt to estimate the probability

$$\pi_{ij}(t_i, t_j) = P[\text{State } j \text{ at time } t_j | \text{State } i \text{ at time } t_i]$$

for $t_i < t_j$, or rate, $\lambda_{ij}$, of transition from one state to another.

In a discrete time framework, homogeneous **Markov Models** are typically used, characterized by a **transition matrix** $P$, with $(i, j)^{th}$ entry $\pi_{ij}$, independent of $t$, with

$$\sum_{j=0}^{M} \pi_{ij} = 1$$

A **multi-state** process is a random process $\{Z(t)\}_{t\geq 0}$ describing the state within which the individual lies at time $t$.

This kind of modelling is very useful for modelling progression to credit default; different states could correspond to different credit ratings.