Censored Data Analysis

Here, with this re-ordering, \( z \) (the vector of missing-data outcomes) evidently consists of the actual (uncensored) failure times (on the log scale) \((z_{m+1}, \ldots, z_n)\).

Rewriting the model in distributional form, defining the complete-data outcome vector as
\[
\begin{align*}
t^* &= (t_1, \ldots, t_m, z_{m+1}, \ldots, z_n) = (z, t) \quad \text{and (as usual in regression) conditioning on the predictor values} \ v_i, \\
(t_i^*|\beta_0, \beta_1, \sigma^2) &\sim \text{indep} \ N(\beta_0 + \beta_1v_i, \sigma^2)
\end{align*}
\]
(\text{for} \ i = 1, \ldots, n); thus
\[
p(t^*|\beta_0, \beta_1, \sigma^2) = c l(\beta_0, \beta_1, \sigma^2|t^*) \quad (25)
\]
\[
= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{-\frac{1}{2\sigma^2}[t_i^* - (\beta_0 + \beta_1v_i)]^2\right\},
\]
from which the augmented log-likelihood is
\[
\log l(\beta_0, \beta_1, \sigma^2|z, t) = c - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^m [t_i - (\beta_0 + \beta_1v_i)]^2
\]
\[
- \frac{1}{2\sigma^2} \sum_{i=m+1}^n [z_i - (\beta_0 + \beta_1v_i)]^2.
\]

With \( \theta = (\beta_0, \beta_1, \sigma^2) \) let's treat this as the log augmented posterior \( p(\theta|z, y) \) (which is like working with a completely flat prior) and use EM\(_B\) to find the MLEs.

As usual the \( E \) step requires computing the expectation of this log posterior with respect to \( p(z|\theta, y) \) (evaluated at \( \theta = \theta^{(j)} \)), which in this case is determined by \( p(z_i|\beta_0, \beta_1, \sigma^2, c_i) \), so let's think about that predictive distribution.
Censored Data Analysis

For \( i = 1, \ldots, m \) we know the actual failure times so there's nothing to predict; for \( i = (m + 1), \ldots, n \) the basic sampling model (23) says that the \( z_i \) (given \( \beta_0, \beta_1, \sigma^2 \)) are \( N(\beta_0 + \beta_1 v_i, \sigma^2) \) but constrained to be \( \geq c_i \); thus in the \( E \) step we have to compute

\[
c - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{m} [t_i - (\beta_0 + \beta_1 v_i)]^2
- \frac{1}{2\sigma^2} \sum_{i=m+1}^{n} [E(z_i^2|\beta_0, \beta_1, \sigma^2, z_i > c_i)]
- 2(\beta_0 + \beta_1 v_i)E(z_i|\beta_0, \beta_1, \sigma^2, z_i > c_i) + (\beta_0 + \beta_1 v_i)^2.
\]

I'll just sketch the details from now on; it can be shown by working with truncated Gaussian distributions that

\[
E(z_i|\beta_0, \beta_1, \sigma^2, z_i > c_i) = \mu_i + \sigma H \left( \frac{c_i - \mu_i}{\sigma} \right),
\]

where \( \mu_i = \beta_0 + \beta_1 v_i \) and \( H(x) = \frac{\phi(x)}{1 - \Phi(x)} \), in which \( \phi(x) \) and \( \Phi(x) \) are the density function and CDF of the standard Gaussian distribution, respectively; similarly it turns out that

\[
E(z_i^2|\beta_0, \beta_1, \sigma^2, z_i > c_i)) = \mu_i^2 + \sigma^2 + \\
\sigma(c_i + \mu_i)H \left( \frac{c_i - \mu_i}{\sigma} \right).
\]

As for the \( M \) step, abbreviating \( E(z_i|\beta_0, \beta_1, \sigma^2, z_i > c_i) \) as \( E(z_i) \), and similarly for \( E(z_i^2) \), the partial derivatives turn out as follows:

\[
\frac{\partial Q}{\partial \beta_0} = 0 \text{ iff } \sum_{i=1}^{m} [t_i - (\beta_0 + \beta_1 v_i)] + \\
\sum_{i=m+1}^{n} [E(z_i) - (\beta_0 + \beta_1 v_i)] = 0;
\]
Censored Data Analysis

\[ \frac{\partial Q}{\partial \beta_1} = 0 \text{ iff } \sum_{i=1}^{m} v_i [t_i - (\beta_0 + \beta_1 v_i)] + \]
\[ \sum_{i=m+1}^{n} v_i [E(z_i) - (\beta_0 + \beta_1 v_i)] = 0; \]
\[ \frac{\partial Q}{\partial \sigma^2} = 0 \text{ iff } 0 = -\frac{n}{\sigma^2} + \frac{\sum_{i=1}^{m} [t_i - (\beta_0 + \beta_1 v_i)]^2}{\sigma^4} + \]
\[ \frac{\sum_{i=m+1}^{n} [E(z_i^2) - 2(\beta_0 + \beta_1 v_i)E(z_i) + (\beta_0 + \beta_1 v_i)^2]}{\sigma^4}. \]  \tag{30}

You can get the \( \beta_k^{(j+1)} \) by replacing \( c_i \) by \( E(z_i|\beta_0^{(j)}, \beta_1^{(j)}, \sigma^{(j)}, z_i > c_i) \) and using least squares (see Tanner (1993) for details); to get \( (\sigma^{(j+1)})^2 \) you can compute

\[ (\sigma^{(j+1)})^2 = \frac{\sum_{i=1}^{m} (t_i - \mu_i^{(j)})^2}{n} + \]
\[ \frac{\sigma^{(j)}^2 \sum_{i=m+1}^{n} \left[ 1 + \left( \frac{c_i - \mu_i^{(j)}}{\sigma^{(j)}} \right) H \left( \frac{c_i - \mu_i^{(j)}}{\sigma^{(j)}} \right) \right]}{n}. \]  \tag{31}

where \( \mu_i^{(j)} = \beta_0^{(j)} + \beta_1^{(j)} v_i \); this defines the iterative EM scheme; 16 iterations from \{starting values derived from the regression pretending that the censored observations were uncensored\} yields the MLEs

\[ \hat{\beta}_0 = -6.019 \quad \hat{\beta}_1 = 4.311 \quad \hat{\sigma} = 0.2592 \]  \tag{32}

to 4 significant figures; as usual with EM, this could presumably be turned into a Gibbs sampler without much more effort.
Mixture Modeling

- Another important class of models that benefits from the missing-data perspective is mixture models.

**Example:** Bowmaker et al. (1985) present data on the peak sensitivity wavelengths for individual microspectrophotometric records on the eyes of a species of monkey; here we have \( n = 48 \) conditionally IID measurements on monkey S14:

```r
eyes <- 500 + c(29.0, 30.0, 32.0, 33.1, 33.4, 33.6, 33.7, 34.1, 34.8, 35.3, 35.4, 35.9, 36.1, 36.3, 36.4, 36.6, 37.0, 37.4, 37.5, 38.3, 38.5, 38.6, 39.4, 39.6, 40.4, 40.8, 42.0, 42.8, 43.0, 43.5, 43.8, 43.9, 45.3, 46.2, 48.8, 48.7, 48.9, 49.0, 49.4, 49.9, 50.6, 51.2, 51.4, 51.5, 51.6, 52.8, 52.9, 53.2)

hist(eyes, nclass = 10, prob = T,
     xlab = 'Peak Sensitivity Wavelength', main = '')

qqnorm(eyes, main = 'Peak Sensitivity Wavelength')
```

There were **biological reasons** for expecting a **multimodal distribution**, and we do see evidence of (at least) bimodality.

What do **bimodal normal qqplots** look like?
y1 <- rnorm(500)

y2 <- rnorm(500, 4)

y <- c(y1, y2)

hist(y, nclass = 40, prob = T, main = '',
     xlab = 'Artificial Data')

qqnorm(y, main = 'Artificial Data')

This is what Gaussian bimodality looks like on a normal qqplot; looking back at the previous qqplot you can see some similarity.

So it looks like it would be reasonable to fit a model in which the sampling distribution is a mixture of $J = 2$ Gaussians, each with its own mean and variance; with $\tau_1 + \tau_2 = 1$ this can be written

$$p(y_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2) = \tau_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{1}{2\sigma_1^2}(y_i - \mu_1)^2\right] + \tau_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{1}{2\sigma_2^2}(y_i - \mu_2)^2\right].$$
Mixture Modeling

That doesn't look particularly easy to work with; maybe there's some latent (missing) data \( z = (z_1, \ldots, z_n) \) that would make things easier.

The obvious thing to try is to make use of the hidden structure of the data: this mixture model says that each observation belongs either to group 1 (\( N(\mu_1, \sigma_1^2) \)) or group 2 (\( N(\mu_2, \sigma_2^2) \)) but we don't know for sure which observation belongs to which group; so let \( z_i = 1 \) if observation \( i \) belongs to group 1 and 2 otherwise.

One (slightly clever) way to write this model hierarchically, which is how the WinBUGS people fit mixture models, is as follows:

\[
\begin{align*}
(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_2) & \sim p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_2) \\
(z_i | \tau_2) & \sim \text{Bernoulli}(\tau_2) \\
(y_i | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, z_i) & \sim \text{indep} N(\mu_{1+z_i}, \sigma_{1+z_i}^2).
\end{align*}
\]

Instead (Wikipedia!) let's regard this as a special case of the more general situation in which the \( y_i \) are draws from multivariate Gaussians (of dimension \( d \)) and write the sampling-distribution part of the model like this (later we'll specialize to the univariate case): with \( \tau = (\tau_1, \tau_2) \), the \( (z_i | \tau) \) are IID with \( p(z_i = 1 | \tau) = \tau_1 \) and \( p(z_i = 2 | \tau) = \tau_2 = 1 - \tau_1 \); then the \( y_i \) (given \( z_i \) and the parameters) are conditionally IID with \( (y_i | z_i = 1, \mu_1, \Sigma_1) \sim \mathcal{N}_d(\mu_1, \Sigma_1) \) and \( (y_i | z_i = 2, \mu_2, \Sigma_2) \sim \mathcal{N}_d(\mu_2, \Sigma_2) \).

Taking \( \theta = (\mu_1, \mu_2, \Sigma_1, \Sigma_2, \tau) \), with this way of writing the model the augmented likelihood function is

\[
l(\theta | y, z) = \prod_{i=1}^{n} \sum_{j=1}^{J} \tau_j I(z_i = j) p(y_i | \mu_j, \Sigma_j), \tag{34}
\]
where \( p(y_i|\mu_j, \Sigma_j) \) is the \( N_d(\mu_j, \Sigma_j) \) density function; in exponential-family format this log-likelihood is

\[
\log l(\theta|y, z) = \sum_{i=1}^{n} \sum_{j=1}^{J} I(z_i = j) \left[ \log \tau_j - \frac{1}{2} \log |\Sigma_j| \right. \\
\left. - \frac{1}{2} (y_i - \mu_j)' \Sigma_j^{-1} (y_i - \mu_j) - \frac{d}{2} \log 2\pi \right],
\]

and this can also be thought of as the log augmented posterior \( \log p(\theta|y, z) \) with a completely flat prior, for the purpose of using the EM algorithm to get the MLE \( \hat{\theta} \).

As usual in EM, along with \( p(\theta|y, z) \) we need the other full-conditional distribution \( p(z|\theta, y) \); here as usual this is

\[
p(z|\theta, y) = c p(z, \theta, y) = c p(\theta) p(z|\theta) p(y|z, \theta); \\
\text{with the flat prior } p(\theta) = 1; \\
p(z|\theta) \text{ is Bernoulli; and } \\
p(y|z_i = j, \theta) \text{ is multivariate Gaussian with parameters } \\
(\mu_j, \Sigma_j), \\
\]

\[
T_{ji} = p(z_i = j|\theta, y_i) = \frac{\tau_j p(y_i|\mu_j, \Sigma_j)}{\tau_1 p(y_i|\mu_1, \Sigma_1) + \tau_2 p(y_i|\mu_2, \Sigma_2)}. 
\]

Switching the iteration index to \( t \) to avoid confusion with the group index \( j \), you'll recall that for the \( E \) step we need to compute \( Q(\theta, \theta^{(t)}) \), the expectation of \( \log p(\theta|y, z) \) with respect to \( p(z|\theta^{(t)}, y) \); but \( z \) enters \( \log p(\theta|y, z) \) only through \( I(z_i = j) \) and the expectation of an indicator function is the probability that it's 1, so here we get

\[
Q(\theta, \theta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} T_{ji}^{(t)} \left[ \log \tau_j - \frac{1}{2} \log |\Sigma_j| \right. \\
\left. - \frac{1}{2} (y_i - \mu_j)' \Sigma_j^{-1} (y_i - \mu_j) - \frac{d}{2} \log 2\pi \right],
\]
Mixture Modeling

where

\[ T_{ji}^{(t)} = \frac{\tau_j^{(t)} p(y_i | \mu_j^{(t)}, \Sigma_j^{(t)})}{\tau_1^{(t)} p(y_i | \mu_1^{(t)}, \Sigma_1^{(t)}) + \tau_2^{(t)} p(y_i | \mu_2^{(t)}, \Sigma_2^{(t)})}. \]  (39)

As usual the \( M \) step consists of maximizing \( Q(\theta, \theta^{(t)}) \) as a function of \( \theta \) for fixed \( \theta^{(t)} \); fortunately, that’s not hard here: considering \( \tau \) first, and recalling the constraint \( \tau_1 + \tau_2 = 1 \), you could compute the partial derivatives and use Lagrange multipliers, but it’s easier just to see if this situation matches one of our basic templates:

\[ \tau^{(t+1)} = \arg\max_{\tau} Q(\theta, \theta^{(t)}) \]  (40)

\[ = \arg\max_{\tau} \left[ \left( \sum_{i=1}^{n} T_{1i}^{(t)} \right) \log \tau_1 + \left( \sum_{i=1}^{n} T_{2i}^{(t)} \right) \log \tau_2 \right] \]

But this is just the usual Bernoulli situation, for which the MLE is

\[ \tau_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{ji}^{(t)}}{\sum_{i=1}^{n} (T_{1i}^{(t)} + T_{2i}^{(t)})} = \frac{1}{n} \sum_{i=1}^{n} T_{ji}^{(t)}, \]  (41)

and this makes excellent intuitive sense: the new estimate of \( \tau_j \) is the mean over all observations of the current estimated probabilities that each observation belongs to group \( j \).

Considering \((\mu_1, \Sigma_1)\) next,

\[ (\mu_1^{(t+1)}, \Sigma_1^{(t+1)}) = \arg\max_{(\mu_1, \Sigma_1)} Q(\theta, \theta^{(t)}) \]  (42)

\[ = \arg\max_{(\mu_1, \Sigma_1)} \sum_{i=1}^{n} T_{1i}^{(t)} \left[ -\frac{1}{2} \log |\Sigma_1| \right. \]

\[ -\frac{1}{2} (y_i - \mu_1)'\Sigma_1^{-1}(y_i - \mu_1) \right]; \]
but maximizing this is just the usual MLE calculation in
the one-sample multivariate Gaussian sampling model
except that the $T_{1i}^{(t)}$ values serve as weights; the answer is a
weighted version of the usual estimates:

$$
\mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1i}^{(t)} y_i}{\sum_{i=1}^{n} T_{1i}^{(t)}} \quad \text{and} \quad (43)
$$

$$
\Sigma_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1i}^{(t)} (y_i - \mu_1^{(t+1)}) (y_i - \mu_1^{(t+1)})'}{\sum_{i=1}^{n} T_{1i}^{(t)}}
$$

and of course the estimates for $(\mu_2^{(t+1)}, \Sigma_2^{(t+1)})$ are the same
except that all of the 1 subscripts are replaced by 2.

- Since we've already specified $p(\theta|y, z)$ and $p(z|\theta, y)$ in this
  EM calculation, as usual it's not hard to go beyond EM
  and write a Gibbs sampler to extract the entire posterior.

- When you do the MCMC, you run into an inherent
  identifiability problem in mixture models, which is called
  the label-switching problem (recall that a model is
  unidentifiable if two or more different parameter
  configurations can lead to exactly the same likelihood):
  $(\mu_1, \mu_2)$ could equally well be the means of (group 1,
  group 2) or (group 2, group 1).

The simplest way to beat this problem here is to
re-parameterize the means: in WinBUGS notation and
format, let the mean of group 1 be $\lambda$ and the mean of
group 2 be $(\lambda + \theta)$, where $\theta$ is constrained to be positive;
with this change a slightly simpler model (with
$\sigma_1^2 = \sigma_2^2 = \sigma^2$) becomes

$$
T[i] \iid j \text{ with probability } P[j], j = 1, 2
$$

$$
\mu[i] = \lambda[T[i]], \quad \lambda[2] = \lambda[1] + \theta \quad (44)
$$

$$
y[i] \sim \text{indep} \mathcal{N}(\mu[i], \sigma^2) .
$$
Mixture Modeling

Note: (1) \( \theta \) is constrained to be positive in the prior with the specification

\[
\theta \sim \text{dnorm}(0.0, 1.0E-6) \ I(0.0, ) ,
\]

which multiplies a diffuse Gaussian prior by the indicator function on \((0, \infty)\).

(2) A Dirichlet(1,1) prior is used on the group membership probabilities, which is equivalent to putting a Uniform(0,1) prior on \(P[1]\) or \(P[2]\).

(3) The data are input in sorted order, so that we can see how the posterior probabilities of being in group 1 or 2 change as a function of observation number \(i\).
The two group means are estimated as 536.7 (with a posterior SD of 0.93) and 548.9 (1.3);

the (assumed-to-be) common $\sigma$ is estimated to be 3.78 (0.63); and

the mixing weights are estimated to be 0.60 (0.09) and 0.40 (0.09) for groups 1 and 2, respectively.
Starting with observation 1 (the smallest) and moving to the right, the first 25 or so observations are estimated to be almost certainly in group 1.
Then from about **observation 26 through 33**, there is **substantial uncertainty** as to **group membership**.
And from about observation 34 to 48 it's pretty clear that they all belong to group 2.
Mixture Modeling

The mixture of two Gaussians with parameters given by the posterior means tracks the data well.

You could take the MCMC draws and plot the estimated densities, one for each row in the MCMC data set, to get an idea of the uncertainty band arising from this approach to density estimation.