EM in correlated BVN variates

Presentation: IBS 2017

Malcolm Hudson

22/06/2017

Censored Linear model and EM algorithm for BVN correlated Competing Risks

Valerie Gares[1]; Malcolm Hudson[2]; Maurizio Manuguerra [3]; Val Gebski [4]

- Goal: *parametric* survival analysis with *correlated* competing risks
- **Competing risks** occur when event of interest is precluded (censored) by occurence of other event type(s) e.g. death from another cause precludes further hazard of event of interest
- survival outcome is composite (first event time, status). Coding status: Δ
  0. censored, no event during period of follow-up;
  1. event of interest;
  2. competing event.

[1]Univ. Toulouse, France
[2]Macquarie University and NHMRC CTC
[3]Macquarie University
[4]NHMRC CTC

Why study bivariate Normal competing risks

Well known parametric, semi-parametric and non-parametric survival approaches are univariate

- for logNormal survival outcomes with censored outcomes, Schmee & Hahn, Aitkin (1981), Buckley-James (1979) provide estimators for censored linear regression

Correlated risks are of interest, but non-identifiability of *joint* survival time distribution with 2 competing risks

- an "identifiability crisis" (Crowder 1991)
- response has been development of many semi-parametric models and Jeong-Fine fully parametric model (Jeong and Fine 2007)

Our approach

- study performance of estimates of β, ρ in this ill-posed problem
- sensitivity analysis to:
  - correlation ρ;
  - parametric assumptions.

Three components

1. *EM algorithm* for censored *BVN* competing risks
   - includes simulation study of hazard ratio estimation sensitivity to ρ in two group survival comparisons
2. Moment calculations for BVN using Stein's identity
3. R-package **bnc** (BVN competing risks) fitting AFT lm's

EM algorithm

In *univariate* survival analysis Aitkin's approach is an EM-algorithm used with a censored regression model to estimate parameters μ, σ^2 of normally distributed survival time data.

With univaritate censored data: *time* $T = \min(Y, C)$ and *status* δ $(1/0)$, where C is censor time.

1. Complete data by imputation (E-step) of residual survival using $E(Y^m \mid Y > \tau,\, \delta = 0)$, for $m = 1, 2$.
2. Follow by ML-estimation based on the *imputed* sufficient statistics $\sum Y_j,\ \sum Y_j^2$

We generalize the (univariate) EM algorithm to *correlated* BVN competing risks, event times (Y_1, Y_2). Observed data is (T, Δ), where $T = \min(Y_1, Y_2, C)$.

EM algorithm

New context (BNC linear model)

- latent variable model $Y \sim \mathrm{BVN}(\mu, \Sigma)$
  with AFT model $\mu = XB$
- today, focus on estimating Σ (**B** straightforward)
- Not the standard mixture model EM of McLachlan Section 5.2
  - because of selection of first occuring event
- Generalizes to 2d the time to event of a single of Aitkin 1981
  - as above

EM algorithm for BVN

Particular example (for clarity)

- *known* means 0 of (log-)Normal latent vars, *no censoring*
- $y = \min(Y_1, Y_2)$ with $Y \sim \text{BVN}(0, \Sigma)$, Σ unknown
- Δ identifies which risk is observed (1 or 2)
- two risk times are never *both* observed
- Goal: the ML estimator of Σ from an random sample of y, Δ
- captures main issue, correlation, but *not* impact of censoring

Likelihoods

- observed data likelihood function

$$
\begin{aligned}
L_{y,\Delta}(\Sigma) &= \prod_{j:\Delta_j=1} f_{Y_1}(y_j)\, F_{2|1}(y_j|y_j) \prod_{j:\Delta_j=2} f_{Y_2}(y_j)\, F_{1|2}(y_j|y_j), \\
&\propto \left|\Sigma\right|^{-\frac{n}{2}} \prod_{j=1}^{n_1} \exp\left(-\frac{1}{2\sigma_1^2} y_j^2\right) \Phi\left(\frac{\frac{y_j}{\sigma_2} - \rho\frac{y_j}{\sigma_1}}{\sqrt{1-\rho^2}}\right) \\
&\quad \prod_{j=n_1+1}^{n_1+n_2} \exp\left(-\frac{1}{2\sigma_2^2} y_j^2\right) \Phi\left(\frac{\frac{y_j}{\sigma_1} - \rho\frac{y_j}{\sigma_2}}{\sqrt{1-\rho^2}}\right) \qquad (1)
\end{aligned}
$$

- direct optimization for Σ available, but problem is ill-posed

- complete data $\ell(\Psi \mid Y)$with $\Psi = \Sigma^{-1}, V = Y' Y, 2 \times 2$

$$
\begin{aligned}
\ell(\Psi \mid Y) &= \frac{n}{2} \log \left|\Psi\right| - \frac{1}{2} \mathrm{tr} \left\{\Psi\, Y' Y\right\} \\
&= \frac{n}{2} \log \left|\Psi\right| - \frac{1}{2} \mathrm{tr} \left\{\Psi\, V\right\} \qquad (2) \\
Q(\Psi, \Psi^0) &= E \left[\ell(\Psi;\ Y) \mid y, \Delta,\ \Psi^0\right] \\
&= \frac{n}{2} \log \left|\Psi\right| - \frac{1}{2} \mathrm{tr} \left\{\Psi\, E^0 V\right\} \qquad (3)
\end{aligned}
$$

- EM algorithm based on a step from the initial choice, matrix Ψ^0
- maximize Q wrt Ψ, easy!
- increases observed Likelihood

EM approach

Complete data Y, $nx2$, has 2-d sufficent statistic $Y^{'}Y$ for Ψ.

For initial estimator Ψ^0, imputed value of this sufficient statistic is

$$E^0\, Y^{'}Y = E^0\,(D^0 Z^{'} Z D^0) \\ = D^0\, E^0\,(Z^{'}Z)\, D^0, \quad (4)$$

where

$$E^0\, f(Y) = E\,\left[f(Y)\,|\,y, \Delta; \Psi^0\right] \quad (5)$$

and D^0 is the diagonal matrix with elements (σ_1^0, σ_2^0).

$E\, Y^{'}Y/n = \Sigma$, so the M-step update is
$\Sigma = \Psi^{-1} = D^0\, E^0\,(Z^{'}Z)\, D^0$.

Since $E^0\, Z^{'}Z$ is evaluated from expectations
$\left[E^0(Z_{j1}^2), E^0(Z_{j1}Z_{j2}), E^0(Z_{j2}^2)\right]$ it suffices to estimate moments of
Z, a standardised bivariate Normal distribution

Moments of conditional distribution under constraints

Let $Z = (Z_1, Z_2) \sim \text{BVN}(0, \Sigma)$ in
standardized form $\sigma_{11} = \sigma_{22} = 1$, $\sigma_{12} = \rho$, so

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

For subject j with first event at time y and $\Delta = 2$, we observe
$Y_2 = \sigma_2 Z_2$, with $Y_2 < Y_1 = \sigma_1 Z_1$. We will need
$E(Z_1^2 | Z_1 > a, Z_2 = b)$ with $b = y/\sigma_2$ and $a = y/\sigma_2$

In general we need to evaluate moments:

- $E(Z_1^m | Z_1 > a, Z_2 = b)$, for $m = 1, 2$
  by a univariate Stein identity
- $E(Z_1^l Z_2^m | Z_1 > a, Z_2 > b)$, for $l + m \leq 2$
  by a bivariate Stein identity

Charles Stein

- Charles Stein, mathematician, probabilist and statistician
  - inadmissability of the multivariate normal mean
  - Stein shrinkage
  - Stein Unbiased Risk Estimator

Stein's identity

- Identity *characterises* the multivariate Normal distribution
  - e.g. prove limit theorems by showing this identity is satisfied, for arbitrary f
- **Univariate:** $Y \sim N(\mu, \sigma^2)$ **iff**

$$E[(Y-\mu)f(Y)] = \sigma^2 E[f'(Y)]$$

  ($=>$ proof: integration by parts)
- **Multivariate:** $Y \sim MVN(\mu, \Sigma)$ **iff**

$$Cov[Y, f(Y)] = \Sigma E[\nabla f(Y)]$$

Univariate Stein for first moment

Let $Z = (Z_1, Z_2) \sim \text{BVN}(0, \Sigma)$, with $\sigma_{11} = \sigma_{22} = 1$, $\sigma_{12} = \rho$.

$$\begin{aligned} E_{10.01} &= E(Z_1 | Z_1 > a, Z_2 = b) \\ &= E[(Z_1 - \rho b) | Z_1 > a, Z_2 = b] + \rho b \end{aligned}$$

$$\begin{aligned} E[(Z_1 - \rho b) | Z_1 > a, Z_2 = b] &= E^{Z_2 = b}[(Z_1 - \rho b)\, H(Z_1 - a)] \\ &= (1 - \rho^2) E^{Z_2 = b}[\delta(Z_1 - a)] \\ &= (1 - \rho^2)\; p_{1|2}(a|b) \end{aligned}$$

Here (6) uses Stein's univariate identity, with the derivative of Heavyside, Dirac's delta. Dirac's delta, in convolution has the *sifting property*[5]

[5] Bracewell 2001; see Mathematica

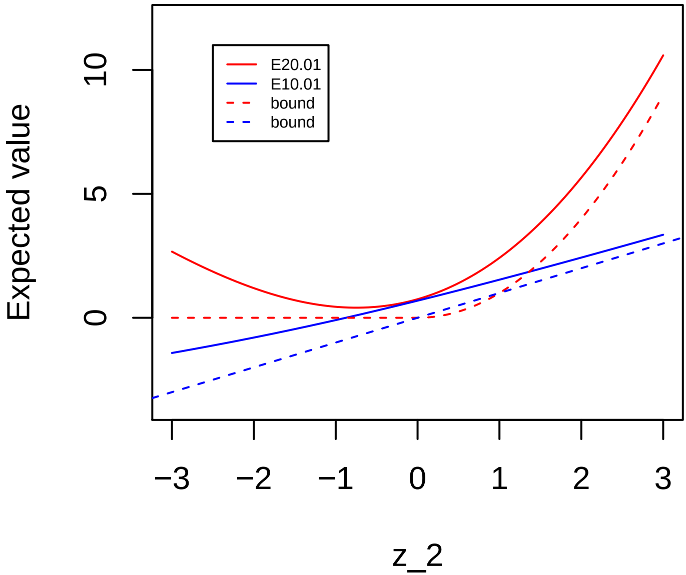
Numerical examples

$$E_{10.01} = E(Z_1 \mid Z_1 > \tau, Z_2 = \tau) \; E_{20.01} = E(Z_1^2 \mid Z_1 > \tau, Z_2 = \tau)$$

```
E10.01 <- function(y) {
    z <- y * (1 - rho)/sdet
    rho * y + sdet * dnorm(z)/pnorm(z, lower = F)
}

E20.01 <- function(y) {
    z <- (y - rho * y)/sdet
    det * (1 + y * dnorm(z)/pnorm(z, lower = F)) + rho * y
}
```

Plots $\rho = 0.5$



R package for BVN correlated Competing Risks

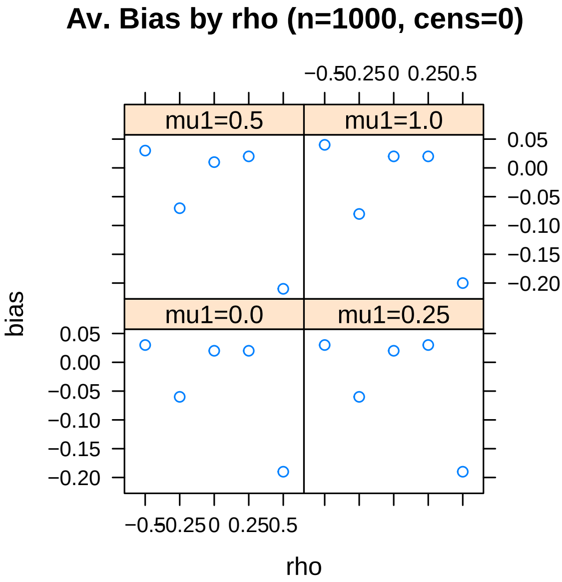
Maurizio Manuguerra [6];
Valerie Gares[7];
Malcolm Hudson[8]

- package BNC (under development)
- *Paper in preparation*

bnc package

- bivariate normal censored (linear model)
- includes code for rho fixed and test code for copula data

[6]Macquarie University
[7]Univ. Toulouse, France
[8]Macquarie University and NHMRC CTC


Av. Bias by rho (n=1000, cens=0)
mu1=0.5
mu1=1.0
mu1=0.0
mu1=0.25
bias
rho
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.5
−0.25
0
0.25
0.5

Conclusion

- Parametric fitting of BVN(μ, Σ) to log-survival data when time to an event of interest is censored by occurence of a second competing risk.
  - with $\mu = XB$ an accelerated failure time model in covariates X
  - bivariate censored linear regression
- Achieved by a novel EM algorithm
  - generalises Aitkin's univariate method
  - EM provides valuable stability in ill-posed estimation (adjusting its startpoint to the extent necessary for consistency with observed data)
  - computations are feasible (only) using Stein's identities
- Accompanying development of the R package **bnc**

Bibliography

Aitkin, M. 1981. "A Note on the Regression Analysis of Censored Data." *Technometrics* 23 (2). American Statistical Association; American Society for Quality: 161–63. `http://www.jstor.org/stable/1268032`.

Buckley, J., and I. James. 1979. "Linear Regression with Censored Data." Journal Article. *Biometrika* 66 (3): 429–36. `http://biomet.oxfordjournals.org/content/66/3/429.full.pdf`.

Crowder, M. 1991. "On the Identifiability Crisis in Competing Risks Analysis." Journal Article. *Scandinavian Journal of Statistics* 18 (3): 223–33. doi:10.2307/4616205.

Jeong, Jong-Hyeon, and Jason P. Fine. 2007. "Parametric Regression on Cumulative Incidence Function." *Biostatistics* 8 (2): 184. doi:10.1093/biostatistics/kxj040.