The Effects of Sharp Selection under Gaussian Assumptions

Assume $X, X'$ i.i.d. $N(0, 1)$. Then for $0 < c < 1$ let $s = (1 - c^2)^{1/2}$ and define

$$U = X, \quad B = cX + sX',$$

so that $U, B \sim N(0, 1)$ are jointly normal, and $c = \text{cor}(U, B)$ as well as $V[B|U] = s^2$. Think of $U$ and $B$ as unilateralism and bilateralism. Assume survival occurs exactly for $B > U$. Thus we consider $(U_s, B_s) \sim L(U, B | B > U)$. We wish to find the post-selection $\text{cor}(U_s, B_s)$ as a function of the pre-selection correlation $c = \text{cor}(U, B)$. To this end we derive the covariance and variances of $U_s$ and $B_s$. We first go after the covariance and deal with the variances later. We obtain the covariance from the variances of $B_s - U_s, B_s + U_s$, which in turn are obtained from their first and second moments. To this end we observe the following:

- $B_s + U_s \sim B + U \sim N(0, (1 + c)^2 + s^2) = N(0, 2(1 + c))$ because the truncation is on $B - U$ which is orthogonal to $B + U$.

- $B_s - U_s \sim N^+(0, (1 - c)^2 + s^2) \sim N^+(0, 2(1 - c))$.

- $B_s + U_s$ and $B_s - U_s$ are stochastically independent.

Thus we know the moments of $B_s + U_s$. For the moments of $B_s - U_s$, let $Z \sim N(0, 2(1 - c))$:

$$E[B_s - U_s] = E[Z|Z > 0] = E[|Z|] = (2/\pi)^{1/2}(2(1 - c))^{1/2} = \frac{2}{\pi^{1/2}}(1 - c)^{1/2}$$

$$E[(B_s - U_s)^2] = E[Z^2|Z > 0] = E[Z^2] = 2(1 - c)$$

For the first moment we used the following fact: $E[X|X > 0] = E[|X|] = (2/\pi)^{1/2}$, which scales up by the standard deviation. — The variances and hence covariance are:

$$V[B_s + U_s] = 2(1 + c)$$

$$V[B_s - U_s] = E[(B_s - U_s)^2] - (E[B_s - U_s])^2 = \left(2 - \frac{4}{\pi}\right)(1 - c)$$

$$C[U_s, B_s] = (V[B_s + U_s] - V[B_s - U_s])/4 = \frac{1}{\pi} + \left(1 - \frac{1}{\pi}\right)c$$

\footnote{Notation: $N^+(0, \sigma^2)$ is the conditional distribution of $Z \sim N(0, \sigma^2)$ given $Z > 0.$}
To get \( \text{cor}(U_s, B_s) \), we need the variances of \( U_s \) and \( B_s \). The first and second moments are:

\[
E[B_s] = (E[B_s + U_s] + E[B_s - U_s]) / 2 = \left( 0 + \frac{2}{\pi^{1/2}} (1 - c)^{1/2} \right) / 2 = + \left( \frac{1 - c}{\pi} \right)^{1/2}
\]

\[
E[U_s] = (E[B_s + U_s] - E[B_s - U_s]) / 2 = \left( 0 - \frac{2}{\pi^{1/2}} (1 - c)^{1/2} \right) / 2 = - \left( \frac{1 - c}{\pi} \right)^{1/2}
\]

\[
E[B_s^2 + U_s^2] = (E[(B_s + U_s)^2] + E[(B_s - U_s)^2]) / 2 = (2(1 + c) + 2(1 - c)) / 2 = 2
\]

\[
E[B_s^2 - U_s^2] = E[(B_s + U_s)(B_s - U_s)] = 0 \quad \text{(from independence)}
\]

\[
E[B_s^2] = E[U_s^2] = 1
\]

\[
V[B_s] = V[U_s] = 1 - \frac{1 - c}{\pi} = \left( 1 - \frac{1}{\pi} \right) + \frac{1}{\pi} c
\]

Thus the correlation is

\[
c_s = \text{cor}[B_s, U_s] = \frac{\frac{1}{\pi} + \left( 1 - \frac{1}{\pi} \right) c}{\left( 1 - \frac{1}{\pi} \right) + \frac{1}{\pi} c} = \frac{1 + (\pi - 1)c}{(\pi - 1) + c}
\]

Examples:

- If \( c = 0 \) (independence) before selection, then \( c_s = 1 / (\pi - 1) = 0.467 \) after selection.

- The correlation after selection is positive, \( c_s > 0 \), iff \( c > -1 / (\pi - 1) = -0.467 \).

- The maximum lift due to selection, \( c_s - c = \max_c \), is at \( c = ((\pi - 1)^2 - 1)^{1/2} - (\pi - 1) = -0.2478083 \). The maximal lift is \( c_s - c = -2 c = 0.4956166 \).
Conditional Expectation and Variance

Preparations: For conditional expectations and variances under selection we need the following definite integrals:

\[
\int_t^\infty \phi(s) ds = 1 - \Phi(t)
\]
\[
\int_t^\infty s\phi(s) ds = \phi(t)
\]
\[
\int_t^\infty s^2\phi(s) ds = t\phi(t) + 1 - \Phi(t)
\]

These can be verified using \(\phi' = -s\phi\). From these we derive the conditional means and variances after selection. The following expression will be needed repeatedly:

\[
\psi(t) = \frac{\phi(t)}{1 - \Phi(t)}
\]

This function has the following properties:

\[
\psi'(t) = \psi(t) (\psi(t) - t) \quad (\forall t), \quad t < \psi(t) < t + \frac{1}{t} \quad (\forall t > 0)
\]

Non-trivial are only the inequalities. The first follows with partial integration from this:

\[
0 < \int_t^\infty (1 - \Phi(s)) ds = -t(1 - \Phi(t)) + \int_t^\infty s\phi(s) ds = -t(1 - \Phi(t)) + \phi(t)
\]

With this in place, we have \(1 - \Phi(t) < \phi(t)/t\), so we get the second inequality:

\[
\int_t^\infty (1 - \Phi(s)) ds < \int_t^\infty \frac{\phi(s)}{s} ds < \frac{1}{t} \int_t^\infty \phi(s) ds = \frac{1 - \Phi(t)}{t}
\]

From the first inequality we see that \(\psi(t) - t > 0\), which holds for all \(t\), not just positive ones, because \(\psi(t) > 0\) always but then \(t < 0\). Hence \(\psi'(t) > 0\), which means \(\psi(t)\) is strictly ascending.

Now we also want to show that \(\psi'(t)\) is strictly ascending, which implies that \(\psi(t)\) is strictly convex. It will follow that \(\psi'(t)\) is in fact a c.d.f. because of its limit behavior near \(\pm\infty\).

The reason for introducing \(\psi(t)\) is that it will repeatedly be used for the following conditional expectations:

\[
E[X' | X' > t] = \psi(t), \quad E[X'^2 | X' > t] = t\psi(t) + 1
\]
The conditional moments of $Y$ after selection:

\[
\begin{align*}
E[Y \mid X = x, Y > x] &= E[cx + sX' \mid cx + sX' > x] \\
&= cx + sE[X' \mid X' > \frac{1-c}{s}x] \\
&= cx + s\psi\left(\frac{1-c}{s}x\right)
\end{align*}
\]

\[
\begin{align*}
E[Y^2 \mid X = x, Y > x] &= E[(cx + sX')^2 \mid cx + sX' > x] \\
&= E[c^2x^2 + 2csxX' + s^2X'^2 \mid cx + sX' > x] \\
&= c^2x^2 + 2csxE[X' \mid cx + sX' > x] + s^2E[X'^2 \mid cx + sX' > x] \\
&= c^2x^2 + 2csx\psi\left(\frac{1-c}{s}x\right) + s^2\left(\frac{1-c}{s}x\psi\left(\frac{1-c}{s}x\right) + 1\right) \\
&= c^2x^2 + 2csx\psi\left(\frac{1-c}{s}x\right) + s(1-c)x\psi\left(\frac{1-c}{s}x\right) + s^2 \\
&= c^2x^2 + s^2 + (1+c)sx\psi\left(\frac{1-c}{s}x\right)
\end{align*}
\]

\[
\begin{align*}
V[Y \mid X = x, Y > x] &= E[Y^2 \mid X = x, Y > x] - (E[Y \mid X = x, Y > x])^2 \\
&= c^2x^2 + s^2 + (1+c)sx\psi\left(\frac{1-c}{s}x\right) - \left(cx + s\psi\left(\frac{1-c}{s}x\right)\right)^2 \\
&= s^2 + (1-c)sx\psi\left(\frac{1-c}{s}x\right) - s^2\psi\left(\frac{1-c}{s}x\right)^2 \\
&= s^2\left(1 + \frac{1-c}{s}x\psi\left(\frac{1-c}{s}x\right) - \psi\left(\frac{1-c}{s}x\right)^2\right)
\end{align*}
\]

Observations:

- The conditional mean after selection bends upwards, from $E[Y \mid X = x] = cx$ near $-\infty$ to $x$ near $+\infty$. Reason: $\psi(t) = 0$ near $-\infty$; $\psi(t) \approx t$ near $+\infty$.

- The conditional variance before selection is $V[Y \mid X = x] = s^2 (\forall x)$. Because $t \mapsto \psi(t)^2 - t\psi(t)$ is a c.d.f., the conditional variance after selection, $V[Y \mid X = x, Y > x]$, is $s^2$ near $-\infty$ and descends to $0$ near $+\infty$.

\[^2\text{We let the angle run from 0 to } \pi, \text{ so that } c \text{ runs from } +1 \text{ to } 0 \text{ to } -1, \text{ and } s \text{ swings from } 0 \text{ to } +1 \text{ to } 0. \text{ Thus } s \text{ is always positive and biases the mean upwards ever more the closer } x \text{ is to } +\infty.\]
Figure 1: Effects of selection according to $B > U$ on association and conditional variance.  
Left: Correlation between $B$ (degree of bilateralism) and $U$ (degree of unilaterism), before and after selection. The graph shows the lift of the correlation caused by selection.  
Right: Graph of the conditional standard deviation of $B$ given $U = u$ and $B > u$ (selection) for $\rho = +0.5, 0, -0.5$.  

\[ \rho = \text{corr}(U,B) \text{ before Selection} \]  
\[ \rho_s = \text{corr}(U,B) \text{ after Selection} \]  
\[ U = u \quad V[\{B \mid U = u, B > u\}]^{\frac{1}{2}} \]  
\[ \rho = +0.5 \quad \rho = 0 \quad \rho = -0.5 \]
Figure 2: Effects of selection according to $B > U$ conditional variance. Same graph as right hand of previous graph, except that the vertical axis is not in multiples of $\sigma = (1 - \rho^2)^{1/2}$ but in absolute terms.