HOA NGUYEN'S PROPOSAL TECHNICAL APPENDIX

1 Simulating correlated random variables

The technique presented in this section can be used to generate Bernoulli random variables which are correlated. Imagine that SNP's in a block are highly correlated with each other but they have low correlation with SNP's outside the block. We can then generate Bernoulli random variables which resemble the structure of haplotype blocks in a genomic region.

In the proposal document, the data are simulated using a modified version of Hudson (2002)'s MS program. The technique presented here can be used as an alternate method for generating block-like haplotypic data. In the derivation below, X_j and X_k represent SNP's. Let ρ_{jk} be the correlation between X_j and X_k. Assume that $X_i \sim Bernoulli(p_i)$. Given X_k, we can generate X_j as follows.

$$Cor(X_k, X_j) = \frac{Cov(X_j, X_k)}{\sqrt{Var(X_j)Var(X_k)}}$$

Since $X_i \sim Bernoulli(p_i)$, $Cov(X_j, X_k) = \rho_{jk}\sqrt{p_j(1-p_j)p_k(1-p_k)}$. Furthermore, $Cov(X_j, X_k) = E[(X_j - \mu_j)(X_k - \mu_k)] = E(X_jX_k) - \mu_j\mu_k = P(X_j = 1, X_k = 1) - p_jp_k$. Thus,

$$\rho_{jk}\sqrt{p_j(1-p_j)p_k(1-p_k)} = P(X_j = 1|X - k = 1)P(X_k = 1) - p_jp_k$$

$$\begin{aligned}
P(X_j = 1|X_k = 1) &= \frac{\rho_{jk}\sqrt{p_j(1-p_j)p_k(1-p_k)} + p_jp_k}{p_k} \\
P(X_j = 0|X_k = 1) &= 1 - P(X_j = 1|X_k = 1) \\
P(X_j = 1|X_k = 0) &= \frac{P(X_k = 0|X_j = 1)P(X_j = 1)}{P(X_k = 0)} \\
&= \frac{[1 - P(X_k = 1|X_j = 1)]P(X_j = 1)}{P(X_k = 0)} \\
&= \frac{[1 - P(X_j = 1|X_jk = 1)P(X_k = 1)/P(X_j = 1)]P(X_j = 1)}{P(X_k = 0)} \\
&= \frac{[1 - P(X_j = 1|X_k = 1)p_k/p_j]p_j}{1 - p_k} \\
P(X_j = 0|X_k = 0) &= 1 - P(X_j = 1|X_k = 0)
\end{aligned}$$

I use a base variable $X_k \sim Bernoulli(p_k)$ and simulate $X_{11}, \ldots, X_{1b}, \ldots, X_{b1}, \ldots, X_{bb}$ such that variables in the same block have the same correlation $\rho_{jk}, j \in \{1, b\}$ with X_k (b is the number of blocks) and within a block, the variables have the same correlation $\rho_j, j \in \{1, b\}$. The simulation of X_j does not require ρ_j; ρ_j is different from ρ_{jk}.

Based on the calculations above, there are constraints imposed upon ρ_{ij} and p_i's.

2 Marginal models give biased β's

2.1 Bias for the no-intercept model

When marginal models are considered, the parameter estimates are biased. Furthermore, when the explanatory variables are not fixed, the variance of the parameter estimates is inflated.

$$
\begin{array}{lll}
\text{Full Model} & : & Y_i = \sum_{j=1}^m \beta_j X_{ij} + \epsilon_i \\
 & & E(Y_i) = \sum_{j=1}^m \beta_j X_{ij} \\
 & & X_{ij} \sim Bernoulli(p_j) \\
 & & \epsilon_i \sim Normal(0, \sigma) \\
\text{Marginal Model} & : & Y_i = \beta_j X_{ij} + \epsilon_i
\end{array}
$$

<u>Bias Calculation</u>

For each marginal model:

$$\hat{\beta}_j = \frac{\sum X_{ij} Y_i}{\sum X_{ij}^2} = \frac{\sum X_{ij} Y_i}{\sum X_{ij}}$$

For simplicity, assume $\beta_j = \beta\, or\, 0$, $p_j = p \,\forall j$. Let $Q = \{j : \beta_j = \beta\}$, $Q^c = \{j : \beta_j = 0\}$, $|Q| = N$.

$$
\begin{aligned}
E(\hat{\beta}_j) &= E(E(\hat{\beta}_j | X)) \\
&= E(E(\frac{\sum_i X_{ij} Y_i}{\sum_i X_{ij}} | X)) \\
&= E(\frac{\sum_i X_{ij} \sum_k \beta_k X_{ki}}{\sum_i X_{ij}}) \\
&= E(\frac{\sum_i X_{ij} \beta_j X_{ij} + \sum_i X_{ij} \sum_{k \neq j} \beta_k X_{ik}}{\sum_i X_{ij}}) \\
&= \beta_j + \frac{E(\sum_i \sum_{k \neq j} X_{ij} X_{ik} \beta_k)}{\sum_i X_{ij}}
\end{aligned}
$$

Thus the bias of β_j is $\dfrac{E(\sum_i \sum_{k \neq j} X_{ij} X_{ik} \beta_k)}{\sum_i X_{ij}}$ for $k \neq j$. If the X_j's are independent, for $j \notin Q$: $\beta_j = 0 \rightarrow E(\hat{\beta}_j) = \dfrac{nNp^2\beta}{np} = N\beta p$. For $j \in Q$: $\beta_j = \beta \rightarrow E(\hat{\beta}_j) = \beta + (N-1)\beta p =$

$\beta(1 - p) + N\beta p$.

Variance Calculation

The calculations below are based on the assumption that X_j's are $\perp$.

$$Var(\hat{\beta}_j) = Var(E(\hat{\beta}_j|X)) + E(Var(\hat{\beta}_j|X))$$

$$\begin{aligned}
E(\hat{\beta}_i|X) &= \beta_j + \frac{(\sum_i \sum_{k\neq j} X_{ij}X_{ik}\beta_k)}{np_j} \\
Var(E(\hat{\beta}_j|X)) &= Var\left(\frac{(\sum_i \sum_{k\neq j} X_{ij}X_{ik}\beta_k)}{np_j}\right) \\
&= Var\left(\frac{U}{V}\right) = E\left(\frac{U^2}{V^2}\right) - E^2\left(\frac{U}{V}\right)
\end{aligned}$$

$U^2 = (\sum_i \sum_{k\neq j} \beta_k X_{ij}X_{ik})^2$. For $j \notin Q$, $\beta_j = 0 \rightarrow$ there are N terms in $\sum_{k\neq j}$. When calculating $E(U^2)$, essential information is the # of terms ending up having p^2, p^3, *or* p^4 after the expectation is taken. In particular, there are nN p^2 terms (the squared terms $X_{ij}^2X_{ik}^2 = X_{ij}X_{ik}$). Now consider the cross terms, there are $\binom{Nn}{2}$ terms in total. p^3 terms must have the i index fixed and k index varies. Thus there are $n\binom{N}{2}$ p^3 terms. The remaining, $\binom{Nn}{2} - n\binom{N}{2}$ would be p^4 terms.

$$\begin{aligned}
E(U^2) &= nN\beta^2p^2 + nN(N-1)\beta^2p^3 + N^2n(n-1)\beta^2p^4 \\
E(V^2) &= E[(\sum_i X_{ij})^2] = np + n(n-1)p^2 \\
E\left(\frac{U^2}{V^2}\right) &\approx \frac{nN\beta^2p^2 + nN(N-1)\beta^2p^3 + N^2n(n-1)\beta^2p^4}{np + n(n-1)p^2} \\
Var(E(\hat{\beta}_j|X))_{j\notin Q} &= E\left(\frac{U^2}{V^2}\right) - N^2\beta^2p^2 \approx \frac{N\beta^2p(1-p)}{1+(n-1)p}
\end{aligned}$$

Similar calculation yields:

$$Var(E(\hat{\beta}_j|X))_{j\in Q} = \frac{(N-1)\beta^2p(1-p)}{1+(n-1)p}$$

$$\begin{aligned}
Var(\hat{\beta}_j|X) &= Var(\frac{\sum_i X_{ij}Y_i}{\sum_i X_{ij}}) \\
&= \frac{\sigma^2\sum_i X_{ij}}{(\sum_i X_{ij})^2} = \frac{\sigma^2\sum_i X_{ij}}{\sum_i X_{ij} + 2\sum_{j\neq l} X_{ij}X_{il}}
\end{aligned}$$

$$E(Var(\hat{\beta}|X)) = \frac{\sigma^2 np}{np + n(n-1)p^2} = \frac{\sigma^2}{1 + (n-1)p}$$

$$\downarrow$$

$$Var(\hat{\beta}_j)_{j \notin Q} = \frac{N\beta^2 p(1-p) + \sigma^2}{1 + (n-1)p}$$

$$Var(\hat{\beta}_j)_{j \in Q} = \frac{(N-1)\beta^2 p(1-p) + \sigma^2}{1 + (n-1)p}$$

Again the given formulae for the bias and the variance in this section are based on the assumption that the X_j's do not have any correlation.

2.2 Bias for the intercept model

Full Model : $Y_i = \sum_{j=1}^{M} \beta_j X_{ij} + \epsilon_i$

$E(Y_i) = \sum_{j=1}^{M} \beta_j X_{ij}$

$X_{ij} \sim Bernoulli(p_j)$

$\epsilon_j \sim Normal(0, \sigma)$

Marginal Model : $Y_i = \beta_{j0} + \beta_{j1} X_{ij} + \epsilon_i$

Least squared estimates : $\hat{\beta}_{1j} = \frac{\sum_i (X_{ji} - \overline{X_j})(Y_i - \overline{Y})}{\sum_i (X_{ji} - \overline{X_j})^2} = \frac{\sum_i Y_i (X_{ji} - \overline{X_j})}{\sum_i (X_{ji} - \overline{X_j})^2}$

$\hat{\beta}_{0j} = \overline{Y} - \hat{\beta}_{1j}\overline{X_j}$

(i) **Assuming the X_j's are independent:**

$$\begin{aligned}
\hat{\beta}_{1j} &= \frac{\sum_i (X_{ji} - \overline{X}_j)(Y_i - \overline{Y})}{\sum_i (X_{1j} - \overline{X}_j)^2} \\
&= \frac{\sum_i (X_{ji}Y_i - \overline{X}_j Y_i - X_{ji}\overline{Y} + \overline{X}_j \overline{Y})}{\sum_i (X_{ji} - \overline{X}_j)^2} \\
&= \frac{\sum_i (X_{ji}Y_i - (1/n \sum_i X_{ji})Y_i - (1/n \sum_i Y_i)X_{ji} + 1/n^2 \sum_i X_{ji} \sum_i Y_i)}{\sum_i (X_{ji}^2 - 2X_{ji}\overline{X}_j + \overline{X}_j^2)} \\
&= \frac{\sum_i X_{ji}Y_i - 1/n \sum_i X_{ji} \sum_i Y_i}{\sum_i X_{ji} - 2/n(\sum_i X_{ji})^2 + 1/n(\sum_i X_{ji})^2} \\
&= \beta_j + \frac{\sum_i X_{ji}\beta_0 + \sum_i \sum_{k \# j} X_{ji}\beta_k X_{ki} - 1/n \sum_i X_{ji} \sum_{k \# j} \beta_k X_{ki} - 1/n \sum_i X_{ji}\beta_0}{\sum_i X_{ji} - 1/n(\sum_i X_{ji})^2} \\
E(\hat{\beta}_{1j})_{j \in Q} &\approx \frac{\beta_0(\frac{(n-1)}{n}np) + nN\beta p^2 - \frac{1}{n}nN\beta p^2}{np - 1/n(np + 2n(n-1)p^2)} \\
&= \frac{\beta_0 + N\beta p}{1 - 2p}
\end{aligned}$$

$$
\begin{aligned}
E(\hat{\beta}_{1j})_{j\in Q} &\approx \beta + \frac{\beta_0 + (N-1)\beta p}{1-2p} \\
&= \frac{\beta_0 + N\beta p}{1-2p} + \beta - \frac{\beta p}{1-2p} \\
&= \frac{\beta_0 + N\beta p}{1-2p} + \beta\frac{1-3p}{1-2p}
\end{aligned}
$$

(ii) **Assuming the X_j's have a correlation structure:**

$$
\begin{aligned}
\hat{\beta}_k &= \frac{\sum_i Y_i(X_{ik} - \overline{X}_k)}{\sum_i (X_{ik} - \overline{X}_k)^2} \\
&= \beta_k + \frac{\sum_i \sum_{j\neq k} \beta_j X_{ij}(X_{ik} - \overline{X}_k)}{\sum_i (X_{ik} - \overline{X}_k)} + \frac{\sum_i \epsilon_i (X_{ik} - \overline{X}_k)}{\sum_i (X_{ik} - \overline{X}_k)^2} \\
&= \beta_k + \frac{\sum_{j\neq k} \beta_j \frac{1}{n} \sum_i (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)}{\frac{1}{n}\sum_i (X_{ik} - \overline{X}_k)^2} + \frac{\sum_i (X_{ik} - \overline{X}_k)}{\sum_i (X_{ik} - \overline{X})^2} \\
E(\hat{\beta}_k) &= \beta_k + \frac{\sum_{j\neq k} Cov(X_k, X_j)}{Var(X_k)} \\
Bias_k &= \frac{\sum_{j\neq k} Cov(X_k, X_j)}{p_k(1-p_k)} \\
&= \sum_{j\neq k} \beta_j Corr(X_k, X_j) \sqrt{\left[\frac{p_j(1-p_j)}{p_k(1-p_k)}\right]} \\
&= \frac{1}{\sigma_k} \sum_{j\neq k} \beta_j \rho_{jk} \sigma_j
\end{aligned}
$$

3 Some notes on the kernel

Normal Reference Rule Bandwidth for estimating the derivatives using the kernel

Optimal bandwidth for estimating the derivatives using the kernel (Fan and Gijbels, 1996):

$$
h_{\nu,opt} = (2\mu + 1)^{\frac{1}{2\nu+5}} \alpha_\nu(K) \left(\int (f^{(\nu+2)}(x))^2 dx \right)^{-\frac{1}{2\nu+5}} n^{-\frac{1}{2\nu+5}}
$$

where $\alpha_\nu(K) = \left(\int u^2 K(u) du\right)^{-2/(2\nu+5)} \left(\int (K^{(\nu)}(u))^2 du\right)^{1/(2\nu+5)}$ These following normal reference bandwidths are computed using the Gaussian kernel.

- for $\hat{f}'$:

$$
h^* = 0.9686 \sigma n^{-1/7}
$$

- for $\hat{f}''$:

$$h^* = 0.9398\sigma n^{-1/9}$$

- for $\hat{f}^{(3)}$:

$$h^* = 0.9289\sigma n^{-1/11}$$

Using these bandwidths, the estimation of the derivatives of a symmetric curve is excellent (simulation results not shown).

Exploration of different gamma distributions with various curvature at the mode show that the kernel does quite well in estimating the third derivative at the mode.

4 Efron's local estimators of the mode and the spread

4.1 bias of the mode

Using a kernel density estimation, bias of the mode $= -\dfrac{h^2\sigma_k^2 f^{(3)}(\theta)}{2f^{(2)}(\theta)}$ where $\sigma_k^2 = \int wK(w)dw$ and θ is the true mode. If we assume that the underlying distribution is the $N(\mu, \sigma)$, the bias of the mode is:$= -\dfrac{h^2\sigma_k^2 \phi^{(3)}(\theta)}{2\phi^{(2)}(\theta)}$

4.2 Investigating $\frac{d^2}{dz^2}log(\hat{f}(z))$, bias of the spread

General setting for kernel density estimators: $\hat{f} = \dfrac{1}{n}\sum_i K_h(X_i - x) = \dfrac{1}{nh}\sum_i K(\dfrac{X_i - x}{h})$.

$$\begin{aligned}
\hat{l} &= log\hat{f} \\
\hat{l}' &= \frac{\hat{f}'}{\hat{f}} = \frac{1/n\sum_i K_h'(X_i - x)}{1/n\sum_i K_h(X_i - x)} \\
\hat{l}'' &= \frac{\hat{f}''\hat{f} - (\hat{f}')^2}{(\hat{f})^2} = func(\sum K_h, \sum K_h', \sum K_h'')
\end{aligned}$$

Partial derivatives:

$$\left(\frac{\partial}{\partial \hat{f}}\hat{l}'', \frac{\partial}{\partial \hat{f}'}\hat{l}'', \frac{\partial}{\partial \hat{f}''}\hat{l}''\right)$$

Define $G = \dfrac{w(x)u(x) - v^2(x)}{u^2(x)}$, where $u(x) = \hat{f}$, $v(x) = \hat{f}'$ and $w(x) = \hat{f}''$. The distribution of G depends on the distributions of (u, v, w). $\left(\dfrac{\partial}{\partial u}G, \dfrac{\partial}{\partial v}G, \dfrac{\partial}{\partial w}G\right) = \left(\dfrac{-wu + 2v^2}{u^3}, -\dfrac{2v}{u^2}, \dfrac{1}{u}\right)$. Let $\partial G_0 = \left(\dfrac{\partial}{\partial u}G, \dfrac{\partial}{\partial v}G, \dfrac{\partial}{\partial w}G\right)_{\mu_u, \mu_v, \mu_w}$

$$G \sim N\left(\partial G_0^T \begin{pmatrix} \mu_u \\ \mu_v \\ \mu_w \end{pmatrix}, \partial G_0^T \sum \partial G_0\right)$$

I will need the distribution of $(\hat{f}, \hat{f}', \hat{f}'')$, or equivalently, the distribution of $(\sum K_h, \sum K_h', \sum K_h'')$.

$$\hat{f}(x) = \frac{1}{n}\sum_i K_h(X_i - x) \sim AN\left(f(x) + \frac{1}{2}\sigma_K^2 h^2 f''(x), \frac{f(x)R(K)}{nh}\right) \quad (1)$$

$$\hat{f}'(x) = \frac{1}{n}\sum_i K_h'(X_i - x) \sim AN\left(f'(x) + \frac{1}{2}\sigma_K^2 h^2 f^{(3)}(x), \frac{f(x)R(K')}{nh^3}\right) \quad (2)$$

$$\hat{f}''(x) = \frac{1}{n}\sum_i K''(X_i - x) \sim AN\left(f'' + \frac{1}{2}\sigma_K^2 h^2 f^{(4)}(x), \frac{f(x)R(K'')}{nh^5}\right) \quad (3)$$

Now, I need the covariances, $cov(\hat{f}, \hat{f}')$, $cov(\hat{f}, \hat{f}'')$, $cov(\hat{f}', \hat{f}'')$.

$$\begin{aligned}
cov(\hat{f}, \hat{f}') &= E\left[(\hat{f}(x) - E\hat{f})(\hat{f}'(x) - E\hat{f}')\right] \\
&= E\left(\hat{f}\hat{f}'\right) - E\hat{f}E\hat{f}' \\
cov(\hat{f}, \hat{f}'') &= E\left[(\hat{f}(x) - E\hat{f})(\hat{f}''(x) - E\hat{f}'')\right] \\
&= E\left(\hat{f}\hat{f}''\right) - E\hat{f}E\hat{f}'' \\
cov(\hat{f}', \hat{f}'') &= E\left[(\hat{f}'(x) - E\hat{f}')(\hat{f}''(x) - E\hat{f}'')\right] \\
&= E\left(\hat{f}'\hat{f}''\right) - E\hat{f}'E\hat{f}''
\end{aligned}$$

$$\begin{aligned}
E\hat{f}\hat{f}' &= \frac{1}{n^2}E\left(\sum_i K_h(X_i - x)\sum_i K_h'(X_i - x)\right) \\
&= \frac{1}{n^2}\left[E(K_h(X_i - x)K_h'(X_i - x)) + \sum_{i \neq j} E(K_h(X_i - x)K_h'(X_i - x))\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}E(K_hK_h') + \frac{\binom{n}{2}}{n^2}E(K_h)E(K_h') \\
&= \frac{1}{n}E(K_hK_h') + \frac{n-1}{n}E(K_h)E(K_h') \\
&= O\left(\frac{1}{n}\right) + \frac{n-1}{n}E(K_h)E(K_h')
\end{aligned}
$$

$$
\begin{aligned}
E\hat{f}\hat{f}'' &= \frac{1}{n^2}E\left(\sum_i K_h(X_i-x)\sum_i K_h''(X_i-x)\right) \\
&= \frac{1}{n^2}\left[E(K_h(X_i-x)K_h''(X_i-x)) + \sum_{i\neq j} E(K_h(X_i-x)K_h''(X_i-x))\right] \\
&= \frac{1}{n}E(K_hK_h'') + \frac{\binom{n}{2}}{n^2}E(K_h)E(K_h'') \\
&= \frac{1}{n}E(K_hK_h'') + \frac{n-1}{n}E(K_h)E(K_h'') \\
&= O\left(\frac{1}{n}\right) + \frac{n-1}{n}E(K_h)E(K_h'')
\end{aligned}
$$

$$
\begin{aligned}
E\hat{f}'\hat{f}'' &= \frac{1}{n^2}E\left(\sum_i K_h'(X_i-x)\sum_i K_h''(X_i-x)\right) \\
&= \frac{1}{n^2}\left[E(K_h'(X_i-x)K_h''(X_i-x)) + \sum_{i\neq j} E(K_h'(X_i-x)K_h''(X_i-x))\right] \\
&= \frac{1}{n}E(K_h'K_h'') + \frac{\binom{n}{2}}{n^2}E(K_h')E(K_h'') \\
&= \frac{1}{n}E(K_h'K_h'') + \frac{n-1}{n}E(K_h')E(K_h'') \\
&= O\left(\frac{1}{n}\right) + \frac{n-1}{n}E(K_h')E(K_h'')
\end{aligned}
$$

$$
(\hat{f}, \hat{f}', \hat{f}'') \sim AN_3\left[\begin{pmatrix} f(x) + \frac{1}{2}\sigma_K^2 h^2 f''(x) \\ f'(x) + \frac{1}{2}\sigma_K^2 h^2 f^{(3)}(x) \\ f''(x) + \frac{1}{2}\sigma_K^2 h^2 f^{(4)}(x) \end{pmatrix}, \sum\right]
$$

$$\sum = \begin{bmatrix} \frac{f(x)R(K)}{nh} & 0 & 0 \\ \cdots & \frac{f(x)R(K')}{nh^3} & 0 \\ \cdots & \cdots & \frac{f(x)R(K'')}{nh^5} \end{bmatrix}$$

Recall that:

$$\begin{aligned} l &= log(f) \\ l' &= \frac{f'}{f} \\ l'' &= \frac{f''f - f'^2}{f^2} \end{aligned}$$

Similarly, with $\hat{f}$ as the kernel estimator,

$$\begin{aligned} \hat{l} &= log(\hat{f}) \\ \hat{l}' &= \frac{\hat{f}'}{\hat{f}} \\ \hat{l}'' &= \frac{\hat{f}''\hat{f} - \hat{f}'^2}{\hat{f}^2} \end{aligned}$$

We would like to know what is the bias for estimating σ^2 using $\frac{d^2}{d} log(\hat{f})$. From the distribution of $(\hat{f}, \hat{f}', \hat{f}'')$, we obtain:

$$\begin{aligned} E(\hat{l}'') &= \frac{-(f'' + 1/2\sigma_K^2 h^2 f^{(4)})(f' + 1/2\sigma_K^2 h^2 f^{(3)}) + 2(f' + 1/2\sigma_K^2 h^2 f^{(3)})}{(f + 1/2\sigma_K^2 h^2 f'')^2} \cdot \\ &\quad \cdot (f + 1/2\sigma_K^2 h^2 f'') - \frac{2(f' + 1/2\sigma_K^2 h^2 f^{(3)})^2}{(f + 1/2\sigma^2 h^2 f'')^2} + \frac{f'' + 1/2\sigma_K^2 h^2 f^{(4)}}{f + 1/2\sigma_K^2 h^2 f''} \\ &= \frac{-(f'' + 1/2\sigma_K^2 h^2 f^{(4)})(f' + 1/2\sigma_K^2 h^2 f^{(3)}) + (f'' + 1/2\sigma_K^2 h^2 f^{(4)})(f + 1/2\sigma_K^2 h^2 f'')}{(f + 1/2\sigma_K^2 h^2 f'')^2} \\ &= \frac{(f'' + 1/2\sigma_K^2 h^2 f^{(4)})((f - f') + 1/2\sigma_K^2 h^2 (f'' - f^{(3)}))}{(f + 1/2\sigma_K^2 h^2 f'')^2} \\ &= \frac{\frac{f'' - f'^2}{f^2} + \frac{f'^2}{f^2} + \frac{1/2\sigma_K^2 h^2 (f^{(4)}(f - f') + f''(f'' - f^{(3)}))}{f^2} + \frac{1/4\sigma_K^4 h^4 f^{(4)}(f'' - f^{(3)})}{f^2}}{\left(1 + 1/2\sigma_K^2 h^2 \frac{f''}{f}\right)^2} \\ &= \frac{\frac{f'' - f'^2}{f^2} + \frac{f'^2}{f^2} + \frac{1/2\sigma_K^2 h^2 (f^{(4)}(f - f') + f''(f'' - f^{(3)}))}{f^2} + \frac{1/4\sigma_K^4 h^4 f^{(4)}(f'' - f^{(3)})}{f^2}}{\left(1 + \sigma_K^2 h^2 \frac{f''}{f} + 1/4\sigma_K^4 h^4 \frac{f''^2}{f^2}\right)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{f'' - f'^2}{f^2} + \frac{f'^2}{f^2} + o(h^2) \\
&= l'' + \frac{f'^2}{f^2} + o(h^2)
\end{aligned}
$$

5 FDR under different variants of G_0

$$
\begin{aligned}
FDP(z) &= \frac{\sum_i (1-H_i) I(Z_i < z)}{\sum_i I(Z_i < z)} \\
FDR(z) = E(FDP(z)) &\approx \frac{E(1/m \sum_i (1-H_i) I(Z_i < z))}{E(1/m \sum_i I(Z_i < z))} \\
&\approx \frac{(1-a)G_0(z)}{G(z)} \approx \frac{(1-a)G_0(z)}{\widehat{G(z)}} = \widehat{R(z)}
\end{aligned}
$$

5.1 biased null distribution is Normal(μ, σ)

Now, assuming that we know the biased null G_0 is N(α, σ^2), denote this distribution by Φ_*. Then

$$
FDR(z) \approx \frac{(1-a)\Phi_*(z)}{G(z)} \approx \frac{(1-a)\Phi_*(z)}{\widehat{G(z)}} \le \frac{\Phi_*(z)}{\widehat{G(z)}}
$$

Recall that z_* is such that $\frac{\Phi(z_*)}{G(z_*)} = \alpha$. Now, actual biased null is $N(\mu, \sigma^2)$, denoted this distribution by Φ_*.

$$
\begin{aligned}
FDR &= \frac{(1-a)G_0(z_*)}{G(z_*)} \\
G_0(z_*) &= \Phi_*(z_*) \\
G(z_*) &= (1-a)\Phi_*(z_*) + aG_1(z_*)
\end{aligned}
$$

Our goal is to get FDR as a function of (μ, σ) and investigate this rate function when there are biases in the estimation of μ and σ. To do that, we first get an expression of the rate function, then carry out a Taylor series expansion for the rate function around $\mu = 0$ and $\sigma = 1$.

$$
\begin{aligned}
\Phi(z_*) &= \int_{-\infty}^{z_*} \frac{1}{\sqrt{2\Pi}} e^{-x^2/2} dx \\
\Phi_*(z_*) &= \int_{-\infty}^{z_*} \frac{1}{\sqrt{2\Pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\frac{z_*-\mu}{\sigma}} \frac{1}{\sqrt{2\Pi}} e^{-t^2/2} dt \\
&= \Phi(\frac{z_*-\mu}{\sigma}) \\
G(z_*) &= (1-a)\Phi_*(z_*) \; + \; aG_1(z_*) \\
&= (1-a)\Phi(\frac{z_*-\mu}{\sigma}) \; + \; aG_1(z_*) \\
r(\cdot) &= \frac{(1-a)\Phi_*(z_*)}{(1-a)\Phi_*(z_*) \; + \; aG_1(z_*)} \\
&= \frac{(1-a)\Phi(\frac{z_*-\mu}{\sigma})}{(1-a)\Phi(\frac{z_*-\mu}{\sigma}) \; + \; aG_1(z_*)}
\end{aligned}
$$

Since $\dfrac{1}{\alpha} \; = \; (1-a) \; + \; a\dfrac{G_1(z_*)}{\Phi(z_*)}$, $G_1(z_*) \; = \; (\dfrac{1}{\alpha} - 1 + a)\dfrac{1}{a}\Phi(z_*)$. We have,

$$
\begin{aligned}
\frac{1}{r} &= 1 \; + \; \frac{a}{1-a}(\frac{1}{\alpha} - 1 + a)\frac{1}{a}\frac{\Phi(z_*)}{\Phi(\frac{z_*-\mu}{\sigma})} \\
r &= \frac{(1-a)\Phi(\frac{z_*-\mu}{\sigma})}{(1-a)\Phi(\frac{z_*-\mu}{\sigma}) + (\frac{1}{\alpha} - 1 + a)\Phi(z_*)}
\end{aligned}
$$

$r = (1-a)\alpha$ if the null is N(0,1).

Bias function:

$$
B(z, \mu, \sigma) \; = \; \frac{(1-a)\Phi(\frac{z-\mu}{\sigma})}{(1-a)\Phi(\frac{z-\mu}{\sigma}) \; + \; (\frac{1}{\alpha} - 1 + a)\Phi(z)} \; - \; \alpha
$$

Now expand the bias function about $(\mu_0, \sigma_0) = (0, 1)$. First partial derivatives:

$$
B_\mu \; = \; \frac{(1-a)(-\frac{1}{\sigma}(\frac{1}{\alpha} - 1 + a)\phi(\frac{z-\mu}{\sigma})\Phi(z)}{[(1-a)\Phi(\frac{z-\mu}{\sigma}) + (\frac{1}{\alpha} - 1 + a)\Phi(z)]^2}
$$

$$B_\sigma \;=\; \frac{(1-a)(-\frac{z-\mu}{\sigma})(\frac{1}{\alpha}-1+a)\phi(\frac{z-\mu}{\sigma})\Phi(z)}{[(1-a)\Phi(\frac{z-\mu}{\sigma})+(\frac{1}{\alpha}-1+a)\Phi(z)]^2}$$

First order Taylor expansion:

$$B(z,\mu,\sigma) \;=\; -a\alpha+\alpha^2\frac{(1-a)(-1)(\frac{1}{\alpha}-1+a)\phi(z)}{\Phi(z)}\mu+\alpha^2\frac{(1-a)(-z)(\frac{1}{\alpha}-1+a)\phi(z)}{\Phi(z)}(\sigma-1)+R(a,\alpha,z)$$

5.2 Biased distribution is a Skewed Normal $SN(\lambda)$

Azzalini (1985) introduced a class of skew-normal distributions which allows the presence of skewness in the normal distribution. A random variable Z is said to have a skew-normal distribution with parameter λ if its density is:

$$\varphi(z;\lambda) \;=\; 2\phi(z)\Phi(\lambda z)$$

where ϕ and Φ are the standard normal density and distribution functions.

Carrying out similar calculations. Rate function and its approximation are:

$$r(\lambda) \;=\; \frac{(1-a)\int_{-\infty}^{z_*}2\phi(z)\Phi(\lambda z)dz}{(1-a)\int_{-\infty}^{z_*}2\phi(z)\Phi(\lambda z)dz+(\frac{1}{\alpha}-1+a)\Phi(z_*)}$$

Expanding the rate function about $\lambda=0$. First order approximation:

$$r(\lambda) \;=\; (1-a)\alpha \;+\; \alpha^2\frac{(1-a)(\frac{1}{\alpha}-1+a)\frac{1}{a}\sqrt{\frac{2}{\pi}}\int_{-\infty}^{z_*}z\phi(z)dz}{\Phi(z_*)}\lambda \;+\; R(a,\alpha,z_*)$$

If the biased distribution of the nulls is a skew-normal instead of the normal, the rate is sensitive to the choice of z_* and λ

6 Simulated z-values directly

In the general setting (as described in the proposal document), we obtain z-values from a transformation of p-values: $Z_i=\Phi^{(-1)}(P_i)$ where P_i's are p-values obtained from testing particular β's in the regression model. In this section, we simulate z-values directly from a known skew distribution. We would like understand better both the parametric and non-parametric estimators of the skewness parameter in the distribution of nulls before we apply the methods to z-values obtained from the regression model.

Suppose that null z-values follow a Gamma distribution; and suppose that alternative z-values follow a normal distribution.

$$G(z) = (1-a)Gamma(\alpha, \beta) + aN(\mu, \sigma).$$

Figure 2 shows an exemplary distribution of a mixture of gamma's and normal's.

6.1 Global estimation of skewness parameter using the skew normal

Figure 3 shows the empirical FDR obtained from 300 simulations. $G(z) = (1-a)Gamma(3,1) + aN(-10,1)$; a=0.1. Using the default $N(0,1)$ to calculate FDR (i.e., no correction) leads to an average empirical FDR of .73. Efron's location-scale correction reduces the empirical FDR to an average of .41. The skew-normal correction gives substantial reduction in FDR control: .18 on average compared to the previous two figures.

6.2 Drawbacks of local method

Understanding the 3^{rd} derivative of the Gamma

Gamma density:

$$f(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$$\begin{aligned}
l(x) &= \alpha log\beta - log\Gamma(\alpha) + (\alpha-1)logx - \beta x \\
l'_x &= \frac{\alpha-1}{x} - \beta \\
l''_x &= -\frac{\alpha-1}{x^2} \\
l'''_x &= 2\frac{(\alpha-1)}{x^3}
\end{aligned}$$

Setting l'_x to 0, the mode $\mu = \frac{\alpha-1}{\beta}$. The third derivative at the mode is then $l'''_x(\mu) = \frac{2\beta^3}{(\alpha-1)^2}$. Judging from this calculation, if β is small (e.g. .1), the third derivative at the mode is close to 0 even though the distribution is asymmetrical.

Judging from the above theoretical derivatives, if we want to pursue further the local method, we should use both the third and the second derivatives. Figure 1 shows a skew distribution:

Gamma(5,.1), but the third theoretical derivative at the mode is quite small: 0.000125. The local method cannot detect skewness by judging the third derivative. The variance in this example is large, which explains why the second derivative is important in estimating the skewness.

Third derivative of the log-likelihood at the mode

Efron (2004) used the second derivative of the log-likelihood at the mode as a measure of spread. As explained above, we need to use both the 2^{nd} and the 3^{rd} derivatives at the mode as a measure of skewness: $\frac{logf^{(3)}}{logf^{(2)}}$. As shown previously, estimating the second derivative of the log-likelihood at the mode requires estimates of the first 3 derivatives of f. Consequently, the third derivative of the log-likelihood at the mode requires the first 4 derivatives of f. This local procedure undoubtedly leads to considerably biased estimate of the skewness.

Power transformation

We carry out the following steps to choose a power transformation for the Z_i's:

(i) Standardize the Z_i's: $W_i = \frac{Z_i - \hat{\mu}_Z}{\hat{\sigma}_Z}$

(ii) For a power transform γ:

$$V_{i,\gamma} = \begin{cases} -|W_i|^\gamma, & W_i < 0 \\ W_i^\gamma, & W_i \geq 0 \end{cases}$$

(iii) Obtain the skewness $\lambda(\gamma)$ for the $V_{i,\gamma}$'s using the local derivatives.

(iv) Choose γ such that $\lambda(\gamma)$ is close to 0.

(v) Once the skewness is removed, i.e. the set $V_{i,\gamma}$ is specified, perform Efron's location-scale correction for the distribution of $V_{i,\gamma}$.

From simulations, we observe that the procedure does not give the desired result, i.e. producing a set of $V_{i,\gamma}$ whose distribution is symmetric. Instead, the chosen power favors distributions with a long and thin tails as shown in Figure 4. These distributions are skew, however, the local third derivative is close to 0. The procedure seems to flatten out the tail and squeeze in the shoulders of the original distribution (see Figure 4).

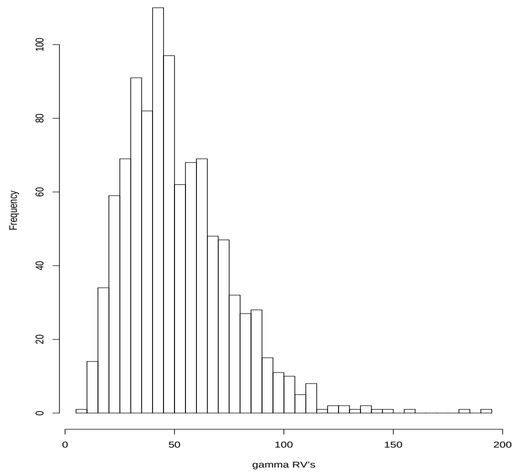

Figure 1: Histogram of Gamma(5,.1) random variables. The local 3^{rd} derivative at the mode is close to 0 even though the distribution is skew.

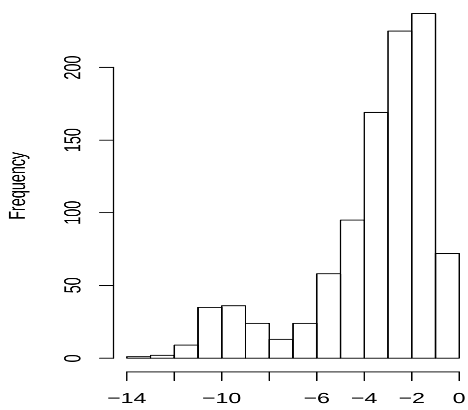

Figure 2: Histogram of an exemplary mixture of nulls (-Gamma(3,1)) and alternatives (N(-10,1)).

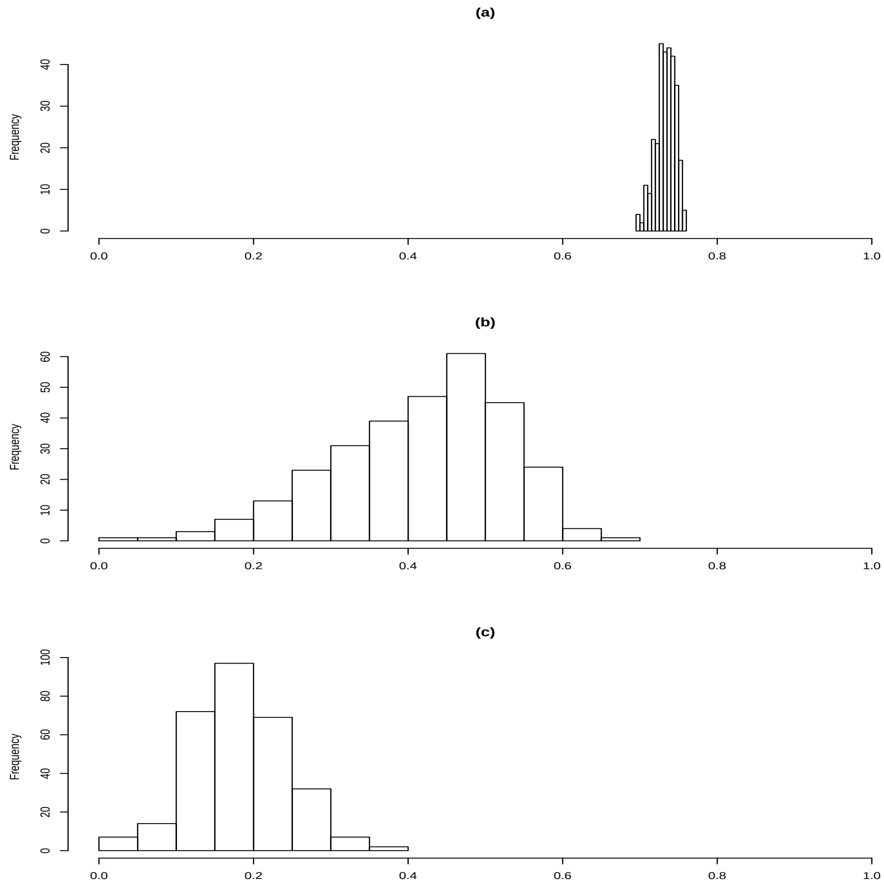

Figure 3: Empirical FDR resulting from (a) No correction (b) Efron's correction (c) Skew normal correction. The histograms are obtained from 300 simulations. Z_i's are simulated from a mixture of gamma's and normal's.

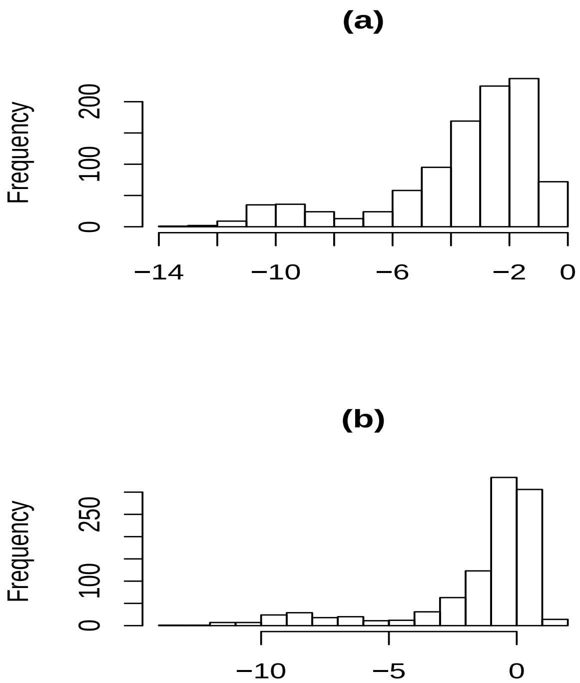

Figure 4: Histogram of (a)Z_i's from a mixture of gamma's and normal's and (b)transformed version $V_{i,\gamma}$'s of Z_i's, γ is chosen such that the third derivative at the mode is close to 0.