Correcting the bias of BGRSS estimation due to truncation

Xing Qiu

December 26, 2017

1 Correcting Bias

1.1 Pre-trimming

By construction, the original trimming is based on a sphere centered at the origin. Namely, for a given trimming proportion c_{trim} (default value: 0.2), for each gene i, we will keep only those pairing genes (k) such that

$$\mathcal{K}_i(c_{\text{trim}}) := \{k = 1, 2, \ldots, i-1,\ i+1, \ldots, m :\ \|R_{ik,\cdot}\|^2 \leqslant R(i, c_{\text{trim}})^2\}, \quad |\mathcal{K}_i(c_{\text{trim}})| = (1-c_{\text{trim}})(m-1). \tag{1}$$

After we have identified these ks, we will use

$$R^{\text{trim}}_{i,\cdot} := \frac{1}{|\mathcal{K}_i|} \cdot \sum_{k \in \mathcal{K}_i} R_{ik,\cdot}, \qquad \widehat{\text{BGRSS}}_i := \sum_{g=1}^{G} \left(R^{\text{trim}}_{i,g}\right)^2 \tag{2}$$

to approximate the oracle $R^*_{i,\cdot}$ and BGRSS*. Here $R^{\text{trim}}_{i,\cdot}$ can be considered as an estimator of $E\left(\mathbf{X}|D_R\right)$, where $\mathbf{X}$ can be interpreted as the list of random vectors $R_{ik,\cdot}$ conditionally on $\boldsymbol{\epsilon}_{i\cdot}$, and D_R is a ball with radius $R = R(i, c_{\text{trim}})$.

Because the trimming set is designed to remove those $R_{ik,\cdot}$ with large radii, it will create a downward bias for $\widehat{\text{BGRSS}}_i$, especially when we *over-trim*; namely, $c_{\text{trim}} > \frac{m_1}{m}$.

In practice, I discovered that conducting a pre-trimming like below can help reduce the bias:

1. Conduct the pre-trimming based on a randomly selected subset of data (to save computational time) with the trimming criterion stated above. Denote the estimated trimmed R as $R^{\text{pre}}_{i,\cdot}$.

2. Use $R^{\text{pre}}_{i,\cdot}$ as the center and then trim $R_{ik,\cdot}$ by $D_{R'}\left(R^{\text{pre}}_{i,\cdot}\right)$, a ball centered at $R^{\text{pre}}_{i,\cdot}$ with a new radius R' such that about c_{trim} points are outside of this ball.

3. Calculate $R^{\text{trim}}_{i,\cdot}$ and then $\widehat{\text{BGRSS}}_i$ based on $D_{R'}\left(R^{\text{pre}}_{i,\cdot}\right)$ instead of D_R.

1.2 Bias-correction

Let $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_m$ be an *i.i.d.* sequence of random vectors following a ,

Let $\mathbf{e}_1 := (1, 0, \ldots, 0)'$ and $\mathbf{X} = (X_1, X_2, \ldots, X_K)' \sim N(\mu \cdot \mathbf{e}_1,\ I_{K\times K})$, for $\mu \geqslant 0$, be a K-dimensional multivariate normal distribution. Note that, by construction, only the first coordinate of $\mathbf{X}$ has a nonzero mean. Denote by $D_R := \left\{\mathbf{y} \in \mathbb{R}^K : \|\mathbf{y}\| \leqslant R\right\}$ a ball with radius R centered at the origin.

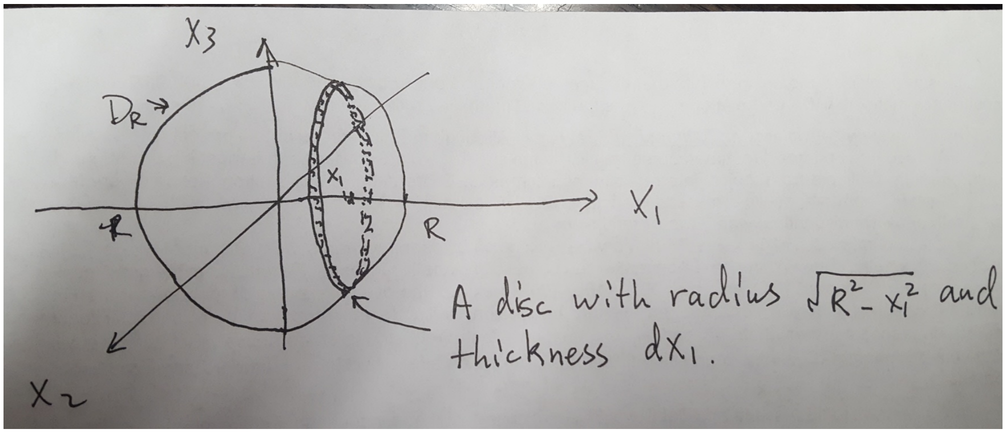


Figure 1: An illustration of the integration on D_R. $E\left(\mathbf{X}|D_R\right)$ can be integrated by small slices of discs with radius $\sqrt{R^2 - x_1^2}$ and thickness $\mathrm{d}x_1$. The probability of this min-disc is $P\left(X_1 \in [x_1, x_1 + \mathrm{d}x_1)\right) \cdot P\left(\sum_{k=2}^{K} X_k^2 \leqslant R^2 - x_1^2\right) = \phi(x_1 - \mu)\mathrm{d}x_1 \cdot F_{\chi^2_{K-1}}(R^2 - x_1^2)$, where $\phi(x_1 - \mu)$ is the density function of X_1, which is a normal random variable with variance 1 and centered at μ; and $F_{\chi^2_{K-1}}(\cdot)$ is the distribution function of $\sum_{k=2}^{K} X_k^2$, which follows a χ^2-distribution with $K-1$ degrees of freedom.

Lemma 1.1. *The expectation of* $\mathbf{X}$ *conditioned on* $\mathbf{X} \in D_R$ *is*

$$E\left(\mathbf{X}|D_R\right) = S_R(\mu) \cdot \mathbf{e}_1, \qquad E\left(X_k|D_R\right) = \begin{cases} S_R(\mu), & k = 1, \\ 0, & k \neq 1. \end{cases} \tag{3}$$

Here function $S_R(\mu) := E\left(X_1|\mathbf{X} \in D_R\right)$ *is defined by*

$$\begin{aligned} S_R(\mu) &:= \frac{\int_{-R}^{R} x \cdot F_{\chi^2_{K-1}}(R^2 - x^2) \cdot \phi(x - \mu)\mathrm{d}x}{\int_{-R}^{R} F_{\chi^2_{K-1}}(R^2 - x^2) \cdot \phi(x - \mu)\mathrm{d}x} \\ &= \frac{\int_{-R}^{R} x e^{-(x-\mu)^2/2} \cdot \gamma\left(\frac{K-1}{2},\, R^2 - x^2\right)\mathrm{d}x}{\int_{-R}^{R} e^{-(x-\mu)^2/2} \cdot \gamma\left(\frac{K-1}{2},\, R^2 - x^2\right)\mathrm{d}x}, \end{aligned} \tag{4}$$

where $F_{\chi^2_{K-1}}(\cdot)$ *is the distribution function of a* χ^2*-distribution with* $K-1$ *degrees of freedom and* $\gamma(\cdot,\cdot)$ *is the lower incomplete gamma function.*

Proof. Please see Figure 1. As a remark, the bottom of $S_R(\mu)$ is $F_{\chi^2_K(\mu^2)}(R^2)$, the probability of a *noncentral* χ^2-distribution with K degrees of freedom being less than R^2. This is because $\sum_{k=1}^{K} X_i^2 \sim \chi^2_K(\mu^2)$. □

Using certain matrix transformation techniques, we have

Corollary 1.2. *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_K)' \sim N(\boldsymbol{\mu},\ \sigma^2 I_{K\times K})$ *be a* K*-dimensional multivariate normal distribution. We have*

$$E\left(\mathbf{X}|D_R\right) = \sigma \cdot S_{R/\sigma}(\|\boldsymbol{\mu}\|/\sigma) \cdot \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}, \qquad \frac{\|E\left(\mathbf{X}|D_R\right)\|}{\sigma} = S_{R/\sigma}(\|\boldsymbol{\mu}\|/\sigma). \tag{5}$$

Proof. Let T be a rotation such that $T\boldsymbol{\mu} = \|\boldsymbol{\mu}\|\mathbf{e}_1$. This rotation always exists (but in general is not unique), and $T^{-1}\mathbf{e}_1 = \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}$.

$\mathbf{Y} := \sigma^{-1}T\mathbf{X}$ will have distribution $N((\|\boldsymbol{\mu}\|/\sigma, 0, \ldots, 0)',\ I_{K\times K})$, and $\sigma^{-1}TD_R = D_{R/\sigma}$. Using Lemma 1.1, we have

$$E\left(\sigma^{-1}T\mathbf{X}|D_R\right) := E\left(\mathbf{Y}|D_R\right) = S_{R/\sigma}(\|\boldsymbol{\mu}\|/\sigma)\cdot\mathbf{e}_1,$$

which implies

$$E\left(\mathbf{X}|D_R\right) = \sigma T^{-1}E\left(\sigma^{-1}T\mathbf{X}|D_R\right) = \sigma\cdot S_{R/\sigma}(\|\boldsymbol{\mu}\|/\sigma)\cdot\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}.$$

□

Proposition 1.3. *The derivative of $S_R(\mu)$ is*

$$\frac{\mathrm{d}S_R(\mu)}{\mathrm{d}\mu} = \frac{H_R(\mu)B_R(\mu) - T_R^2(\mu)}{B_R^2(\mu)} = \mathrm{var}\left(X_1|D_R\right), \tag{6}$$

where

$$T_R(\mu) := \int_{-R}^{R} xe^{-(x-\mu)^2/2}\cdot\gamma\left(\frac{K-1}{2},\ R^2 - x^2\right)\mathrm{d}x, \quad B_R(\mu) := \int_{-R}^{R} e^{-(x-\mu)^2/2}\cdot\gamma\left(\frac{K-1}{2},\ R^2 - x^2\right)\mathrm{d}x. \tag{7}$$

are the top and bottom of $S_R(\mu)$, respectively; and

$$H_R(\mu) := \int_{-R}^{R} x^2e^{-(x-\mu)^2/2}\cdot\gamma\left(\frac{K-1}{2},\ R^2 - x^2\right)\mathrm{d}x. \tag{8}$$

Proof. The derivatives of $T_R(\mu)$ and $B_R(\mu)$ are

$$\begin{aligned}
\frac{\mathrm{d}T_R(\mu)}{\mathrm{d}\mu} &= \int_{-R}^{R} x(x-\mu)e^{-(x-\mu)^2/2}\cdot\gamma\left(\frac{K-1}{2},\ R^2 - x^2\right)\mathrm{d}x \\
&= \int_{-R}^{R} x^2e^{-(x-\mu)^2/2}\cdot\gamma\left(\frac{K-1}{2},\ R^2 - x^2\right)\mathrm{d}x - \mu T_R(\mu) \\
&= H_R(\mu) - \mu T_R(\mu). \\
\frac{\mathrm{d}B_R(\mu)}{\mathrm{d}\mu} &= \int_{-R}^{R} (x-\mu)e^{-(x-\mu)^2/2}\cdot\gamma\left(\frac{K-1}{2},\ R^2 - x^2\right)\mathrm{d}x \\
&= T_R(\mu) - \mu B_R(\mu).
\end{aligned} \tag{9}$$

Therefore, the derivative of $S_R(\mu)$ is

$$\begin{aligned}
\frac{\mathrm{d}S_R(\mu)}{\mathrm{d}\mu} &= \frac{T_R'(\mu)B_R(\mu) - T_R(\mu)B_R'(\mu)}{B_R(\mu)^2} \\
&= \frac{\left(H_R(\mu) - \mu T_R(\mu)\right)B_R(\mu) - T_R(\mu)\left(T_R(\mu) - \mu B_R(\mu)\right)}{B_R^2(\mu)} \\
&= \frac{H_R(\mu)B_R(\mu) - T_R^2(\mu)}{B_R^2(\mu)}.
\end{aligned}$$

Notice that the conditional density of X_1 given $\mathbf{X}\in D_R$ is

$$f_{X_1}\left(x|D_R\right) := \frac{F_{\chi^2_{K-1}}(R^2 - x^2)\cdot\phi(x-\mu)}{P\left(D_R\right)} = \frac{e^{-(x-\mu)^2/2}\cdot\gamma\left(\frac{K-1}{2},\ R^2 - x^2\right)}{B_R(\mu)}, \tag{10}$$

we have

$$\begin{aligned}\operatorname{var}\left(X_1|D_R\right) &= E\left(X_1^2|D_R\right) - \left(E\left(X_1|D_R\right)\right)^2\\ &= \frac{H_R(\mu)}{B_R(\mu)} - \frac{T_R^2(\mu)}{B_R^2(\mu)} = S_R'(\mu).\end{aligned}$$

□

Lemma 1.4. *Function $S_R(\mu)$ has the following properties*

1. *$S_R(\mu)$ is a smooth function of μ.*
2. *$\dfrac{S_R(\mu)}{\mathrm{d}\mu} > 0$. Consequently, $S_R(\mu)$ is strictly increasing w.r.t. μ.*
3. *$S_R(0) = 0$; $S_R'(0) = 1$; $\lim_{\mu\to+\infty} S_R(\mu) = R$.*

Proof. Property 1 is trivial. Property 2 is a consequence of Proposition 1.3. Property 3a is trivial because $x \cdot e^{-(x)^2/2} \cdot \gamma\left(\frac{K-1}{2}, R^2 - x^2\right)$ is an odd function of x, so $T_R(0) = 0$. Property 3b holds because ... Property 3c holds because ... □

Theorem 1.5. *For every $\hat{\mu}^{\text{trim}} \in [0, R)$, there exists a* unique *$\mu^c \in [0, \infty)$, such that $\hat{\mu}^{\text{trim}} = S_R(\mu^c)$. Equivalently, we have $\mu^c = S_R^{-1}\left(\hat{\mu}^{\text{trim}}\right)$. Furthermore, the following Newton-Raphson method converges to this μ^c:*

1. *Let $\mu_0 = \hat{\mu}^{\text{trim}}$.*
2. *For $k = 0, 1, 2, \ldots,$, let*

$$\mu_{k+1} = \mu_k - \frac{S_R(\mu_k) - \hat{\mu}^{\text{trim}}}{S_R'(\mu_k)}. \tag{11}$$

Proof. Needs a little more work. □

Theorem 1.6 (Exactness)**.** *Let X_{ij} be a triangular array such that ...*

Theorem 1.7 (Oracle)**.** *Based on our model and let both $n, m \to \infty$. We also need to assume that $\frac{m_1}{m} \leqslant c_{\text{trim}}$, the trimming proportion. when $m \to \infty$, we have for each $i \in \mathcal{S}^0$*

$$\mu^c = S_R^{-1}\left(\hat{\mu}^{\text{trim}}\right) \to \mu \quad a.s. \tag{12}$$

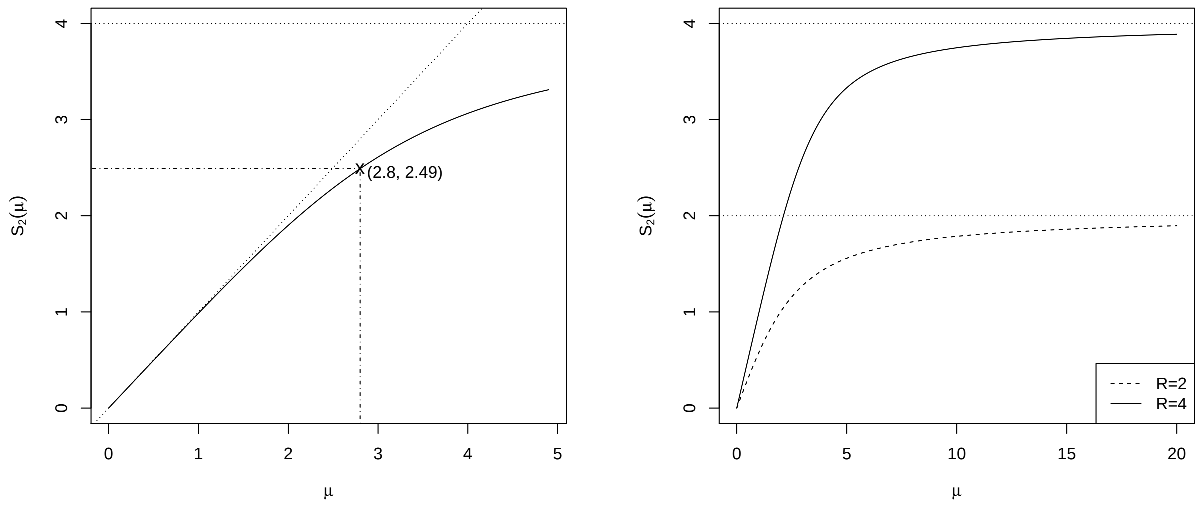
Proof. The strategy of the proof:

1. Those $R_{ik,\cdot}$ paired with the NDEGs ($k \in \mathcal{S}^1$) will vanish to ∞ when $\min_g N_g \to \infty$.
2. With probability one, the remaining terms will be those from the null distribution, trimmed by R.
3. The sample conditional expectation of those remaining $R_{ik,\cdot}$ will converge strongly to $S_R(\mu)$ due to SLLN, roughly with rate $O(m^{-1/2})$ due to CLT. In other words, we have

$$\hat{\mu}^{\text{trim}} = S_R(\mu) + O(m^{-1/2}). \tag{13}$$

4. $S_R^{-1}\left(\hat{\mu}^{\text{trim}}\right) = \mu + \frac{1}{S_R'(\mu)} \cdot O(m^{-1/2})$, which is the desired result.

□



(a)

(b)

Figure 2: Illustrations of $S_R(\mu)$. (a): An illustration of function $S_R(\mu)$. The true μ is 2.8. The observed trimmed sample mean is $S_2(2.8) + O(m^{-1/2}) = 2.49 + O(m^{-1/2})$. (b): An illustration of function $S_R(\mu)$ converging to R when μ is large.

1.3 Using bias correction in Super-δ method

As a reminder, in Super-δ, trimming was applied to $R_{ik,g}$:

$$\begin{aligned} R_{ik,g} &:= \sqrt{N_g}\left(\bar{\delta}_{ik,g\cdot} - \bar{\bar{\delta}}_{ik,\cdot\cdot}\right) = \sqrt{N_g}\left(\bar{Y}_{ig} - \bar{Y}_{kg} - \bar{\bar{Y}}_i + \bar{\bar{Y}}_k\right) \\ &= \sqrt{N_g}\underbrace{\left(\mu_{i,g} + \bar{\epsilon}_{i,g\cdot} - (\bar{\mu}_{i,\cdot} + \bar{\bar{\epsilon}}_{i,\cdot\cdot})\right)}_{\text{the oracle part}} - \sqrt{N_g}\left(\mu_{k,g} + \bar{\epsilon}_{k,g\cdot} - (\bar{\mu}_{k,\cdot} + \bar{\bar{\epsilon}}_{k,\cdot\cdot})\right) \\ &= R^*_{i,g} + \sqrt{N_g}\left(\mu_{k,g} - \bar{\mu}_{k,\cdot} + \bar{\epsilon}_{k,g\cdot} - \bar{\bar{\epsilon}}_{k,\cdot\cdot}\right). \end{aligned} \tag{14}$$

Due to the "Vanishing Theorem" (needs more work), those $R_{ik,g}$ paired with DEGs will vanish to infinity, therefore be removed by the trimming procedure. For $k \in \mathcal{S}^0$, $\mu_{k,g} - \bar{\mu}_{k,\cdot} = 0$, so

$$R_{ik,\cdot} | \boldsymbol{\epsilon}_i \sim N\left(R^*_{i,\cdot}, \sigma^2\left(I_{G\times G} - \frac{\mathbf{N}\mathbf{N}'}{N}\right)\right), \quad \mathbf{N} := (N_1, N_2, \ldots, N_G). \tag{15}$$

It is easy to see that $\langle R_{ik,\cdot}, \mathbf{N}/N\rangle = \langle R^*_{i,\cdot}, \mathbf{N}/N\rangle = 0$. Let T_G be a $(G-1)\times G$-dimensional semi-orthogonal matrix such that $T_G\left(I_{G\times G} - \frac{\mathbf{N}\mathbf{N}'}{N}\right)T'_G = I_{(G-1)\times(G-1)}$. $\tilde{R}_{ik,\cdot} := T_G R_{ik,\cdot}$ will preserve the length of $R_{ik,\cdot}$ and have the following distribution

$$\begin{aligned} &\tilde{R}_{ik,\cdot} := T_G R_{ik,\cdot} \sim N\left(\tilde{R}^*_{i,\cdot}, \sigma^2 I_{(G-1)\times(G-1)}\right), \quad \tilde{R}^*_{i,\cdot} := T_G R^*_{i,\cdot}. \\ &\|\tilde{R}_{ik,\cdot}\| = \|R_{ik,\cdot}\|, \qquad \|\tilde{R}^*_{i,\cdot}\| = \|R^*_{ik,\cdot}\|. \end{aligned} \tag{16}$$

Inspired by Equation (5), we propose the following bias adjusted estimator for $\mathrm{BGRSS}_i :=$

$\|R_{ik,\cdot}\|^2$.

$$\widehat{\text{BGRSS}}_i^c := \hat{\sigma}_\epsilon^2 \cdot \left[S^{-1}_{R_i/\hat{\sigma}_\epsilon} \left(\frac{\|R^{\text{trim}}_{i,\cdot}\|}{\hat{\sigma}_\epsilon} \right) \right]^2. \tag{17}$$

1.4 A discussion on the DEGs

In Theorem 1.7, we proved the oracle property for NDEGs. In this section, we briefly mention the asymptotic properties for the DEGs.

1. For $i \in \mathcal{S}^1$, its pairing with NDEGs or most other DEGs will all vanish to infinity, due to the fact that the oracle $R_{i,\cdot} \to \infty$ as $\min_g N_g \to \infty$.
2. Unfortunately, when normalized by $N^{-1/2}$, those $R_{ik,\cdot}$ paired with DEGs are not going to be separated from those paired with NDEGs.
3. The statistical power will be at least bounded from below by an inequality (needs more work)

$$\mu^c \geqslant (1 - c_{\text{trim}}) \cdot \mu. \tag{18}$$

1.5 An efficient algorithm for high-dimensional Newton-Raphson's method

1. For gene i, calculate R_i and $\hat{\mu}_i^{\text{trim}}$.
2. Create $P_i = \{x_{i1}, x_{i2}, \ldots, x_{i,101}\}$, a partition of $[-R_i, R_i]$ into 200 evenly spaced intervals, then evaluate $F_{\chi^2_{K-1}}(R^2 - x^2)$ on this grid. Note that we only need to do it once for all iterations in applying Newton-Raphson method for this gene.
3. Now apply the Newton-Raphson method for this gene. We will use $\hat{\mu}_i^{\text{trim}}$ as the initial value $\mu_{i,0}$.
4. In each step, calculate $\phi(x - \mu_k)$, then compute $B_{R_i}(\mu_{i,k})$, $T_{R_i}(\mu_{i,k})$, and $H_{R_i}(\mu_{i,k})$.
5. Using these terms to compute $S_{R_i}(\mu_{i,k})$ and $S'_{R_i}(\mu_{i,k})$, and use Equation (11) to compute $\mu_{i,k+1}$.
6. Check the error term, $|S_{R_i}(\mu_{i,k+1}) - S_{R_i}(\mu_{i,k})|$. If it is smaller than a prespecified threshold, stop. Alternatively, if k is greater than a pre-specified maximum iteration number, stop.

It only takes about 1.6 seconds to calculate 10,000 μ by the above algorithm on my laptop computer.

1.6 Simulation results

Simulation settings

- $m = 5,000$ genes; 20% of them are true DEGs.
- $n = 100$ samples divided into $G = 3$ groups (20, 30, 40 samples in each group).
- Other specifications please see `test_oracle2.R`.

- Run Super-delta with **over-trimming: 30%** of genes are removed in the trimming step.
- Two approaches are used: With or without bias-correction. Both approaches are compared with the oracle F-test.
- p-values computed from the F-tests are adjusted by the BH-procedure, and the significant genes are selected at FDR=0.05 level.

	Oracle	Uncorrected	Bias.corrected
Type.I.err	0.0063	0.0015	0.0060
FDR	0.0320	0.0096	0.0309
Power	0.7570	0.6190	0.7520

Table 1: Type I error, FDR, and statistical power of three Super-delta approaches. The statistical performance of Super-delta with bias correction is very close to that of the oracle F-tests. Without bias correction, Super-delta is overly conservative due to **over-trimming**.

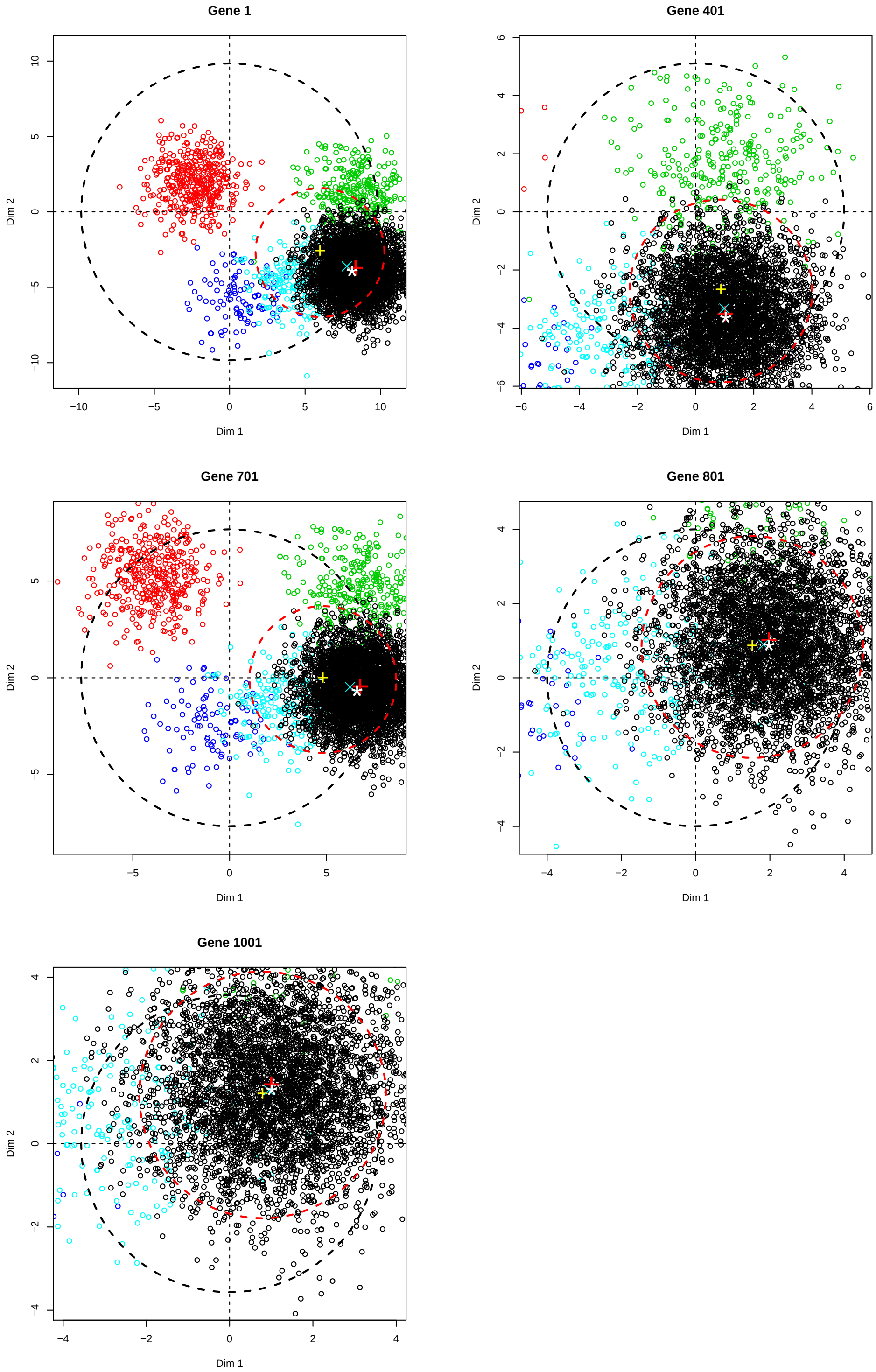


Figure 3: An illustration of the pre-trim and bias-correction. We use five representative genes (the first four are DEGs with different differential patterns; the last one is an NDEG. The red crosses represent the oracle $R^*_{i,\cdot}$; yellow plus signs represent the first trim center (the original method), cyan "x" represent trimmed estimator $R^{\text{trim}}_{i,\cdot}$, which is estimated from the second trimming, but before bias-correction; white stars represent bias-corrected estimator. We can see that white stars are the best estimators of the oracle $R^*_{i,\cdot}$.

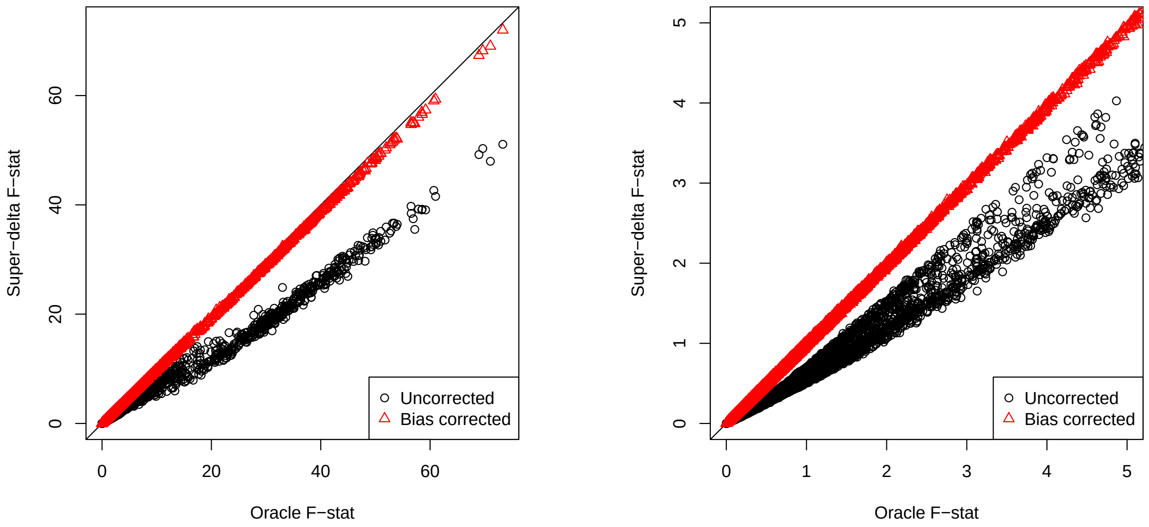


Figure 4: A plot of the F-statistics produced by the two Super-delta approaches and the oracle F-statistics. Left panel illustrates all F-statistics; right panel shows only those with relatively small values (essentially those associated with the NDEGs). We see that with **over-trimming**, the original Super-delta method under-estimates F-statistics. With bias-correction, we can recover the oracle F-statistics very well even with over-trimming. A closer comparison shows that for NDEGs (see the right panel), Super-delta with bias correction works **almost perfectly**, which is predicted by the **Oracle Property** Theorem. For DEGs (larger red triangles in the left panel), there are still some very small biases that are not fully compensated by the bias correction algorithm. These small remaining biases are all *downward* biases, which explains why the statistical power of Super-delta with bias correction is *slightly* less powerful than that of the oracle F-tests (last row of Table 1).

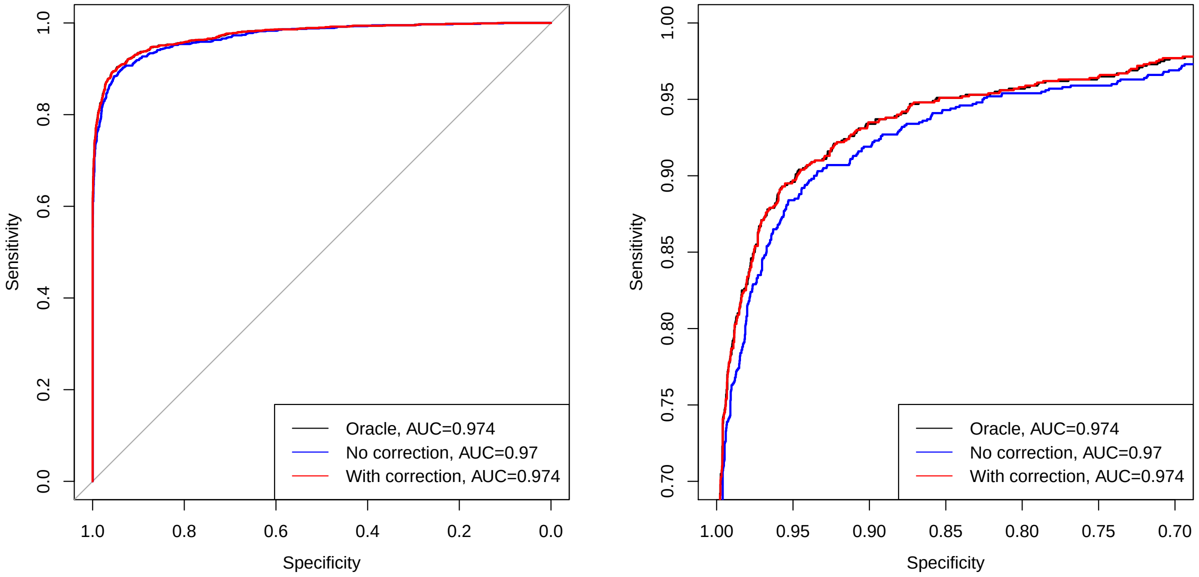


Figure 5: ROC analysis. While all three methods seem to have similar performance in terms of ROC curves, a closer look (the right panel) reveals that the performance of Super-delta without correction is slightly poorer than its counterpart with bias-correction, which in turn is virtually equivalent to that of the oracle F-test.