Comment: FDP vs FDR and the Effect of Conditioning

Armin Schwartzman

Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, MA, 02115


To link to this article: http://dx.doi.org/10.1080/01621459.2012.712876
Comment: FDP vs FDR and the Effect of Conditioning

Armin SCHWARTZMAN

Accurate estimation of the false discovery proportion (FDP) and false discovery rate (FDR) is a problem that has been eagerly waiting for a solution. Fan, Han, and Gu have found one. They have achieved this by both clarifying the concept of the problem and providing a feasible algorithmic solution. In this comment, I discuss some of the central concepts involving FDP and its estimation, originally introduced by Storey (2002), and as estimators of FDP(0) = V(t)/R(t) rather than FDR(t) = E[V(t)/R(t)], particularly in high correlation situations. This understanding is crucial because it allows Fan, Han, and Gu to adapt the idea of the conditional FDR estimator and produce an estimator that is consistent for FDP under some set of asymptotic conditions. In contrast, Fan, Han, and Gu suggest that estimating FDR may be more difficult, as they propose an asymptotic approximation to it but leave some of the details for future work.

1. FDP VERSUS FDR

FDR estimation, originally introduced by Storey (2002), was first seen as a tool for FDR control (Genovese and Wasserman 2004; Storey, Taylor, and Siegmund 2004) but then gained interest on its own mainly thanks to Efron (2007). In Fan, Han, and Gu’s notation, ignoring the estimation of the number of null hypotheses $p_0$ (which is normally assumed to be close to the known number of tests $p$), the original FDR estimator by Storey (2002) is essentially $E[V(t)]/R(t) = p_0 R(t)$. Because it has the random variable $R(t)$ in the denominator, this estimator is biased and highly variable as an estimator of FDR(0) mainly thanks to Efron (2007). In Fan, Han, and Gu’s notation, ignoring the estimation of the number of null hypotheses $p_0$, the original FDR estimator $E[V(t)]/R(t)$ is consistent for FDP under some set of asymptotic conditions. In contrast, Fan, Han, and Gu suggest that estimating FDR may be more difficult, as they propose an asymptotic approximation to it but leave some of the details for future work.

2. THE EFFECT OF CONDITIONING

The essence of Fan, Han, and Gu’s estimator is in its proper use of conditioning. Without the formalities of the proofs given in the Appendix, the expressions in Theorem 1 and Proposition 2 can be explained briefly as follows. Recall Fan, Han, and Gu’s principal factor approximation (PFA) model [Equation (10)]

$$Z_i = \mu_i + \eta_i + K_i, \quad i = 1, \ldots, n$$

with $\eta_i = b_i^T W \sim N(0, \sigma_i^2)$ and $K_i \sim N(0, 1 - \sigma_i^2)$, where the $K_i$’s are independent of the factors $W$. Now write

$$R(t) = \sum_{i=1}^p [1(Z_i \geq z_{t/2}) + 1(Z_i \leq z_{t/2})],$$

where $z_{t/2} = \Phi^{-1}(t/2)$ (note that $z_{t/2} < 0$ for $t < 0.5$). Then

$$E[R(t) | W] = \sum_{i=1}^p P(Z_i \geq z_{t/2} | W) + P(Z_i \leq z_{t/2} | W)$$

$$= \sum_{i=1}^p P(-K_i \leq z_{t/2} + \eta_i + \mu_i | W)$$

we observe that the desired property (7) follows from (10) and (12).

REFERENCES


\[ + P(K_i \leq z_{i/2} - \eta_i - \mu_i | W) \]
\[ = \sum_{i=1}^{p} \Phi[a_i(z_{i/2} + \eta_i + \mu_i)] + \Phi[a_i(z_{i/2} - \eta_i - \mu_i)]. \tag{2} \]

where \( a_i = [\text{var}(K_i)]^{-1/2} = (1 - \sigma_i^2)^{-1/2} \). Similarly, write
\[ V(t) = \sum_{i \in \text{true null}} [I(Z_i \geq -z_{i/2}) + 1(Z_i \leq z_{i/2})]. \tag{3} \]

Then
\[ E[V(t) | W] = \sum_{i \in \text{true null}} \Phi[a_i(z_{i/2} + \eta_i)] + \Phi[a_i(z_{i/2} - \eta_i)]. \]

In this notation, Proposition 2 establishes the asymptotic equivalence
\[ \frac{R(t)}{p} \approx \frac{E[R(t) | W]}{p}, \quad \frac{V(t)}{p_0} \approx \frac{E[V(t) | W]}{p_0}, \]
which in Theorem 1 becomes
\[ \text{FDP}(t) = \frac{V(t)}{R(t)} \approx \frac{E[V(t) | W]}{E[R(t) | W]}, \]

because \( p_0/p \to 1 \). The key idea of Fan, Han, and Gu’s estimator is to combine these approximations to establish
\[ \text{FDP}(t) \approx \frac{E[V(t) | W]}{R(t)}. \tag{4} \]

In essence, this is the same concept of conditioning as in Friguet, Kloareg, and Causeur’s estimator. However, it is the details that make all the difference. Less importantly, Friguet, Kloareg, and Causeur used a conservative estimate of the true null set to estimate the numerator \( E[V(t) | W] \). Instead, Fan, Han, and Gu use the asymptotic approximation
\[ \text{FDP}(t) \approx \text{FDP}_{A}(t) = \frac{E_0[R(t) | W]}{R(t)} \]

where
\[ E[V(t) | W] \approx E_0[R(t) | W] = \sum_{i=1}^{p} \Phi[a_i(z_{i/2} + \eta_i)] + \Phi[a_i(z_{i/2} - \eta_i)]. \]

More importantly, the conditioning factors \( W \) have a different interpretation. Friguet, Kloareg, and Causeur assumed that the factor model holds exactly and estimate the unknown factors by the expectation–maximization algorithm. For Fan, Han, and Gu, the factor model is only an approximation to the arbitrary covariance matrix of the data. Therefore, the strength of Fan, Han, and Gu’s estimator is the ability to estimate the approximating factors, which they achieve via \( L_1 \)-regression.

Fan, Han, and Gu also compare their estimator with that of Efron (2007). It is worth exploring the relation between the two a bit further. Schwartzman (2010) showed that Efron’s conditional FDR estimator can be obtained from a Gram–Charlier A series expansion approximation to the empirical cdf of the true null random variables \( Z_i \) for large \( p_0 \):
\[ \frac{1}{p_0} \sum_{i \in \text{true null}} 1(Z_i \leq x) \approx \Phi(x) - \sum_{j=1}^{\infty} \frac{A_i}{\sqrt{j!}} (-1)^j \phi^{(j-1)}(x), \]

where \( A_1, A_2, \ldots \) are uncorrelated random variables with mean 0 and variance \( \text{var}(A_i) = \alpha_i = \sum_{j \neq i} \rho_{ij}^2 / (p(p-1)) \), \( \rho_{ij} = \text{cor}(Z_i, Z_j) \), and \( \phi^{(j)}(x) \) denotes the \( j \)th derivative of the standard normal density function \( \phi(x) \). Indeed, using this expansion in the definition of \( V(t) \) (3) gives
\[ E[V(t) | A] \approx p_0 \left[ t + 2 \sum_{k=1}^{\infty} \frac{A_{2k}}{(2k)!} \phi^{(2k-1)}(z_{t/2}) \right], \]

where \( A = (A_1, A_2, \ldots) \), yielding
\[ \frac{E[V(t) | A]}{R(t)} \approx \frac{p_0 t}{R(t)} \left[ 1 + 2 \sum_{t=1}^{\infty} \frac{A_{2k}}{(2k)!} \phi^{(2k-1)}(z_{t/2}) \right]. \tag{5} \]

Efron’s estimator is obtained taking only the \( k = 1 \) term in the above expansion and defining \( A = A_2 \).

In comparison with (4), it becomes clear how Efron and Fan, Han, and Gu use conditioning to capture the dependence between the variables \( Z_i \). Efron’s approach is to first reduce the set of \( Z_i \)’s to their empirical cdf, which is equivalent to their order statistics, and then use the derivatives of the normal density (related to the Hermite polynomials) as a set of basis functions, conditioning on their random coefficients. In contrast, Fan, Han, and Gu use the eigenvectors of the covariance matrix as a set of basis functions for establishing the PFA to the vector of variables \( (Z_1, \ldots, Z_p) \) and condition on their random coefficients. Because the two sets of basis functions are different, the two estimators can be made to coincide [Equation (27) of Fan, Han, and Gu] only when the amount of dependence in the data, measured here by the \( \sigma_i \), is small.

Fan, Han, and Gu’s approach has several advantages with respect to Efron’s. First, the set of order statistics is not a sufficient statistic for a set of correlated random variables. Thus, by approximating the data directly via PFA, Fan, Han, and Gu gain estimation efficiency, as observed in Fan, Han, and Gu’s Figure 2. Second, while Efron has suggested an estimator for \( A \), more terms in the expansion would be needed to make (5) accurate, and the positivity constraints on the coefficients make estimation difficult (Jondeau and Rockinger 2001). In contrast, Fan, Han, and Gu have been able to provide a procedure based on \( L_1 \)-regression to consistently estimate the conditioning factors in the PFA.

Finally, the Gram–Charlier expansion, and thus Efron’s estimator, is intimately linked to the normal distribution. Its generalization is limited to a small set of nonnormal distributions, such as \( \chi^2 \) (Schwartzman 2010). Derivation (2) suggests that Fan, Han, and Gu’s approach could also possibly be generalized to nonnormal distributions. The normality is crucial in that, to go from row 2 to row 3 of (2), normality guarantees that the residual variables \( K_i \) are independent of the factors \( W \). The expression in row 2 could potentially be used as the numerator in the FDP estimator (4) in nonnormal situations provided that the dependence between the residual variables and the random factors can be specified.

### 3. FDP VERSUS FDR AGAIN

As mentioned at the beginning of this comment, Fan, Han, and Gu are able to estimate FDP, but leave the precise estimation of FDR for future work. It is interesting and surprising
that FDP may be easier to estimate than FDR, since FDP is a random variable, while FDR is a parameter of the distribution. However, this appears to be true in high correlation situations. To understand this, Figure 1 shows the effect of correlation on the distribution of FDP in the equal correlation scenario, where the covariance matrix of the data has diagonal entries equal to 1 and off-diagonal entries equal to ρ. Here I use the same parameters as in Figure 2 of Fan, Han, and Gu: ρ = 1000, p1 = 50, n = 100, t = 0.005, and βi = 1 for i ∈ {false null}, based on 1000 simulations. These distributions could also be obtained from the expression for FDP given in Example 1 of Fan, Han, and Gu.

In Figure 1, as correlation increases, the distribution of FDP becomes more skewed, slowly splitting into two components. The equal correlation model is equivalent to a one-factor model; as correlation increases, the common factor W ~ N(0, 1) gets a larger weight and is easier to estimate. In the extreme case of perfect correlation, all the observed Z_i’s are either equal to W or equal to W + μ. Then the FDP is either zero if |W| < |z_i/| or p0/p = 0.95 if |z_i/| < min |W|, |W + μ| or 1 if |W + μ| < |z_i/| < |W|. In this extreme situation with large p0, the FDP can be estimated perfectly from the data. However, there is no information to estimate FDR.

Both Friguet, Kloareg, and Causeur take advantage of the presence of strong common factors in high correlation situations to estimate FDP. However, Friguet, Kloareg, and Causeur noted that when correlation is low, the common factors are harder to estimate and thus their conditional estimator does not correlate with FDP. In contrast, judging from Fan, Han, and Gu’s results for the independent Cauchy scenario in their Figure 2, their estimator is adaptive to the data and appears to able to estimate the FDP even under these more challenging circumstances.

In the context of Figure 2 of Fan, Han, and Gu, it is easy to explain why the unconditional FDR estimator p0R(t)/R(t), shown in green, is almost perfectly negatively correlated with FDP, a phenomenon also observed by Efron (2007) and Friguet, Kloareg, and Causeur (2009). Defining S(t) = R(t) − V(t) as in Table 1 of Fan, Han, and Gu, we may write FDP(t) = 1 − S(t)/R(t).

Solving for R(t) and replacing in the unconditional estimator gives that

$$\frac{p_0 t}{R(t)} = \frac{p_0 t}{S(t)} \{1 - \text{FDP}(t)\} \approx \frac{p_0 t}{p_1} \{1 - \text{FDP}(t)\},$$

which is a linearly decreasing function of FDP. The approximation above reflects the fact that, in the simulation scenario of Figure 2, μ = √nβ1 = 10 is strong enough that the nonnull cases are essentially always detected, so that S(t) ≈ p1. The intercept and negative slope p0t/p1 = 0.095 correspond precisely to the green graph observed there.

To close, I want to congratulate Fan, Han, and Gu for an important contribution to the field of large-scale multiple testing. Because their results are mainly theoretical, it is in the practical implementation details where more work is needed, especially in terms of how to choose the number of factors in the PFA model. I look forward to seeing their method implemented in practice in the search for scientific discoveries.

REFERENCES


