Published in final edited form as:

J Am Stat Assoc. 2012 ; 107(499): 1039–1041. doi:10.1080/01621459.2012.712876.

Comment on “Estimating False Discovery Proportion Under Arbitrary Covariance Dependence” by Fan et al.

Armin Schwartzman [Assistant Professor]

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

Armin Schwartzman: armins@hsph.harvard.edu

Abstract

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 via conditioning in contrast with the estimators by Efron (2007) and Friguet et al. (2009).

1 FDP vs. FDR

The concept of FDR estimation, originally introduced by Storey (2002), was first seen as a tool for FDR control (Storey et al., 2004; Genovese and Wasserman, 2004), but then gained interest on its own mainly thanks to Efron (2007). In Fan et al.’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 $E[V(t)]/R(t) = p_0 t / R(t)$. Because it has the random variable $R(t)$ in the denominator, this estimator is biased and highly variable as an estimator of $FDR(t) = E[V(t)/R(t)]$, particularly in the presence of correlation (Schwartzman and Lin, 2011).

As discussed by Fan et al., two recent estimators have attempted to incorporate the effect of correlation explicitly. Efron’s estimator is essentially $E[V(t) | \hat{A}]/R(t)$, where $\hat{A}$ is a random variable that estimates the amount of dispersion between the test statistics in a given experiment (Efron, 2007). Friguet et al.’s estimator is essentially $E[V(t) | \hat{W}]/R(t)$, where $\hat{W}$ is an estimate of the realized random factors in a factor analysis model in a given experiment (Friguet et al., 2009). Both estimators are presented as conditional FDR estimators that correct the original FDR estimator $E[V(t)]/R(t)$ by a multiplicative factor. However, in their simulations, both authors realize that in high correlation situations, their estimators actually correlate with the realized FDP.

Fan et al.’s first important conceptual realization is that the conditional estimators above may be better thought of as estimators of $FDP(t) = V(t)/R(t)$ rather than $FDR(t) = E[V(t)/R(t)]$, particularly in high correlation situations. This understanding is crucial because it allows Fan et al. 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 et al.

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 et al.'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 et al.'s principal factor approximation (PFA) model (Eq. (10))

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

with $\eta_i = \mathbf{b}_i^T \mathbf{W} \sim N(0, \sigma_i^2)$ and $K_i \sim N(0, 1 - \sigma_i^2)$, where the K_i 's are independent of the factors $\mathbf{W}$. Now write

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

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

$$\begin{aligned} E[R(t) | \mathbf{W}] &= \sum_{i=1}^p P(Z_i \geq -z_{t/2} | \mathbf{W}) \\ &\quad + P(Z_i \leq z_{t/2} | \mathbf{W}) \\ &= \sum_{i=1}^p P(-K_i \leq z_{t/2} + \eta_i + \mu_i | \mathbf{W}) \\ &\quad + P(K_i \leq z_{t/2} - \eta_i - \mu_i | \mathbf{W}) \\ &= \sum_{i=1}^p \Phi[a_i(z_{t/2} + \eta_i + \mu_i)] \\ &\quad + \Phi[a_i(z_{t/2} - \eta_i - \mu_i)], \end{aligned} \quad (2)$$

where $a_i = [\text{var}(K_i)]^{-1/2} = (1 - \sigma_i^2)^{-1/2}$. Similarly, write

$$V(t) = \sum_{i \in \{\text{true null}\}} [1(Z_i \geq -z_{t/2}) + 1(Z_i \leq z_{t/2})]. \quad (3)$$

Then

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

In this notation, Proposition 2 establishes the asymptotic equivalence

$$\frac{R(t)}{p} \doteq \frac{E[R(t)|\mathbf{W}]}{p}, \quad \frac{V(t)}{p_0} \doteq \frac{E[V(t)|\mathbf{W}]}{p_0},$$

which in Theorem 1 becomes

$$\text{FDP}(t) = \frac{V(t)}{R(t)} \doteq \frac{E[V(t)|\mathbf{W}]}{E[R(t)|\mathbf{W}]}$$

because $p_0/p \rightarrow 1$. The key idea of Fan et al.'s estimator is to combine these approximations to establish

$$\text{FDP}(t) \doteq \frac{E[V(t)|\mathbf{W}]}{R(t)}. \quad (4)$$

In essence, this is the same concept of conditioning as in Friguet et al.'s estimator. However, it is the details that make all the difference. Less importantly, Friguet et al. use a conservative estimate of the true null set to estimate the numerator $E[V(t)|\mathbf{W}]$. Instead, Fan et al. use the asymptotic approximation

$$\text{FDP}(t) \doteq \text{FDP}_A(t) = \frac{E_0[R(t)|\mathbf{W}]}{R(t)}$$

where

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

More importantly, the conditioning factors $\mathbf{W}$ have a different interpretation. Friguet et al. assume that the factor model holds exactly, and estimate the unknown factors by the EM algorithm. For Fan et al., the factor model is only an approximation to the arbitrary covariance matrix of the data. Therefore, the strength of Fan et al.'s estimator is the ability to estimate the approximating factors, which they achieve via L_1 -regression.

Fan et al. also compare their estimator with that of Efron (2007). It is worth exploring the relation between the two a bit further. Schwartzman (2010) shows 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 :

$$\frac{1}{p_0} \sum_{i \in \{\text{true null}\}} 1(Z_i \leq x) \doteq \Phi(x) - \sum_{j=1}^{\infty} \frac{A_j}{\sqrt{j!}} (-1)^j \phi^{(j-1)}(x),$$

where $A_1, A_2, \dots$ are uncorrelated random variables with mean 0 and variance $\text{var}(A_j) = \alpha_j = \sum_{i \neq i'} \rho_{ii'}^j / [p(p-1)]$, $\rho_{ii'} = \text{cor}(Z_i, Z_{i'})$, 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

$$E[V(t)|\mathbf{A}] \doteq p_0 \left[t + 2 \sum_{k=1}^{\infty} \frac{A_{2k}}{\sqrt{(2k)!}} \phi^{(2k-1)}(z_{t/2}) \right]$$

where $\mathbf{A} = (A_1, A_2, \dots)$, yielding

$$\frac{E[V(t)|\mathbf{A}]}{R(t)} \doteq \frac{p_0 t}{R(t)} \left[1 + \frac{2}{t} \sum_{k=1}^{\infty} \frac{A_{2k}}{\sqrt{(2k)!}} \phi^{(2k-1)}(z_{t/2}) \right]. \quad (5)$$

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 et al. 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 et al. use the eigenvectors of the covariance matrix as a set of basis functions for establishing the PFA to the vector of variables $(Z_1, \dots, Z_p)$ and condition on their random coefficients. Because the two sets of basis functions are different, the two estimators can be made to coincide (Eq. (27) of Fan et al.) only when the amount of dependence in the data, measured here by the σ_i , is small.

Fan et al.'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 et al. gain estimation efficiency, as observed in Fan et al.'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 et al. 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 non-normal distributions, such as χ^2 (Schwartzman, 2010). Derivation (2) suggests that Fan et al.'s approach could also possibly be generalized to non-normal 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 $\mathbf{W}$. The expression in row 2 could potentially be used as the numerator in the FDP estimator (4) in non-normal situations provided that the dependence between the residual variables and the random factors can be specified.

3 FDP vs. FDR again

As mentioned at the beginning of this comment, Fan et al. 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 below 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 et al.: $p = 1000$, $p_1 = 50$, $n = 100$, $t = 0.005$ and $\beta_i = 1$ for $i \in \{\text{false null}\}$, based on 1000 simulations. These distributions could also be obtained from the expression for FDP given in Example 1 of Fan et al.

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 $W \sim N(0, 1)$ 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 $W + \mu$. Then the FDP is either zero if $|W| < |z_{t/2}|$ or $p_0/p = 0.95$ if $|W| > |z_{t/2}|$. In this extreme situation the FDP can be estimated perfectly from the data. However, there is no information to estimate FDR.

Both Friguet et al. and Fan et al. take advantage of the presence of strong common factors in high correlation situations to estimate FDP. However, Friguet et al. note 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 et al.'s results for the independent Cauchy scenario in their Figure 2, their estimator is adaptive to the data and appears to be able to estimate the FDP even under these more challenging circumstances.

In the context of Figure 2 of Fan et al., it is easy to explain why the unconditional FDR estimator $p_0 t / R(t)$, shown in green, is almost perfectly negatively correlated with FDP, a phenomenon also observed by Efron (2007) and Friguet et al. (2009). Defining $S(t) = R(t) - V(t)$ as in Table 1 of Fan et al., we may write $\text{FDP}(t) = 1 - S(t)/R(t)$. 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, $\mu = \sqrt{n}\beta_1 = 10$ is strong enough that the non-null cases are essentially always detected, so that $S(t) \approx p_1$. The intercept and negative slope $p_0 t / p_1 = 0.095$ correspond precisely to the green graph observed there.

To close, I want to congratulate Fan et al. for an important contribution to the field of large-scale multiple testing. I look forward to seeing their method implemented in practice in the search for scientific discoveries.

Acknowledgments

This work was partially supported by NIH grant PO1-CA134294.

References

- Efron B. Correlation and Large-Scale Simultaneous Hypothesis Testing. *J Amer Statist Assoc.* 2007; 102:93–103.
- Friguet C, Kloareg M, Causeur D. A Factor Model Approach to Multiple Testing Under Dependence. *J Amer Statist Assoc.* 2009; 104:1406–1415.
- Genovese CR, Wasserman L. A stochastic process approach to false discovery control. *Ann Statist.* 2004; 32:1035–1061.
- Jondeau E, Rockinger M. Gram-Charlier densities. *Journal of Economic Dynamics & Control.* 2001; 25:1457–1483.
- Schwartzman A. Comment on "Correlated z-values and the accuracy of large-scale statistical estimates" by Bradley Efron. *J Amer Statist Assoc.* 2010; 105:1059–1063.
- Schwartzman A, Lin X. The effect of correlation in false discovery rate estimation. *Biometrika.* 2011; 98:199–214. [PubMed: 23049127]
- Storey JD. A direct approach to false discovery rates. *J R Statist Soc B.* 2002; 64:479–498.
- Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Statist Soc B.* 2004; 66:187–205.

Figure 1. Boxplots of simulated FDP values as a function of the equal correlation parameter ρ . The dashed line is the mean FDP, i.e. the FDR.