Open Journal of Statistics, **2012, 2, 163-171**
http://dx.doi.org/10.4236/ojs.2012.22018 Published Online April 2012 (http://www.SciRP.org/journal/ojs)

A Tight Prediction Interval for False Discovery Proportion under Dependence

Shulian Shang, Mengling Liu, Yongzhao Shao*
Division of Biostatistics, New York University School of Medicine, New York, USA
Email: *shaoy01@nyu.edu

Received February 1, 2012; revised March 5, 2012; accepted March 16, 2012

ABSTRACT

The false discovery proportion (FDP) is a useful measure of abundance of false positives when a large number of hypotheses are being tested simultaneously. Methods for controlling the expected value of the FDP, namely the false discovery rate (FDR), have become widely used. It is highly desired to have an accurate prediction interval for the FDP in such applications. Some degree of dependence among test statistics exists in almost all applications involving multiple testing. Methods for constructing tight prediction intervals for the FDP that take account of dependence among test statistics are of great practical importance. This paper derives a formula for the variance of the FDP and uses it to obtain an upper prediction interval for the FDP, under some semi-parametric assumptions on dependence among test statistics. Simulation studies indicate that the proposed formula-based prediction interval has good coverage probability under commonly assumed weak dependence. The prediction interval is generally more accurate than those obtained from existing methods. In addition, a permutation-based upper prediction interval for the FDP is provided, which can be useful when dependence is strong and the number of tests is not too large. The proposed prediction intervals are illustrated using a prostate cancer dataset.

Keywords: Multiple Testing; False Discovery Proportion; False Discovery Rate; Weak Dependence; Correlated Test Statistics; High-Dimensional Data Analysis; Prediction Interval; Upper Prediction Bound; Permutation-Based Method

1. Introduction

When a large number of hypotheses are tested simultaneously, a direct measure of the abundance of false positive findings is the false discovery proportion (FDP), defined as FDP, or $Q = \frac{V}{R \vee 1}$, where R denotes the total number of rejections, V denotes the number of rejections of true null hypotheses, and $R \vee 1 = \max\{R, 1\}$. Motivated by various genetic and genomic studies and other applications, many useful procedures have been proposed to control the expected value of FDP, namely the false discovery rate (FDR) [1-5]. Indeed, it is well known that controlling FDR has power advantages over the traditional way of controlling family-wise type I error [1,2]. Suppose a study is properly designed to control the FDR at 5%. If such a study is independently repeated many times, the average of the FDPs in these repeated studies can be expected to be no more than 5%. However, for a particular study (without repetition), the FDP is more directly relevant than FDR. Therefore, when a study is designed to control FDR under common designs, it is still very much desirable to assess FDP, e.g. to construct a prediction interval for the FDP. One can also consider designing a study controlling FDP instead of FDR. This approach has been less successful since FDP is a random variable and is less straightforward to control than the FDR. Indeed, researchers have proposed various procedures aimed at controlling the FDP [6-13], from which confidence envelopes for the FDP can be obtained simultaneously for all possible rejection regions. However, confidence envelopes from the existing FDP controlling procedures are often too conservative for predicting a tight range for the FDP. In particular, when weak correlations exist among test statistics, methods for constructing tight prediction interval for the FDP are still limited.

In the multiple testing context, test statistics are often correlated, e.g. in microarray experiments and functional magnetic resonance imaging, correlations arise due to biological, spatial, temporal or technical factors. A major challenge for predicting FDP is to account for unknown correlations between test statistics. It has been shown via numerical studies that when test statistics are correlated, the variability of FDP can increase dramatically [14-17]. This can also be seen from the variance formula derived

*Corresponding author.

in the next section (Formula (2)). Permutation-based methods are often considered in the presence of dependency, e.g. [15]. Permutation-based methods have several limitations. For instance, they are not applicable when no group structure (e.g., groups of cases and controls) is present as in some imaging studies [9]. Additionally, if the purpose of testing is to detect differences in means, then a permutation-based test can have an inflated Type I error rate by picking up signals due to unequal variances or skewness of two distributions [18]. Pan [3] and Xie *et al.* [19] also pointed out that permutation-based procedures tend to overestimate FDR. Finally, a permutation-based approach is generally very computationally intensive; it often becomes not feasible when the number of tests is large.

Other works on FDP have been proposed with efforts to accommodate the correlations among test statistics. For example, Ge *et al.* [12,20] proposed a formula for the upper prediction bound of the FDP assuming that test statistics under true null hypotheses are independent and also proposed a permutation algorithm to obtain a simultaneous upper prediction band of the FDP. Under the assumption that *p*-values are independent or follow a conditional equicorrelated multivariate normal model, Roquain and Villers [21] provided exact calculations for the cumulative distribution function (CDF) and moments of FDP for the step-up and step-down procedures. Ghosal and Roy [22] proposed a nonparametric Bayesian procedure to obtain the posterior distribution of FDP under the intraclass or autoregressive correlation structure. In all these studies on FDP under dependence, the correlation among test statistics is either ignored or assumed to follow some parametric models. A flexible semiparametric approach to modeling dependency among test statistics has not emerged.

In this paper, we first derive an explicit formula for the variance of the FDP under a semiparametric weak dependence assumption among the test statistics. The variance formula is easily interpretable and elucidates the effect of correlation on the variability of FDP. Using the variance formula, we obtain an upper prediction interval for the FDP. This approach is semiparametric in nature because only the average of the pairwise Pearson correlation between test statistics needs to be estimated. The formula-based prediction interval is easy to evaluate even when testing a vast number of hypotheses where no other methods are computationally feasible. Simulation studies indicate that the formula-based prediction interval has good coverage probabilities under weak to moderate dependence. In many situations, as illustrated, the prediction interval is quite short (tight) and generally more accurate than competitors. In addition, we discuss a permutation-based upper prediction interval for FDP which is useful under strong dependence. We illustrate the proposed prediction intervals using a prostate cancer dataset.

2. Methods

2.1. Notation

Consider testing m hypotheses simultaneously. Rejections are made based on p-values, with a fixed rejection region $[0,\alpha)$ for some α. Denote the rejection status of the ith test by $R_i(\alpha) = I(p_i < \alpha)$, where p_i denotes the p-value of the ith test and $I(\cdot)$ is an indicator function. Denote the power of the ith test as $1-\beta_i$.

Let V and U be the total number of incorrect and correct rejections, respectively. The total number of rejections or discoveries is $R = \sum_{i=1}^{m} R_i(\alpha)$. Let M_0 denote the index set of the m_0 tests for which null hypotheses are true and M_1 the index set of $m_1 = m - m_0$ tests for which alternative hypotheses are true. The proportion of true null hypotheses is $\pi_0 = \frac{m_0}{m}$. When test statistics are dependent, as in [23], we have

$$\begin{aligned}\mathrm{Var}(V) &= m_0\alpha(1-\alpha) + \sum_{i,j\in M_0, i\neq j} \theta_V^{ij}\alpha(1-\alpha)\\ &= m_0\alpha(1-\alpha)\left\{1+(m_0-1)\bar{\theta}_V\right\},\end{aligned}$$

where $\theta_V^{ij} = \mathrm{corr}\left\{R_i(\alpha), R_j(\alpha)\right\}$ for $i, j \in M_0$, and $\bar{\theta}_V = \frac{\sum_{i,j\in M_0, i\neq j}\theta_V^{ij}}{m_0(m_0-1)}$ is the average correlation coefficient. Similarly, for the correct rejections,

$$\begin{aligned}\mathrm{Var}(U) &= \sum_{i\in M_1}\beta_i(1-\beta_i)\\ &\quad + \sum_{i,j\in M_1, i\neq j}\theta_U^{ij}\sqrt{\beta_i(1-\beta_i)\beta_j(1-\beta_j)},\end{aligned}$$

where $\theta_U^{ij} = \mathrm{corr}\left\{R_i(\alpha), R_j(\alpha)\right\}$ for $i, j \in M_1$. Denote $\bar{\beta} = \frac{1}{m_1}\sum_{i\in M_1}\beta_i$. If effect sizes are all equal, *i.e.* $\beta_i = \bar{\beta}$ for all i, we can obtain a simplified formula $\mathrm{Var}(U) = m_1\bar{\beta}(1-\bar{\beta})\left\{1+(m_1-1)\bar{\theta}_U\right\}$, where $\bar{\theta}_U = \frac{\sum_{i,j\in M_1, i\neq j}\theta_U^{ij}}{m_1(m_1-1)}$. Additionally, let $\theta_{UV}^{ij} = \mathrm{corr}\left\{R_i(\alpha), R_j(\alpha)\right\}$ for $i \in M_1$, $j \in M_0$. Denote the average correlation $\bar{\theta}_{UV} = \frac{\sum_{i\in M_1}\sum_{j\in M_0}\theta_{UV}^{ij}}{m_0 m_1}$.

Table 1 summarizes the outcomes of m tests and their expected values.

2.2. Formula-Based Prediction Interval

2.2.1. Derivation of Prediction Interval

Under some general regularity conditions including weak

Table 1. Outcome and expected outcome of testing *m* hypotheses.

	Outcome		
	Reject H_0	Accept H_0	Total
H_0 is true	V	$m_0 - V$	m_0
H_1 is true	U	$m_1 - U$	m_1
Total	R	$m - R$	m
	Expected Outcome		
	Reject H_0	Accept H_0	Total
H_0 is true	$m_0\alpha$	$m_0(1-\alpha)$	m_0
H_1 is true	$m_1(1-\bar{\beta})$	$m_1\bar{\beta}$	m_1
Total	$m_0\alpha + m_1(1-\bar{\beta})$	$m_0(1-\alpha)+m_1\bar{\beta}$	m

dependence among test statistics, Farcomeni [24] proved that the FDP, $\left\{Q_\alpha = \frac{V_\alpha}{R_\alpha \vee 1}, \alpha \in (0,1)\right\}$, as a stochastic process indexed by α, is an asymptotically Gaussian process (see Theorem 2 of [24]). In particular, for a fixed α, the FDP has an asymptotically normal distribution under weak dependence as discussed in Farcomeni [24]. More specifically, assuming that $0 < \pi_0 < 1$, when test statistics are independent or weakly dependent, $\bar{\theta}_V \to 0$, $\bar{\theta}_U \to 0$ and $\bar{\theta}_{UV} \to 0$ as $m \to \infty$, we show that FDP follows a Normal distribution asymptotically with special mean and variance (see Appendix A (A.1)). No assumptions about higher-order correlation terms are required.

When effect sizes are all equal, explicit forms for the mean (μ_Q) and variance (σ_Q^2) of the FDP can be easily obtained using the delta method and are given by

$$\mu_Q \approx \frac{\pi_0\alpha}{\pi_0\alpha + (1-\pi_0)(1-\bar{\beta})}, \tag{1}$$

$$\sigma_Q^2 \approx \frac{\pi_0(1-\pi_0)^2\alpha(1-\alpha)(1-\bar{\beta})}{\left\{\pi_0\alpha + (1-\pi_0)(1-\bar{\beta})\right\}^4}\Sigma, \tag{2}$$

where

$$\Sigma = \frac{1}{m}\left(1-\bar{\beta} + \frac{\pi_0}{1-\pi_0}\omega\bar{\beta}\right) + \left(\pi_0 - \frac{1}{m}\right)(1-\bar{\beta})\bar{\theta}_V + \pi_0\omega\bar{\beta}\bar{\theta}_U - 2\pi_0\sqrt{\omega\bar{\beta}(1-\bar{\beta})}\bar{\theta}_{UV} \tag{3}$$

and $\omega = \frac{\alpha}{1-\alpha}$. For a moderate to large sample size n, the average Type II error $\bar{\beta} \to 0$. Then

$$\sigma_Q^2 \to c\left\{\frac{1}{m} + \left(\pi_0 - \frac{1}{m}\right)\bar{\theta}_V + \pi_0\omega\bar{\beta}\bar{\theta}_U - 2\pi_0\sqrt{\omega\bar{\beta}}\bar{\theta}_{UV}\right\} \tag{4}$$

where $c = \frac{\pi_0(1-\pi_0)^2\alpha(1-\alpha)}{(\pi_0\alpha + 1 - \pi_0)^4}$. From (4), it is evident that when all the test statistics are independent, σ_Q^2 is inversely proportional to m; when some test statistics are dependent, correlations also contribute to the variance. The rejection threshold α in multiple testing is typically less than 0.05 and thus ω is small, making the last two terms of (4) small. The average correlation among true null test statistics, which is represented by $\bar{\theta}_V$, can have a large influence on the variance of FDP.

When all the parameters are known, a prediction interval could be derived based on the asymptotic distribution of FDP. We shall discuss the estimation of parameters in details next section. In multiple testing we are primarily concerned about high FDPs so an upper prediction interval is of interest. A $100(1-\gamma)\%$ upper prediction interval for FDP is given by $\left[0, \mu_Q + z_\gamma\sigma_Q\right]$, where z_γ is the $100(1-\gamma)$th quantile of the standard normal distribution.

With finite sample size in practice, the distribution of FDP under dependence is often skewed, suggesting transformations such as the log-transformation to be practically useful. Moreover, by the delta method (see Appendix A for more details), when the FDP is asymptotically normal, Y = log(FDP) is also asymptotically normal, *i.e.*, $Y \sim N(\mu_Y, \sigma_Y^2)$ asymptotically, where formulas for μ_Y and σ_Y^2 are derived in Appendix A (A.2). Thus it is not surprising that Y = log(FDP) is closer to normal than the FDP itself, particularly under weak dependence. The approximate mean and variance of Y are:

$$\mu_Y \approx \log(\mu_Q) \approx \log\left\{\frac{\pi_0\alpha}{\pi_0\alpha + (1-\pi_0)(1-\bar{\beta})}\right\} \tag{5}$$

$$\sigma_Y^2 \approx \frac{(1-\pi_0)^2(1-\alpha)(1-\bar{\beta})}{\pi_0\alpha\left\{\pi_0\alpha + (1-\pi_0)(1-\bar{\beta})\right\}^2}\Sigma \tag{6}$$

where Σ is in (3). Applying the exponential transformation, a $100(1-\gamma)\%$ upper prediction interval for the FDP can be constructed as $\left[0, \exp(\mu_Y + z_\gamma\sigma_Y)\right]$.

2.2.2. Estimation

To calculate the formula-based prediction interval $\left[0, \exp(\mu_Y + z_\gamma\sigma_Y)\right]$, we need to estimate necessary parameters in μ_Y and σ_Y first. We adopt the estimator for π_0 proposed by Storey [2]: $\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m}$ for $\lambda \in [0,1]$ with the choice of $\lambda = 0.5$. We use the method of moment to estimate $\bar{\beta}$. Because $E[R] = m_0\alpha + m_1(1-\bar{\beta})$, plugging in the estimate of π_0 and the observed total number of rejections R, we have $\hat{\beta} = 1 - \frac{R - m\hat{\pi}_0\alpha}{m(1-\hat{\pi}_0)}$. The resulting estimator $\hat{\mu}_Q$ is

essentially the same as the FDR estimator proposed by Storey [2]: $\hat{\mu}_Q = \frac{m\hat{\pi}_0\alpha}{R}$. However, the objective here is to obtain a predictive interval or equivalently an upper bound for FDP since we mainly care about large values of FDP.

The correlation θ^{ij} between the *i*th and *j*th rejection indicators is

$$\theta^{ij} = \frac{E\left[R_i(\alpha)R_j(\alpha)\right] - E\left[R_i(\alpha)\right]E\left[R_j(\alpha)\right]}{\sqrt{\mathrm{Var}\left(R_i(\alpha)\right)\mathrm{Var}\left(R_j(\alpha)\right)}}.$$

We here consider one-sided *z*-test for two-group comparison to illustrate the estimation of correlation. Two-sided *z*-test and *t*-test are given in Appendix B. Following the notation defined in Section 2.1, we have

$$\theta_V^{ij} = \frac{\Psi\left(-z_\alpha, -z_\alpha; \rho^{ij}\right) - \alpha^2}{\alpha(1-\alpha)}, \quad (7)$$

$$\theta_U^{ij} = \frac{\Psi\left(z_{\bar{\beta}}, z_{\bar{\beta}}; \rho^{ij}\right) - \left(1-\bar{\beta}\right)^2}{\bar{\beta}\left(1-\bar{\beta}\right)}, \quad (8)$$

$$\theta_{UV}^{ij} = \frac{\Psi\left(-z_\alpha, z_{\bar{\beta}}; \rho^{ij}\right) - \alpha\left(1-\bar{\beta}\right)}{\sqrt{\alpha(1-\alpha)\bar{\beta}\left(1-\bar{\beta}\right)}}, \quad (9)$$

where Ψ is the CDF of the standard bivariate normal distribution, and ρ^{ij} denotes the Pearson correlation between the *i*th and *j*th test statistics. We propose the following procedure to estimate the average correlations. In practice, when *m* is very large ($m > 2000$), we propose to run the procedure on a random subset of *m* tests to save computation time.

1) Estimate the correlations between test statistics ρ^{ij} using the sample correlations. As in [25], an empirical Bayes shrinkage estimator of sample correlations can be used.

2) For the *i*th test with *z*-score z_i, estimate the conditional probability of the corresponding hypothesis being a true alternative hypothesis:

$$P\left(i \in M_1 \middle| z_i\right) = \frac{\phi\left(z_i - \mu_z\right)\left(1-\pi_0\right)}{\phi\left(z_i - \mu_z\right)\left(1-\pi_0\right) + \phi\left(z_i\right)\pi_0}$$

where $\phi(z)$ is the density function of $N(0,1)$ and μ_z is the mean of *z*-scores when the alternative hypothesis is true. From the estimate $\hat{\beta}$ we can get $\hat{\mu}_z = z_\alpha - \Phi^{-1}\left(\hat{\beta}\right)$.

3) Predict whether the tests belong to M_1 or M_0 by generating Bernoulli random variables with the estimated probability $P\left(i \in M_1 \middle| z_i\right)$.

4) After the identity of each test is predicted, the correlation θ^{ij} between any two rejection indicators can be calculated from Formulae (7)-(9).

5) Estimate the average correlations $\bar{\theta}_V$, $\bar{\theta}_U$ and $\bar{\theta}_{UV}$ using the respective sample means of pairwise correlations.

6) Repeat steps 3 to 5 for a few times, and take the average of these estimates of average correlations.

2.3. Permutation-Based Prediction Interval

The permutation-based procedure proposed by Korn *et al.* [6,7] can be adapted to construct an upper prediction interval for the FDP under general dependence. The method can be expected to be robust because it does not depend on parametric or weak dependence assumptions, but it requires very intensive computation which may not be feasible for testing a very large number of hypotheses. Let n_1 and n_2 be the sample sizes of the two groups and suppose that unpaired *t*-test is performed. First, permute the group labels and calculate the two-sample *t*-statistic *p*-values for all *m* tests under the permutated labels. If the number of possible permutations $\frac{(n_1+n_2)!}{n_1!n_2!}$ is too large, perform $w = 500$ or 1000 random permutations. Second, for each permutation, order the *p*-values from smallest to largest. Let $p_{(1)}^j, p_{(2)}^j, \cdots, p_{(m)}^j$ denote the ordered *p*-values for the *j*th permutation, $j = 1,2,\cdots,w$. Write the ordered *p*-values in a $w \times m$ matrix:

$$\begin{pmatrix} p_{(1)}^1 & p_{(2)}^1 & \cdots & p_{(m)}^1 \\ p_{(1)}^2 & p_{(2)}^2 & \cdots & p_{(m)}^2 \\ \cdots & & & \\ p_{(1)}^w & p_{(2)}^w & \cdots & p_{(m)}^w \end{pmatrix}.$$

Third, order the *p*-values within each column and put the smallest *p*-values on top. Denote the resulting matrix by *S*, and its element at the *j*th row and *l*th column by S_l^j where $j = 1,2,\cdots,w$ and $l = 1,2,\cdots,m$.

To construct a $100(1-\gamma)\%$ upper prediction interval for the FDP, we first find an upper bound $\tilde{V}$ for *V*. Find the $[\gamma w]$th row of the matrix *S* where [a] denotes the closest integer smaller than or equal to a. The upper bound for *V* can be estimated as: $\tilde{V}(\alpha) = \sum_{l=1}^m I\left(S_l^{[\gamma w]} \le \alpha\right)$. By construction $S_1^j < S_2^j < \cdots < S_m^j$ for $j = 1,2,\cdots,w$. Using Korn's controlling procedure, $S_{k+1}^{[\gamma w]}$ is the threshold for at most *k* false discoveries with 1-γ probability. Hence, *k* is the $100(1-\gamma)\%$ upper bound for *V* at threshold $S_{k+1}^{[\gamma w]}$. Given a rejection region $[0,\alpha)$, the definition of $\tilde{V}(\alpha)$ implies that $S_{\tilde{V}(\alpha)}^{[\gamma w]} \le \alpha < S_{\tilde{V}(\alpha)+1}^{[\gamma w]}$. Then

$$P\left(V(\alpha) \le \tilde{V}(\alpha)\right) > P\left(V\left(S_{\tilde{V}(\alpha)+1}^{[\gamma w]}\right) \le \tilde{V}(\alpha)\right) \ge 1-\gamma.$$

Therefore, $\tilde{V}(\alpha)$ is a conservative $100(1-\gamma)\%$ upper bound for $V(\alpha)$ at a fixed α. Then the upper bound for the FDP can be calculated as $\frac{\tilde{V}}{R \vee 1}$. Since the permutation approach preserves the correlation structure of the data, it works under potentially strong dependence structure.

3. Numerical Studies

3.1. Simulation: Formula-Based Upper Prediction Intervals

In this section, we evaluate the performance of the formula-based prediction interval under various correlation structures via simulations and compare it with the method of Ge *et al.* [20]. We considered m = 10,000 hypotheses to be tested using one-sided z-test and π_0 = 0.7. Results from two-sided z-tests were similar and not reported. For true null and true alternative hypotheses, z scores were generated from $N(0,\Sigma_n)$ and $N(\mu_a,\Sigma_a)$ respectively, where μ_a was 2.1, 2.7 or 4.3 and the correlation matrix will be specified below. The diagonal entries of both Σ_n and Σ_a were set to be 1. The threshold α was fixed to be 0.0085.

Two scenarios were considered: null test statistics were moderately dependent and alternative test statistics were weakly dependent; both null and alternative test statistics were weakly dependent. A proportion of test statistics were set to be correlated, with blockwise dependence or unstructured sparse dependence. In the blockwise dependence structure, tests were correlated within blocks and independent across blocks with the block-size of 50. We set 25% null test statistics to be correlated with correlation 0.8 and 5% alternative test statistics to be correlated with correlation 0.2; or 5% null test statistics to be correlated with correlation 0.2 and 5% alternative test statistics to be correlated with correlation 0.5.

We evaluated the performance of the proposed prediction intervals using the true correlations between test statistics. **Table 2** shows results from 1000 replications. When null test statistics are moderately correlated (upper panel), the coverage probabilities of our prediction intervals are close to the nominal levels. The interval with log transformation is more accurate than the one without transformation (results not shown). In comparison, Ge's intervals have the problem of under-coverage because the required independence assumption is violated. When both null and alternative test statistics are weakly correlated (lower panel), our prediction intervals have good coverage probabilities and are tighter than Ge's. The estimates of the standard deviation of FDP are very close to the true values in both scenarios.

For the general sparse dependence structure, we set aside a small proportion of test statistics to be correlated and the rest of test statistics independent. We first generated a lower triangular matrix A with diagonal entries equal to 1 and lower off-diagonal entries simulated from $N(0.1, 0.1)$. The covariance matrix was computed as AA^{T}, and was normalized to be a correlation matrix for the dependent test statistics. Null tests and alternative tests can be correlated but with no dependence structure assumed. For the two scenarios, we set 750 null and 50 alternative test statistics to be correlated; or 100 null and 400 alternative test statistics to be correlated.

Table 2. Estimates of σ_Q, upper limits (UL) of prediction intervals and coverage probabilities (CP) (all in %) under blockwise dependence.

True		Estimates		Conf. level	FB		Ge	
FDR	σ_Q	$\widehat{\text{FDR}}$	$\hat{\sigma}_Q$		UL	CP	UL	CP
2.0	0.54	2.0	0.54	90	2.8	89.8	2.5	81.1
				95	3.1	94.4	2.6	83.0
3.0	0.84	3.0	0.83	90	4.3	90.0	3.7	81.3
				95	4.7	94.2	3.9	84.1
5.0	1.32	5.1	1.31	90	7.1	90.8	6.3	83.7
				95	7.8	94.5	6.5	85.2
2.0	0.27	2.0	0.26	90	2.4	89.2	2.5	94.4
				95	2.5	94.6	2.6	96.8
3.0	0.38	3.0	0.39	90	3.6	90.7	3.7	95.7
				95	3.7	95.5	3.9	97.2
5.0	0.66	5.1	0.66	90	6.0	92.4	6.3	96.5
				95	6.3	96.6	6.5	98.3

σ_Q: sd(FDP); FB: formula-based prediction interval with log transformation; Ge: Ge's prediction interval; Upper panel: 25% null test statistics are correlated with ρ_v = 0.8, 5% alternative test statistics are correlated with ρ_u = 0.2; Lower panel: 5% null test statistics are correlated with ρ_v = 0.2, 5% alternative test statistics are correlated with ρ_u = 0.5.

Results from 1000 replications are shown in **Table 3**. The upper panel shows the situation where more null test statistics are correlated. Our prediction intervals have good coverage probabilities, while Ge's intervals undercover the true FDP. The lower panel shows the situation where more alternative test statistics are correlated. Our prediction intervals cover the true FDP well while Ge's intervals are conservative.

3.2. Comparison with Simultaneous Prediction Band

We considered two-group mean comparison in the context of gene expression study and simulated the expression data to assess the performance of formula-based and permutation-based prediction intervals. We compared with the method of Ge *et al.* [20] and the simultaneous prediction band method of Meinshausen [11]. We set m = 5000 and π_0 = 0.7. The total sample size was set to be

Table 3. Estimates of σ_Q, upper limits (UL) of prediction intervals and coverage probabilities (CP) (all in %) under sparse dependence.

True		Estimates		Conf. level	FB		Ge	
FDR	σ_Q	$\widehat{\text{FDR}}$	$\hat{\sigma}_Q$		UL	CP	UL	CP
2.0	0.54	2.0	0.53	90	2.8	92.5	2.5	87.8
				95	3.1	94.7	2.6	89.6
3.0	0.79	3.0	0.79	90	4.1	90.0	3.6	88.6
				95	4.4	93.9	3.8	90.3
5.0	1.24	5.1	1.22	90	6.8	90.6	6.2	88.1
				95	7.4	93.8	6.5	89.7
2.0	0.26	2.0	0.27	90	2.4	90.7	2.5	95.2
				95	2.5	95.1	2.6	97.3
3.0	0.43	3.0	0.42	90	3.6	91.0	3.7	93.8
				95	3.8	95.0	3.8	96.8
5.0	0.70	5.1	0.71	90	6.1	91.2	6.3	97.0
				95	6.4	95.4	6.5	98.6

σ_Q: sd(FDP); FB: formula-based prediction interval with log transformation; Ge: Ge's prediction interval; Upper panel: 750 (11%) null test statistics are correlated with $\bar{\rho}_v$ = 0.33, 50 (1.7%) alternative test statistics are correlated with $\bar{\rho}_u$ = 0.16, and $\bar{\rho}_{uv}$ = 0.096; Lower panel: 100 (1.4%) null test statistics are correlated with $\bar{\rho}_v$ = 0.19, 400 (13%) alternative test statistics are correlated with $\bar{\rho}_u$ = 0.13, and $\bar{\rho}_{uv}$ = 0.05.

100, 150 or 200 and equally divided between two groups. The data were generated from $N(\mu_n, \Sigma_n)$ or $N(\mu_a, \Sigma_a)$ for the two groups, where μ_n = 0 and μ_a = 0.6. The diagonal entries of both Σ_n and Σ_a were 1. Blockwise correlation structure was used and block-size was 50. We set 20% null genes to be correlated with correlation 0.8 within block, and 3.3% alternative genes to be correlated with correlation 0.2 within block.

One-sided *t*-test was performed and the threshold α was fixed at 0.01. For calculating the formula-based interval, correlations between test statistics were estimated from correlations between gene expression levels. Pairwise Peason correlations were calculated from 500 randomly chosen genes across all the subjects, after subtracting off each gene's mean within each group as in [26]. Sample correlations were then shrunk using the empirical Bayes method [25] to correct the well known inflation of variability in correlation estimates. The correlations θ between rejection status were then calculated using the procedure in Section 2.2.2, repeating the procedure for 3 times. For calculating the permutation-based prediction intervals, a total of 500 randomly chosen permutations of the groups were used.

Coverage probabilities of four prediction intervals are given in **Table 4** (200 replications). The formula-based prediction intervals cover the true FDP well and are slightly conservative. Since the sample correlations are still over-dispersed after shrinkage, the variance of FDP is over-estimated. The permutetion-based interval is more conservative than the formula-based ones. In contrast, Ge's prediction interval is too liberal. The simultaneous prediction bands are about twice as high as the formula-based intervals. Hence it is not very useful when point-wise intervals are needed. In terms of computational efficiency, when the sample size was 100, the central processor unit (CPU) time for calculating the formula-based and permutation-based prediction intervals in one run of simulation was 81 seconds and 20 minutes respectively, on a 2.66 GHz processor with 4 GB of memory. The permutation-based approach will become more computationally intensive as *m* gets larger.

We have also varied *m*, rejection region and correlation structures. When the dependence is weak, our formula-based prediction interval works well in various scenarios. It is the tightest one among all intervals that we study.

3.3. A Real Data Example

The study in Wang *et al.* [27] measured 13,935 mRNA gene expression levels in 125 lymphoblastoid cell lines derived from 62 aggressive and 63 nonaggressive prostate cancer patients. The purpose is to identify candidate genes whose expression levels are associated with aggressive phenotype of prostate cancer. Two sample two-

Table 4. Comparison of the upper limits (UL) and coverage probabilities (CP) of four prediction intervals (all in %).

True		Estimates		*n*	Conf. level	FB		Per		Ge		MN	
FDR	σ_Q	$\widehat{\text{FDR}}$	$\hat{\sigma}_Q$			UL	CP	UL	CP	UL	CP	UL	CP
3.2	1.22	3.1	1.31	100	90	5.3	92.0	6.5	95.5	4.0	85.0	10.0	100.0
					95	6.2	94.5	7.6	99.0	4.2	86.0	11.2	100.0
2.5	0.90	2.5	1.02	150	90	4.3	93.0	5.3	97.5	3.3	86.5	8.2	100.0
					95	5.0	95.0	6.2	98.0	3.4	89.5	9.1	100.0
2.0	0.74	2.0	0.87	200	90	4.0	94.0	5.0	99.0	3.1	86.0	7.3	100.0
					95	4.6	96.5	5.8	100.0	3.2	88.5	8.0	100.0

σ_Q: sd(FDP); FB: formula-based prediction interval with log transformation; Per: permutation-based upper prediction bounds; Ge: Ge's prediction interval; MN: simultaneous prediction bands using Meinshausen's permutation algorithm [11]; *n*: sample size.

sided t tests were performed for each gene, and the proportion of true null hypotheses was estimated to be $\hat{\pi}_0 = 0.60$. The FDR controlling procedure in [2] was used to control FDR at 3% or 5%, rejecting 1708 or 2208 hypotheses respectively. Sample correlations were estimated from 2000 randomly sampled genes repeating the estimation procedure for 3 times. The correlation is weak and the average of estimated ρ_V is very close to 0. The estimated $\overline{\theta}_V$ is 0.0059 at $\alpha = 0.006$. The formula-based upper prediction intervals with log transformation (FB), permutation-based intervals (Per) and simultaneous prediction bands (MN) are shown in **Table 5**. When FDR is controlled at 5%, with 90% probability the actual FDP is as high as 12.7% (FB). Hence with the correlations in this dataset, FDP could far exceed its mean with high probability. Since the purpose of the study is to identify target genes for a large-scale validation study, a smaller rejection region may be more appropriate to avoid excessive false positives. The permutation-based approach gives more conservative intervals than the formula-based one. The simultaneous prediction bands are high and too conservative for fixed rejection regions.

4. Discussion

It is feasible to construct a tight prediction interval for the FDP without specifying a parametric correlation structure for test statistics. When the dependence is weak, we derived a prediction interval for the FDP based on the variance formula which takes correlations into consideration. This formula-based approach is computationally efficient even when the number of tests is very large. The prediction interval could help investigators decide what rejection regions are suitable for a particular study to control FDR. If the upper limit of prediction interval is unacceptably high, then selecting a smaller rejection region might be more appropriate. We also discussed a permutation procedure which can be employed to find a prediction interval for the FDP without assuming weak dependence. This approach can be computationally quite intensive, especially when the number of tests is large.

Table 5. Comparison of upper limits (UL) of prediction intervals for the prostate cancer data (all in %).

Method	$\widehat{\mathrm{FDR}}(\hat{\sigma}_0)$	90% UL	95% UL
FB	3.0 (2.72)	9.6	13.3
	5.0 (3.63)	12.7	16.5
Per	3.0	10.1	13.4
	5.0	15.5	20.2
MN	3.0	22.2	26.8
	5.0	29.5	36.2

FB: formula-based prediction intervals with log transformation; Per: permutation-based prediction intervals; MN: simultaneous prediction bands using Meinshausen's permutation algorithm.

5. Acknowledgements

We would like to thank the reviewers for their careful reading of our paper. This research was partially supported by a Stony Wold-Herbert Foundation grant and the NYU Cancer Center Supporting Grant (2P30 CA16087-23), and by the NYU NIEHS Center Grant (5P30 ES00260-44).

REFERENCES

[1] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate—A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society Series B-Methodological*, Vol. 57, No. 1, 1995, pp. 289-300.

[2] J. D. Storey, "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, Vol. 64, No. 3, 2002, pp. 479-498. doi:10.1111/1467-9868.00346

[3] W. Pan, "On the Use of Permutation in and the Performance of a Class of Nonparametric Methods to Detect Differential Gene Expression," *Bioinformatics*, Vol. 19, No. 11, 2003, pp. 1333-1340. doi:10.1093/bioinformatics/btg167

[4] J. D. Storey, J. E. Taylor and D. Siegmund, "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, Vol. 66, No. 1, 2004, pp. 187-205. doi:10.1111/j.1467-9868.2004.00439.x

[5] X. D. Zhang, P. F. Kuan, M. Ferrer, X. Shu, Y. C. Liu, A. T. Gates, P. Kunapuli, E. M. Stec, M. Xu, S. D. Marine, *et al.*, "Hit Selection with False Discovery Rate Control in Genome-Scale Rnai Screens," *Nucleic Acids Research*, Vol. 36, No. 14, 2008, pp. 4667-4679. doi:10.1093/nar/gkn435

[6] E. L. Korn, J. F. Troendle, L. M. McShane and R. Simon, "Controlling the Number of False Discoveries: Application to High-Dimensional Genomic Data," *Journal of Statistical Planning and Inference*, Vol. 124, No. 2, 2004, pp. 379-398. doi:10.1016/S0378-3758(03)00211-8

[7] E. L. Korn, M. C. Li, L. M. McShane and R. Simon, "An Investigation of Two Multivariate Permutation Methods for Controlling the False Discovery Proportion," *Statistics in Medicine*, Vol. 26, No. 24, 2007, pp. 4428-4440. doi:10.1002/sim.2865

[8] C. R. Genovese and L. Wasserman, "A Stochastic Process Approach to False Discovery Control," *Annals of Statistics*, Vol. 32, No. 3, 2004, pp. 1035-1061. doi:10.1214/009053604000000283

[9] C. R. Genovese and L. Wasserman, "Exceedance Control of the False Discovery Proportion," *Journal of the American Statistical Association,* Vol. 101, No. 476, 2006, pp. 1408-1417. doi:10.1198/016214506000000339

[10] M. J. van der Laan, S. Dudoit and K. S. Pollard, "Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives," *Statistical Applications in Genetics and Molecular Biology*, Vol. 3, No. 1, 2004, Article 15.

[11] N. Meinshausen, "False Discovery Control for Multiple Tests of Association under General Dependence," *Scandinavian Journal of Statistics*, Vol. 33, No. 2, 2006, pp. 227-237. doi:10.1111/j.1467-9469.2005.00488.x

[12] Y. C. Ge, S. C. Sealfon and T. P. Speed, "Multiple Testing and Its Applications to Microarrays," *Statistical Methods in Medical Research*, Vol. 18, No. 6, 2009, pp. 543-563. doi:10.1177/0962280209351899

[13] A. Farcomeni, "Generalized Augmentation to Control the False Discovery Exceedance in Multiple Testing," *Scandinavian Journal of Statistics*, Vol. 36, No. 3, 2009, pp. 501-517.

[14] Q. Yang, J. Cui, I. Chazaro, L. A. Cupples and S. Demissie, "Power and Type I Error Rate of False Discovery Rate Approaches in Genome-Wide Association Studie," *BMC Genetics*, Vol. 6, Suppl. 1, 2005.

[15] Y. Pawitan, S. Calza and A. Ploner, "Estimation of False Discovery Proportion under General Dependence," *Bioinformatics*, Vol. 22, No. 24, 2006, pp. 3025-3031. doi:10.1093/bioinformatics/btl527

[16] R. Heller, "Correlated Z-Values and the Accuracy of Large-Scale Statistical Estimates Comment," *Journal of the American Statistical Association*, Vol. 105, No. 491, 2010, pp. 1057-1059. doi:10.1198/jasa.2010.tm10240

[17] A. Schwartzman and X. H. Lin, "The Effect of Correlation in False Discovery Rate Estimation," *Biometrika*, Vol. 98, No. 1, 2011, pp. 199-214. doi:10.1093/biomet/asq075

[18] Y. F. Huang, H. Y. Xu, V. Calian and J. C. Hsu, "To Permute or Not to Permute," *Bioinformatics*, Vol. 22, No. 18, 2006, pp. 2244-2248. doi:10.1093/bioinformatics/btl383

[19] Y. Xie, W. Pan and A. B. Khodursky, "A Note on Using Permutation-Based False Discovery Rate Estimates to Compare Different Analysis Methods for Microarray Data," *Bioinformatics*, Vol. 21, No. 23, 2005, pp. 4280-4288. doi:10.1093/bioinformatics/bti685

[20] Y. C. Ge and X. Li, "Control of the False Discovery Proportion for Independently Tested Null Hypotheses," *Journal of Probability and Statistics*, 2012, in Press.

[21] E. Roquain and F. Villers, "Exact Calculations for False Discovery Proportion with Application to Least Favorable Configurations," *Annals of Statistics,* Vol. 39, No. 1, 2011, pp. 584-612. doi:10.1214/10-AOS847

[22] S. Ghosal and A. Roy, "Predicting False Discovery Proportion under Dependence," *Journal of the American Statistical Association*, Vol. 106, No. 495, 2011, pp. 1208-1218. doi:10.1198/jasa.2011.tm10488

[23] Y. Shao and C. H. Tseng, "Sample Size Calculation with Dependence Adjustment for FDR-Control in Microarray Studies," *Statistics in Medicine*, Vol. 26, No. 23, 2007, pp. 4219-4237. doi:10.1002/sim.2862

[24] A. Farcomeni, "Some Results on the Control of the False Discovery Rate under Dependence," *Scandinavian Journal of Statistics*, Vol. 34, No. 2, 2007, pp. 275-297. doi:10.1111/j.1467-9469.2006.00530.x

[25] B. Efron, "Empirical Bayes Estimates for Large-Scale Prediction Problems," *Journal of the American Statistical Association*, Vol. 104, No. 487, 2009, pp. 1015-1028. doi:10.1198/jasa.2009.tm08523

[26] B. Efron, "Correlation and Large-Scale Simultaneous Significance Testing," *Journal of the American Statistical Association*, Vol. 102, No. 477, 2007, pp. 93-103. doi:10.1198/016214506000001211

[27] L. Wang, H. Tang, V. Thayanithy, S. Subramanian, A. L. Oberg, J. M. Cunningham, J. R. Cerhan, C. J. Steer and S. N. Thibodeau, "Gene Networks and MicroRNAs Implicated in Aggressive Prostate Cancer," *Cancer Research*, Vol. 69, No. 24, 2009, pp. 9490-9497. doi:10.1158/0008-5472.CAN-09-2183

Appendix A. Asymptotic Distribution of the FDP

A.1. Under the general weak dependence assumptions the FDP is asymptotically normal (e.g. Theorem 2 of Farcomeni [24]). Thus in this appendix, we assume that some weak dependence assumption is satisfied so that the FDP is asymptotically normal. In particular, we also assume approximate joint normality of V and R. As in [23], when test statistics under H_0 are weakly dependent and $\bar{\theta}_V \to 0$ as $m \to \infty$, the asymptotic normality implies $\frac{V}{m} \sim N\left(\pi_0\alpha, \pi_0\alpha(1-\alpha)\left\{\frac{1}{m} + \bar{\theta}_V\left(\pi_0 - \frac{1}{m}\right)\right\}\right)$.

Similarly, when test statistics under H_1 are weakly dependent and $\bar{\theta}_U \to 0$ as $m \to \infty$, asymptotic normality implies

$$\frac{U}{m} \sim N\left((1-\pi_0)(1-\bar{\beta}), (1-\pi_0)\bar{\beta}(1-\bar{\beta})\left\{ \frac{1}{m} + \bar{\theta}_U \left(1-\pi_0 - \frac{1}{m} \right) \right\} \right).$$

In addition, if $\bar{\theta}_{UV} \to 0$ as $m \to \infty$, then $\mathrm{Cov}(U,V) = m_0 m_1 \sqrt{\alpha(1-\alpha)\bar{\beta}(1-\bar{\beta})}\bar{\theta}_{UV} \to 0$ as $m \to \infty$, so U and V are asymptotically uncorrelated.

When V and R have a joint Normal distribution, for $R > V > 0$, FDP has an approximate Normal distribution. For large m, $\mathrm{FDP} = \frac{V}{R} = \frac{\sigma_V}{\sigma_R} \frac{Z_1 + \mu_V/\sigma_V}{Z_2 + \mu_R/\sigma_R}$, where μ_V, σ_V^2, μ_R, σ_R^2 are the asymptotic mean and variance of V and R, respectively. (Z_1, Z_2) are jointly Normally distributed with mean 0, variance 1 and correlation $\rho(V,R)$. Then by Taylor expansion,

$$\begin{aligned} \frac{Z_1 + \frac{\mu_V}{\sigma_V}}{Z_2 + \frac{\mu_R}{\sigma_R}} &= \frac{\frac{Z_1}{\mu_R/\sigma_R} + \frac{\mu_V/\sigma_V}{\mu_R/\sigma_R}}{\frac{Z_2}{\mu_R/\sigma_R} + 1} \\ &\approx \left(\frac{Z_1}{\mu_R/\sigma_R} + \frac{\mu_V/\sigma_V}{\mu_R/\sigma_R} \right)\left(1 - \frac{Z_2}{\mu_R/\sigma_R} \right) \\ &\approx \frac{1}{\mu_R/\sigma_R} Z_1 - \frac{\mu_V/\sigma_V}{(\mu_R/\sigma_R)^2} Z_2 + \frac{\mu_V/\sigma_V}{\mu_R/\sigma_R} \end{aligned}$$

Therefore, FDP has an approximate Normal distribution. Assume that effect sizes are all equal. The variance of FDP can be derived using the delta method.

$$\begin{aligned} \sigma_Q^2 &\approx \frac{1}{\mu_R^2}\sigma_V^2 + \frac{\mu_V^2}{\mu_R^4}\sigma_R^2 - 2\frac{\mu_V}{\mu_R^3}\mathrm{Cov}(V,R) \\ &\approx \frac{\pi_0(1-\pi_0)^2\alpha(1-\alpha)(1-\bar{\beta})}{\left\{ \pi_0\alpha + (1-\pi_0)(1-\bar{\beta}) \right\}^4} \Sigma \end{aligned}$$

where Σ is in (3).

A.2. By the delta method, $\log(\mathrm{FDP}) - \log(\mu_Q) \to N\left(0, \frac{\sigma_Q^2}{\mu_Q^2}\right)$ in distribution. The mean and variance of $Y = \log(\mathrm{FDP})$ are:

$$\mu_Y \approx \log(\mu_Q) \approx \log\left\{ \frac{\pi_0\alpha}{\pi_0\alpha + (1-\pi_0)(1-\bar{\beta})} \right\}$$

$$\sigma_Y^2 \approx \frac{(1-\pi_0)^2(1-\alpha)(1-\bar{\beta})}{\pi_0\alpha\left\{ \pi_0\alpha + (1-\pi_0)(1-\bar{\beta}) \right\}^2} \Sigma$$

where Σ is in (3).

Appendix B. Correlation Formulas for θ^{ij}

For a two sample two-sided z-test, the common power is

$$1-\bar{\beta} = P(|z_i| > z_{\alpha/2}) = 1 - \Phi(z_{\alpha/2} - \mu_z) + \Phi(-z_{\alpha/2} - \mu_z).$$

From the observed R, $\hat{\beta}$ can be estimated as in Section 2.2.2, and then $\hat{\mu}_z$ is estimated from the above equation. For $i, j \in M_0$, the correlation is $\theta_V^{ij} = \frac{C_1 - \alpha^2}{\alpha(1-\alpha)}$, where

$$C_1 = 2\left\{ \Psi(-z_{\alpha/2}, -z_{\alpha/2}; \rho^{ij}) + \Phi(-z_{\alpha/2}) - \Psi(z_{\alpha/2}, -z_{\alpha/2}; \rho^{ij}) \right\}.$$

For $i, j \in M_1$, $\theta_U^{ij} = \frac{C_2 - (1-\bar{\beta})^2}{\bar{\beta}(1-\bar{\beta})}$, where

$$\begin{aligned} C_2 = {} & \Psi(\tilde{z}_{\alpha/2}^+, \tilde{z}_{\alpha/2}^+; \rho^{ij}) + \Psi(\tilde{z}_{\alpha/2}^-, \tilde{z}_{\alpha/2}^-; \rho^{ij}) \\ & + 2\left\{ \Phi(\tilde{z}_{\alpha/2}^-) - \Psi(-\tilde{z}_{\alpha/2}^+, \tilde{z}_{\alpha/2}^-; \rho^{ij}) \right\}, \end{aligned}$$

$\tilde{z}_{\alpha/2}^+ = -z_{\alpha/2} + \mu_z$ and $\tilde{z}_{\alpha/2}^- = -z_{\alpha/2} - \mu_z$. For other situations, the correlation is $\theta_{UV}^{ij} = \frac{C_3 - \alpha(1-\bar{\beta})}{\sqrt{\alpha(1-\alpha)\bar{\beta}(1-\bar{\beta})}}$, where

$$\begin{aligned} C_3 = {} & \Psi(\tilde{z}_{\alpha/2}^+, -z_{\alpha/2}; \rho^{ij}) + \Psi(\tilde{z}_{\alpha/2}^-, -z_{\alpha/2}; \rho^{ij}) + \Phi(-z_{\alpha/2}) \\ & - \Psi(-\tilde{z}_{\alpha/2}^+, -z_{\alpha/2}; \rho^{ij}) + \Phi(\tilde{z}_{\alpha/2}^-) - \Psi(\tilde{z}_{\alpha/2}^-, z_{\alpha/2}; \rho^{ij}). \end{aligned}$$

If t-tests are performed, we could convert the t statistics to z-scores by a bijective quantile transformation as in [26]. $z_i = \Phi^{-1}(G_{n-2}(t_i))$, $i = 1, 2, \cdots, m$, where G_{n-2} is the CDF of t distribution with $n-2$ degrees of freedom. Then all the previous procedures apply.

