Web Appendix for "False Discovery Control in Large-Scale Spatial Multiple Testing"

Wenguang Sun
University of Southern California, Los Angeles, USA
Brian J. Reich
North Carolina State University, Raleigh, USA
T. Tony Cai
University of Pennsylvania, Philadelphia, USA
Michele Guindani
UT MD Anderson Cancer Center, Houston, USA
Armin Schwartzman
North Carolina State University, Raleigh, USA

1. Proofs of Theorems 2 and 4, and the lemmas

1.1. Proof of Lemma 1

Let G_0, G_1, g_0 and g_1 be defined as in Section 7.1. By Fubini's Theorem, the mFDR of $\boldsymbol{\delta} = [I\{T(s) < t\} : s \in S]$ is

$$\mathrm{mFDR}(t) = \frac{E\left[\int_S \{1-\theta(s)\}\delta(s)d\nu(s)\right]}{E\left\{\int_S \delta(s)d\nu(s)\right\}} = \frac{G_0(t)}{G_0(t)+G_1(t)}.$$

The derivative of $\mathrm{mFDR}(t)$ is

$$\frac{d}{dt}\mathrm{mFDR}(t) = \frac{g_0(t)G_1(t) - g_1(t)G_0(t)}{\{G_0(t)+G_1(t)\}^2} = \frac{g_0(t)\int_0^t \frac{g_1(x)}{g_0(x)}g_0(x)dx - g_1(t)G_0(t)}{\{G_0(t)+G_1(t)\}^2}.$$

If $\boldsymbol{T}$ satisfies the MRC, then $g_1(t)/g_0(t)$ decreases in t. Therefore

$$\frac{d}{dt}\mathrm{mFDR}(t) > \frac{g_0(t)\int_0^t \frac{g_1(t)}{g_0(t)}g_0(x)dx - g_1(t)G_0(t)}{\{G_0(t)+G_1(t)\}^2} = 0$$

and the result follows.

1.2. Proof of Theorem 2

(a). Let $\boldsymbol{\vartheta} = (\vartheta_1, \cdots, \vartheta_K)$ and $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_K)$ be the true states and decisions for clusters $C_1, \cdots, C_K$. The posterior risk is

$$\begin{aligned} E_{\boldsymbol{\vartheta}|\boldsymbol{X}^n}\left[L_w(\boldsymbol{\vartheta}, \boldsymbol{\Delta})\right] &= \sum_{k=1}^{K} \lambda w_k \Delta_k P(\vartheta_k = 0|\boldsymbol{X}^n) + w_k(1-\Delta_k)P(\vartheta_k = 1|\boldsymbol{X}^n) \\ &= \sum_{k=1}^{K} \Delta_k \left\{\lambda w_k P(\vartheta_k = 0|\boldsymbol{X}^n) - w_k P(\vartheta_k = 1|\boldsymbol{X}^n)\right\} + \sum_{k=1}^{K} w_k P(\vartheta_k = 1|\boldsymbol{X}^n). \end{aligned}$$

Therefore the optimal classification rule is $\boldsymbol{\Delta}_{OR} = \{\Delta_{OR}^k : k = 1, \cdots, K\}$, where

$$\Delta_{OR}^k = I\left[\frac{P_\Psi(\vartheta_k = 0|\boldsymbol{X}^n)}{P_\Psi(\vartheta_k = 1|\boldsymbol{X}^n)} < \frac{1}{\lambda}\right] = I\left[T_{OR}(C_k) < (1+\lambda)^{-1}\right].$$

Next we show that $\boldsymbol{T}_{OR}^{\mathcal{C}} = \{T_{OR}(C_k) : k = 1, \cdots, K\}$ satisfies the GMRC. Let $p_k = P(\vartheta_k = 1|\boldsymbol{X}^n)$ and $G_{jk}(t) = P(T_{OR}(C_k) < t|\vartheta_k = j)$, $j = 0, 1$. Next, let $g_{jk}(t)$ be the derivative of $G_{jk}(t)$. The goal is to show that

$$\{\sum_{k=1}^{K} w_k p_k g_{1k}(t)\}/\{\sum_{k=1}^{K} w_k(1-p_k)g_{0k}(t)\}$$

decreases in t. Consider a weighted classification problem with loss function

$$L = \frac{1-t}{t}\sum_{k=1}^{K} w_k(1-\vartheta_k)\Delta_k + \sum_{k=1}^{K} w_k \vartheta_k(1-\Delta_k).$$

Suppose $\boldsymbol{\Delta} = \{I(T_{OR}(C_k) < c) : k = 1, \cdots, K\}$ is used in the weighted classification. Then the classification risk is

$$R = \frac{1-t}{t}\sum_{k=1}^{K}(1-p_k)w_k G_{0k}(c) + \sum_{k=1}^{K} w_k p_k\{1 - G_{1k}(c)\}.$$

The optimal threshold t^* which minimizes the risk satisfies:

$$\frac{1-t}{t}\sum_{k=1}^{K}(1-p_k)w_k g_{0k}(t^*) - \sum_{k=1}^{K} p_k w_k g_{1k}(t^*) = 0.$$

We have shown that the optimal cutoff should be

$$t^* = 1/(\frac{1-t}{t} + 1) = t.$$

Therefore

$$\frac{\sum_{k=1}^{K} w_k p_k g_{1k}(t)}{\sum_{k=1}^{K} w_k(1-p_k)g_{0k}(t)} = t^{-1} - 1$$

and the result follows.

(b). The mFCR of $\mathbf{\Delta} = \{I(T_k < t) : k = 1, \cdots, K\}$ is

$$\text{mFCR}(t) = \frac{\sum_{k=1}^{K} w_k(1-p_k)G_{0k}(t)}{\sum_{k=1}^{K} w_k\{(1-p_k)G_{0k}(t) + p_k G_{1k}(t)}.$$

It can be shown similarly as Lemma 1 that the mFCR increases in t under the GMRC. The rest of the proof follows similar lines as the proof of Theorem 1(c). The only difference is that we need to consider a weighted loss function instead and show that $t^c_{OR}(\alpha) > \alpha$. Finally we argue by contradiction that $\mathbf{\Delta}_{OR}$ must have the smallest wMDR.

1.3. Proof of Lemma 2

Let $\epsilon > 0$. Define $A^+_\epsilon = [A_l - \epsilon, A_u + \epsilon]$ and $A^-_\epsilon = [A_l + \epsilon, A_u - \epsilon]$. The goal is to show that $\theta(s)$ and $\theta^m(s)$ would only differ on a negligible area with overwhelming probability. According to the definitions of $\mu(s)$ and $\mu^m(s)$, we claim that one of the following two events must occur if $\theta(s) \neq \theta^m(s)$

- Event A: $\mu(s)$ and $\mu^m(s)$ are not close, i.e. $|\mu(s) - \mu^m(s)| > \epsilon$;
- Event B: $\mu(s)$ and $\mu^m(s)$ are close but $\mu(s)$ is close to one of the ending points of the indifference region, i.e. $\mu(s) \in (A \setminus A^-_\epsilon) \cup (A^+_\epsilon \setminus A)$.

If this claim seems obscure, consider its contrapositive statement which is easy to verify. Next let C denote the event $\theta(s) \neq \theta^m(s)$, then $C \subset (A \cup B)$. It follows that $P(C) \leq P(A) + P(B)$, which gives the following inequality

$$P\left[\theta(s) \neq \theta^m(s)\right] \leq P\left[\mu(s) \in (A \setminus A^-_\epsilon) \cup (A^+_\epsilon \setminus A)\right] + P\left(|\mu(s) - \mu^m(s)| > \epsilon\right).$$

For any small $\eta > 0$, Condition 1 implies that there exists an $\epsilon > 0$ such that

$$P\left[\mu(s) \in (A \setminus A^-_\epsilon) \cup (A^+_\epsilon \setminus A)\right] < \eta/2.$$

Condition 2 implies that for this ϵ, there exists a partition $S = \cup_{i=1}^m S_i$ such that

$$\int_S P\left[|\mu(s) - \mu^m(s)| > \epsilon\right] d\nu(s) < \eta/2.$$

Therefore $\int_S P\{\theta(s) \neq \theta^m(s)\} d\nu(s) < \eta$ and the result follows.

1.4. Proof of Theorem 4

Suppose r hypotheses are rejected. Then the FCR of Procedure 3 is

$$\begin{aligned}
\text{FCR} &= E\left\{\frac{\sum_{k=1}^{K} w_k(1-\vartheta_k)\Delta_k}{\sum_{k=1}^{K} w_k\Delta_k} I(\sum_k \Delta_k > 0)\right\} \\
&= E\left\{\frac{I(\sum_k \Delta_k > 0)}{\sum_{k=1}^{K} w_k\Delta_k} \sum_{k=1}^{K} w_k\Delta_k E(1-\vartheta_k|\boldsymbol{X})\right\} \\
&= E\left\{\frac{\sum_{k=1}^{r} w_{(k)}T^c_{(k)}}{\sum_{k=1}^{r} w_{(k)}} I(\sum_k \Delta_k > 0)\right\},
\end{aligned}$$

which is always less than α due to the operation characteristic of Procedure 3.

2. Technical details in computation

We describe the MCMC algorithm for Gaussian random field. Denote the complete data set as $\hat{\boldsymbol{\beta}} = [\hat{\beta}(s_1), ..., \hat{\beta}(s_n)]^T$. In matrix notation, the model for $\hat{\boldsymbol{\beta}}$ is written $\hat{\boldsymbol{\beta}} \sim \mathrm{N}[\boldsymbol{\beta}, \Sigma_1(r, \rho_1)]$ and $\boldsymbol{\beta} \sim \mathrm{N}[\bar{\beta}\mathbf{1}, \sigma_2^2\Sigma_2(\rho_2)]$, where $\boldsymbol{\beta} = [\beta(s_1), ..., \beta(s_n)]^T$ is the vector of true slopes, $\Sigma_1(r, \rho_1)$ is the error covariance matrix with (i, j) element $w(s_i)w(s_j)[(1-r)I(s_i = s_j)+r\exp(-||s_i - s_j||/\rho_1)]$, $\mathbf{1}$ is the column vector of ones, and $\Sigma_2(\rho_2)$ is the correlation matrix with (i, j) element $\exp(-||s_i - s_j||/\rho_2)$. The Metropolis within Gibbs algorithm is used to draw posterior samples. This begins with an initial value for each model parameter, and then parameters are updated one-at-a-time, conditionally on all other parameters. The full conditional distributions of $\boldsymbol{\beta}$, $\bar{\beta}$ and σ_2^2 are conjugate, and these parameters are simply updated by drawing from the full conditional distributions:

$$\begin{aligned}
\boldsymbol{\beta}|\text{rest} &\sim \mathrm{N}\left[\Delta\left\{\Sigma_1(r,\rho_1)^{-1}\hat{\boldsymbol{\beta}} + \sigma_2^{-2}\bar{\beta}\Sigma_2(\rho_2)^{-1}\mathbf{1}\right\}, \Delta\right] \\
\bar{\beta}|\text{rest} &\sim \mathrm{N}\left[\frac{\sigma_2^{-2}\mathbf{1}^T\Sigma_2(\rho_2)^{-1}\boldsymbol{\beta}}{\sigma_2^{-2}\mathbf{1}^T\Sigma_2(\rho_2)^{-1}\mathbf{1} + \sigma_0^{-2}}, \frac{1}{\sigma_2^{-2}\mathbf{1}^T\Sigma_2(\rho_2)^{-1}\mathbf{1} + \sigma_0^{-2}}\right] \\
\sigma_2^{-2}|\text{rest} &\sim \mathrm{Gamma}\left(n/2 + a, (\boldsymbol{\beta} - \bar{\beta}\mathbf{1})^T\Sigma_2(\rho_2)^{-1}(\boldsymbol{\beta} - \bar{\beta}\mathbf{1})/2 + b\right)
\end{aligned}$$

where $\Delta = [\Sigma_1(r, \rho_1)^{-1} + \sigma_2^{-2}\Sigma_2(\rho_2)^{-1}]^{-1}$ and priors are denoted $\bar{\beta} \sim \mathrm{N}(0, \sigma_0^2)$ and $\sigma_2^{-2} \sim \mathrm{Gamma}(a, b)$.

The full conditional distributions of the spatial correlation parameters r, ρ_1, and ρ_2 are not conjugate, and therefore these parameters are updated using Metropolis-Hastings updates. To update r at the j^{th} MCMC iteration, we generate a candidate $r^{(c)} \sim \mathrm{Beta}[C_r r^{(j-1)}, C_r(1 - r^{(j-1)})]$, where $r^{(j-1)}$ is the value at MCMC iteration $j-1$ and C_r is a tuning parameter. The acceptance ratio assuming prior $r \sim \mathrm{Beta}(a_r, b_r)$ is

$$R = \left\{\frac{\phi[\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \Sigma_1(r^c, \rho_1)]}{\phi[\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \Sigma_1(r^{(j-1)}, \rho_1)]}\right\}\left\{\frac{\mathrm{B}[r^c|a_r, b_r]}{\mathrm{B}[r^{(j-1)}|a_r, b_r]}\right\}\left\{\frac{\mathrm{B}[r^{(j-1)}|C_r r^c, C_r(1-r^c)]}{\mathrm{B}[r^c|C_r r^{(j-1)}, C_r(1-r^{(j-1)})]}\right\},$$

where $\phi(\mathbf{y}|\boldsymbol{\mu}, \Sigma)$ is the multivariate normal density function with variate $\mathbf{y}$, mean $\boldsymbol{\mu}$, and covariance Σ and $\mathrm{B}(y|a, b)$ is the $\mathrm{Beta}(a, b)$ density function. The candidate is accepted with probability $\min\{R, 1\}$. If the candidate is accepted, then $r^{(j)} = r^c$, otherwise the previous value is retained, $r^{(j)} = r^{(j-1)}$. Similarly, the candidate distribution and acceptance ratio for ρ_1 are $\rho_1^{(c)} \sim \mathrm{Beta}[C_1\rho_1^{(j-1)}, C_1(1 - \rho_1^{(j-1)})]$ and

$$R = \left\{\frac{\phi[\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \Sigma_1(r, \rho_1^c)]}{\phi[\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \Sigma_1(r, \rho_1^{(j-1)})]}\right\}\left\{\frac{\mathrm{B}[\rho_1^c|a_1, b_1]}{\mathrm{B}[\rho_1^{(j-1)}|a_1, b_1]}\right\}\left\{\frac{\mathrm{B}[\rho_1^{(j-1)}|C_1\rho_1^c, C_1(1-\rho_1^c)]}{\mathrm{B}[\rho_1^c|C_1\rho_1^{(j-1)}, C_1(1-\rho_1^{(j-1)})]}\right\},$$

and the candidate distribution and acceptance ratio for ρ_2 are $\rho_2^{(c)} \sim \mathrm{Beta}[C_2\rho_2^{(j-1)}, C_2(1 - \rho_2^{(j-1)})]$ and

$$R = \left\{\frac{\phi[\boldsymbol{\beta}|\bar{\beta}\mathbf{1}, \sigma_2^2\Sigma_2(\rho_2^c)]}{\phi[\boldsymbol{\beta}|\bar{\beta}\mathbf{1}, \sigma_2^2\Sigma_2(\rho_2^{(j-1)})]}\right\}\left\{\frac{\mathrm{B}[\rho_2^c|a_2, b_2]}{\mathrm{B}[\rho_2^{(j-1)}|a_2, b_2]}\right\}\left\{\frac{\mathrm{B}[\rho_2^{(j-1)}|C_2\rho_2^c, C_2(1-\rho_2^c)]}{\mathrm{B}[\rho_2^c|C_2\rho_2^{(j-1)}, C_2(1-\rho_2^{(j-1)})]}\right\}.$$

In addition to $\boldsymbol{\beta}$ at the n measurement locations, our multiple comparisons algorithms require samples from the true slopes at prediction locations $\mathbf{t}_1, ..., \mathbf{t}_N$, i.e., $\boldsymbol{\beta}_p = [\beta(\mathbf{t}_1), ..., \beta(\mathbf{t}_N)]^T$.

To obtain these predictions, we sample $\boldsymbol{\beta}_p|\boldsymbol{\beta}, \bar{\beta}, \sigma_2, \rho_2$ following the usual conditional distribution induced by the multivariate normal distribution at each MCMC iteration. This produces samples from the posterior predictive distribution $\boldsymbol{\beta}_p|\hat{\boldsymbol{\beta}}$, marginally over all model parameters.

The tuning parameters C_r, C_1, and C_2 are selected to give acceptance probability around 0.4. For the ozone data, we generate 25,000 samples and discard the first 10,000 as burn-in. For the simulation study, we generate 10,000 samples and discard the first 2,500 as burn-in. Convergence is monitored using trace plots of several representative parameters.

3. Simulation results on misspecified models

In this section, we conduct simulations to study sensitivity to covariance misspecification. We simulated data as described in Section 5.2, but fit the model with the wrong spatial covariance function. For data generated with exponential covariance, we fit the Matérn with smoothness fixed at $\kappa = 2.5$; for data generated from the Matérn with smoothness $\kappa = 2.5$, we fit the exponential. The results are summarized in Figure 1. We find that when a Matérn correlation is fit to data generated with exponential correlation the FDR is slightly too high. Therefore, it appears there is some sensitivity to model misspecification and that we must be careful to conduct exploratory analysis to ensure that the spatial model fits the data reasonably well.

4. Miscellanea

4.1. On using alternative power measures as optimization criteria

The false negative rate (FNR, Sarkar, 2002; Genovese and Wasserman, 2002) and average power (AP, Efron, 2007) are alternative power measures. The FNR in our case is the expected proportion of locations containing signals among all non-rejected locations. The AP is the expected proportion of signal area that are correctly identified. The optimality result in our paper can be extended to their "marginal" versions

$$\text{mFNR} = \frac{E\{\nu(R^c \cap S_1)\}}{E\{\nu(R^c)\}} \text{ and } \text{mAP} = \frac{E\{\nu(R \cap S_1)\}}{E\{\nu(S_1)\}},$$

where $R^c = \{s : \delta(s) = 0\}$ is the non-rejection area and $S_1 = \{s : \theta(s) = 1\}$ is the non-null area. Specifically, it follows from some algebra that MDR, mFNR and mAP are monotonic functions of the threshold t when the monotone ratio condition holds (the proof of the MDR case is trivial; the proofs of the cases of the mFNR and mAP follow similar lines as those in Lemma 1). This result on monotonicity implies that if a testing procedure at mFDR level α maximizes the MDR, then it also minimizes/maximizes the mFNR/mAP. Hence all theoretical results in the paper can apply directly to the setting where the mFNR or mAP is chosen as the optimization criterion.

4.2. On using a universal threshold t

We have assumed that our testing rule $\delta(s) = I[T(s) < t]$ has a universal threshold t. This does not affect the generality of the decision rule. Note that $\delta(s) = I[T(s) < t]$ can be flexibly modified via two equivalent schemes: varying $T(s)$ or varying t. Specifically, if it is desirable to change t to $\tilde{t}$, then we can always vary $T(s)$ to achieve the same effect, e.g.

write $\delta(s) = I[T(s) < \tilde{t}]$ as $\delta(s) = I[\tilde{T}(s) < t]$, where $\tilde{T}(s) = (t/\tilde{t})T(s)$. In the testing rule $\delta(s) = I[T(s) < t]$, we deliberately use a fixed t and let $T(s)$ vary to incorporate all relevant information (either in the sample or past data). This provides great convenience in methodological developments since we can thus work under the classical two-step strategy in multiple testing: first rank the hypotheses according to $T(s)$ and then choose a cutoff t along the rankings. *This strategy only makes sense when a universal t is used.*

4.3. On differentiability of $G_0(t)$ and $G_1(t)$.

The definition of the monotone ratio condition (3.3) requires that the functions $G_0(t)$ and $G_1(t)$ are differentiable. As pointed out by a reviewer, the assumption needs to be checked in practice. In this section, we closely examine this condition and discuss its impact in the context of a Gaussian random field.

Firstly, the derivatives are always well-defined. Note that (i) according to the definition, $G_j(t)$ are non-decreasing functions defined on $t \in [0, 1]$, $j = 0, 1$, and (ii) $G_j(t)$ map sets of measure 0 to sets of measure 0. Based on (i) and (ii), we claim that $G_j(t)$ are differentiable almost everywhere on $[0, 1]$ (expect for a countable number of points). See Theorem 7.18 on page 146 in Rudin (1987). For the countable number of points we can let $g_j(t) = \frac{d}{dt}G_j(t) =$ constant. Therefore $g_j(t)$ are always well-defined.

Secondly, in a Gaussian random field, for the countable number of points where g_j are arbitrarily defined, there is no impact on the result. Specifically, let $x^n = (x_1, \cdots, x_n)$ be the observed data points. Consider the oracle test statistic $T_{OR}(s) = T^s_{OR}(x^n) = P[\theta(s) = 0|x^n]$, where $T^s_{OR}(\cdot)$ is a function which maps the observed data to a real number. Let $\mu_s \equiv \mu(s)$. Then $(X_1, \cdots, X_n, \mu_s)$ has a joint multivariate normal distribution. Recall that A is the indifference region and $\theta(s) = I[\mu(s) \in A^c]$. It follows that

$$T_{OR}(s) = P[\mu(s) \in A|x_1, \cdots, x_n] = \frac{\int_A f(x_1, \cdots, x_n, \mu_s)d\mu_s}{f(x_1, \cdots, x_n)}.$$

It is clear that there is no point mass for $T_{OR}(s)$ and we conclude that a set with measure zero has no effect on the result. This argument works for other test statistics such as the p-value, or the original observation $X(s)$ as well. So there should be no worries about the differentiability condition.

References

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics 35*(4), 1351–1377.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B* ***64***, 499–517.

Rudin, W. (1987). *Real and complex analysis.* Tata McGraw-Hill Education.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist. 30*, 239–257.

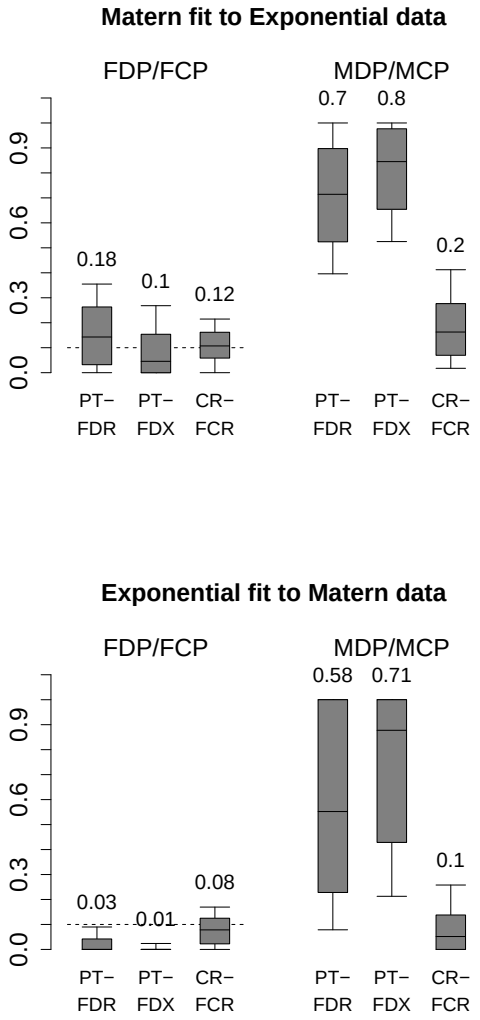

Fig. 1. False discovery control in misspecified models.