Controlling the False Discovery Rate via symmetrized data aggregation based on SLOPE

Yihe Guo, Xuemin Zi
Tianjin University of Technology and Education, Tianjin, 300222, China
E-mail: 3024766577@qq.com

Abstract. The symmetrized data aggregation (SDA) method has obvious advantages in controlling false discovery rate (FDR) under dependence, but it is affected by data sparsity level, signal amplitude and feature correlation structures. The FDR control becomes less accurate due to the additional estimation errors. In this paper, we expand the symmetrized data aggregation (SDA) filter by using sorted L-one penalized estimation (SLOPE) method, and in comparison with SDA, SDA-SLOPE method can estimate FDR more accurately and improve the true discovery rate (TDR). Through simulation study, it is found that SDA-SLOPE can adapt to the changes of data sparsity, signal amplitude and feature correlation, and the robustness and effectiveness of this method in FDR control are verified.

1. Introduction
The conventional FDR methods in multiple testing, such as the Benjamini-Hochberg (BH) procedure [1] and the adaptive p-value procedure [2], implement effective FDR control under the assumption that the data are independent. However, in dependent settings, those methods will become quite conservative and cause the loss of the true discovery rate (TDR). Fan (2012) proposed a method to estimate false discovery proportion under arbitrary covariance dependence [3]. Fan and Han (2017) [4] and Li and Zhong (2017) [5] found the signal-to-noise ratio (SNR) can be improved by aggregating the weak signals of each individual, but their methods depend heavily on the accuracy of parameter estimation and asymptotic normality of test statistics. Bogdan (2014) was inspired by BH procedure [1] and proposed a new variable selection method—sorted L-one penalized estimation (SLOPE) [6-8], which can control the false discovery rate (FDR) while adapting to the feature correlation and improve the prediction accuracy. Similar to the spirit of knockoff filter in Barber and Candes (2015) [9], Du et al. (2020) proposed a new information pooling strategy—the symmetrized data aggregation (SDA) that utilizes the covariance structures [10]. Through sample splitting, variable selection and information pooling, SDA can control FDR under general correlation. In this paper, the SDA filter is expanded by using SLOPE method, and the new SDA-SLOPE method achieves more accurate FDR control while improving the TDR significantly.

2. Model and problem formulation
Suppose $p$ streams of observations $\xi_i = \{\xi_{i1}, \xi_{i2}, \ldots, \xi_{ip}\}$, $i = 1, \ldots, n$, is distributed along $p$-dimensions with mean $\mu = (\mu_1, \cdots, \mu_p)^T$ and covariance matrix $\Sigma$. The null and alternative hypothesis for the $j$th stream are:

$H_{0j} : \mu_j = 0$ versus $H_{1j} : \mu_j \neq 0$, $j = 1, \ldots, p$. 
Therefore, we refer to the data streams with some deviation from reference level as "non-nulls" or "alternatives" (i.e. \( j \in H_1 \)), while the streams on the reference level as "nulls" (i.e. \( j \in H_0 \)). The mean 
\[
\bar{\xi} = n^{-1} \sum_{i=1}^{n} \xi_i, \quad \text{asymptotically follows a multivariate normal distribution.}
\]

Similar to Bogdan (2015) [6], the problem of multiple hypothesis under correlation conditions can be transformed into a variable selection problem in linear regression. Through a "whitening" transformation, our multiple testing problem can be presented in the form of a regression equation

\[
Y = X\mu + z
\]

where \( X = \Sigma^{-1/2} \in \mathbb{R}^{p \times p} \) can be seen as the design matrix, \( Y \) is the response variables, and \( Z \sim N(0, n^{-1}I_p) \) is a \( p \)-dimension vector of random errors.

### 2.1. Symmetrized Data Aggregation

The symmetrized data aggregation (SDA) is a new information pooling strategy. Firstly, The SDA divides the data into two independent parts \( D_1 \) and \( D_2 \), similar to the method of Wasserman and Roeder (2009) [10], sets \( n_1 = \frac{2}{3}n \) for part \( D_1 \) and \( n_2 = n - n_1 \) for part \( D_2 \). Then the two sub datasets are used to construct a pair of statistics that satisfy the symmetry property, and both of which follow the distribution of \( N(0,1) \) for the nulls, and are calculated to test \( H_0 \):

\[
\left( T_{1j}, T_{2j} \right) = \left( \sqrt{n_1} \xi_1, \sqrt{n_2} \xi_2 \right)
\]

Then aggregating the information of the two symmetric statistics to construct symmetric statistics \( W_j = T_{1j}T_{2j} \). Clearly, \( T_{1j} \) and \( T_{2j} \) tend to have large absolute values for most of the alternatives with the same sign, so \( W_j \) is likely to be a large and positive value, while \( W_j \) is asymptotically symmetric for the nulls. The False discovery proportion (FDP) can be defined as

\[
\text{FDP} = \frac{\# \{ j : W_j > -t \}}{\# \{ j : W_j > -t \} + 1}
\]

Since if the signal is not too weak, the number of \( W_j \) for the alternatives below \(-t\) is usually very small, thus \( \# \{ j : W_j \leq -t \} \) can be seen as the approximation to the number of false positives due to symmetry of the nulls. Therefore, the fraction above can be seen as the estimate of FDP.

### 2.2. SDA algorithm based on SLOPE

Our algorithm is a modified SDA procedure with the help of SLOPE, referred to as SDA-SLOPE method. After sample splitting, to extract information from \( D_1 \), we propose to use SLOPE to select variables. SLOPE provides a useful tool here because this method can adapt to the sparsity and correlation structure of data, and can implement more accurately FDR control. Let \( \xi = n^{-1} \xi_i, y = X\xi \). The regression coefficient is obtained by solving the following objective function

\[
\min \| y - X\beta \|^2 + \sigma \cdot \sum_{l=1}^{p} \lambda_{BH}(l) |\beta_l|
\]

SLOPE method uses the sorted L-one norm as the penalty term, which is equivalent to a convex optimization program. It estimates the regression parameters and also implements the selection of variables. The regression coefficient obtained by SLOPE method is \( \hat{\mu}_1 = (\hat{\mu}_{11}, \cdots, \hat{\mu}_{1p}) \), then let \( S = \{ j : \hat{\mu}_{1j} \neq 0 \} \) denote the subset of coordinates selected by SLOPE.

Next, in order to obtain a more accurate estimator in the next step, by using the coordinates in the subset \( S \) to construct \( T_{2j} \), which is equivalent to a variable selection step can significantly increase the signal-to-noise ratio (SNR) of \( D_2 \). Let \( \xi = n^{-1} \sum_{l=1}^{p} \lambda_{BH}(l) |\beta_l| \), \( y = X\xi \), \( e_j = (0, \cdots, 0, 1, 0, \cdots, 0)' \), and \( X_S = (X_j : j \in S) \). The least squares estimates (LSEs) obtained by using the \( D_2 \) data set are given by \( \hat{\beta}_2 = (\hat{\beta}_{21}, \cdots, \hat{\beta}_{2p})' \)

\[
\hat{\beta}_{2j} = \begin{cases} 
\left( e_j' X_S' X_S^{-1} X_S y_2 \right)^{-1}, & j \in S; \\
0, & j \in S^c
\end{cases}
\]
The two parts are used to construct a pair of statistics respectively, both of which follow the distribution of \( N(0,1) \) under the null, and are calculated to test \( H_0 \):

\[
\left( T_{1j}, T_{2j} \right) = \left( \frac{\sqrt{n_j} \hat{\mu}_j}{\sigma_{S,j}}, \frac{\sqrt{n_j} k_j}{\sigma_{S,j}} \right)
\]  

(7)

then aggregating the information of the two symmetric statistics to construct symmetric statistic \( W_j = T_{1j} T_{2j} \). According to the symmetry of the statistic \( W_j \) for the nulls, we can judge whether to reject the null hypothesis by selecting a threshold

\[
L = \inf \left\{ t > 0 : \frac{\#(j: W_j \leq -t)}{\#(j: W_j \leq -t)} \leq \alpha \right\}
\]  

(8)

and reject the null hypothesis if \( W_j \geq L \). If the set above is empty, set \( L = +\infty \).

3. Simulation

Let \( n = 200 \) and \( p = 1000 \), the data streams \( \xi_i \) are generated by the multivariate normal distribution, where \( \mu = [\mu_i, i = 1, \ldots, p] \) is the \( p \)-dimension mean vector of \( \xi_i \), the covariance matrix is \( \Sigma = (\sigma_{ij})_{p \times p}, \sigma_{ij} = \rho^{|i-j|}, i, j = 1, 2, \ldots, p \). We randomly select a proportion \( k \) of the data streams as the true signals with magnitude \( A \). Fixed \( p = 2000, n = 200 \), sparsity level \( k = 0.1 \), signal magnitude \( A = 2.5 \) and feature correlation \( \rho = 0.5 \). Under the nominal level \( \alpha = 0.2 \), figure1 displays the result of FDP and true discovery proportion (TDP) obtained by SDA-SLOPE and SDA method respectively.

![Figure 1: (a): The curve of FDP(t) represents the estimate of FDP curve and the curve of FDP(t) represents the true FDP against t for the method of SDA-SLOPE; (b) the corresponding FDP estimate and the true FDP for the SDA method; (c): the true TDP for SDA-SLOPE and SDA respectively.](image)

As is shown in panel (a) and (b) in figure1, the estimated FDP procedure for the SDA-SLOPE method approximates the true FDP procedure fairly accurately. However, the estimate FDP obtained by SDA method that is actually smaller than the true FDP, this is because SDA method uses a fixed threshold, which makes it unable to accurately estimate FDP. The panel (c) in figure1 compares the TDP procedures of SDA-SLOPE and SDA, at the nominal level \( \alpha = 0.2 \), the TDP of SDA-SLOPE is 0.85 with the threshold \( t = 0.05 \), which is much higher than that of SDA is 0.68 with threshold \( t = 0.37 \).

3.1. Effect of sparsity level

Next, we compare the effect of sparse level on the two methods. For each run, we fix \( n = 200, p = 1000, A = 2.5, \rho = 0.5 \) and set the sparsity level \( k = 0.01, 0.02, \ldots, 0.2 \), and apply the two methods to the simulation data. Figure2 shows the FDR and TDR obtained through 500 trials under the nominal level \( \alpha = 0.2 \).
Figure 2: The FDR and TDR comparison for varying sparse level $k$ in the two methods.

As we can see from the figure 2, both two methods approximately control TDR at the level of 0.8. And SDA-SLOPE successfully control FDR at the nominal level, while SDA method fail to accurately control FDR at nominal level under various sparse levels. This is because SDA-SLOPE method uses adaptive threshold to select variables, which can achieve more accurate control of FDR while adapting to data sparsity.

3.2. Effect of signal amplitude

For this experiment, the effect of different signal amplitudes on SDA-SLOPE method and SDA method were tested under the nominal level $\alpha = 0.2$. Here $n = 200$, $p = 1000$, $k = 0.1$, $\rho = 0.5$, and vary $A = 0.5, 1, 1.5, \cdots, 5$. Figure 3 shows the FDR and TDR obtained by 500 trials.

Figure 3: The FDR and TDR comparison for varying signal amplitudes $A$ in the two methods.

In figure 3, it is shown that both two methods can control FDR at the nominal level, and SDA-SLOPE controls the FDR more accurately near the nominal level. Clearly, the TDR of SDA slightly increases as the signal amplitude increase, and TDR of SDA-SLOPE is significantly higher than that of SDA. This is because SDA-SLOPE uses an adaptive threshold when selecting variables, which can better adapt to the size of unknown data and has higher prediction accuracy.

3.3. Effect of feature correlation

Finally, we turn to study the influence of feature correlation on the two methods. We fix $n = 200$, $p = 1000$, $A = 2.5$, $k = 0.1$, and vary the magnitude of feature correlation $\rho$ from independence ($\rho = 0$) to strong dependence ($\rho = 0.9$). Figure 4 shows the FDR and TDR obtained by 500 trials under the nominal level $\alpha = 0.2$. 
Figure 4: The FDR and TDR comparison for varying feature correlation \( \rho \) in the two methods.

From figure 4, it is shown that both two methods can control FDR at the nominal level, and SDA-SLOPE achieves a more conservative value, with the TDR of both is about 0.8. Since the two methods consider the correlation structures, different magnitudes of feature correlation have little influence on the two methods.

4. Conclusion
In comparison with SDA, SDA-SLOPE method can better adapt to the sparsity of data and the magnitude of signal amplitude. It can realize more accurate FDR control and improve the TDR significantly. Through simulation study, the validity and robustness of this method are verified.

Acknowledgments
This paper was supported by the NNSF of China, grants No. 11771332 and 11771220, Youth Program of Tianjin Natural Science Foundation, Grant No. 18JCQNJC70100.

References
[1] Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing Journal of the Royal Statistical Society 57(1) 289-300.
[2] Benjamini Y and Hochberg Y 1997 Multiple hypotheses testing with weights Scandinavian Journal of Statistics 24(3) 407-418.
[3] Fan J, Han X and Gu W 2012 Estimating false discovery proportion under arbitrary covariance dependence Journal of the American statistical association 107 1019–1035.
[4] Fan J and Han X 2017 Estimation of the false discovery proportion with unknown dependence Journal of the Royal Statistical Society: Series B 79 1143–1164.
[5] Li J and Zhong P 2017 A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence The Annals of Statistics 45 557–590.
[6] Tibshirani R 1996 Regression shrinkage and selection via the lasso Journal of the Royal Statistical Society: Series B (Methodological) 58(1) 267–288.
[7] Bogan M, Ewout V, Sabatti C, et al 2015 SLOPE-adaptive variable selection via convex optimization. Annals of Applied Statistics 9(3) 1103-1140.
[8] Su W, Candes and Emmanuel J 2015 SLOPE is adaptive to unknown sparsity and asymptotically minimax The Annals of Statistics 44(3) 1038-1068.
[9] Barber R, Candes and Emmanuel J 2014 Controlling the false discovery rate via knockoffs The Annals of Statistics 43(5) 2055–2085.
[10] Du L, Xu G, Sun W and Zou C 2020 False discovery rate control under general dependence by symmetrized data aggregation arXiv:2002.11992 [stat.ME].
[11] Wasserman, Larry and Roeder 2009 High-dimensional variable selection The Annals of Statistics 37 2178–2201.
[12] Benjamini Y and Yekutieli D 2001 The control of the false discovery rate in multiple testing under dependency *The Annals of Statistics* **29**(4) 1165-1188.