Identification of Differentially Expressed Genes via Knockoff Statistics in Single-cell RNA Sequencing Data Analysis

by

Lixia Yi

B.S. in Statistics, Shandong University, China, 2017

M.S. in Statistics, University of Wisconsin-Madison, 2018

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Lixia Yi

It was defended on

May 24th 2024

and approved by

Linxi Liu, Ph.D., Department of Statistics

Lucas K. Mentch, Ph.D., Department of Statistics

Yu Cheng, Ph.D., Department of Statistics

Zihuai He, Ph.D., Department of Neurology & Neurological Science, Stanford University

Identification of Differentially Expressed Genes via Knockoff Statistics in Single-cell RNA Sequencing Data Analysis

Lixia Yi, PhD

University of Pittsburgh, 2024

Model-X knockoffs [Candès et al., 2018] is a recent statistical framework that allows scientists to discover true effects while controlling the false discovery rate (FDR) with finite sample guarantee by creating a synthetic copy of the original variables—knockoffs—as control. The framework works under arbitrary dimensional settings, but with the increase of dimensions, it becomes increasingly difficult to create knockoffs due to the computational cost. The missingness of data, which is common in many high-dimensional datasets, adds another layer of difficulty for knockoff construction. We propose knockoff constructions based on a latent factor model that are able to handle the missing data, and are faster than the out-of-box method in Candès et al. [2018]. We apply our approach to differentially expressed gene analysis with single-cell RNA sequencing data to verify the FDR control and cross-reference the discovered genes with findings from other studies.

Keywords Model-X knockoffs; variable selection; false discovery rate; single-cell RNA sequencing; high-dimensionality.

Table of Contents

List of Tables

List of Figures

1.0 Background

Variable selection is an important topic in statistical research. Traditionally, it is done by, for example, inspecting the magnitude of the fitted coefficients in a linear regression model or cross validated Lasso [Tibshirani, 1996], or utilizing importance scores from a random forest [Breiman, 2001]. Model-X knockoffs [Candès et al., 2018] is a recent statistical framework that allows scientists to discover true effects while controlling the expected proportion of false discoveries. This framework does not require any knowledge of $Y|X$—how the response depends on the explanatory variables. Instead, the false discovery rate (FDR) control entirely relies on the accurate knowledge of the feature distribution. Furthermore, it does not impose any additional requirements onto the model used and can leverage almost any feature importance measures to select variables. Due to this flexibility, the model-X knockoffs procedure is applied in a wide variety of fields including but not limited to neural networks [Lu et al., 2018], time series modeling [Fan et al., 2020], biology [Gao et al., 2018] and genetics [Sesia et al., 2019, 2021]. Additionally, scientists have applied knockoffs to select variables with different controls such as per family error rate and k family-wise error rate [Ren et al., 2023], and directional FDR [Barber and Candès, 2019]. There are also multiple papers that have discussed the usage of feature importance measures that are different from Lasso coefficient-difference [Barber and Candès, 2019; Gimenez et al., 2019; Janson and Su, 2016], which was first introduced in Candès et al. [2018] and is the most commonly applied statistics in the model-X knockoff literature.

The robustness of the framework for FDR control under approximate knowledge of the feature distribution was also discussed in Barber et al. [2020], while multiple authors discussed the power of the framework under various settings [Ke et al., 2020; Liu and Rigollet, 2019; Spector and Janson, 2022; Weinstein et al., 2017].

However, aside from the benefits of the model-X knockoff framework, it also involves several challenges that need to be addressed. For example, unlike in the earlier work by Barber and Candès [2015], the algorithm controls the FDR by creating a random copy of the feature variables. Therefore, the variables that are selected are inherently random and will create difficulties for scientists to draw consistent conclusions on the results. There are various works to derandomize the variable selection results by Emery and Keich [2019]; Nguyen et al. [2020]; Ren et al. [2023];

Su et al. [2015] and Gimenez and Zou [2019]. Another challenge that is often mentioned is how to accurately model the feature distribution in order to create those synthetic copies. While the identical copy of the features could be trivially used, it will have no power. Therefore, other than the default Gaussian knockoff, scientists have approached alternative knockoff constructions via hidden markov models [Sesia et al., 2019], Bayesian networks [Gimenez et al., 2019], Metropolis–Hastings-like algorithms [Bates et al., 2021], and also via neural networks such as auto-encoders [Liu and Zheng, 2018], generative adversarial networks [Jordon et al., 2018], and moment-matching networks [Romano et al., 2020], which demonstrated promising empirical results.

Single-cell RNA sequencing (scRNA-seq), the data that we want to analyse, is a high-throughput RNA sequencing technology at the single-cell level that provides high resolution gene expression data. It allows scientists to develop a better understanding into a wide range of research topics. There exists an abundant literature on the various applications, and both technical and analytical challenges in regard of scRNA-seq data research. It is difficult to provide a comprehensive summary of the literature on these subjects, and it is also out of the scope of this thesis. Excellent reviews could be found in the works of such as Jovic et al. [2022]; Kolodziejczyk et al. [2015]; Lähnemann et al. [2020] and Stegle et al. [2015].

Instead, in this chapter, I will only focus on the advances and challenges of gene expression studies from the differential expressed genes (DEG) analysis perspective, serving as a motivation for the thesis. Gene expression technology is commonly used in molecular biology research to answer diverse biological questions and it is often through comparing the gene expression changes under different treatment conditions, in other words, by identifying DEGs. Gene expression studies, depending on how the data is collected, could be categorized into randomized designed experiments as well as observational studies. The emergence of (bulk) RNA-seq technologies, which reads the expression levels accross a large population of input cells, combined with new statistical tools [Anders and Huber, 2010; Ritchie et al., 2015; Robinson et al., 2009] have provided researchers with more detailed insights into certain problems, such as studying the selection pressure that applies to gene expression levels between the same tissue taken from different species. Subsequent technological advancements have introduced scRNA-seq [Tang et al., 2009], offering an even more granular view of gene expressions on a single-cell level, and providing possible solutions to questions that bulk RNA-seq is unable to answer. For example, with scRNA-seq, scientists now may probe into more complex tissues such as brain tissues which comprises many different cell types. This granularity

makes it possible to determine whether differences in gene expression are due to the prevalence of certain cell types, a determination that was not feasible with bulk RNA-seq. Furthermore, scRNA-seq may provide insights into the stochastic nature of gene expression, further enhancing our understanding of cellular processes. While from a computational perspective, statistical tools for bulk RNA-seq can be used for scRNA-seq DEG analysis, there are unique challenges that need to be addressed. First, the scRNA-seq data is typically a high-dimensional dataset. In many previous works, inference on high-dimensional data would involve at least one of the following: Assuming a linear model with homoscedastic error terms, assuming sparse signal, or only provide asymptotic guarantees [Fan and Lv, 2010; Javanmard and Montanari, 2014a,b; Lockhart et al., 2014; Meinshausen and Bühlmann, 2010; van de Geer et al., 2014; Zhang and Zhang, 2014]. In comparison, the model-X framework can possibly accomodate for any model for both the response and covariates, does not impose any sparsity assumptions, and provides finite sample guarantees. Hence its attractiveness in practice. However, while the model-X knockoff procedure is designed with high-dimensional applications in mind, it cannot avoid the increase in computational complexity that comes along with the increase in observations and variables, thus urging a way to simplify the procedure. Second, a large fraction of the data is of zero or low read counts. Jiang et al. [2022] provided an in depth discussion on the mechanisms that lead to the zero-inflation ,which is also referred as dropout in the literature, in scRNA-seq data. From a DEG analysis perspective, there are test designed with zero-inflation in mind, such as SCDE [Kharchenko et al., 2014] and MAST [Finak et al., 2015]. But this large fraction of zeros will also cause difficulties when constructing the knockoff variables since in practice, they are often considered as missing values. A reasonable approach would be to recover the missing values before constructing the knockoffs. There are in general two different school of thoughts that are developed independently by the biology community and statistics community respectively. On one hand, there are heuristic imputation methods that are developed tailored for scRNA-seq data, which include, but is not limited to, model-based imputation methods [Huang et al., 2018; Li and Li, 2018], data smoothing methods [Gong et al., 2018; Van Dijk et al., 2018], and deep learning methods [Eraslan et al., 2019]. On the other hand, there are low-rank matrix completion methods [Candès and Recht, 2012; Candès and Plan, 2010; Candès and Tao, 2010; Hastie et al., 1999, 2015; Mazumder et al., 2010] whose statistical inference and uncertainty quantification under sub-Gaussian or sub-exponential noise are studied in depth [Chen et al., 2019, 2020; Farias et al., 2022]. In this thesis, we choose the latter approach for imputation

for the following reasons: Despite the scRNA-seq data specific imputation methods are dominated by negative binomial models, the log transformed count data can be realistically modelled with a Gaussian model [Finak et al., 2015] or as Grün et al. [2014] mentioned, the noise in single-cell transcriptomics could be considered as log-normal. Furthermore, we found out that the scRNA-seq specific imputation methods do not translate into powerful knockoffs in our simulations. Finally, low-rank imputation methods are more suitable under the context of knockoff construction as is provides opportunities for faster knockoff constructions, especially Gaussian knockoffs, and they are able to better capture the joint correlation structure of the data.

2.0 Identification of Differentially Expressed Genes via Knockoff Statistics in Single-cell RNA Sequencing Data Analysis

2.1 Introduction

The recent development of single-cell RNA-sequencing (scRNA-seq) technologies has taken transcriptomic studies to a new frontier. It allows genome-wide profiling of gene expression levels at the single-cell resolution. With scRNA-seq data, an important statistical task is to identify differentially expressed genes (DEGs) in case-control studies, as results from DEG analysis can contribute to a more comprehensive understanding of the disease mechanism and new discovery of potential risk factors. Given the high variety and the large amount of cells being analyzed in the single-cell experiment, the new technology may also lead to more powerful DEG identification within the same cell type or tissue, or within the same developmental state.

From DEG analysis, one may expect to discover a short-list of genes as candidates for the follow-up investigations. Methodologically, most existing methods for DEG identification, such as MAST [Finak et al., 2015] and edgeR [Robinson et al., 2009], test for *marginal independence*. More specifically, they test whether the expression level of a gene varies across the case and the control group, *only* conditional on a set of covariates representing potential confounding effects, such as gender, age, and batch effect. When a number of genes are under consideration simultaneously, to ensure reproducibility, the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] is usually applied to control the false discovery rate (FDR). However, without appropriately adjusting for the correlation among gene expressions, this type of approach may lead to a number of suspicious findings due to co-expression or unmeasured confounders, and consequently an inflated FDR. In this work, we will study a novel method that performs the DEG analysis that carefully accounts for the correlation structure. This way, we may eliminate the questionable findings, and further shorten the list of genes for downstream analysis.

Our proposed method is under the knockoff framework [Barber and Candès, 2015; Candès et al., 2018]. The knockoff filter is a recently introduced statistical method for multiple testing with guaranteed FDR control, and is particularly powerful for high-dimensional conditional inference. When applying the method, a group of synthetic expressions, called *knockoff variables*, will be

constructed to serve as "negative controls". The inference is made by contrasting the original variables to their knockoffs. Under the knockoff framework, the adjustment for correlation structure is mainly achieved by constructing knockoffs satisfying the exchangeability condition. This implies that the choice of test statistics can be very flexible—it allows us to either implement a new test specifically designed for single-cell data, or incorporate the existing tests, such as the commonly used likelihood-ratio test and Wilcoxon signed-rank test [Ge et al., 2021].

For scRNA-seq data, a noticeable feature is a large fraction of genes—usually more than 90%—with zero or low read counts. This is mainly due to the low transcript capture and limit of sequencing efficiency of current technologies. Statistically, the uncaptured expressions can be viewed as missing values. In this work, we view the uncaptured expressions as data missing completely at random (MCAR).

Specifically, we use $\boldsymbol{G}^* \in \mathbb{R}^{n\times p}$ to denote the underlying true expression *without* any missingness for n individuals across p genes, and

$$\boldsymbol{G} = \boldsymbol{G}^* + \boldsymbol{E}, \tag{2.1}$$

the realization corrupted by measurment errors $\boldsymbol{E}$. The observed log-normalized gene expressions in a scRNA-seq experiment with uncaptureness is denoted by $\boldsymbol{G}^{\text{obs}} \in \mathbb{R}^{n\times p}$. Let Ω be the collection of indices (i, j) for which the expression level is strictly positive, i.e., $\Omega = \{(i, j) : G^{\text{obs}}_{ij} > 0\}$. It can be alternatively viewed as the set corresponding to non-missing expressions. The observed data $\boldsymbol{G}^{\text{obs}}$ and the complete data $\boldsymbol{G}$ are connected by a "projection" operation P_Ω, which projects the complete matrix onto the expressed set with missing entries replaced by 0 such that $\boldsymbol{G}^{\text{obs}} = P_\Omega(\boldsymbol{G})$. The MCAR assumption implies that $(i, j) \in \Omega$ in the scRNA-seq experiment with probability $\theta \in (0, 1]$ *independently*. The measurement error is modeled by a normal distribution, and for all expressed genes with $(i, j) \in \Omega$

$$G^{\text{obs}}_{ij} = G^*_{ij} + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma^2_j) \text{ independently},$$

where e_{ij} are the entries in $\boldsymbol{E}$. In order to reasonably describe correlations among gene expression and efficiently generate knockoffs, we introduce the following latent factor model,

$$\boldsymbol{G}^* = \boldsymbol{A}\boldsymbol{B}^\top \quad \text{and} \quad \boldsymbol{G} = \boldsymbol{A}\boldsymbol{B}^\top + \boldsymbol{E}, \tag{2.2}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times r}$ is a random matrix of latent factors and $\boldsymbol{B} \in \mathbb{R}^{p\times r}$ is a matrix of deterministic factor loadings. Usually, r is assumed to be much smaller than p and n. For the rest of the chapter,

we will use $G_i^\top = (G_{i1}, \ldots, G_{ip})^\top$ to denote the ith rows of $\boldsymbol{G}$. We also assume rows of $\boldsymbol{A}$ and consequentely $\boldsymbol{G}$ are identically distributed. In this case, we use $G^\top = (G_1, \ldots, G_p)^\top$ to denote the population level gene expressions.

The intuition behind such a model is that we assume a large proportion of covariance among gene expressions (columns) can be captured by a low-rank structure. The low-dimensional factors $\boldsymbol{A}$ may contain several observed covariates, such as gender, age, ethnic group, sample ID, and batch effect. We will introduce an algorithm to recover the remaining latent factors given the observed ones. Conditional on the low-rank structure, we assume the randomness only comes from the measurement error and is independent. The model has been widely used under the context of missing value imputations [Cai et al., 2010; Candès and Tao, 2010; Hastie et al., 2015; Kapur et al., 2016; Mazumder et al., 2010; Mongia et al., 2019]. Under the knockoff framework, we can also take advantage the of the low-rank structure to efficiently generate knockoff variables and make conditional inference for a large set of genes simultaneously.

We would like to point out, the new multiple testing method is not only applicable to scRNA-seq data, it can also be applied to any missing value problems as long as model (2.2) is sufficient to characterize the correlation and the data is MCAR.

The rest of chapter is organized as the following. In Section 2.2, we first provide a brief overview of the knockoff framework. Then we introduce a new algorithm for missing data imputation and latent factors recovery in Section 2.3, and several different ways to generate knockoff variables based on the imputation results in Section 2.4. For scRNA-seq data, we will introduce two variable selection procedures based on a specific types of test statistics in Section 2.5. In Section 2.6 and Section 2.7, we illustrate the FDR control and power of the proposed method on both synthetic and real signals on real data sets.

2.2 Review of the knockoff framework

The knockoff filter was first introduced by Barber and Candès [2015] for linear models to select outcome associated variables while controlling for the false discovery rate (FDR). It was later extended to a wider range of models including nonlinear models, and to high-dimensional settings by Candès et al. [2018]. The knockoff procedure imposes a distributional assumption on the ex-

planatory variables, instead of the conditional distribution of the outcome. In this way, exact FDR control can be achieved. Since explanatory variables are viewed as random, the method is called model-X knockoffs.

Under the knockoff framework, we simultaneously test for the null hypotheses corresponding to conditional independence,

$$H_j : G_j \perp\!\!\!\perp Y \mid G_{-j}$$

for each $j \in [p] \coloneqq \{1, \ldots, p\}$, where G_{-j} denotes expressions for all genes except for gene j. For single-cell data, the null hypothesis implies that the expression level for gene j is *not* associated with the treatment given expressions for all the other genes. The knockoff filter provides theoretical guarantees that the selection of a subset of genes $\widehat{S}$ has an FDR controlled at a prespecified significance level q,

$$\text{FDR} \coloneqq \mathbb{E}\left(\frac{|\widehat{S} \cap \mathcal{H}_0|}{|\widehat{S}| \vee 1}\right) \leq q,$$

where $\mathcal{H}_0 \subset [p]$ is the set of null hypotheses. To achieve this, the critical step is to construct a group of knockoff variables $\widetilde{G}$, such that they are exchangeable to the originals. In specific, the following exchangeability needs to be satisfied for any subset $S \subset [p]$:

$$\left(G, \widetilde{G}\right)_{\text{swap}(S)} \overset{d}{=} \left(G, \widetilde{G}\right), \tag{2.3}$$

where $(G, \widetilde{G})_{\text{swap}(S)}$ means for each $j \in S$, swap G_j with $\widetilde{G}_j$, and $\overset{d}{=}$ indicates equality in distribution. More specifically, for any $S \subset \mathcal{H}_0$,

$$\left(G, \widetilde{G}\right)_{\text{swap}(S)} \mid y \overset{d}{=} \left(G, \widetilde{G}\right) \mid y.$$

In addition, the knockoff variables need to be independent from the response given the original data: $\widetilde{G} \perp\!\!\!\perp Y \mid G$. With appropriately constructed knockoff variables, knockoff statistics W could be computed for each variable where a large positive W_j would provide evidence against the null hypothesis. The choice of knockoff statistics can be very flexible as long as they only depend on original and knockoff variables and the response via some function w_j,

$$W_j = w_j([G, \widetilde{G}], y),$$

and satisfy a flip-sign property, meaning that the sign of W_j should change whenever we swap variables,

$$w_j([G, \widetilde{G}]_{\text{swap}(j)}, y) = -w_j([G, \widetilde{G}], y), \quad \text{for } j \in S.$$

Note that by construction, the knockoff statistics for null variables should follow a symmetric distribution around 0 whereas for non-null variables they are supposed to be positive with larger probability. Taking advantage of this property, the false discovery proportion (FDP) could be estimated via

$$\widehat{\mathrm{FDP}}(t) = \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}}.$$

And for any target FDR level q, we could select a set of variables $\widehat{S} = \{j : W_j \geq \tau\}$ where

$$\tau = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\},$$

such that $\widehat{\mathrm{FDP}}(\tau) \leq q$. Note that in the case the above set is empty, $\tau = +\infty$ such that $\widehat{S} = \emptyset$ and $\widehat{\mathrm{FDP}}(+\infty) = 0$.

To apply the model-X knockoff filter to single-cell RNA sequencing (scRNA-seq) data, we mainly face the following two difficulties. First, it is challenging to construct a group of knockoff variables satisfying the exchangeability condition (2.3), due to the high-dimensionality and large proportion of missing values in scRNA-seq data. Second, it is unclear what is an appropriate choice of knockoff statistics. In the following sections, we will provide solutions to overcome these two difficulties. In particular, in order to deal with large number of missingness in scRNA-seq data, we introduce a modification of the softImpute algorithm that takes advantage of a low-rank approximation to impute missing values, and estimate the distribution of gene expressions after the imputation.

2.3 Imputation

The large number of missingness raises several issues for the downstream data analysis. First, if the missing values are directly dropped, only a very limited amount of data will be available afterwards given the high missing rate, and consequently one may expect a significant power loss. Second, the missing values may affect the estimation of the distribution G. In particular, the correlation among gene expressions can be underestimated when missingness is present. Therefore, we suggest to impute the missing values.

2.3.1 Imputation for scRNA-seq data

Fortunately, there is already a large collection of existing literature on missing value imputation, some of them being referred as matrix completion. A commonly used strategy is to take advantage of a low-rank approximation. Along this direction there are a number of important works posing missing value imputation under the context of convex optimization, and introduced efficient algorithms for solving the convex problems [Candès and Tao, 2010; Hastie et al., 1999, 2015; Mazumder et al., 2010]. For scRNA-seq data, we will adapt the algorithm proposed in Hastie et al. [2015]. First, we provide a brief overview of the original algorithm, softImpute.

Assume that $\boldsymbol{G}^{\text{obs}}$ is centered in the sense that the column means $\boldsymbol{m} = (m_j)_{1\le j\le p}$ of the *expressed* part are subtracted from the observed expressions. More explicitly, m_j is calculated as

$$m_j = \frac{\sum_{i=1}^{n} G_{ij}^{\text{obs}} \mathbb{1}_{\{G_{ij}^{\text{obs}} \text{ is expressed}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{G_{ij}^{\text{obs}} \text{ is expressed}\}}}.$$

In the same fashion, $\boldsymbol{G}$ is assumed to be centered and $\boldsymbol{G}$ can then be recovered by solving the following optimization problem

$$\text{minimize}_{\boldsymbol{A},\boldsymbol{B}} \ \frac{1}{2}\left\|P_{\Omega}(\boldsymbol{G}-\boldsymbol{A}\boldsymbol{B}^{\top})\right\|_F^2 + \frac{\lambda}{2}\left(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2\right), \tag{2.4}$$

where $\|\cdot\|_F$ is the Frobenius norm. $\boldsymbol{A}$ and $\boldsymbol{B}$ are two matrices with of dimension $n \times r$ and $p \times r$ respectively. And the missing values are imputed by setting $\boldsymbol{G}^{\text{imp}} \leftarrow P_{\Omega}(\boldsymbol{G}^{\text{obs}}) + P_{\Omega}^{\perp}(\boldsymbol{A}\boldsymbol{B}^{\top})$, where $P_{\Omega}^{\perp}$ is the projection onto the unexpressed set.

In scRNA-seq data, we usually have some additional information collected in the experiment, such as gender, age, batch label, and cellular detection rate (CDR), which is the proportion of genes detected in each cell. These variables can usually explain a proportion of variance. Moreover, relying solely on information internal to the imputed data may introduce inflated correlation between the genes and cells [Andrews and Hemberg, 2019]. When some of the additional variables are identified as confounders, conditioning on them can help alleviate confounding effects. Therefore, we hope to introduce an imputation algorithm by considering this set of covariates. The new algorithm will be a modification of softImpute.

Let's define $\boldsymbol{A} = [\boldsymbol{X}, \boldsymbol{A}']$, where $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ corresponds to a group of k centered covariates observed in the single-cell sequencing experiment, which are also viewed as known factors, and $\boldsymbol{A}' \in \mathbb{R}^{n\times(r-k)}$ $(r \ge k)$ is the matrix of unknown factors. The optimization problem for missing

value imputation then becomes

$$\text{minimize}_{\boldsymbol{A}',\boldsymbol{B}} \ \frac{1}{2}\left\|P_\Omega(\boldsymbol{G}-\boldsymbol{A}\boldsymbol{B}^\top)\right\|_F^2+\frac{\lambda}{2}\left(\|\boldsymbol{A}'\|_F^2+\|\boldsymbol{B}\|_F^2\right). \tag{2.5}$$

We will apply Algorithm 1, which is a modification of the Algorithm 5.1 in Hastie et al. [2015], to update $\boldsymbol{A}'$ and $\boldsymbol{B}$ iteratively.

Algorithm 1: sc-softImpute

1. Initialize $\ell=0$, $\boldsymbol{A}'_0=\boldsymbol{U}\boldsymbol{D}$, $\boldsymbol{A}_0=[\boldsymbol{X},\boldsymbol{A}'_0]$ and $\boldsymbol{B}_0={\boldsymbol{G}^{\text{obs}}}^\top\boldsymbol{A}_0(\boldsymbol{A}_0^\top\boldsymbol{A}_0)^{-1}$, where $\boldsymbol{U}\boldsymbol{D}$ is obtained by performing singular value decomposition (SVD) of $\boldsymbol{G}^{\text{obs}}$ and taking the leading $r-k$ factors.
2. Fix $\boldsymbol{B}_\ell$ and update $\boldsymbol{A}_{\ell+1}$ by

$$\boldsymbol{A}'_{\ell+1}\leftarrow\boldsymbol{G}^{\text{imp}}\boldsymbol{B}_{A',\ell}(\boldsymbol{B}_{A',\ell}^\top\boldsymbol{B}_{A',\ell}+\lambda\boldsymbol{I})^{-1} \tag{2.6}$$

   where $\boldsymbol{G}^{\text{imp}}\leftarrow P_\Omega(\boldsymbol{G}^{\text{obs}})+P_\Omega^\perp([\boldsymbol{X},\boldsymbol{A}'_\ell]\boldsymbol{B}_\ell^\top)-\boldsymbol{X}\boldsymbol{B}_{X,\ell}^\top$ and $\boldsymbol{B}_\ell=[\boldsymbol{B}_{X,\ell},\boldsymbol{B}_{A',\ell}]$. Then let $\boldsymbol{A}_{\ell+1}\leftarrow[\boldsymbol{X},\boldsymbol{A}'_{\ell+1}]$.
3. Fix $\boldsymbol{A}_{\ell+1}$ and update $\boldsymbol{B}_{\ell+1}$ by

$$\boldsymbol{B}_{\ell+1}\leftarrow{\boldsymbol{G}^{\text{imp}}}^\top\boldsymbol{A}_{\ell+1}(\boldsymbol{A}_{\ell+1}^\top\boldsymbol{A}_{\ell+1}+\lambda\boldsymbol{I})^{-1} \tag{2.7}$$

   where $\boldsymbol{G}^{\text{imp}}\leftarrow P_\Omega(\boldsymbol{G}^{\text{obs}})+P_\Omega^\perp(\boldsymbol{A}_{\ell+1}\boldsymbol{B}_\ell^\top)$.
4. $\ell\leftarrow\ell+1$.
5. Repeat step 2 to 4 until convergence.

It can be shown that Algorithm 1 will converge and the proof closely follows that of Theorem 3 in Hastie et al. [2015]. For simplicity, let's denote the objective function as

$$F(\boldsymbol{A}',\boldsymbol{B})\coloneqq\frac{1}{2}\left\|P_\Omega(\boldsymbol{G}-[\boldsymbol{X},\boldsymbol{A}']\boldsymbol{B}^\top)\right\|_F^2+\frac{\lambda}{2}\left(\|\boldsymbol{A}'\|_F^2+\|\boldsymbol{B}\|_F^2\right),$$

such that (2.5) can be rewritten as

$$\text{minimize}_{\boldsymbol{A}',\boldsymbol{B}}F(\boldsymbol{A}',\boldsymbol{B}).$$

For the objective function $F(\boldsymbol{A}',\boldsymbol{B})$, we can show its value decreases in each iteration, as summarized in the following Theorem 1. In combination with the fact that $F(\boldsymbol{A}',\boldsymbol{B})$ has a lower bound, we know the algorithm will converge.

Theorem 1. *Let $\{(\boldsymbol{A}'_\ell, \boldsymbol{B}_\ell)\}$ be the iterates generated by Algorithm 1. Then the function values are monotonically decreasing,*

$$F(\boldsymbol{A}'_\ell, \boldsymbol{B}_\ell) \geq F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_\ell) \geq F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_{\ell+1}), \quad \ell \geq 1.$$

In the final step of the imputation, we will add noise to the imputed values by sampling $e_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_j^2)$. We estimate σ_j^2 by

$$\hat{\sigma}_j^2 = \left\| P_{\Omega_j}(\boldsymbol{G}^{\text{obs}} - \boldsymbol{A}_\ell \boldsymbol{B}_\ell^\top) \right\|_2^2 \Big/ \left(\sum_{i=1}^{n} \mathbb{1}_{\{G_{ij}^{\text{obs}} \text{ is expressed}\}} - r - 1 \right), \tag{2.8}$$

where P_{Ω_j} is the projection of the matrix onto the expressed set for variable j and $\|\cdot\|_2$ is the ℓ_2 norm. The idea of adding noise to the recovered values in $\Omega^{\complement}$ stems from the model (2.1). Since $\boldsymbol{A}_\ell \boldsymbol{B}_\ell^\top$ is an estimator for $\boldsymbol{G}^*$, to keep consistency between the unobserved values as well as the observed ones, the final output therefore should be

$$\boldsymbol{G} \leftarrow P_\Omega(\boldsymbol{G}^{\text{obs}}) + P_\Omega^\perp(\boldsymbol{A}_\ell \boldsymbol{B}_\ell^\top) + P_\Omega^\perp(\boldsymbol{E}),$$

where the entries of $\boldsymbol{E}$ are randomly sampled e_{ij}'s.

2.3.2 Theoretical intuition for imputation

Farias et al. [2022] developed theoretical support for noisy matrix completion of low-rank matrices given only partial and corrupted entries as described in (2.1) that provide some intuition for the algorithm in this section. Assume that

- $\boldsymbol{G}^* \in \mathbb{R}^{n \times p}$ is a rank r matrix where $n \leq p$ and $\boldsymbol{G}^* = \boldsymbol{U}^* \boldsymbol{D}^* \boldsymbol{V}^{*\top}$ is the SVD of $\boldsymbol{G}^*$.
- Each index (i, j) belongs to the observed set Ω independently with a probability θ.
- $\|\boldsymbol{U}^*\|_{2,\infty} \leq \sqrt{\mu r / n}$ and $\|\boldsymbol{V}^*\|_{2,\infty} \leq \sqrt{\mu r / p}$ where $\|\cdot\|_{2,\infty}$ indicates the largest ℓ_2 norm of all rows of a matrix.
- The entries of $\boldsymbol{E}$ are independent, mean-zero, sub-exponential random variables with variances σ_{ij}^2 such that $\inf\{t > 0 : \mathbb{E}(\exp(|e_{ij}|/t)) \leq 2\} \leq L$, and are independent from Ω.
- $n\theta \gg \kappa^4 \mu^2 r^2 \log^3 p$ and $L \log(p) \sqrt{p/\theta} \ll \sigma_{\min} / \sqrt{\kappa^4 \mu r \log p}$ where $\sigma_{\min}$ and $\sigma_{\max}$ are smallest and largest singular values in $\boldsymbol{D}^*$ respectively and $\kappa = \sigma_{\max}/\sigma_{\min}$.

Let $\boldsymbol{G}^d = \boldsymbol{A}^d \boldsymbol{B}^{d\top}$ where $\boldsymbol{A}^d = \boldsymbol{A}_\ell \left(I_r + \frac{\lambda}{\theta}(\boldsymbol{A}_\ell^\top \boldsymbol{A}_\ell)^{-1}\right)^{1/2}$ and
$\boldsymbol{B}^d = \boldsymbol{B}_\ell \left(I_r + \frac{\lambda}{\theta}(\boldsymbol{B}_\ell^\top \boldsymbol{B}_\ell)^{-1}\right)^{1/2}$. The $\boldsymbol{A}_\ell$ and $\boldsymbol{B}_\ell$ here are the iterates from the *original* softImpute algorithm and $\boldsymbol{G}^d$ is the de-biased estimator for G^*. Then for every (i, j), we have

$$\sup_{t\in\mathbb{R}} \left| P\left\{ \frac{G_{ij}^d - G_{ij}^*}{s_{ij}} \le t \right\} - \Phi(t) \right| \lesssim s_{ij}^{-3} \frac{L^2\mu^3 r^3}{n^2\theta} + s_{ij}^{-1}\left(\frac{L^2 \log^3(p)\mu r \kappa^5}{\theta\sigma_{\min}} + \frac{L\log^2(p)\mu^2 r^2 \kappa^4}{\theta n} \right) + \frac{1}{n^{10}},$$

where $\Phi(\cdot)$ is the CDF of the standard Gaussian, and $s_{ij} > 0$ is defined as

$$s_{ij}^2 := \frac{\sum_{l=1}^n \sigma_{lj}^2 \left(\sum_{k=1}^r U_{ik}^* U_{lk}^*\right)^2 + \sum_{l=1}^p \sigma_{il}^2 \left(\sum_{k=1}^r V_{lk}^* V_{jk}^*\right)^2}{\theta}.$$

The results in Farias et al. [2022] is a direct improvement over Chen et al. [2019]. In fact, if we assume homogeneous Gaussian noise and a square matrix, the results in the first remark will reduce to Theorem 2 in Chen et al. [2019]. Similar discussions not involving de-biasing could be found in Chen et al. [2020] under the homogeneous Gaussian noise and square matrix assumption.

The results described above can not be directly applied to sc-softImpute but we can achieve theoretical results that are close. Adjust Algorithm 1 where it will instead optimize

$$\text{minimize}_{\boldsymbol{A}', \boldsymbol{B}} \ \frac{1}{2}\left\| P_\Omega(\boldsymbol{G} - \boldsymbol{A}\boldsymbol{B}^\top) \right\|_F^2 + \frac{\lambda}{2}\left(\|\boldsymbol{A}'\|_F^2 + \|\boldsymbol{B}_A\|_F^2 \right).$$

Algorithm 2: sc-softImpute-debias

1. Initialize $\ell = 0$, $\boldsymbol{A}'_0 = \boldsymbol{U}\boldsymbol{D}$, $\boldsymbol{A}_0 = [\boldsymbol{X}, \boldsymbol{A}'_0]$ and $\boldsymbol{B}_0 = \boldsymbol{G}^{\text{obs}\top}\boldsymbol{A}_0(\boldsymbol{A}_0^\top\boldsymbol{A}_0)^{-1}$, where $\boldsymbol{U}\boldsymbol{D}$ is obtained by performing singular value decomposition (SVD) of $\boldsymbol{G}^{\text{obs}}$ and taking the leading $r-k$ factors.
2. Fix $\boldsymbol{B}_\ell$ and update $\boldsymbol{A}_{\ell+1}$ by

$$\boldsymbol{A}'_{\ell+1} \leftarrow \boldsymbol{G}^{\text{imp}}\boldsymbol{B}_{A',\ell}(\boldsymbol{B}_{A',\ell}^\top\boldsymbol{B}_{A',\ell} + \lambda\boldsymbol{I})^{-1} \tag{2.9}$$

   where $\boldsymbol{G}^{\text{imp}} \leftarrow P_\Omega(\boldsymbol{G}^{\text{obs}}) + P_\Omega^\perp([\boldsymbol{X}, \boldsymbol{A}'_\ell]\boldsymbol{B}_\ell^\top) - \boldsymbol{X}\boldsymbol{B}_{X,\ell}^\top$ and $\boldsymbol{B}_\ell = [\boldsymbol{B}_{X,\ell}, \boldsymbol{B}_{A',\ell}]$. Then let $\boldsymbol{A}_{\ell+1} \leftarrow [\boldsymbol{X}, \boldsymbol{A}'_{\ell+1}]$.
3. Fix $\boldsymbol{A}_{\ell+1}$ and update $\boldsymbol{B}_{\ell+1}$ by

$$\boldsymbol{B}_{\ell+1} \leftarrow \boldsymbol{G}^{\text{imp}\top}\boldsymbol{A}_{\ell+1}(\boldsymbol{A}_{\ell+1}^\top\boldsymbol{A}_{\ell+1} + \lambda\boldsymbol{I}_A)^{-1}$$

   where $\boldsymbol{G}^{\text{imp}} \leftarrow P_\Omega(\boldsymbol{G}^{\text{obs}}) + P_\Omega^\perp(\boldsymbol{A}_{\ell+1}\boldsymbol{B}_\ell^\top)$, and $\boldsymbol{I}_A$ is an identity matrix where the first k diagonal elements are set to 0.
4. $\ell \leftarrow \ell + 1$.
5. Repeat step 2 to 4 until convergence.
6. Let $\boldsymbol{G}^d = \left[\boldsymbol{X}, \boldsymbol{A}^d\right]\left[\boldsymbol{B}_{X,\ell}, \boldsymbol{B}^d\right]^\top$ where $\boldsymbol{A}^d = \boldsymbol{A}'_\ell\left(\boldsymbol{I}_{r-k} + \frac{\lambda}{\theta}(\boldsymbol{A}_\ell^{'\top}\boldsymbol{A}'_\ell)^{-1}\right)^{1/2}$ and $\boldsymbol{B}^d = \boldsymbol{B}_{A',\ell}\left(\boldsymbol{I}_{r-k} + \frac{\lambda}{\theta}(\boldsymbol{B}_{A',\ell}^\top\boldsymbol{B}_{A',\ell})^{-1}\right)^{1/2}$.

It could be shown with slight modifications to Theorem 1 that Algorithm 2 converges, as detailed in Appendix A.2. While a full proof for debiasing is outside the scope of this work, an intuitive explanation is that $\boldsymbol{B}_{X,\ell}$ is an unbiased estimation, since it is not penalized. On the other hand, $\boldsymbol{A}'_\ell$ and $\boldsymbol{B}_{A',\ell}$ are biased estimations, hence undergo a debiasing process following the methodology outlined by Farias et al. [2022]. However, since the debiasing process increases the variance of the recovered matrix, it reduces the power of the constructed knockoffs. For further details on the impact of this trade-off, simulation results are included in Appendix A.3.

2.4 Knockoff construction

In this section, we will discuss several methods for constructing knockoff variables based on the imputation result. This includes a construction based on the spiked covariance model in Section 2.4.2, and another one based on the low-rank approximation in Section 2.4.3. By taking advantage of the latent factor model (2.2), these two constructions provide a more computationally efficient way for handling high-dimensional single-cell data, in comparison to the standard Gaussian knockoff method discussed in Section 2.4.1. Additionally, we will introduce a method for efficiently constructing multiple groups of knockoff variables in Section 2.4.4.

2.4.1 Review of Gaussian knockoffs

As discussed in Candès et al. [2018], when G is assumed to be Gaussian, to construct knockoff variables for G that satisfy the exchangeability condition (2.3), it would suffice to match the first two moments of $(G, \widetilde{G})_{\text{swap}(S)}$ and $(G, \widetilde{G})$ for any subset $S \subset [p]$ and then sample $\widetilde{G}$ from the conditional multivariate Gaussian distribution $\widetilde{G} \mid G$. Matching the first moment is easy to achieve. Matching the second moment is equivalent to finding a diagonal matrix $\text{diag}\{\boldsymbol{s}\}$ such that $\text{Cov}\left((G, \widetilde{G})\right)$ is positive semidefinite, where

$$\text{Cov}\left((G, \widetilde{G})\right) = \begin{bmatrix} \boldsymbol{\Sigma}_G & \boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma}_G \end{bmatrix}.$$

This indicates that $\boldsymbol{s} \in \mathbb{R}^p$ should satisfy

$$s_j \geq 0 \text{ and } \text{diag}\{\boldsymbol{s}\} \preceq 2\boldsymbol{\Sigma}_G. \tag{2.10}$$

Without loss of generality, we may assume the diagonal entries of $\boldsymbol{\Sigma}_G$ equal to 1 for the remainder of this subsection. In Candès et al. [2018], the authors introduced the following two methods for finding $\boldsymbol{s}$: The semidefinite program (SDP) construction that solves the convex program

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_j |1 - s_j^{\text{SDP}}| \\ \text{subject to} \quad & s_j^{\text{SDP}} \geq 0 \\ & \text{diag}\{\boldsymbol{s}^{\text{SDP}}\} \preceq 2\boldsymbol{\Sigma}_G. \end{aligned}$$

And the approximate semidefinite program (ASDP) construction whose purpose is to accelerate the SDP construction for high-dimensional problems. It is a two-step procedure: First, choose an

approximation $\boldsymbol{\Sigma}_{\text{approx}}$ for $\boldsymbol{\Sigma}_G$ and solve

$$\begin{aligned} &\text{minimize} && \textstyle\sum_j |1 - s_j| \\ &\text{subject to} && s_j \geq 0 \\ & && \text{diag}\{\boldsymbol{s}\} \preceq 2\boldsymbol{\Sigma}_{\text{approx}}. \end{aligned}$$

Then solve

$$\begin{aligned} &\text{maximize} && \gamma \\ &\text{subject to} && \text{diag}\{\gamma\boldsymbol{s}\} \preceq 2\boldsymbol{\Sigma}_G. \end{aligned}$$

Finally, set $\boldsymbol{s}^{\text{ASDP}} = \gamma\boldsymbol{s}$.

However, both approaches are difficult to implement for scRNA-seq data without further adjustments, primarily due to the high-dimensionality. Even though the low expression rate has been addressed in Section 2.3 by recovering the incomplete data, the large n and p remains to be an obstacle as it will take a lot of computational resources to estimate $\boldsymbol{\Sigma}_G$ and solve the optimization problem in order to obtain $\boldsymbol{s}$. Even with the application of ASDP, where $\boldsymbol{\Sigma}_{\text{approx}}$ could be constructed as a block diagonal matrix by splitting $\boldsymbol{\Sigma}_G$ into multiple more computationally feasible blocks and running multiple SDPs in parallel, the second step of ASDP could still be slow.

2.4.2 Knockoff construction based on the spiked covariance model

In model (2.2), if we further impose another normality assumption for the random latent factors $\boldsymbol{A}$, it then implies a spiked covariance model [Johnstone, 2001] for the data $\boldsymbol{G}$. Efficient algorithms are available for knockoff generation based on such a model. To be more specific, we would like to assume rows of the matrix $\boldsymbol{A}$ are independent and identically distributed according to the following multivariate normal model

$$A \sim \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A). \tag{2.11}$$

Then model (2.2) can be equivalently written as

$$G \mid A = \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{B}\boldsymbol{a}_i, \boldsymbol{D}), \tag{2.12}$$

where G and A are $p \times 1$ and $r \times 1$ random vectors respectively, $\boldsymbol{a}_i^\top = (a_{il})_{1\leq l\leq r}^\top$ are the i-th row of $\boldsymbol{A}$, and $\boldsymbol{D} = \text{diag}((\sigma_j^2)_{1\leq j\leq p})$. Without loss of generality, we may assume $\boldsymbol{\mu}_A = 0$. After some

calculations, we are able to demonstrate that (2.11) and (2.12) together imply the covariance of G is

$$\boldsymbol{\Sigma}_G = \boldsymbol{D} + \boldsymbol{B}\boldsymbol{\Sigma}_A\boldsymbol{B}^\top. \tag{2.13}$$

The decomposition (2.13) is also called the spiked covariance model in the literature. It assumes the covariance can be captured by a low-rank component $\boldsymbol{B}\boldsymbol{\Sigma}_A\boldsymbol{B}^\top$ and a diagonal component $\boldsymbol{D}$, therefore achieving a faster estimation of the covariance. Such a model is frequently imposed for high-dimensional data.

In practice, $\boldsymbol{A}$ and $\boldsymbol{B}$ can be recovered by applying Algorithm 1, $\boldsymbol{\Sigma}_A$ can be estimated using the empirical covariance of $\boldsymbol{A}$, and σ_j^2 can be estimated by (2.8) as in Section 2.3. We will denote our estimate for $\boldsymbol{\Sigma}_G$ based on the model (2.13) as $\widehat{\boldsymbol{\Sigma}}_G$.

To construct knockoff variables, the spiked covariance model also allows us to skip the step of solving SDP or ASDP, which is usually the most time consuming part for high-dimensional data. Under model (2.13), it is easy to show that

$$\boldsymbol{s}^{\text{decomp}} = (2D_{jj})_{1\le j\le p}$$

satisfies (2.10) and will lead to valid construction of knockoffs. Alternatively, for low-dimensional data, after estimating $\boldsymbol{\Sigma}_G$, we can still solve the ASDP to find $\boldsymbol{s}$. In the rest of the chapter, we will denote the knockoff variables constructed by solving the ASDP with $\boldsymbol{\Sigma}_G$ as asdp knockoffs, and, on top of that, knockoffs constructed by directly providing $\boldsymbol{s}^{\text{decomp}}$ as decomp knockoffs.

Could there be a more optimal $\boldsymbol{s}$? In terms of solving an optimization problem, possibly yes, since there might be some more room to stretch into in $2\boldsymbol{B}\boldsymbol{\Sigma}_A\boldsymbol{B}^\top$. But the search for the optimal $\boldsymbol{s}$ would defeat the purpose of saving computational resources as this would again require solving an SDP or ASDP program for a $p \times p$ matrix. It must also be emphasized that while $\boldsymbol{s}^{\text{decomp}}$ is not "optimal" in terms of optimization, it does not necessarily mean that it leads to a less powerful knockoff construction. For example, Gimenez and Zou [2019] have pointed out that there exist alternate convex optimization problems to construct knockoffs and have also pointed out that SDP knockoff constructions will maximize some diagonal terms at the cost of others. In fact, this newly proposed knockoff construction could non-trivially become one of the most effective ways to generate knockoff variables for scRNA-seq data, and high-dimensional data in general, due to its relative simplicity. We will discuss more on this topic in Section 2.4.6 and Section 2.6.

2.4.3 Knockoff construction based on the low-rank decomposition

A novel approach introduced and discussed in works like Fan et al. [2020] and Zhu et al. [2021] is the other direction that we wish to focus on. Given the model (2.2), knockoff variables could be constructed as

$$\widetilde{\boldsymbol{G}} = \boldsymbol{A}\boldsymbol{B}^\top + \widetilde{\boldsymbol{E}}, \tag{2.14}$$

where entries of $\widetilde{\boldsymbol{E}}$ are independently sampled from $\mathcal{N}(0, \sigma_j^2), 1 \leq j \leq p$. Such a construction (2.14) satisfies the exchangeability condition (2.3), but it requires the knowledge of $\boldsymbol{G}^*$ and the distribution of $\boldsymbol{E}$—both of which are only available in an oracle scenario. In practice, both parts need to be estimated such that the knockoff variables are constructed as

$$\widetilde{\boldsymbol{G}} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{B}}^\top + \widetilde{\boldsymbol{E}}^{\text{est}},$$

where each $\tilde{e}_{ij}^{\text{est}}$ is independently sampled from $\mathcal{N}(0, \hat{\sigma}_j^2)$, and $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{B}}$ are estimated low-rank structures. Note that this low-rank construction can be directly constructed based on the results from Algorithm 1.

The advantage of the low-rank (LR) knockoff construction lies in its convenience and speed. It is immediately available after recovering the missing gene expressions and avoids the hurdles of covariance estimation and solving SDPs. However, it should be emphasized that the exchangeability (2.3) is no longer strictly satisfied in practice as the estimation of $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{E}}^{\text{est}}$ now depends on the data $\boldsymbol{G}^{\text{obs}}$. In Fan et al. [2020], the authors provided proofs for asymptotic FDR control and power analysis under the homoscedastic sub-Gaussian error assumption and Lasso coefficient-difference knockoff statistics. Though the asymptotic properties of the knockoff variables under the heteroscedastic case and more general knockoff statistics remains to be explored, this construction provides compelling empirical FDR control and power for variable selection.

2.4.4 Multiple knockoffs and e-BH procedure

Variables selected by the knockoff framework are inherently random due to the stochastic nature of knockoff variables. Previous studies, such as those by Gimenez and Zou [2019], He et al. [2021] and Ren and Barber [2022], have shown that by constructing multiple groups of knockoff variables, one can stabilize the selection. The construction of multiple Gaussian knockoffs has been studied by

Gimenez and Zou [2019] in detail. Recently, Ren and Barber [2022] have shown that one can translate the knockoff procedure to an e-BH procedure [Wang and Ramdas, 2022] for derandomization by combining multiple single knockoffs.

For scRNA-seq data, we will generalize constructions introduced in Section 2.4.2 and Section 2.4.3 to the multiple knockoffs case. To do so, we need to first generalize the exchangeability condition (2.3): Let $G^0 = G$ denote the original variables and $\left(G^1, \ldots, G^M\right)$ denote M group of knockoff variables. To simplify the notation, we assume that the random variables are organized as row vectors. The exchangeability is defined as

$$\left(G^0, G^1, \ldots, G^M\right)_{\text{swap}(\pi)} \stackrel{d}{=} \left(G^0, G^1, \ldots, G^M\right) \quad \text{for any } \pi, \tag{2.15}$$

where $\pi = (\pi_j)_{1 \le j \le p}$ is a collection of p permutations over the set of integers $\{0, \ldots, M\}$ and $\left(G^0, G^1, \ldots, G^M\right)_{\text{swap}(\pi)}$ is understood as the j-th variable from group m is $G_j^{\pi_j(m)}$. For example, with $p = 2$, $M = 2$, and only the first variable being permuted,

$$\left(G_1^0, G_2^0, G_1^1, G_2^1, G_1^2, G_2^2\right)_{\text{swap}(\pi_1)} \stackrel{d}{=} \left(G_1^{\pi_1(0)}, G_2^0, G_1^{\pi_1(1)}, G_2^1, G_1^{\pi_1(2)}, G_2^2\right).$$

Note that when $M = 1$, the exchangeability condition (2.15) boils down to (2.3). Similar to the single knockoff case, the knockoff variables should also be conditionally independent of Y: $\left(G^1, \ldots, G^M\right) \perp\!\!\!\perp Y \mid G^0$. When G is Gaussian, a sufficient condition is to let the joint distribution $\left(G^0, G^1, \ldots, G^M\right)$ follow a multivariate Gaussian distribution with matching mean for each G^m and covariance

$$\boldsymbol{\Sigma}_M = \begin{bmatrix} \boldsymbol{\Sigma}_G & \boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}\} & \cdots & \boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma}_G & \cdots & \boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}\} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}\} & \cdots & \boldsymbol{\Sigma}_G \end{bmatrix},$$

where $\boldsymbol{\Sigma}_M$ has $M + 1$ blocks column and row-wise and $\boldsymbol{s}$ is chosen such that $\boldsymbol{\Sigma}_M$ is positive semidefinite. When diagonal elements of $\boldsymbol{\Sigma}_G$ have been scaled to 1, a proper $\boldsymbol{s}$ could be obtained by solving an SDP

$$\begin{aligned} \text{minimize} \quad & \sum_j |1 - s_j^{\text{SDP}}| \\ \text{subject to} \quad & s_j^{\text{SDP}} \ge 0 \\ & \text{diag}\{\boldsymbol{s}^{\text{SDP}}\} \preceq \frac{M+1}{M} \boldsymbol{\Sigma}_G, \end{aligned}$$

or an ASDP as described in Section 2.4.1 with only slight modifications.

For the construction based on the spiked covariance model, while we can still avoid solving the SDP by directly providing

$$\boldsymbol{s}^{\text{multi-decomp}} = \frac{M+1}{M}(D_{jj})_{1\leq j\leq p}, \tag{2.16}$$

generating multiple decomp (multi-decomp) knockoffs requires sampling from an $M \times p$ dimensional multivariate normal distribution $\left(G^1, \ldots, G^M\right) \mid G^0$, which is computationally prohibitive. This, however, can be mitigated by a clever technique proposed in He et al. [2024]. Notice that the Gaussian knockoff variables in a single knockoff setting can be written as

$$\widetilde{G}^\top = \left(\boldsymbol{I} - \text{diag}\{\boldsymbol{s}\}\boldsymbol{\Sigma}_G^{-1}\right) G^\top + V^\top,$$

where $V \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$ and $\boldsymbol{C} = 2\text{diag}\{\boldsymbol{s}\} - \text{diag}\{\boldsymbol{s}\}\boldsymbol{\Sigma}_G^{-1}\text{diag}\{\boldsymbol{s}\}$. In a similar fashion, the multiple knockoff variables can be written as

$$\left(G^1, \ldots, G^M\right)^\top = \begin{bmatrix} \boldsymbol{I} - \text{diag}\{\boldsymbol{s}\}\boldsymbol{\Sigma}_G^{-1} \\ \vdots \\ \boldsymbol{I} - \text{diag}\{\boldsymbol{s}\}\boldsymbol{\Sigma}_G^{-1} \end{bmatrix} G^\top + V_M^\top,$$

where $V_M \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_M)$ and

$$\boldsymbol{C}_M = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{C} - \text{diag}\{\boldsymbol{s}\} & \cdots & \boldsymbol{C} - \text{diag}\{\boldsymbol{s}\} \\ \boldsymbol{C} - \text{diag}\{\boldsymbol{s}\} & \boldsymbol{C} & \cdots & \boldsymbol{C} - \text{diag}\{\boldsymbol{s}\} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C} - \text{diag}\{\boldsymbol{s}\} & \boldsymbol{C} - \text{diag}\{\boldsymbol{s}\} & \cdots & \boldsymbol{C} \end{bmatrix}.$$

Instead of sampling from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_M)$, an $M \times p$ dimensional multivariate normal distribution, we can take advantage of the structure of $\boldsymbol{C}_M$ and use the following sampling method:

1. Sample V_1 from $\mathcal{N}\left(\boldsymbol{0}, \boldsymbol{C} - \frac{M-1}{M}\text{diag}\{\boldsymbol{s}\}\right)$.
2. Sample $V_{2,m}$ i.i.d. from $\mathcal{N}(\boldsymbol{0}, \text{diag}\{\boldsymbol{s}\})$, for $m = 1, \ldots, M$.
3. Calculate $V_2^m = V_{2,m} - \bar{V}_2$, where $\bar{V}_2 = \frac{1}{M}\sum_{m=1}^M V_{2,m}$.
4. Calculate $V^m = V_1 + V_2^m$ and let $G^{m\top} = \left(\boldsymbol{I} - \text{diag}\{\boldsymbol{s}\}\boldsymbol{\Sigma}_G^{-1}\right) G^\top + V^{m\top}$.

The computational complexity is reduced from $O(M^3p^3)$ to $O(p^3)$.

For the low-rank approximation based method, to construct multiple knockoffs, it suffices to sample from independent normal distributions repeatedly. Specifically, the m-th group of knockoffs can be obtained as

$$\boldsymbol{G}^m = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{B}}^\top + \widetilde{\boldsymbol{E}}^{\text{est},m}, \quad \tilde{e}_{ij}^{\text{est},m} \sim \mathcal{N}(0, \hat{\sigma}_j^2) \text{ independently.}$$

This way, the low-rank construction skips the step of sampling from an ultra-high-dimensional normal distribution, and is more efficient in practice.

We will defer the description of the e-BH procedure to Section 2.5.1 since it is in fact a variable selection method. For the e-BH procedure according to Ren and Barber [2022], it derandomizes knockoffs with e-values [Vovk and Wang, 2021], and there is no assumption on the dependence structure of those e-values [Wang and Ramdas, 2022]. Therefore, no simultaneous multiple knockoff generation as described in Gimenez and Zou [2019] is required, but multiple single-knockoff results are combined. This allows us to derandomize the variable selection of not only decomp and LR knockoffs, but any knockoff construction. However, multiple knockoffs are still favored in the case of sparse signals or stricter FDR targets due to lower detection thresholds $\lceil \frac{1}{qM} \rceil$.

2.4.5 Knockoff variable rescaling

Recall that we assume both $\boldsymbol{G}$ and $\boldsymbol{G}^{\text{obs}}$ are centered in the previous sections. While the exchangeability between $\boldsymbol{G}$ and $\widetilde{\boldsymbol{G}}$ is clear, the same exchangeability will not hold for the original log-normalized gene expression data $\boldsymbol{G}^{\text{obs}}$. As suggested in Andrews and Hemberg [2019], the un-imputed data should be used to calculate DEG tests, and therefore the knockoff statistics $W = (W_1, \ldots, W_p)$ should as well. This requires us to rescale the knockoff variables or there will inevitably be a mismatch between the first and second moments that will violate the exchangeability property (2.3). To address this issue and to mimic the observed gene expression data, the knockoffs are handled in two parts:

- For the observed expressions, we add column means of the expressed parts $\boldsymbol{m}$ back to $\widetilde{\boldsymbol{G}}$.
- For the unexpressed part, the corresponding values of knockoff variables are set to be 0.

The idea is that each entry of the observed data $\boldsymbol{G}^{\text{obs}}$ can be characterized by a conditional distribution $G_{ij}|Z_{ij}$, where Z_{ij} is a random variable indicating whether G_{ij} is expressed. When $Z_{ij} = 0$, $G_{ij} = 0$; when $Z_{ij} = 1$, G_{ij} follows a continuous distribution. To ensure exchangeability, for the

knockoff part, we can find a trivial construction for Z_{ij} by setting $\widetilde{Z}_{ij} = Z_{ij}$. Due to the MCAR assumption, if we mask knockoff variables according to $\widetilde{Z}_{ij}$, the exchangeability still holds. In summary, we have the following lemma:

Lemma 1. *Let $\widetilde{\boldsymbol{G}}^{obs} = P_\Omega(\widetilde{\boldsymbol{G}})$ be the adjusted knockoff variables, then*

$$\left(\boldsymbol{G}^{obs}, \widetilde{\boldsymbol{G}}^{obs}\right)_{swap(S)} \stackrel{d}{=} \left(\boldsymbol{G}^{obs}, \widetilde{\boldsymbol{G}}^{obs}\right)$$

for any $j \in S$.

Certainly, there may be more sensible ways to preserve the exchangeability property that draw inspiration from a deeper understanding on the dependencies of genes within scRNA-seq data sets. However, since MCAR is assumed in this work, these kind of considerations are not necessary.

2.4.6 Computational complexity

For different knockoff construction approaches discussed so far, their computation mainly falls into the following four steps or part of them: imputation, covariance estimation, solving SDP or ASDP, and sampling. We will discuss the cost of each step in the following:

1. *Imputation.* The computational complexity for sc-softImpute, citing the results in Hastie et al. [2015], is $O(2|\Omega|r^2 + nr^3 + pr^3)$ for each iteration, where $|\Omega|$ is the total number of expressed genes in our data. The number of iterations required for convergence has not been discussed but in practice, the algorithm would provide marginal improvements after limited iterations and it would be acceptable to stop early.
2. *Covariance estimation.* The covariance reconstruction is $O(p^2 r)$ and $r \ll n$. In contrast to directly calculating the correlation, which is $O(np^2)$, our proposed method is much faster.
3. *Solving SDP or ASDP.* The complexity of solving an SDP would be $O(4p^3 \times \sqrt{p}\log(1/\epsilon))$ where $O(\sqrt{p}log(1/\epsilon_1))$ is the number of iterations required for convergence under a desired tolerance $\epsilon_1 > 0$ [Benson et al., 2000]. Solving an ASDP would be faster, as $\boldsymbol{\Sigma}_G$ could be split into multiple $p' \times p'$ block matrices such that the estimated complexity would be $O(4p'^{2.5} p \log(1/\epsilon_1) + p^3 \log(1/\epsilon_2))$ where $O(p^3 \log(1/\epsilon_2))$ is for searching for a maximum γ via binary search under the tolerance $\epsilon_2 > 0$.
4. *Sampling.* The complexity of sampling knockoff variables from a conditional multivariate Gaussian distribution is estimated to be of $O(p^3)$—mainly to calculate the inverse of the covariance

Table 2.1: A breakdown of the computational complexity of each knockoff construction.

	asdp	decomp	LR	multi-decomp	multi-LR
Imputation	✓	✓	✓	✓	✓
Covariance estimation	✓	✓		✓	
Solving SDP or ASDP	✓				
Sampling from multivariate normal	✓	✓		✓	

matrix. If we take advantage of the low-rank approximation in (2.13) and apply the Sherman-Morrison-Woodbury formula, the complexity would be estimated at $O(r^3 + pr^2 + p^2 r) = O(p^2 r)$.

As summarized in Table 2.1, the decomp-knockoff construction avoids solving an SDP or an ASDP as a whole since we are directly providing $\boldsymbol{s}^{\text{decomp}}$, and this will considerably speed up the whole process. Whereas multi-decomp knockoffs are not substantially slower to construction compared to decomp-knockoff. The LR-knockoff construction is the fastest as it is able to completely avoid the covariance reconstruction process and it samples knockoff variables from independent Gaussian distributions. Multi-LR knockoffs are only slightly slower than LR knockoffs to construct, as it repeats the sampling process M times but does not sample from a conditional multivariate Gaussian distribution. When it comes to variable selection, multi-LR knockoffs might lose the edge in terms of the cost for calculating knockoff statistics. In conclusion, when p is large, the proposed methods in this section might become the only feasible knockoff constructions in comparison to the ones introduced in Candès et al. [2018].

2.5 Variable Selection

Under the knockoff framework, the validity of conditional inference is mainly guaranteed by the exchangeability condition. When it holds, we do have some flexibility in terms of the choice of knockoff statistics. Conventionally, Lasso coefficients are used to construct knockoff statistics under

different settings, but under the context of differential analysis for scRNA-seq data, it would be preferred to use models that are specifically designed for this purpose. While FDR control would have theoretical guarantee, good knockoff statistics matter when a high power is desired as well. A wide variety of models could be used to analyze scRNA-seq data, including but not limited to methods that are initially developed for differential analysis of bulk RNA-seq data [Love et al., 2014; Robinson et al., 2009], zero-inflated models specifically for scRNA-seq data [Finak et al., 2015; Kharchenko et al., 2014; Risso et al., 2018], and nonparametric methods [Hollander et al., 2013; Li and Tibshirani, 2013]. In this chapter, we want to focus on the knockoff statistics W that is defined based on p-values obtained by applying three types of widely used tests for scRNA-seq data: the Wilcoxon rank sum test (WRT) [Hollander et al., 2013], the Model-based Analysis of Single-cell Transcriptomics (MAST) model [Finak et al., 2015], and logistic regression test (LRT). We would like to first briefly describe the three methods.

The WRT is a non-parametric two-sample test. For each variable, the observations are separated into two groups with n_1 and n_2 samples each, where $n_1 + n_2 = n$, and then ranked jointly. The test statistics equals to the sum of the ranks in the first group, which is then standardized and approximately calibrated by a standard Gaussian distribution,

$$Z_{\text{WRT}} = \frac{\sum \text{ranks}_{\text{group 1}} - n_1(n_1 + n_2 + 1)/2}{(n_1 n_2 (n_1 + n_2 + 1)/12)^{1/2}}.$$

The null hypothesis is rejected when $|Z_{\text{WRT}}| > z_{1-\alpha/2}$.

MAST is a hurdle model designed by Finak et al. [2015] to address the bimodal expression distribution and zero-inflation of gene expression. The expression rate $\mathbb{P}(Z_{ij} = 1)$ and the level of expression, conditioning on the gene being expressed, $G_{ij}|Z_{ij} = 1$, are modeled with a logistic regression and Gaussian linear model, respectively,

$$\begin{aligned}
&\text{logit}(\mathbb{P}(Z_{ij} = 1|X = \boldsymbol{x}_i)) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j^D, \\
&G_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j^C + \epsilon_{ij},\ \epsilon_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_j^2) \text{ and } Z_{ij} = 1.
\end{aligned}$$

Here, X are covariates, which can but do not necessarily need to be the group of k covariates in Section 2.3, and $\boldsymbol{x}_i^\top$ is the i-th row of $\boldsymbol{X}$. The coefficients of the logistic model are regularized using a Bayesian approach and the heterogeneous gene-specific variances of the linear model are shrinked to a global estimate of the variance using an empirical Bayes method. Testing for differential expression is carried out using the likelihood ratio test for each gene.

LRT was initially considered when microarray gene expression assays were developed [Shevade and Keerthi, 2003; Xing et al., 2001], and attracted further attention after larger sample sizes become available due to scRNA-seq techniques [Ntranos et al., 2019]. In this work, we will consider a multivariate logistic regression, where the p-values are calculated via likelihood ratio tests. We use the the R package `Seurat` (version `4.0.4`) [Satija et al., 2015] developed and maintained by the Satija lab to implement all tests.

The knockoff statistics for our variable selection is defined as $W_j = p'_j - \tilde{p}'_j$ where $p'_j = -\log(p_j)$ and $\tilde{p}'_j = -\log(\tilde{p}_j)$. p_j and $\tilde{p}_j$ are p-values calculated on the log-normalized data $\boldsymbol{G}^{\text{obs}}$ and the rescaled knockoffs $\widetilde{\boldsymbol{G}}^{\text{obs}}$ respectively. Note that if a variable is independent from the model, its corresponding p-values for original and knockoff counterpart follow the same distribution such that the knockoff statistic can be either positive or negative with equal probability. Furthermore, Bonferroni corrected p-values will also be considered. In fact, to ensure simpler presentation, only corrected p-values are included in the main text. Results related to non-corrected p-values are presented in Appendix A.5.

For single knockoff, all genes whose knockoff statistics $W_j \geq \tau$ are selected, where

$$\tau = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\},$$

such that the FDP is expected to be controlled below q.

For multiple knockoffs, we apply the modified version described by He et al. [2021]. The corresponding p-values for each of $\boldsymbol{G}^0, \boldsymbol{G}^1, \ldots,$ and $\boldsymbol{G}^M$ are now denoted as p_j^m and $p_j'^m = -\log(p_j^m)$ for $j = 1, \ldots, p$ and $m = 0, 1, \ldots, M$, where $\boldsymbol{G}^0 = \boldsymbol{G}^{\text{obs}}$, and $\boldsymbol{G}^m$ denote the rescaled multi-LR or multi-decomp knockoffs. The knockoff statistic for each feature $1 \leq j \leq p$ now is redefined as

$$W_j = \tau_j \mathbb{1}_{\{\kappa_j = 0\}},$$

where $\tau_j = p_j'^{(0)} - \underset{1 \leq m \leq M}{\text{median}}\, p_j'^{(m)}$ is the difference between the largest transformed p-value and the median of the remaining ones and $\kappa_j = \underset{0 \leq m \leq M}{\arg\max}\, p_j'^m$. The threshold for deciding whether the null hypothesis should be rejected is then defined as

$$\tau = \min\left\{t > 0 : \frac{\frac{1}{M} + \frac{1}{M}\#\{j : \kappa_j \geq 1, \tau_j \geq t\}}{\#\{j : \kappa_j = 0, \tau_j \geq t\}}\right\}$$

such that $\widehat{S} = \{j : W_j \geq \tau\}$.

2.5.1 e-BH procedure

For the e-BH procedure, we generate M copies of knockoffs $\widetilde{\boldsymbol{G}}^{\text{obs},1}, \ldots, \widetilde{\boldsymbol{G}}^{\text{obs},M}$ and calculate their respective importance statistics $\left\{W_j^m\right\}_{1 \le j \le p}$. Here, M does not necessarily need to be the same as in Section 2.4.4 and it should be clear depending on the context. Then, we can calculate their e-values via

$$e_j^m = p \times \frac{\mathbb{1}_{\{W_j^m \ge \tau^m\}}}{1 + \#\{j : W_j \le -\tau^m\}},$$

where

$$\tau^m = \min\left\{t > 0 : \frac{1 + \#\{j : W_j^m \le -t\}}{\#\{j : W_j^m \ge t\}} \le q_{\text{kn}}\right\}.$$

For derandomization, take the average of the M set of e-values $e_j^{\text{avg}} = \sum_{m=1}^{M} e_j^m / M$. Finally, select all variables with an average e-value greater than or equal to $e_{[\hat{k}]}^{\text{avg}}$, where $\hat{k} = \max_{1 \le j \le p}\{j : e_{[j]}^{\text{avg}} \ge p/(q_{\text{ebh}} j)\}$ and $e_{[j]}^{\text{avg}}$ is the j-th ordered average e-value, from the *largest to the smallest*, such that $\widehat{S} = \{j : e_j^{\text{avg}} \ge e_{[\hat{k}]}^{\text{avg}}\}$.

It should be noted that the knockoffs mentioned at the beginning of the subsection can be *any* valid knockoffs. In this chapter, we only consider decomp and LR knockoffs and we denote the e-BH procedures using the respective knockoffs as e-decomp and e-LR. Furthermore, notice that there are two different qs: q_{kn} and q_{ebh}. Here, q_{ebh} is used to control the false discovery rate, whereas q_{kn} is in fact a parameter that need to be tuned. As discussed in Ren and Barber [2022], it is preferred to choose $q_{\text{kn}} < q_{\text{ebh}}$ when $M > 1$, and we will use the suggested setting $q_{\text{kn}} = q_{\text{ebh}}/2$.

Why do we introduce both multiple knockffs *and* e-BH for derandomization? After all, the e-BH procedure does not require us to simultaneously generate multiple knockoff and it is more flexible to work with. There is, however, one issue: The e-values we calculate are based on single knockoffs, and at least $\lceil \frac{1}{q_{\text{kn}}} \rceil$ signal variables are required in order to select any variables. Compared to the multiple knockoff procedure, the fact that $q_{\text{kn}} < q_{\text{ebh}} = q$ will only exacerbate this shortcoming. Hence there is a tradeoff between the low detection threshold of multiple knockffs and the flexibilty of e-BH.

2.5.2 Q-values for knockoffs and e-BH

Another important measure of statistical significance regarding FDR would be q-values [Storey and Tibshirani, 2003]. In this section, we will define the q-values under the context of model-X

knockoff framework and e-BH procedure, beginning with the definition of the q-value for usual FDR control based on ordered p-values:

$$q\text{-value} = \min_{t \geq p\text{-val}} \widehat{\text{FDR}}(t),$$

where p-val is the p-value of the hypothesis under consideration and $\widehat{\text{FDR}}(t)$ is the estimated FDR if we are to reject all tests with p-value less than t. According to He et al. [2021], the q-value for a knockoff statistics $W_j > 0$ is defined as

$$\text{knockoff-}q_j = \min_{t \leq W_j} \frac{1 + \#\{\ell : W_\ell \leq -t\}}{\#\{\ell : W_\ell \geq t\}},$$

where $\frac{1+\#\{\ell:W_\ell \leq -t\}}{\#\{\ell:W_\ell \geq t\}}$ is an estimate of the proportion of false discoveries if we are to select all genes with knockoff statistics greater than $t > 0$. The q-value for a gene with knockoff statistics $W_j \leq 0$ is defined as knockoff-$q_j = 1$ as the gene will never be selected.

Similarly, the q-values for multiple knockoffs can be defined as

$$\text{multi-knockoff-}q_j = \min_{t \leq \tau_j} \frac{\frac{1}{M} + \frac{1}{M}\#\{\ell : \kappa_\ell \geq 1, \tau_\ell \geq t\}}{\#\{\ell : \kappa_\ell = 0, \tau_\ell \geq t\}}.$$

The q-value for a gene where $\kappa_j \neq 0$ is defined as multi-knockoff-$q_j = 1$.

Calculating q-values for the e-BH procedure is more difficult compared to the ones for (multiple) knockoffs, which have a straightforward equation, since the knockoff e-values depend on the target FDR q. Of course, we can still estimate these q-values by adhering to the definition and calculate them in intervals $\left(\frac{k-1}{N}, \frac{k}{N}\right]$, where $k = 1, \ldots, N$. At any given $q = \frac{k}{N}$, the variable is then assigned with the smallest possible value of $\frac{k}{N}$ it remains to be selected.

2.6 Simulations

In this section, we compare the performance of the various knockoff constructions under different scenarios. We begin with a brief discussion on the choice for λ. Then, in Section 2.6.1 we discuss the necessity of imputation. In Section 2.6.2, we compare the performance of different knockoff constructions against the construction introduced in Candès et al. [2018] based on synthetic signals. Furthermore, as a baseline, we also compare knockoff variable selection with the

Benjamini-Hochberg (BH) procedure [Benjamini and Hochberg, 1995], which is a canonical approach to FDR control. Assume there are l hypotheses under consideration. The procedure is a step-up one with ordered p-values $p_{(1)}, \ldots, p_{(l)}$ as input. To control FDR at level q, let

$$\hat{k} = \max_{1 \le j \le l} \{j : p_{(j)} \le qj/l\},$$

then all hypotheses with p-values no larger than $p_{(\hat{k})}$ are rejected.

In the imputation step, there is an important tuning parameter λ (see Algorithm 1 in Section 2.3), which can be viewed as the tuning parameter for a ridge regression when we iteratively update A and B. In ridge regression, increasing the value of λ would lead to a stronger shrinkage effect in the coefficients estimated. Likewise, increasing the value of λ would decrease the rank of the solution obtained by applying Algorithm 1. However, unlike cross-validation in ridge regression, there is no sound way to pick a good λ. As described in Hastie et al. [2015], one could begin with $\lambda_{\max}$, corresponding to the largest singular value of $\boldsymbol{G}^{\text{obs}}$, and select any value in $[0, \lambda_{\max})$. Intuitively the choice of λ reflects a bias-variance tradeoff: by decreasing λ, the accuracy of the matrix recovery would increase as more information is used, until at some point, the algorithm will "over-recover". Therefore, it is suggested to use a non-zero but small λ to reach the full potential of the variable selection process.

2.6.1 Benefits of imputation

We benefit from the imputation mainly from the following two perspectives. First, due to the nature of scRNA-seq data, there are a lot of missing values, which would lead to an unstable estimation of the covariance structure and consequently knockoff variables of low quality. We compare the difference of covariance estimation with and without imputation based on the dataset from scREAD, a scRNA-seq database [Jiang et al., 2020]. Specifically, we use the dataset sampled from the superior parietal lobe region [Alsema et al., 2020]. The full dataset includes 15,141 microglia transcriptomes and 16,767 genes, though we focus on the first 2000 genes for simplicity. In sc-softImpute, centered CDR and gender, and 20 leading principal components of the centered $\boldsymbol{G}^{\text{obs}}$ are used to initialize $\boldsymbol{X}$ and $\boldsymbol{A}_0'$ respectively, and $\lambda = 25.2$ is used in the imputation algorithm. Genes with less than or equal to 23 expressions as well as genes with no variation in gender are excluded since they will fail to provide estimations for the covariance in the case where no imputation is involved. Therefore, a subset of 1949 genes out of the first 2000 genes, and 15,141 observations

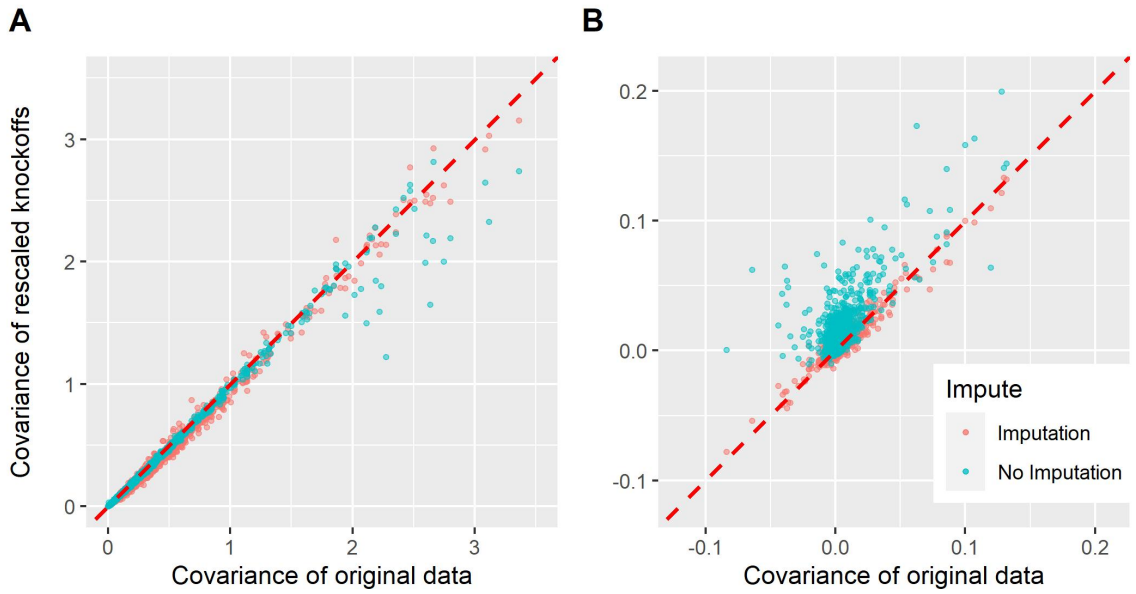

Figure 2.1: Comparison of knockoff covariance and original covariance. In the left panel (A) are the estimated variances of the rescaled decomp-knockoff variables against the estimated variances of the original variables. In the right panel (B) are 5000 randomly sampled values from each of the upper triangles of the estimated covariance matrices. The randomly sampled indices are shared between comparison.

are used for the estimation, and the same subset is used for the estimation without imputation. As shown in Figure 2.1, we first estimated the covariance with imputation (denoted as $\widehat{\boldsymbol{\Sigma}}_G$) and without imputation (denoted as $\widehat{\boldsymbol{\Sigma}}_{G^{\text{obs}}}$), then generated decomp-knockoff variables by using the two estimated covariance matrices respectively. Since no imputation was involved in the latter case, there are no estimators for the $\boldsymbol{A}$ and $\boldsymbol{B}$ as in equation (2.13) readily available. Therefore, $\boldsymbol{A}$ is constructed using CDR, gender, and 20 leading principal components of $\boldsymbol{G}^{\text{obs}}$, then it is centered and scaled, and rows of $\boldsymbol{B}$ are estimated by using the coefficients of linear regression with the expressed parts of each gene as response according to equation (2.12). Both sets of knockoffs are rescaled accordingly and denoted as 'Imputation' (colored in orange) and 'No Imputation' (colored

in cyan) in the figure, respectively. In the left panel, we plot the estimated variances of the rescaled knockoff variables against the estimated variances of the original variables. On the right, we plot 5000 randomly sampled off-diagonal elements of the covariance matrix of the rescaled knockoff variables against the corresponding covariance of the original variables. From the figure, we can see that the covariance of the knockoffs generated using the $\widehat{\boldsymbol{\Sigma}}_{G^{\text{obs}}}$ tend to be further away from the diagonal line than its counterpart, which suggests that imputing the missing data would enable us to generate knockoff variables better preserving the exchangeability condition. Moreover, carrying out the imputation process allows us for different knockoff constructions such as the LR construction and the multi-LR construction, both of which come with relatively no additional computational cost, and could serve as alternatives from decomp knockoffs when computing budgets are tight. Of course, in the case of multi-LR knockoffs, since $M + 1$ sets of p-values will be computed, the cost of calculating knockoff statistics will be higher.

2.6.2 Comparison of knockoff constructions under synthetic signals

In this subsection, we conduct a range of simulations to test the power and FDR control of the proposed knockoff constructions (asdp, decomp, LR, multi-decomp and multi-LR) and e-BH procedures (e-decomp and e-LR) based on the same dataset described in Section 2.6.1. First, we will lay out the details of the simulation setting.

Imputation. In practice, the choice of covariates $\boldsymbol{X}$ used in imputation would largely depend on domain knowledge and the availability of data. Given our data set, centered CDR and gender, and 20 leading principal components of the centered $\boldsymbol{G}^{\text{obs}}$ are used to initialize $\boldsymbol{X}$ and $\boldsymbol{A}_0'$ respectively. Genes with less than or equal to $r + 1$ ($r = 22$) expressions are excluded since they will fail to provide estimations for σ_j^2. This will leave us with a subset of 1951 genes out of the set of the first 2000 genes, and 15,141 observations.

Knockoff construction. Based on the imputation results, construction of decomp knockoffs and asdp knockoffs are implemented using the R package `knockoff` (version `0.3.3`) under R version `4.1.0` with the reconstructed covariance as described in Section 2.4.2. We will set $\boldsymbol{s}^{\text{decomp}} = (1.95D_{jj})_{1\le j\le p}$ instead of $(2D_{jj})_{1\le j\le p}$ to generate decomp knockoffs. Because in practice, keeping $2\boldsymbol{\Sigma}_G - \text{diag}\{\boldsymbol{s}^{\text{decomp}}\}$ in equation (2.10) strictly positive definite would allow us to avoid possible numerical errors without compromising the power. Similarly, $\boldsymbol{s}^{\text{multi-decomp}} = (1.15D_{jj})_{1\le j\le p}$ in-

stead of $(1.2D_{jj})_{1\leq j\leq p}$ when $M = 5$ sets of knockoffs are generated for multi-decomp knockoffs. Both LR knockoffs and multi-LR knockoffs are sampled as described in Section 2.4.3 and 2.4.4. $M = 5$ sets of knockoffs are generated for multi-LR knockoffs. Each of the constructions is only sampled once.

Signal generation. After centering and scaling $\boldsymbol{G}^{\text{obs}}$, we randomly generate 50 signals in each repetition by repeatedly sampling from a Gaussian distribution $\mathcal{N}(0, \text{sgn}^2 \times \frac{2\log(p)}{n})$ and only keeping those with absolute values greater than $\text{sgn} \times \sqrt{\frac{2\log(p)}{n}}$. sgn is used to indicate the signal strength and is set to be sgn = 3. Signal locations are randomly chosen. We use a logit link $g(x) = \log(\frac{x}{1-x})$ and

$$g(\mu_i) = \beta_1 G_1 + \ldots \beta_p G_p, \text{ where } \mu_i = \mathbb{P}(Y = 1|G),$$

to generate the response.

Variable selection. WRT, MAST, and LRT are used to calculate the Bonferroni corrected knockoff statistics. We incorporate CDR and gender as covariates in MAST and LRT. We consider target FDRs $q = 0.1, 0.05$ and 0.01 for comparison. M = 5 copies of knockoffs are generated for both e-decomp and e-LR.

In Figure 2.2, the results of 20 repetitions of different knockoff constructions under two FDR targets $q = 0.1$ and $q = 0.01$ are presented. The first thing one might notice in the figure is how knockoff statistics that are constructed with LRT p-values yield clearly superior results than knockoff statistics constructed using MAST or WRT p-values, achieving more powerful variable selection under controlled FDR. This is unsurprising since the signals are generated via the logistic model in the simulations, naturally making LRT the most effective at identifying these signals. Following LRT, it could be observed that knockoff statistics constructed using MAST yield more powerful results than the non-parametric counterpart. Overall, while we do have some flexibility in terms of the choice of knockoff statistics under the knockoff framework, the way it is constructed matters a lot for powerful variable selection.

Aside from the difference between knockoff statistics, it could be observed that decomp knockoffs largely perform the best among all proposed knockoff constructions when $q = 0.1$. As we have mentioned in Section 2.4.2, asdp knockoffs are not necessarily more powerful than decomp knockoffs. Based on our simulations, it might even be preferable to opt for decomp knockoffs over asdp knockoffs in practice for their ease in construction and likely higher power. LR knockoffs have a slightly worse power under LRT and WRT compared to decomp knockoffs, but they perform

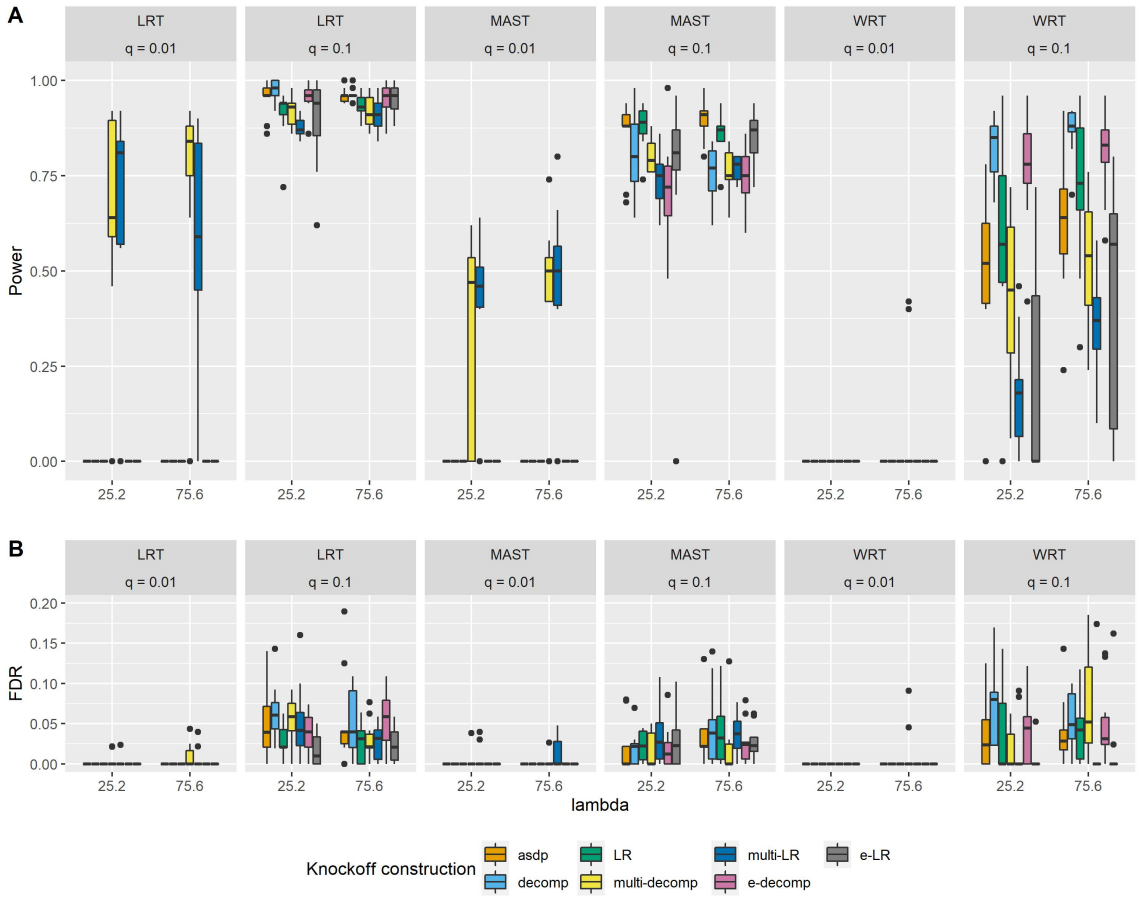

Figure 2.2: Comparison of knockoff constructions and BH procedure under q = 0.1. Panel A and B are box-plots of the FDP and power over 20 repetitions respectively. Simulations are carried out according to the details described in Section 2.6.2 using 1951 genes and 15,141 observations, sgn = 3 and $\lambda_{\max} = 252.07$.

similarly to asdp knockoffs and the FDR is well controlled. The loss in power for LR knockoffs is compensated by the fact that they are the fastest to construct. For multi-decomp and multi-LR knockoffs, the worse power in comparison is expected as Gimenez and Zou [2019] have pointed out that "sampling multi-knockoffs impose a more stringent constraint to construct the knockoff

Table 2.2: Comparison of knockoff constructions and BH procedure under q = 0.05. This table shows the FDR and power of simulations where sgn = 3, target FDR $q = 0.05$ and Bonferroni corrected p-values are used to calculate knockoff statistics. The average FDP and power over 20 simulations for each setting are shown in each column with the standard deviation in the parentheses. The BH procedure uses uncorrected p-values.

		LRT		MAST		WRT	
Knockoff Construction	λ	FDR	Power	FDR	Power	FDR	Power
asdp	0.0	0.03 (0.04)	0.89 (0.08)	0.01 (0.03)	0.59 (0.32)	0.05 (0.11)	0.33 (0.28)
asdp	25.2	0.03 (0.03)	0.87 (0.14)	0.02 (0.04)	0.72 (0.22)	0.01 (0.02)	0.16 (0.27)
asdp	50.4	0.06 (0.04)	0.93 (0.05)	0.03 (0.05)	0.73 (0.15)	0.08 (0.11)	0.54 (0.19)
asdp	75.6	0.04 (0.04)	0.94 (0.05)	0.05 (0.07)	0.81 (0.12)	0.05 (0.08)	0.38 (0.28)
decomp	0.0	0.04 (0.03)	0.96 (0.03)	0.03 (0.04)	0.71 (0.22)	0.09 (0.14)	0.74 (0.27)
decomp	25.2	0.05 (0.03)	0.95 (0.04)	0.02 (0.05)	0.65 (0.20)	0.05 (0.09)	0.62 (0.34)
decomp	50.4	0.05 (0.05)	0.96 (0.03)	0.02 (0.04)	0.57 (0.29)	0.12 (0.13)	0.69 (0.23)
decomp	75.6	0.05 (0.06)	0.95 (0.03)	0.02 (0.04)	0.55 (0.17)	0.07 (0.10)	0.68 (0.26)
LR	0.0	0.03 (0.02)	0.87 (0.08)	0.02 (0.03)	0.76 (0.12)	0.00 (0.01)	0.11 (0.20)
LR	25.2	0.03 (0.02)	0.90 (0.07)	0.02 (0.02)	0.78 (0.20)	0.01 (0.03)	0.38 (0.28)
LR	50.4	0.03 (0.03)	0.94 (0.08)	0.05 (0.05)	0.72 (0.21)	0.01 (0.03)	0.29 (0.29)
LR	75.6	0.03 (0.03)	0.91 (0.07)	0.04 (0.05)	0.78 (0.14)	0.02 (0.02)	0.45 (0.34)
multi-decomp	0.0	0.03 (0.03)	0.88 (0.05)	0.02 (0.02)	0.74 (0.07)	0.00 (0.02)	0.06 (0.09)
multi-decomp	25.2	0.04 (0.03)	0.92 (0.04)	0.02 (0.02)	0.76 (0.07)	0.01 (0.03)	0.20 (0.17)
multi-decomp	50.4	0.02 (0.02)	0.92 (0.05)	0.03 (0.04)	0.74 (0.09)	0.02 (0.04)	0.32 (0.16)
multi-decomp	75.6	0.03 (0.02)	0.92 (0.03)	0.01 (0.02)	0.76 (0.09)	0.07 (0.10)	0.35 (0.20)
multi-LR	0.0	0.02 (0.02)	0.83 (0.06)	0.02 (0.02)	0.69 (0.09)	0.02 (0.07)	0.04 (0.09)
multi-LR	25.2	0.03 (0.02)	0.86 (0.07)	0.02 (0.03)	0.71 (0.11)	0.01 (0.06)	0.07 (0.09)
multi-LR	50.4	0.04 (0.02)	0.90 (0.06)	0.01 (0.02)	0.76 (0.11)	0.00 (0.02)	0.12 (0.13)
multi-LR	75.6	0.03 (0.03)	0.91 (0.05)	0.05 (0.06)	0.74 (0.06)	0.01 (0.02)	0.17 (0.13)
e-decomp	0.0	0.03 (0.03)	0.89 (0.21)	0.00 (0.01)	0.12 (0.30)	0.02 (0.07)	0.16 (0.33)
e-decomp	25.2	0.04 (0.05)	0.88 (0.21)	0.00 (0.00)	0.04 (0.17)	0.02 (0.05)	0.36 (0.41)
e-decomp	50.4	0.03 (0.02)	0.94 (0.05)	0.00 (0.02)	0.03 (0.14)	0.04 (0.07)	0.44 (0.41)
e-decomp	75.6	0.03 (0.02)	0.94 (0.05)	0.00 (0.00)	0.04 (0.17)	0.06 (0.08)	0.48 (0.41)
e-LR	0.0	0.01 (0.02)	0.39 (0.44)	0.00 (0.01)	0.04 (0.17)	0.00 (0.00)	0.00 (0.00)
e-LR	25.2	0.01 (0.02)	0.59 (0.45)	0.01 (0.02)	0.19 (0.33)	0.00 (0.00)	0.00 (0.00)
e-LR	50.4	0.02 (0.02)	0.84 (0.29)	0.00 (0.01)	0.29 (0.40)	0.00 (0.00)	0.00 (0.00)
e-LR	75.6	0.02 (0.02)	0.92 (0.06)	0.01 (0.02)	0.15 (0.31)	0.00 (0.00)	0.00 (0.00)
asdp_cov		0.02 (0.03)	0.77 (0.22)	0.02 (0.03)	0.57 (0.27)	0.01 (0.02)	0.25 (0.38)
BH		0.13 (0.12)	0.99 (0.01)	0.12 (0.11)	0.96 (0.03)	0.29 (0.21)	0.94 (0.03)

conditional distribution", and multiple knockoffs do not guarantee higher power. However, it is an acceptable tradeoff for the low detection threshold $\lceil \frac{1}{qM} \rceil$ of multi-knockoffs. Notice that only when using multi-decomp and multi-LR knockoffs, it is possible to select meaningful variables under stringent target FDR since other single knockoffs methods require at least 100 signal variables to reach the expected detection threshold whereas multi-knockoffs require only 20 signals. This property will be useful if it is expected that the signal in the data of interest is sparse. The improvement in stability of multi-knockoffs, in the sense that the same signals are consistently identified by different multi-knockoffs, cannot be analyzed under our current setting, but it is discussed in detail by Gimenez and Zou [2019]. For the e-BH results, we can see that both e-decomp and e-LR perform similarly compared to their respective single knockoff counterparts both in terms of FDR control and power, if λ and the knockoff statistics are chosen properly. However, when we use WRT to calculate knockoff statistics, variable selection results using e-LR have a large variance. This is because derandomized knockoffs tend to either almost always select or almost never select signal variables. In this specific case, many of the individual simulations simply did not select any of the signal variables. The same tendency will also occur in multi-knockoffs and e-decomp results, though it is not reflected in the figure.

In Table 2.2, the knockoff statistics and constructions remain the same as the ones used in Figure 2.2 but variables are selected under a different target FDR level $q = 0.05$. The observations are more or less parallel to the ones made in Figure 2.2. To add value to the table, we include simulation results of the BH procedure and compare our proposed methods with the standard knockoff generator by Candès et al. [2018]. A set of knockoffs is constructed using the observed and centered log-normalized gene expression $\boldsymbol{G}^{\text{obs}}$, meaning that instead of using the decomposed covariance as in equation (2.13), the empirical covariance is estimated using *all* data. Knockoffs are generated by solving an ASDP and sampling from a conditional multivariate Gaussian distribution, which are then rescaled accordingly. The results are labeled as *asdp_cov* knockoffs. It should be noted that the empirical covariance is estimated using a shrinkage covariance estimator introduced in Schäfer and Strimmer [2005] due to the sparsity of scRNA-seq data, which is arguably an improved but more complex covariance estimator. Compared to the results for asdp_cov knockoffs and BH procedure, the proposed variable selection methods are clearly superior: asdp_cov knockoffs demonstrate a far lower power with larger standard deviation, mostly because it occurs that no variables are selected over many of the repetitions, whereas the BH procedure, while being more powerful, fails at con-

trolling the FDR below the target $q = 0.05$, not even with LRT p-values. Furthermore, in e-decomp and e-LR, high variances in terms of power that are similar to the case observed in Figure 2.2 are no longer limited to WRT-based knockoffs statistics. This is because now $q_{\text{kn}} = 0.025$ when calculating the e-values with each of the single knockoff copies, and it becomes harder to consistently detect the signal variables. Combine the small target FDR with less powerful knockoffs and knockoff statistics, many of the repetitions will suffer from low power. The only cases that perform well are those in which knockoffs statistics are calculated based on LRT.

2.6.3 The knockoff filter is robust to confounding effects

In this subsection, we will use simulation studies to illustrate how our approach can make adjustments to potential confounders, especially the batch effect. Since scRNA-seq data is collected in different batches, batch effects can be a potential source of variation and experimental noise that scientists are widely aware of [Leek et al., 2010; Stegle et al., 2015; Tung et al., 2017]. For reference, we also compare our method with the BH procedure. Our simulation is based on a real dataset from Yang et al. [2022]. The full dataset includes 143,793 single-nucleus transcriptomes and 23,537 genes. We will work on a subset with 12,193 samples that are collected from hippocampus tissues and astrocyte cell type. In the following, we will first lay out the details for the simulation setting that is different from the setting in Section 2.6.2.

Imputation. Centered gender, and 20 leading principal components of the centered $\boldsymbol{G}^{\text{obs}}$ are used to initialize $\boldsymbol{X}$ and $\boldsymbol{A}_0'$ respectively. On top of gender and principal components, one of the variations of batch and CDR, respectively, are included as covariates $\boldsymbol{X}$:

- batch, permuted batch or no batch, and
- CDR or no CDR.

After excluding genes with less than or equal to 26 expressions, which is the maximum value of $r + 1$ ($r = 25$) out of all possible combinations, the same subset of 2275 genes out of the set of the first 3000 genes and 12,193 observations is used for all simulation variations.

Knockoff construction. Only decomp knockoffs are presented in this subsection since it is still the best performing knockoff construction. Including other knockoff constructions does not provide additional insight on top of what we have already learned in Section 2.6.2.

Signal generation. As in Section 2.6.2, 50 signals are generated via the logistic model and sgn = 3. Except that the signal variables are controlled, meaning that for each repetition, the signal locations are the same across all variations, and batch is included in the signal generation. Namely,

$$g(\mu_i) = \beta_1 G_1 + \ldots \beta_p G_p + 2 \times \text{sgn} \times \sqrt{\frac{2\log(p)}{n}} \times \text{batch},$$

where batch are the batch labels in numeric values.

Variable selection. LRT and MAST results are considered for BH procedure and only LRT p-values, which are Bonferroni corrected, are considered for the knockoff filter. To calculate LRT p-values, gender, age, and the same variation of batch and CDR are incorporated as covariates of the logistic regression model. For the knockoff filter, we also consider the comparison for whether or not CDR was incorporated in the imputation step. The target FDR is set as $q = 0.1$.

In Table 2.3, we provide simulation results for the BH procedure. It can be clearly seen that the FDR control largely depends on which set of covariates is *explicitly* included in the model. In the absence of CDR, the FDR can be out of control even when we make adjustments with respect to the batch effect. When batch labels are permuted or removed from the test, we observe either a small further inflation in FDR when LRT is used, or a larger inflation in the case of MAST. When CDR is included as a covariate in the test, however, such inflations can not be observed. This suggests that CDR might be another confounding variable, and can partly explain the variation of batch effect. In practice, batch effect can usually be easily identified as a potential confounder in scRNA-seq experiments, while it is more tricky to identify CDR. In summary, the FDR control by BH procedure is very sensitive to the choice of covariates (potential confounders). If a confounder is unobserved or is not appropriately incorporated, the FDR can be uncontrolled.

In Table 2.4, we summarize the simulation results for the proposed method. Based on the results, our method is quite robust to potential confounders: no matter whether batch effect or CDR is included in the imputation and calculation of LRT p-values, the FDR is under control in all cases. We also observe that when batch labels are permuted or removed, their effect on the variable selection is rather minimal when λ is chosen properly. It is still beneficial to include CDR when calculating p-values, as we can see it generally leads to a higher power.

A possible explanation for these observations is that the latent factors $\boldsymbol{A}$ can explain the variations of both batch effect and CDR. For our approach, by recovering latent factors and generating knockoff variables preserving the exchangeability condition, we can implicitly make adjustments

Table 2.3: BH procedure and confounding effects. This table shows the FDR and power of simulations where sgn = 3, target FDR $q = 0.1$ and non-corrected p-values based on LRT and MAST are used to for BH procedure. The average FDP and Power over 20 simulations for each setting are shown in each column with the standard deviation in the parentheses.

		no CDR		with CDR	
Test statistics	batch	FDR	Power	FDR	Power
LRT	original	0.14 (0.07)	0.99 (0.01)	0.08 (0.03)	0.99 (0.01)
	permute	0.17 (0.10)	0.99 (0.01)	0.09 (0.04)	0.99 (0.01)
	remove	0.18 (0.12)	0.99 (0.01)	0.10 (0.03)	0.99 (0.01)
MAST	original	0.22 (0.27)	0.97 (0.03)	0.03 (0.02)	0.96 (0.03)
	permute	0.31 (0.32)	0.97 (0.02)	0.02 (0.02)	0.96 (0.03)
	remove	0.27 (0.29)	0.98 (0.02)	0.03 (0.02)	0.96 (0.03)

with respect to batch effect and CDR. This also suggests that, if a confounder is unobserved (such as when CDR is not successfully identified), but the variation of which can be explained by the latent factors, our method can still adjust for such a confounder without explicitly including it in the model. However, when latent factors cannot fully explain the variation of an unobserved confounder, we would still fail to make the adjustment.

2.7 Application to scRNA-seq data

In this section, we apply the proposed methods to the scRNA-seq data set previously studied in Section 2.6.3, but we are now focusing on observed outcomes (Alzheimer's disease and no cognitive impairment) instead of synthetical ones. We will identify DEGs based on date combined from hippocampus and superior frontal cortex, cell-type by cell-type. Specifically, we will carry out our analysis separately for astrocytes ($n = 22695,\ p = 11933$), oligodendrocytes ($n = 34774,\ p = 11769$), pericytes ($n = 27195,\ p = 11250$), microglias ($n = 3373,\ p = 6565$) and neurons ($n = 2941,\ p = 8377$), under the knockoff framework. We begin with a description of the details of our approach:

Table 2.4: Knockoff procedure and confounding effects. This table shows the FDR and power of simulations where sgn = 3, target FDR $q = 0.1$ and Bonferroni corrected p-values are used to calculate knockoff statistics. Only decomp knockoffs and LRT based knockoff statistics are included. The average FDP and power over 20 simulations for each setting are shown in each column with the standard deviation in the parentheses.

Imputation		no CDR		no CDR		with CDR	
LRT		no CDR		with CDR		with CDR	
batch	λ	FDR	Power	FDR	Power	FDR	Power
original	0.00	0.00 (0.00)	0.18 (0.21)	0.00 (0.00)	0.19 (0.22)	0.00 (0.00)	0.17 (0.22)
	22.2	0.00 (0.01)	0.47 (0.29)	0.00 (0.01)	0.57 (0.22)	0.00 (0.01)	0.57 (0.23)
	44.4	0.00 (0.01)	0.57 (0.30)	0.01 (0.01)	0.67 (0.19)	0.01 (0.01)	0.67 (0.19)
	66.6	0.01 (0.01)	0.62 (0.28)	0.01 (0.01)	0.74 (0.15)	0.01 (0.01)	0.74 (0.15)
permute	0.00	0.00 (0.02)	0.20 (0.25)	0.00 (0.00)	0.23 (0.24)	0.00 (0.00)	0.23 (0.26)
	22.2	0.01 (0.02)	0.48 (0.26)	0.00 (0.01)	0.59 (0.21)	0.00 (0.01)	0.59 (0.23)
	44.4	0.00 (0.01)	0.61 (0.27)	0.00 (0.00)	0.71 (0.17)	0.00 (0.00)	0.71 (0.17)
	66.6	0.00 (0.01)	0.65 (0.24)	0.00 (0.01)	0.79 (0.12)	0.00 (0.01)	0.79 (0.12)
remove	0.00	0.00 (0.00)	0.17 (0.20)	0.00 (0.00)	0.19 (0.21)	0.00 (0.00)	0.19 (0.23)
	22.2	0.00 (0.01)	0.46 (0.30)	0.00 (0.01)	0.53 (0.27)	0.00 (0.01)	0.53 (0.27)
	44.4	0.00 (0.01)	0.56 (0.28)	0.00 (0.01)	0.66 (0.21)	0.00 (0.01)	0.66 (0.21)
	66.6	0.00 (0.01)	0.62 (0.26)	0.01 (0.01)	0.74 (0.17)	0.01 (0.01)	0.74 (0.17)

Imputation. Gender, batch, and CDR are used as covariates for imputation. Additionally, 115 leading principal components are used as an initialization for astrocytes, oligodendrocytes, and pericytes, while 50 leading principal components are used for microglias and neurons. The number of principal components is chosen to be approximately equal to 10% of the dimension, which is understood as the number of genes after excluding those ones with too few number of observed expressions (less than number of covariates + 1).

Knockoff construction. For derandomization purposes, we consider constructing multiple knock-

Table 2.5: Number of DEGs selected under the knockoff framework. This table shows the number of DEGs selected under the knockoff framework, using multiple knockoffs or e-BH procedure. Respectively, λ's are set to be 77.4, 19.4, 34.0, 90.5 and 58.5 for each of the cell types, from the left to the right, which equal to $0.1\lambda_{\max}$. The target FDR is set at $q = 0.1$.

Method	Test	Astrocyte	Microglia	Neuron	Oligo-dendrocyte	Pericyte
multi-decomp	LRT	502	38	22	1	3
	MAST	531	37	15	2	2
	WRT	398	16	12	0	3
multi-LR	LRT	518	28	19	1	3
	MAST	555	25	23	2	2
	WRT	333	19	26	0	2
e-decomp	LRT	135	92	0	0	0
	MAST	92	0	0	0	0
	WRT	0	0	21	0	0
e-LR	LRT	63	0	0	0	0
	MAST	449	26	0	0	0
	WRT	0	0	0	0	0

offs. Namely, we construct both multi-decomp and multi-LR knockoffs with $M = 10$. $\boldsymbol{s}^{\text{multi-decomp}}$ in (2.16) is set to be $(D_{jj})_{1\le j\le p}$ to avoid potential numerical errors.

Variable selection. We calculate Bonferroni corrected p-values for WRT, MAST, and LRT, and calculate knockoff statistics based on them. In MAST and LRT, we include gender, age, batch, and CDR as covariates. The target FDR is $q = 0.1$. We consider all three tests as it is unknown which one can better capture the association in real data. For e-BH procedure, we generate M = 10 groups of single knockoffs for both decomp and LR constructions. We will compare our methods to the BH procedure. For the latter one, in order to reduce the number of selections, it is standard in practice to filter genes by log fold change (logFC) before applying the BH procedure, which is calculated as the logarithm of the average *non*-normalized expression between the two groups.

In Table 2.5, we present the number of genes selected by using different knockoff constructions in combination with different knockoff statistics. Table 2.6 shows the number of genes selected using

Table 2.6: Number of DEGs selected using BH procedure. This table shows the number of DEGs selected using BH procedure, after screening for genes with logFC $>$ lfc where lfc $= 0, 0.05, 0.1, 0.25, 0.5$. The target FDR is set at 0.1 and 0.05.

CellID	Sample Size	logFC	Number of genes	LRT		MAST		WRT	
				0.1	0.05	0.1	0.05	0.1	0.05
Astrocyte	22695	0	11933	5006	4284	4642	4009	5685	4941
		0.05	3837	3682	3576	3529	3356	3683	3586
		0.1	1609	1597	1594	1599	1594	1604	1601
		0.25	227	226	226	227	227	227	227
		0.5	29	29	29	29	29	29	29
Microglia	3373	0	6565	881	644	569	474	1395	1029
		0.05	4236	1062	770	677	516	1672	1223
		0.1	2498	1204	924	795	591	1630	1331
		0.25	445	430	418	397	371	435	425
		0.5	49	49	49	49	49	49	49
Neuron	2941	0	8377	653	467	402	320	1468	1098
		0.05	4942	833	590	481	364	1794	1335
		0.1	2698	1037	706	573	436	1883	1514
		0.25	380	332	320	308	282	377	375
		0.5	29	27	27	27	27	29	29
Oligo-dendro-cyte	34773	0	11769	4138	3432	4161	3445	8834	8168
		0.05	3313	2989	2801	2974	2779	2704	2611
		0.1	1159	1146	1146	1141	1136	904	868
		0.25	129	129	129	129	129	128	127
		0.5	11	11	11	11	11	11	11
Pericyte	27196	0	11250	5335	4660	6766	5974	5371	4683
		0.05	3737	3580	3502	3651	3586	3561	3460
		0.1	1624	1607	1606	1616	1614	1622	1622
		0.25	344	344	343	344	344	344	344
		0.5	61	61	61	61	61	61	61

BH procedure after filtering genes with different thresholds for logFC. Notice that when filtering by logFC > 0, all genes are included, implying that we directly apply the BH procedure for FDR control.

Based on results summarized in Table 2.5 and Table 2.6, we can clearly see that the BH procedure without logFC screening identifies far more number of genes compared to that by using the knockoff procedure, regardless of cell type. Even in the absence of the AD DEGs ground truth, this disparity raises concerns about the potential inflation of FDR for the BH procedure. For the proposed knockoff procedure, the less number of identified DEGs might be due to the fact that we perform conditional inference under the knockoff framework instead of marginal one. Conceptually, when there are co-expressed genes, or genes from the same pathway, only the one that has an indepedent effect on the disease will be selected.

We will focus on astrocytes for the remainder of the section for illustration purpose, as for this cell type both methods have a relatively large number of discoveries. And we will compare the knockoff-selected genes with the BH-selected genes.

Figure 2.3 is a scatterplot comparing the multi-decomp, e-decomp, multi-LR, and e-LR q-values against the BH q-values (logFC > 0.17). The logFC is chosen such that the number of genes selected using the BH procedure (545) is close to the number of genes selected using knockoffs. It is important to note that the two-step procedure involving logFC screening and BH procedure is not statistically rigorous, as it may suffer from selection bias. However, we admit that this is a practical compromise, given that without logFC screening, as demonstrated in Table 2.6, the number of selected genes would be impractically large for follow-up studies. We compare our results with a genome-wide association studies (GWAS) of AD [Yang et al., 2022] and use the top 45 risk genes to annotate our discoveries. DEG identification based on scRNA-seq data is usually affected by the large variance of the data, and may vary significantly due to differences in the cell types, unobserved confounders, and the type of tests applied. In contrast, top risk genes from GWAS are generally replicable, and is highly likely to be differently expressed. This is the reason why we use GWAS results as a benchmark. As shown in the figure, we can see that our proposed methods are able to identify some of the risk genes alongside the BH procedure, including WWOX, CLU, and SLC24A4, while also exclusively identifying APOE, one of the most discussed genes related to AD.

The observations in this section highlight several benefits of using our proposed methods. First, conditional independence test by using knockoffs can help reduce the number of discoveries. Our proposed knockoff construction also has a potential to protect us from FDR inflation due to unobserved confounders, while the BH procedure can be more vulnerable in that case. Second, we are able to avoid the ad-hoc nature of logFC screening before applying the BH procedure. As

shown in this section, the choice of the threshold significantly changes the number of DEGs identified. Finally, despite the inherent randomness of the knockoff framework, it allows us to apply stabilizing procedures such as multiple knockoffs or e-BH to improve replicability of the discoveries. The BH procedure on the other hand, can have a FDP with high variance when being applied to dependent p-values. And it is known that the BH procedure could produce highly skewed FDP distributions [Efron, 2012] that may lead to misleading conclusions. These advantages overall suggest that our knockoff-based methods are more effective and reliable alternatives for variable selection with guaranteed FDR control.

2.8 Discussion

In this chapter, we introduced a knockoff construction based on the spiked covariance model to address the high-dimensionality and high-missingness in scRNA-seq data. Building upon the model-X knockoff framework introduced by Candès et al. [2018], we assume that the correlation among variables can be mainly captured by a low-rank structure, and make additional assumptions on the distribution of the latent random factors. This approach leads to the construction of more efficient and more powerful knockoffs. Additionally, we explored recent advancements in the knockoff literature, including the low-rank knockoff construction [Fan et al., 2020; Zhu et al., 2021], and two stabilizing methods: multiple knockoffs [Gimenez and Zou, 2019] and e-BH procedure [Ren and Barber, 2022; Wang and Ramdas, 2022]. Importantly, while our focus is on scRNA-seq data, the methodology described in this chapter can be applied to any high-dimensional dataset with missing values, provided that the model in equation (2.2) is true and the missingness pattern is MCAR.

When applied to scRNA-seq data, we demonstrated that our knockoff approach can provide a shorter list of discoveries that tend to have independent effects on the disease. It does not require an ad-hoc logFC screening step either. The method can be more robust compared to the BH procedure, especially in the presence of unaddressed confounding factors.

We recognize the recent work on *Clipper* by Ge et al. [2021], which is a statistical framework for variable selection with FDR control in high-throughput data analysis. Motivated by the work by Candès et al. [2018] and Gimenez and Zou [2019], *Clipper* shares a lot of similarities with the knockoff framework, among which a noticable feature is that the method does not rely on p-values.

The type-I error control is achieved by running permutations instead of constructing knockoffs hence the method is easier to implement. However, by using permutations, *Clipper* still performs marginal inference and is not able to account for dependencies among genes. Furthermore, it does not consider any potential confounding factors. We believe that our method provides more flexibility and is likely to be more powerful than *Clipper*. Additionally, although we are using p-values to calculate the knockoff statistics, the validity of the FDR control depends on the flip-sign property and is not affected by model misspecifications.

For future work, it would be interesting to explore knockoff construction that can capture both column and row dependencies in scRNA-seq data. The individual cell donor is another major source of variation, as multiple cells may come from the same donor. In fact, the (biological) variation explained by the individual can be far greater than the (technical) variation due to batch [Tung et al., 2017]. In this work, we used knockoffs to capture the column-wise covariance among gene expressions, but an interesting but more difficult question is: Using knockoffs, can we capture the row-wise covariance among the cells, and will such a construction achieve better FDR control?

It remains unclear whether the conditional inference under the knockoff framework is still valid after we impute the missing values. In this work, we rely on imputation to recover the low-rank component, estimate the covariance matrix, and generate knockoffs, but the imputed values are not used when calculating test statistics. In future, we would like to investigate whether imputing missing values can help us recover the true covariance matrix when the model described by (2.11) and (2.12) holds, and whether the exchangeability holds for decomp- or LR- knockoffs introduced in Section 2.4.

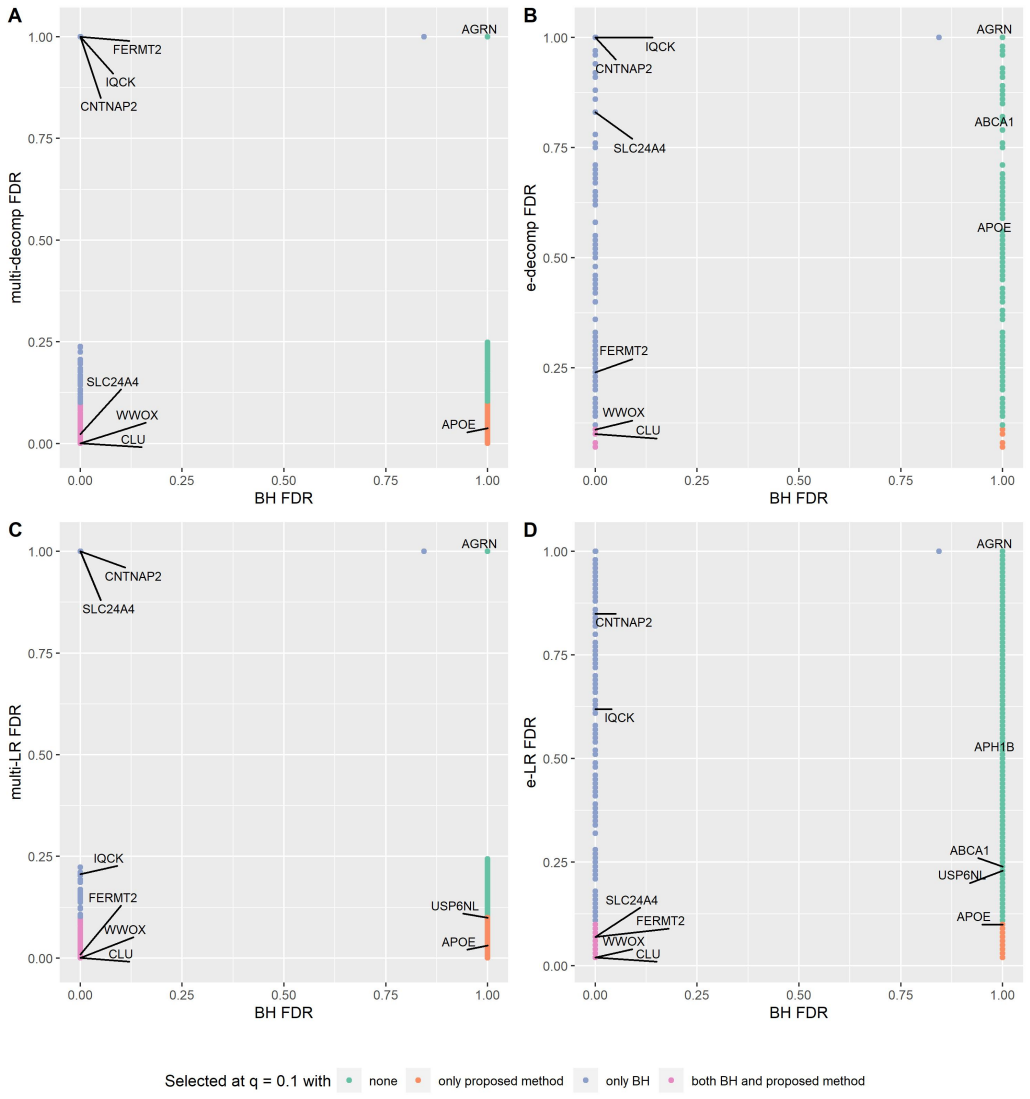

Figure 2.3: Comparison of q-values between proposed knockoff methods and BH procedure (logFC > 0.17). Comparison of the q-values (FDR) between multi-decomp knockoffs (panel A), e-decomp procedure(panel B), multi-LR knockoffs (panel C) and e-LR procedure (panel D), and BH procedure. MAST is used to calculate the knockoff statistics and p-values, and logFC > 0.17 in the BH procedure. The target FDR is set at $q = 0.1$ for coloring purposes.

3.0 Appendix

A.1 Proof for Theorem 1

Theorem 1. *Let $\{(\boldsymbol{A}'_\ell, \boldsymbol{B}_\ell)\}$ be the iterates generated by Algorithm 1. Then the function values of*

$$F(\boldsymbol{A}', \boldsymbol{B}) \coloneqq \frac{1}{2}\left\|P_\Omega(\boldsymbol{G} - [\boldsymbol{X}, \boldsymbol{A}']\boldsymbol{B}^\top)\right\|_F^2 + \frac{\lambda}{2}\left(\|\boldsymbol{A}'\|_F^2 + \|\boldsymbol{B}\|_F^2\right),$$

are monotonically decreasing,

$$F(\boldsymbol{A}'_\ell, \boldsymbol{B}_\ell) \geq F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_\ell) \geq F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_{\ell+1}), \quad \ell \geq 1.$$

Proof. To begin with, we need to define two surrogate functions

$$\begin{aligned} Q_A(\boldsymbol{Z}_1|\boldsymbol{A}', \boldsymbol{B}) \coloneqq & \frac{1}{2}\left\|P_\Omega(\boldsymbol{G} - [\boldsymbol{X}, \boldsymbol{Z}_1]\boldsymbol{B}^\top) + P_\Omega^\perp([\boldsymbol{X}, \boldsymbol{A}']\boldsymbol{B}^\top - [\boldsymbol{X}, \boldsymbol{Z}_1]\boldsymbol{B}^\top)\right\|_F^2 \\ & + \frac{\lambda}{2}\left(\|\boldsymbol{Z}_1\|_F^2 + \|\boldsymbol{B}\|_F^2\right), \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} Q_B(\boldsymbol{Z}_2|\boldsymbol{A}', \boldsymbol{B}) \coloneqq & \frac{1}{2}\left\|P_\Omega(\boldsymbol{G} - [\boldsymbol{X}, \boldsymbol{A}']\boldsymbol{Z}_2^\top) + P_\Omega^\perp([\boldsymbol{X}, \boldsymbol{A}']\boldsymbol{B}^\top - [\boldsymbol{X}, \boldsymbol{A}']\boldsymbol{Z}_2^\top)\right\|_F^2 \\ & + \frac{\lambda}{2}\left(\|\boldsymbol{A}'\|_F^2 + \|\boldsymbol{Z}_2\|_F^2\right). \end{aligned} \tag{A.2}$$

It should be noted that $Q_A(\boldsymbol{Z}_1|\boldsymbol{A}', \boldsymbol{B}) \geq F(\boldsymbol{Z}_1, \boldsymbol{B})$ and $Q_B(\boldsymbol{Z}_2|\boldsymbol{A}', \boldsymbol{B}) \geq F(\boldsymbol{A}', \boldsymbol{Z}_2)$ where the equality holds at $\boldsymbol{Z}_1 = \boldsymbol{A}'$ and $\boldsymbol{Z}_2 = \boldsymbol{B}$ respectively. In step 3, by fixing $\boldsymbol{A}_{\ell+1}$, (2.7) is the solution of the ridge regression problem

$$\boldsymbol{B}_{\ell+1} = \arg\min Q_B(\boldsymbol{Z}_2|\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_\ell),$$

as we can see that the first part of equation (A.2) can be transformed into

$$\frac{1}{2}\left\|P_\Omega(\boldsymbol{G}) + P_\Omega^\perp([\boldsymbol{X}, \boldsymbol{A}'_{\ell+1}]\boldsymbol{B}_\ell^\top) - [\boldsymbol{X}, \boldsymbol{A}'_{\ell+1}]\boldsymbol{Z}_2^\top\right\|_F^2.$$

Thus we can establish the inequality $F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_\ell) \geq F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_{\ell+1})$ via

$$F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_\ell) = Q_B(\boldsymbol{B}_\ell|\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_\ell) \geq Q_B(\boldsymbol{B}_{\ell+1}|\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_\ell) \geq F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_{\ell+1}).$$

Similarly, in step 2, we could conclude that $F(\boldsymbol{A}'_\ell, \boldsymbol{B}_\ell) \geq F(\boldsymbol{A}'_{\ell+1}, \boldsymbol{B}_\ell)$. Notice that compared to step 3, we have an additional correction term '$-\boldsymbol{Q}\boldsymbol{B}_{X,\ell}^\top$'. That is because the columns $\boldsymbol{X}$ in $\boldsymbol{A}_\ell$ are

fixed and should not be updated in this step of the iteration. Since the first part of equation (A.1) could be transformed to

$$\frac{1}{2}\left\|P_\Omega(\boldsymbol{G})+P_\Omega^\perp([\boldsymbol{X},\boldsymbol{A}'_\ell]\boldsymbol{B}_\ell^\top)-\boldsymbol{X}\boldsymbol{B}_{X,\ell}^\top-\boldsymbol{Z}_1\boldsymbol{B}_{A',\ell}^\top\right\|_F^2,$$

such that the solution (2.6) solves

$$\boldsymbol{A}'_{\ell+1}=\arg\min Q_A(\boldsymbol{Z}_1|\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell).$$

In conclusion,

$$F(\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)=Q_A(\boldsymbol{A}'_\ell|\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)\geq Q_A(\boldsymbol{A}'_{\ell+1}|\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)\geq F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell),$$

and the proof is completed. □

A.2 Proof of convergence for Algorithm 2

Similar to Algorithm 1, it can be shown that Algorithm 2 will converge. For simplicity, let's denote the objective function as

$$F(\boldsymbol{A}',\boldsymbol{B})\coloneqq\frac{1}{2}\left\|P_\Omega(\boldsymbol{G}-[\boldsymbol{X},\boldsymbol{A}']\boldsymbol{B}^\top)\right\|_F^2+\frac{\lambda}{2}\left(\|\boldsymbol{A}'\|_F^2+\|\boldsymbol{B}_A\|_F^2\right),$$

such that the objective function

$$\text{minimize}_{\boldsymbol{A}',\boldsymbol{B}}\ \frac{1}{2}\left\|P_\Omega(\boldsymbol{G}-\boldsymbol{A}\boldsymbol{B}^\top)\right\|_F^2+\frac{\lambda}{2}\left(\|\boldsymbol{A}'\|_F^2+\|\boldsymbol{B}_A\|_F^2\right)$$

can be rewritten as

$$\text{minimize}_{\boldsymbol{A}',\boldsymbol{B}}F(\boldsymbol{A}',\boldsymbol{B}).$$

For the objective function $F(\boldsymbol{A}',\boldsymbol{B})$, we can show its value decreases in each iteration, as summarized in the following Theorem 2. In combination with the fact that $F(\boldsymbol{A}',\boldsymbol{B})$ has a lower bound, we know the algorithm will converge.

Theorem 2. *Let $\{(\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)\}$ be the iterates generated by Algorithm 1. Then the function values are monotonically decreasing,*

$$F(\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)\geq F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell)\geq F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_{\ell+1}),\quad \ell\geq 1.$$

Proof. To begin with, we need to define two surrogate functions

$$\begin{aligned} Q_A(\boldsymbol{Z}_1|\boldsymbol{A}',\boldsymbol{B}) \coloneqq & \frac{1}{2}\left\|P_\Omega(\boldsymbol{G}-[\boldsymbol{X},\boldsymbol{Z}_1]\boldsymbol{B}^\top)+P_\Omega^\perp([\boldsymbol{X},\boldsymbol{A}']\boldsymbol{B}^\top-[\boldsymbol{X},\boldsymbol{Z}_1]\boldsymbol{B}^\top)\right\|_F^2 \\ & +\frac{\lambda}{2}\left(\|\boldsymbol{Z}_1\|_F^2+\|\boldsymbol{B}\|_F^2\right), \end{aligned} \tag{A.3}$$

and

$$\begin{aligned} Q_B(\boldsymbol{Z}_2|\boldsymbol{A}',\boldsymbol{B}) \coloneqq & \frac{1}{2}\left\|P_\Omega(\boldsymbol{G}-[\boldsymbol{X},\boldsymbol{A}']\boldsymbol{Z}_2^\top)+P_\Omega^\perp([\boldsymbol{X},\boldsymbol{A}']\boldsymbol{B}^\top-[\boldsymbol{X},\boldsymbol{A}']\boldsymbol{Z}_2^\top)\right\|_F^2 \\ & +\frac{\lambda}{2}\left(\|\boldsymbol{A}'\|_F^2+\|\boldsymbol{Z}_{2,A}\|_F^2\right), \end{aligned} \tag{A.4}$$

where $\boldsymbol{Z}_2=[\boldsymbol{Z}_{X,2},\boldsymbol{Z}_{A,2}]$. It should be noted that $Q_A(\boldsymbol{Z}_1|\boldsymbol{A}',\boldsymbol{B})\geq F(\boldsymbol{Z}_1,\boldsymbol{B})$ and $Q_B(\boldsymbol{Z}_2|\boldsymbol{A}',\boldsymbol{B})\geq F(\boldsymbol{A}',\boldsymbol{Z}_2)$ where the equality holds at $\boldsymbol{Z}_1=\boldsymbol{A}'$ and $\boldsymbol{Z}_2=\boldsymbol{B}$ respectively. In step 3, by fixing $\boldsymbol{A}_{\ell+1}$, (2.7) is the solution of the ridge regression problem

$$\boldsymbol{B}_{\ell+1}=\arg\min Q_B(\boldsymbol{Z}_2|\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell),$$

as we can see that the first part of equation (A.4) can be transformed into

$$\frac{1}{2}\left\|P_\Omega(\boldsymbol{G})+P_\Omega^\perp([\boldsymbol{X},\boldsymbol{A}'_{\ell+1}]\boldsymbol{B}_\ell^\top)-[\boldsymbol{X},\boldsymbol{A}'_{\ell+1}]\boldsymbol{Z}_2^\top\right\|_F^2.$$

Thus we can establish the inequality $F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell)\geq F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_{\ell+1})$ via

$$F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell)=Q_B(\boldsymbol{B}_\ell|\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell)\geq Q_B(\boldsymbol{B}_{\ell+1}|\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell)\geq F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_{\ell+1}).$$

Similarly, in step 2, we could conclude that $F(\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)\geq F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell)$. Notice that compared to step 3, we have an additional correction term '$-\boldsymbol{Q}\boldsymbol{B}_{X,\ell}^\top$'. That is because the columns $\boldsymbol{X}$ in $\boldsymbol{A}_\ell$ are fixed and should not be updated in this step of the iteration. Since the first part of equation (A.3) could be transformed to

$$\frac{1}{2}\left\|P_\Omega(\boldsymbol{G})+P_\Omega^\perp([\boldsymbol{X},\boldsymbol{A}'_\ell]\boldsymbol{B}_\ell^\top)-\boldsymbol{X}\boldsymbol{B}_{X,\ell}^\top-\boldsymbol{Z}_1\boldsymbol{B}_{A',\ell}^\top\right\|_F^2,$$

such that the solution (2.9) solves

$$\boldsymbol{A}'_{\ell+1}=\arg\min Q_A(\boldsymbol{Z}_1|\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell).$$

In conclusion,

$$F(\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)=Q_A(\boldsymbol{A}'_\ell|\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)\geq Q_A(\boldsymbol{A}'_{\ell+1}|\boldsymbol{A}'_\ell,\boldsymbol{B}_\ell)\geq F(\boldsymbol{A}'_{\ell+1},\boldsymbol{B}_\ell),$$

and the proof is completed. □

A.3 Simulations with debiased knockoffs

In this section, we replicate the simulations previously conducted in Section 2.6.2 and Section 2.6.3. The only difference is that we use Algorithm 2 instead of Algorithm 1 to recover the missing values and estimate $\boldsymbol{A}$ and $\boldsymbol{B}$. As shown in Figure A.1 and Table A.1, the simulation results corresponding to the ones in Section 2.6.2 demonstrate similar power with proper FDR control. In other words, there is no obvious performance advantage of biased knockoffs over debiased knockoffs for this specific simulation setting, and vice versa. However, as shown in Table A.2, which corresponds to Table 2.4 from Section 2.6.3, while debiased knockoffs demonstrate similar robustness against potential confounders, their power significantly lags behind that of biased knockoffs. To avoid numerical errors, we added a small penalty—a diagonal matrix of 0.05 times the average of the diagonal values of $\boldsymbol{A}_\ell^{'\top}\boldsymbol{A}_\ell'$ or $\boldsymbol{B}_{A',\ell}^\top\boldsymbol{B}_{A',\ell}$—to $\boldsymbol{A}_\ell^{'\top}\boldsymbol{A}_\ell'$ and $\boldsymbol{B}_{A',\ell}^\top\boldsymbol{B}_{A',\ell}$ respectively when calculating their inverse. Notice that when $\lambda = 0$, the knockoffs are unbiased, hence both algorithms will produce identical outcomes. The power increases with larger a λ, but even the best case scenarios when $\lambda > 0$ is smaller than the power shown in Table 2.4. Therefore, we find the debiased knockoffs less appealing in practice and leave it here in the appendix as an interesting finding.

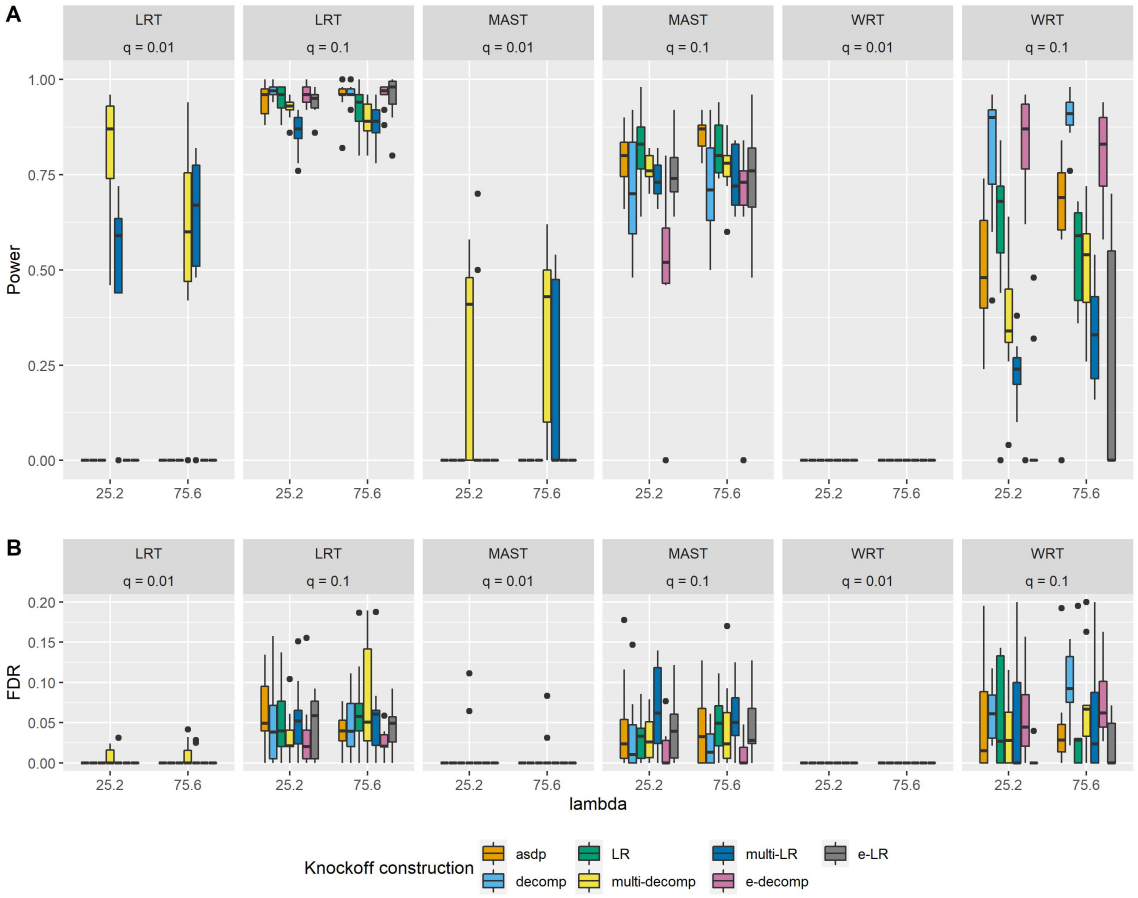

Figure A.1: Comparison of debiased knockoff constructions and BH procedure under q = 0.1. Panel A and B are box-plots of the FDP and power over 20 repetitions respectively. Simulations are carried out according to the details described in Section 2.6.2 using 1951 genes and 15,141 observations, sgn = 3 and $\lambda_{\max} = 252.07$, except that knockoffs are generated using the recovered matrix from Algorithm 2.

Table A.1: Comparison of debiased knockoff constructions and BH procedure under q = 0.05. Table of the results where knockoffs are generated using the recovered matrix from Algorithm 2. sgn = 3, target FDR $q = 0.05$ and Bonferroni corrected p-values are used to calculate knockoff statistics.

		LRT		MAST		WRT	
Knockoff Construction	λ	FDR	Power	FDR	Power	FDR	Power
asdp	0.0	0.03 (0.02)	0.88 (0.08)	0.03 (0.03)	0.74 (0.23)	0.01 (0.03)	0.19 (0.24)
asdp	25.2	0.03 (0.04)	0.87 (0.12)	0.03 (0.03)	0.66 (0.17)	0.02 (0.06)	0.15 (0.21)
asdp	50.4	0.05 (0.06)	0.93 (0.05)	0.02 (0.04)	0.74 (0.22)	0.06 (0.12)	0.38 (0.25)
asdp	75.6	0.04 (0.03)	0.92 (0.06)	0.03 (0.04)	0.72 (0.22)	0.07 (0.10)	0.38 (0.27)
decomp	0.0	0.04 (0.03)	0.94 (0.04)	0.01 (0.02)	0.70 (0.27)	0.05 (0.09)	0.66 (0.28)
decomp	25.2	0.04 (0.04)	0.94 (0.08)	0.01 (0.03)	0.62 (0.20)	0.07 (0.08)	0.75 (0.24)
decomp	50.4	0.03 (0.03)	0.95 (0.03)	0.02 (0.04)	0.58 (0.25)	0.15 (0.16)	0.79 (0.14)
decomp	75.6	0.04 (0.06)	0.95 (0.04)	0.01 (0.02)	0.56 (0.25)	0.14 (0.15)	0.81 (0.17)
LR	0.0	0.03 (0.03)	0.89 (0.08)	0.03 (0.04)	0.74 (0.16)	0.01 (0.02)	0.08 (0.19)
LR	25.2	0.04 (0.04)	0.92 (0.05)	0.02 (0.03)	0.74 (0.16)	0.06 (0.14)	0.24 (0.27)
LR	50.4	0.03 (0.03)	0.89 (0.13)	0.03 (0.04)	0.76 (0.22)	0.07 (0.16)	0.37 (0.33)
LR	75.6	0.06 (0.06)	0.91 (0.10)	0.03 (0.03)	0.63 (0.30)	0.09 (0.17)	0.33 (0.23)
multi-decomp	0.0	0.03 (0.02)	0.87 (0.06)	0.02 (0.03)	0.75 (0.09)	0.00 (0.00)	0.06 (0.08)
multi-decomp	25.2	0.04 (0.03)	0.92 (0.04)	0.03 (0.04)	0.69 (0.11)	0.06 (0.13)	0.18 (0.16)
multi-decomp	50.4	0.07 (0.06)	0.90 (0.05)	0.03 (0.03)	0.72 (0.10)	0.05 (0.13)	0.19 (0.14)
multi-decomp	75.6	0.06 (0.05)	0.88 (0.05)	0.04 (0.06)	0.72 (0.09)	0.08 (0.1)	0.25 (0.21)
multi-LR	0.0	0.02 (0.02)	0.84 (0.08)	0.02 (0.03)	0.68 (0.16)	0.00 (0.00)	0.01 (0.04)
multi-LR	25.2	0.04 (0.04)	0.85 (0.06)	0.03 (0.02)	0.66 (0.12)	0.10 (0.17)	0.13 (0.09)
multi-LR	50.4	0.04 (0.03)	0.89 (0.04)	0.02 (0.02)	0.71 (0.1)	0.13 (0.15)	0.14 (0.11)
multi-LR	75.6	0.03 (0.03)	0.88 (0.06)	0.03 (0.03)	0.68 (0.09)	0.13 (0.19)	0.15 (0.08)
e-decomp	0.0	0.02 (0.02)	0.91 (0.05)	0.00 (0.01)	0.16 (0.33)	0.01 (0.02)	0.30 (0.41)
e-decomp	25.2	0.02 (0.03)	0.92 (0.05)	0.00 (0.01)	0.04 (0.17)	0.02 (0.03)	0.38 (0.43)
e-decomp	50.4	0.02 (0.03)	0.92 (0.22)	0.00 (0.01)	0.04 (0.17)	0.05 (0.14)	0.38 (0.43)
e-decomp	75.6	0.02 (0.02)	0.95 (0.04)	0.00 (0.00)	0.00 (0.00)	0.03 (0.05)	0.42 (0.43)
e-LR	0.0	0.01 (0.01)	0.59 (0.44)	0.01 (0.03)	0.22 (0.38)	0.00 (0.00)	0.00 (0.00)
e-LR	25.2	0.02 (0.03)	0.68 (0.36)	0.00 (0.00)	0.05 (0.21)	0.00 (0.00)	0.00 (0.00)
e-LR	50.4	0.03 (0.03)	0.88 (0.22)	0.00 (0.01)	0.28 (0.39)	0.00 (0.00)	0.00 (0.00)
e-LR	75.6	0.02 (0.02)	0.83 (0.29)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
asdp_cov		0.02 (0.03)	0.77 (0.22)	0.02 (0.03)	0.57 (0.27)	0.01 (0.02)	0.25 (0.38)
BH		0.13 (0.12)	0.99 (0.01)	0.12 (0.11)	0.96 (0.03)	0.29 (0.21)	0.94 (0.03)

Table A.2: Debiased knockoff procedure and confounding effects. Table of the results where knockoffs are generated using the recovered matrix from Algorithm 2. sgn = 3, target FDR $q = 0.1$ and Bonferroni corrected p-values are used to calculate knockoff statistics. Only decomp knockoffs and LRT based knockoff statistics are included. The average FDP and power over 20 simulations for each setting are shown in each column with the standard deviation in the parentheses.

Imputation		no CDR		no CDR		with CDR	
LRT		no CDR		with CDR		with CDR	
batch	λ	FDR	Power	FDR	Power	FDR	Power
original	0.00	0.00 (0.00)	0.18 (0.21)	0.00 (0.00)	0.19 (0.22)	0.00 (0.00)	0.17 (0.22)
	22.2	0.00 (0.01)	0.31 (0.27)	0.00 (0.01)	0.33 (0.26)	0.00 (0.01)	0.40 (0.25)
	44.4	0.01 (0.01)	0.41 (0.26)	0.00 (0.01)	0.42 (0.24)	0.00 (0.01)	0.47 (0.24)
	66.6	0.01 (0.02)	0.43 (0.29)	0.00 (0.01)	0.48 (0.23)	0.00 (0.01)	0.56 (0.22)
permute	0.00	0.00 (0.02)	0.20 (0.25)	0.00 (0.00)	0.23 (0.24)	0.00 (0.00)	0.23 (0.26)
	22.2	0.00 (0.02)	0.39 (0.24)	0.00 (0.00)	0.42 (0.24)	0.00 (0.01)	0.42 (0.26)
	44.4	0.00 (0.01)	0.44 (0.24)	0.00 (0.01)	0.48 (0.22)	0.00 (0.00)	0.52 (0.22)
	66.6	0.01 (0.02)	0.49 (0.26)	0.00 (0.01)	0.52 (0.23)	0.00 (0.01)	0.58 (0.22)
remove	0.00	0.00 (0.00)	0.17 (0.20)	0.00 (0.00)	0.19 (0.21)	0.00 (0.00)	0.19 (0.23)
	22.2	0.00 (0.01)	0.30 (0.27)	0.00 (0.01)	0.35 (0.27)	0.00 (0.01)	0.38 (0.26)
	44.4	0.01 (0.01)	0.39 (0.25)	0.01 (0.01)	0.41 (0.25)	0.00 (0.01)	0.46 (0.25)
	66.6	0.01 (0.03)	0.40 (0.27)	0.01 (0.01)	0.47 (0.27)	0.00 (0.01)	0.52 (0.27)

A.4 A closer look on the e-BH results from Section 2.6.2

As shown in Section 2.6.2, while multiple knockoffs and e-BH both address the randomness of the knockoff procedure, under certain scenarios, the results of e-BH will dramatically degrade. Especially when the target FDR q is set to be small. In this section, we want to take a closer look on the e-BH procedure and explain why it happens.

Under the e-BH procedure, variables with large e-values are more likely to be selected. Considering that the e-values is calculated according to

$$e_j^m = p \times \frac{\mathbb{1}_{\{W_j^m \geq \tau^m\}}}{1 + \#\{j : W_j \leq -\tau^m\}},$$

where

$$\tau^m = \min\left\{t > 0 : \frac{1 + \#\{j : W_j^m \leq -t\}}{\#\{j : W_j^m \geq t\}} \leq q_{\text{kn}}\right\},$$

it is clear that the e-value is large if and only if the variable is selected, and since we take the average of e-values $e_j^{\text{avg}} = \sum_{m=1}^{M} e_j^m / M$ for the sake of derandomization, it means that only variables which are consistently being selected during the individual knockoff precedures will have a large e-value. To illustrate our point, Figure A.3 shows one of the simulations out of the 20 repetitions for e-decomp. There are multiple factors that affect the e-values and whether the variable is selected: The test statistics, the q_{kn} chosen for individual knockoff precedures, and, consequently, the number of times a variable is being selected during each of the knockoff precedure. Therefore, we should consider the e-BH procedure as an aggregate of independent knockoff procedures. This leads to the question: How can we have a more powerful e-BH procedure? While we know that variables with larger e-values are more likely to be selected, and variables that are being consistently selected will have a larger e-value, it does not necessarily mean that optimizing the individual knockoff procedures such that as many variables allowed under the target FDR as possible will lead to the best results. Aside from the LRT test statistic results where we are taking advantage of our knowledge of the signal generation, Figure A.2 demonstrates that the choice of q_{kn} is an important parameter that needs to be chosen carefully. When q_{kn} is too small, the individual knockoff procedures are unable to detect the signals, and when q_{kn} is too large, the stability of the e-BH selection deteriorates because the e-BH procedure is not able to differentiate between the many signals (see also Figure A.3). Therefore, while based on the results in Section 2.6.2, choosing $q_{kn} = q/2$ does not work well, Ren and Barber [2022] correctly pointed out that the choice of q_{kn} greatly affects the power. And like the authors,

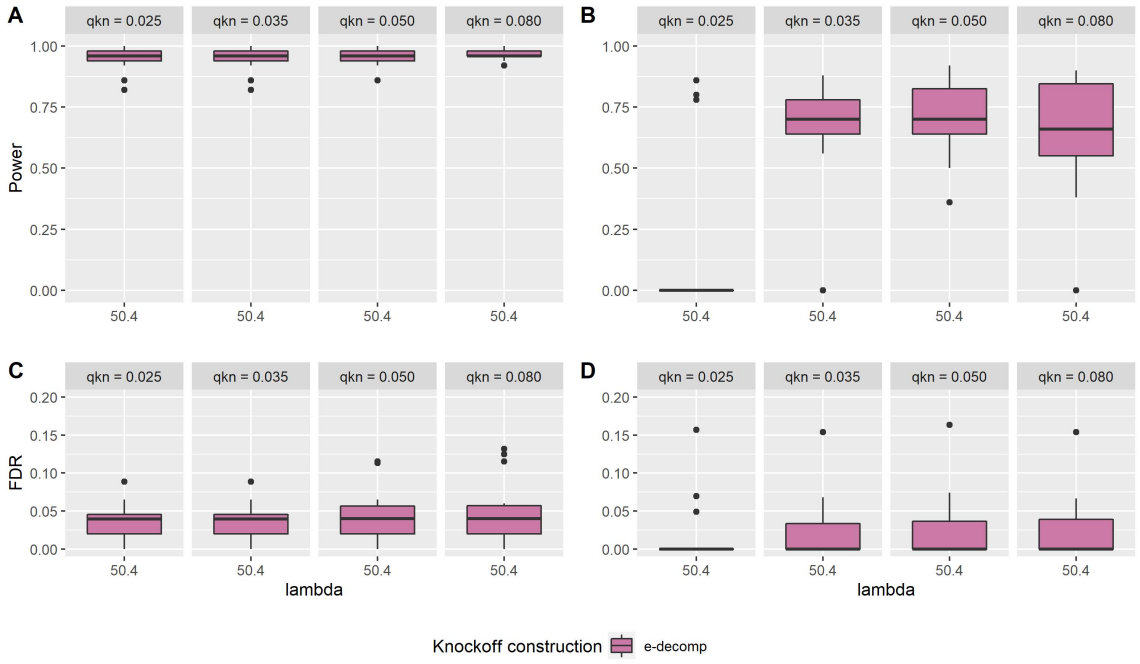

Figure A.2: Comparison of different q_{kn} under q = 0.1. Panel in the first row (A and B) and second row (C and D) are box-plots of the FDP and power over 20 repetitions respectively. Panel A and C are results based on LRT, panel B and D are results based on MAST. Simulations are carried out according to the e-decomp procedure described in Section 2.6.2 using 1951 genes and 15,141 observations, sgn = 3 and $\lambda = 50.4$.

we do not have a concrete answer for how we can choose q_{kn} such that the e-BH procedure is more powerful.

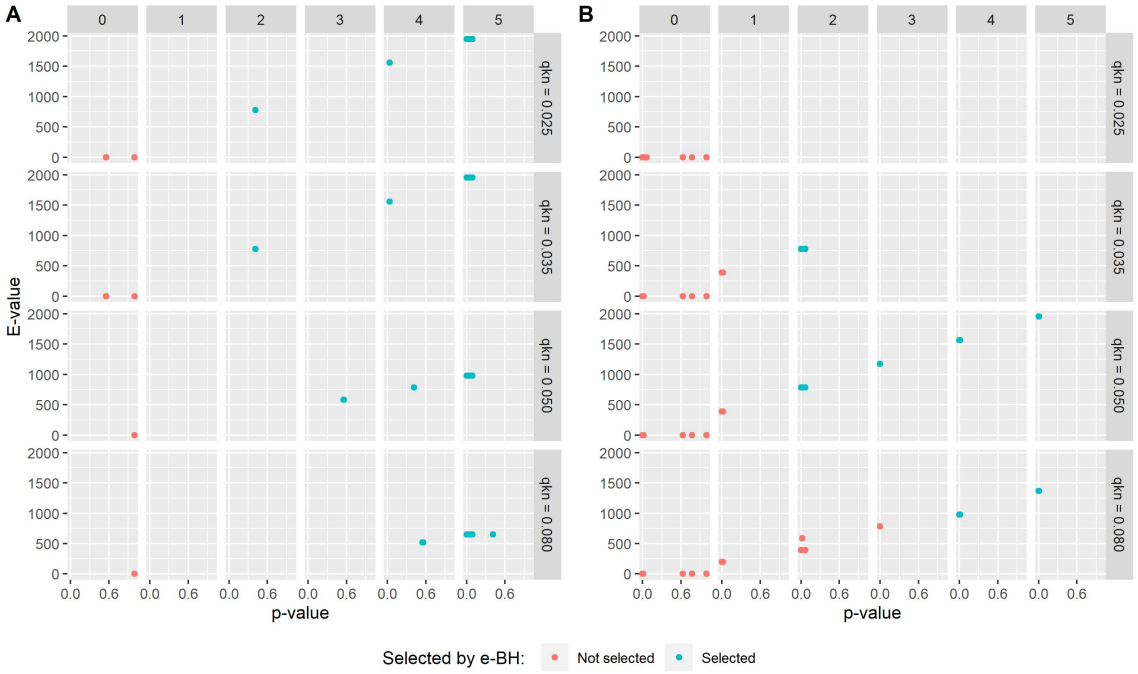

Figure A.3: Relation of e-values and independent knockoff variable selections under q = 0.1 and different q_{kn}. Panel A and B show the e-values calculated with five sets of knockoff against the p-values of the original variables. The columns indicate the number of times a variable is selected by the individual knockoffs and the rows indicate the q_{kn} used for variable selection. Panel A is the result based on LRT, and panel B is the result based on MAST. Simulations are carried out according to the e-decomp procedure described in Section 2.6.2 using 1951 genes and 15,141 observations, sgn = 3 and $\lambda = 50.4$.

A.5 Comparison of knockoff statistics calculated using corrected and non-corrected p-values

Instead of non-corrected p-values, using Bonferroni corrected p-values provide us with better controlled FDR and higher power. While the FDR is guaranteed to be controlled for any knockoff statistics as long as it satisfies the flip-sign property, different knockoff statistics will select variables with vastly different power. As shown in Table A.3, variable selection is much more powerful if Bonferroni corrected p-values are used to calculate the knockoff statistics regardless of the test used to calculate p-values.

Furthermore, on top of asdp_cov-knockoff results, the results where the covariances were estimated using the recovered data $\boldsymbol{G}$ instead of $\boldsymbol{G}^{\text{obs}}$ are also included in the tables and denoted as asdp_impcov knockoffs. It could be observed that asdp_cov knockoffs perform better regardless of knockoff statistics or rescaling method.

Table A.3: Comparison of knockoff constructions with non-corrected and Bonferroni corrected p-values, and BH procedure under q = 0.1

		MAST				WRT			
		p-value		bonf. corrected p-value		p-value		bonf. corrected p-value	
Knockoff Construction	λ	FDR	Power	FDR	Power	FDR	Power	FDR	Power
asdp	0.0	0.09 (0.07)	0.35 (0.15)	0.03 (0.05)	0.82 (0.12)	0.13 (0.16)	0.30 (0.17)	0.06 (0.12)	0.58 (0.23)
asdp	25.2	0.13 (0.14)	0.32 (0.20)	0.04 (0.05)	0.85 (0.09)	0.11 (0.12)	0.27 (0.18)	0.03 (0.03)	0.48 (0.24)
asdp	50.4	0.10 (0.09)	0.40 (0.13)	0.05 (0.07)	0.85 (0.07)	0.21 (0.17)	0.39 (0.16)	0.11 (0.13)	0.67 (0.20)
asdp	75.6	0.10 (0.08)	0.47 (0.11)	0.06 (0.07)	0.87 (0.08)	0.15 (0.14)	0.37 (0.14)	0.09 (0.09)	0.67 (0.15)
decomp	0.0	0.07 (0.07)	0.42 (0.13)	0.04 (0.04)	0.83 (0.11)	0.21 (0.19)	0.66 (0.14)	0.13 (0.14)	0.81 (0.19)
decomp	25.2	0.11 (0.08)	0.43 (0.18)	0.04 (0.07)	0.79 (0.09)	0.17 (0.10)	0.68 (0.13)	0.11 (0.10)	0.81 (0.15)
decomp	50.4	0.14 (0.11)	0.47 (0.14)	0.06 (0.10)	0.75 (0.17)	0.21 (0.16)	0.70 (0.08)	0.17 (0.15)	0.80 (0.10)
decomp	75.6	0.10 (0.06)	0.46 (0.10)	0.03 (0.04)	0.74 (0.09)	0.18 (0.14)	0.70 (0.08)	0.12 (0.12)	0.87 (0.06)
LR	0.0	0.09 (0.07)	0.37 (0.12)	0.03 (0.04)	0.84 (0.08)	0.11 (0.13)	0.20 (0.15)	0.01 (0.03)	0.28 (0.24)
LR	25.2	0.11 (0.10)	0.39 (0.14)	0.03 (0.03)	0.86 (0.07)	0.13 (0.09)	0.34 (0.17)	0.04 (0.05)	0.55 (0.26)
LR	50.4	0.12 (0.11)	0.43 (0.09)	0.06 (0.07)	0.84 (0.08)	0.17 (0.20)	0.38 (0.16)	0.08 (0.16)	0.53 (0.21)
LR	75.6	0.11 (0.10)	0.43 (0.12)	0.05 (0.06)	0.85 (0.07)	0.13 (0.11)	0.39 (0.15)	0.04 (0.04)	0.67 (0.19)
multi-LR	0.0	0.09 (0.07)	0.40 (0.08)	0.04 (0.04)	0.75 (0.06)	0.06 (0.14)	0.09 (0.10)	0.06 (0.13)	0.12 (0.11)
multi-LR	25.2	0.10 (0.08)	0.39 (0.09)	0.04 (0.03)	0.76 (0.07)	0.01 (0.02)	0.17 (0.08)	0.01 (0.03)	0.27 (0.17)
multi-LR	50.4	0.07 (0.07)	0.42 (0.06)	0.02 (0.03)	0.80 (0.06)	0.03 (0.05)	0.20 (0.08)	0.03 (0.05)	0.35 (0.17)
multi-LR	75.6	0.11 (0.09)	0.47 (0.08)	0.06 (0.07)	0.78 (0.04)	0.04 (0.05)	0.25 (0.11)	0.04 (0.09)	0.38 (0.15)
asdp_impcov	0.0	0.02 (0.07)	0.04 (0.08)	0.03 (0.08)	0.33 (0.24)	0.03 (0.15)	0.01 (0.04)	0.00 (0.00)	0.00 (0.00)
asdp_impcov	25.2	0.02 (0.08)	0.07 (0.11)	0.02 (0.05)	0.46 (0.22)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
asdp_impcov	50.4	0.03 (0.06)	0.12 (0.15)	0.03 (0.05)	0.61 (0.18)	0.00 (0.00)	0.00 (0.00)	0.00 (0.02)	0.01 (0.05)
asdp_impcov	75.6	0.05 (0.08)	0.16 (0.13)	0.04 (0.04)	0.69 (0.14)	0.02 (0.09)	0.01 (0.03)	0.07 (0.18)	0.03 (0.06)
asdp_cov		0.05 (0.07)	0.21 (0.17)	0.02 (0.03)	0.74 (0.19)	0.01 (0.05)	0.00 (0.02)	0.01 (0.02)	0.30 (0.42)
BH		0.18 (0.13)	0.97 (0.02)	N/A	N/A	0.39 (0.21)	0.96 (0.03)	N/A	N/A

NOTE: Table of simulation results where sgn = 3, target FDR $q = 0.1$. Non-corrected and Bonferroni corrected p-values are used to calculate knockoff statistics. The average FDP and power over 20 simulations for each setting are shown in each column with the standard deviation in the parentheses. The BH procedure only uses uncorrected p-values and is here for reference.

A.6 Comparison of rescaling methods

Three rescaling methods were initially considered:

1. Add the mean $\boldsymbol{m}$ to all generated knockoffs $\widetilde{\boldsymbol{G}}$ (colored in orange).
2. Add the mean $\boldsymbol{m}$ to the observed expression in $\widetilde{\boldsymbol{G}}$ and keep the value of the unexpressed parts as generated (colored in green).
3. Add the mean $\boldsymbol{m}$ to the observed expression in $\widetilde{\boldsymbol{G}}$ and set the values of the unexpressed parts to 0 (colored in blue).

While it is clear through Section 2.4.5 that the last method is the correct one, it might be interesting to discuss the two other methods. The idea behind rescaling the knockoffs has always been to match the distribution of the original log-normalized data $\boldsymbol{G}^{\text{obs}}$, that will be used in generating the knockoff statistics, and to satisfy the exchangeability property. Therefore, since the data was centered when generating the knockoff variables, the knockoffs need to be de-centered afterwards. However, simply de-centering them by adding the column mean $\boldsymbol{m}$ back is incorrect and leads us with three different methods. We use the same data as described in Section 2.6.1 to illustrate the difference between the three rescaling methods. As shown in Figure A.4, in panel A, we plot the estimated variances of the rescaled knockoff variables against the estimated variance of the original variables while comparing between asdp, decomp and asdp_cov-knockoff constructions. In panel B, we plot 5000 randomly sampled off-diagonal elements from the upper triangle of the covariance matrix of the rescaled knockoff variables against their corresponding covariances of the orignal variables. As shown in panel A and B, the covariance of knockoffs is largely overestimated for asdp and decomp-knockoff constructions, with some exception of the orange colored results. In general, we can conclude that the exchangeability is clearly violated for the first and second rescaling method. If we focus on asdp_cov knockoffs, we may also notice that the covariance of the knockoffs and original variables align with each other approximately in the case where the second and third rescaling method is applied. And this is due to the fact that asdp_cov knockoffs are constructed based off the un-imputed data, hence the knockoff values for the unexpressed parts where close to 0 to begin with.

As shown in Table A.4, the second rescaling method (expresc) is more powerful than the thrid and correct one (expresc0) regardless of the knockoff construction. However, since we know that it

is wrong, the results are not included in the main result.

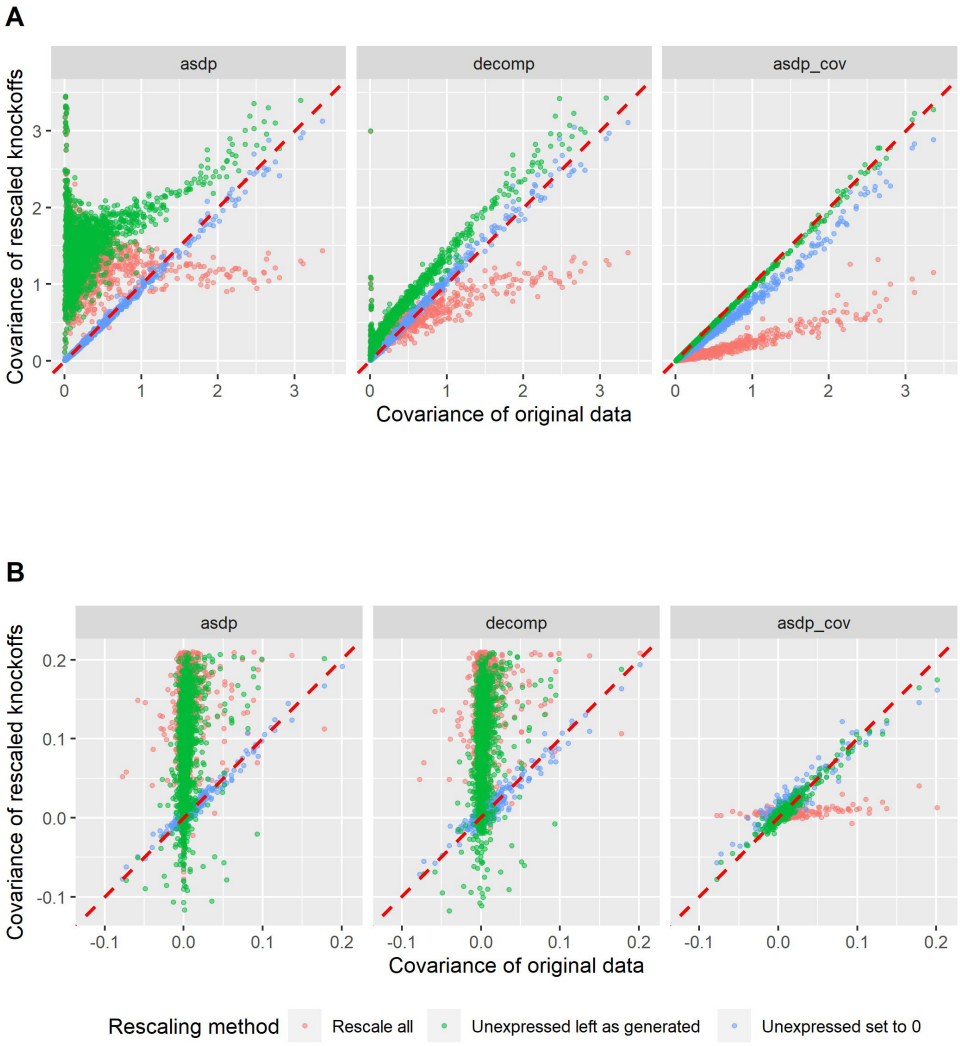

Figure A.4: Additional comparison of knockoff covariance and original covariance. In panel A are the estimated variances of the rescaled knockoff variables against the estimated variance of the original variables. In panel B are 5000 randomly sampled values from each of the upper triangles of the estimated covariance matrices. Rescaling results for asdp, decomp and asdp_cov knockoffs are shown and the randomly sampled indices are fixed for comparison.

Table A.4: Comparison of rescaling methods under q = 0.1. Table of simulation results where sgn = 3, target FDR $q = 0.1$ and Bonferroni corrected p-values were used to calculate knockoff statistics. The average FDP and power over 20 simulations for each setting are shown in each column with the standard deviation in the parentheses. Notice that the BH procedure uses uncorrected p-values and unaffected by the rescaling methods.

		MAST				WRT			
		expresc		expresc0		expresc		expresc0	
Knockoff Construction	λ	FDR	Power	FDR	Power	FDR	Power	FDR	Power
asdp	0.0	0.04 (0.05)	0.93 (0.07)	0.03 (0.05)	0.82 (0.12)	0.07 (0.07)	0.85 (0.13)	0.06 (0.12)	0.58 (0.23)
asdp	25.2	0.05 (0.06)	0.95 (0.04)	0.04 (0.05)	0.85 (0.09)	0.10 (0.15)	0.89 (0.05)	0.03 (0.03)	0.48 (0.24)
asdp	50.4	0.06 (0.08)	0.92 (0.03)	0.05 (0.07)	0.85 (0.07)	0.15 (0.13)	0.87 (0.10)	0.11 (0.13)	0.67 (0.20)
asdp	75.6	0.08 (0.10)	0.93 (0.03)	0.06 (0.07)	0.87 (0.08)	0.13 (0.12)	0.89 (0.06)	0.09 (0.09)	0.67 (0.15)
decomp	0.0	0.04 (0.04)	0.91 (0.06)	0.04 (0.04)	0.83 (0.11)	0.06 (0.06)	0.78 (0.26)	0.13 (0.14)	0.81 (0.19)
decomp	25.2	0.06 (0.06)	0.94 (0.04)	0.04 (0.07)	0.79 (0.09)	0.15 (0.16)	0.84 (0.16)	0.11 (0.10)	0.81 (0.15)
decomp	50.4	0.07 (0.11)	0.94 (0.04)	0.06 (0.10)	0.75 (0.17)	0.23 (0.20)	0.86 (0.10)	0.17 (0.15)	0.80 (0.10)
decomp	75.6	0.06 (0.07)	0.93 (0.03)	0.03 (0.04)	0.74 (0.09)	0.14 (0.12)	0.88 (0.07)	0.12 (0.12)	0.87 (0.06)
LR	0.0	0.03 (0.04)	0.93 (0.04)	0.03 (0.04)	0.84 (0.08)	0.09 (0.10)	0.85 (0.12)	0.01 (0.03)	0.28 (0.24)
LR	25.2	0.04 (0.04)	0.92 (0.05)	0.03 (0.03)	0.86 (0.07)	0.10 (0.08)	0.90 (0.07)	0.04 (0.05)	0.55 (0.26)
LR	50.4	0.06 (0.06)	0.93 (0.06)	0.06 (0.07)	0.84 (0.08)	0.11 (0.12)	0.84 (0.25)	0.08 (0.16)	0.53 (0.21)
LR	75.6	0.08 (0.08)	0.92 (0.04)	0.05 (0.06)	0.85 (0.07)	0.06 (0.04)	0.90 (0.08)	0.04 (0.04)	0.67 (0.19)
multi-LR	0.0	0.06 (0.05)	0.91 (0.06)	0.04 (0.04)	0.75 (0.06)	0.15 (0.19)	0.82 (0.11)	0.06 (0.13)	0.12 (0.11)
multi-LR	25.2	0.03 (0.04)	0.94 (0.03)	0.04 (0.03)	0.76 (0.07)	0.09 (0.09)	0.89 (0.07)	0.01 (0.03)	0.27 (0.17)
multi-LR	50.4	0.07 (0.08)	0.93 (0.03)	0.02 (0.03)	0.80 (0.06)	0.07 (0.05)	0.91 (0.04)	0.03 (0.05)	0.35 (0.17)
multi-LR	75.6	0.07 (0.07)	0.94 (0.03)	0.06 (0.07)	0.78 (0.04)	0.12 (0.10)	0.89 (0.05)	0.04 (0.09)	0.38 (0.15)
cov		0.03 (0.03)	0.86 (0.16)	0.02 (0.03)	0.74 (0.19)	0.03 (0.04)	0.74 (0.22)	0.01 (0.02)	0.30 (0.42)
BH		0.18 (0.13)	0.97 (0.02)	0.18 (0.13)	0.97 (0.02)	0.39 (0.21)	0.96 (0.03)	0.39 (0.21)	0.96 (0.03)

Bibliography

Astrid M. Alsema, Qiong Jiang, Laura Kracht, Emma Gerrits, Marissa L. Dubbelaar, Anneke Miedema, Nieske Brouwer, Elly M. Hol, Jinte Middeldorp, Roland van Dijk, Maya Woodbury, Astrid Wachter, Simon Xi, Thomas Möller, Knut P. Biber, Susanne M. Kooistra, Erik W. G. M. Boddeke, and Bart J. L. Eggen. Profiling microglia from alzheimer's disease donors and non-demented elderly in acute human postmortem cortical tissue. *Frontiers in Molecular Neuroscience*, 13, 2020.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

Tallulah S. Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7(1740), 2019.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Rina Foygel Barber and Emmanuel J. Candès. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

Rina Foygel Barber, Emmanuel J. Candès, and Richard J. Samworth. Robust inference with knockoffs. *The Annals of Statistics*, 48(3):1409–1431, 2020.

Stephen Bates, Emmanuel J. Candès, Lucas Janson, and Wenshuo Wang. Metropolized knockoff sampling. *Journal of the American Statistical Association*, 116(535):1413–1427, 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Steven J. Benson, Yinyu Ye, and Xiong Zhang. Solving large-scale sparse semidefinite programs for combinatorial optimization. *SIAM Journal on Optimization*, 10(2):443–461, 2000.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111—-119, 2012.

Emmanuel J. Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98 (6):925–936, 2010.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Emmanuel J. Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Kristen Emery and Uri Keich. Controlling the fdr in variable selection via multiple knockoffs. *arXiv preprint arXiv:1911.09442*, 2019.

Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, 2019.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

Yingying Fan, Jinchi Lv, Mahrad Sharifvaghefi, and Yoshimasa Uematsu. IPAD: Stable interpretable forecasting with knockoffs inference. *Journal of the American Statistical Association*, 115(532):1822–1834, 2020.

Vivek Farias, Andrew A. Li, and Tianyi Peng. Uncertainty quantification for low-rank matrix completion with heterogeneous and sub-exponential noise. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 1179–1189, Virtual Conference, 2022.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, 2015.

Chao Gao, Hanbo Sun, Tuo Wang, Ming Tang, Nicolaas I. Bohnen, Martijn L. T. M. Müller, Talia Herman, Nir Giladi, Alexandr Kalinin, Cathie Spino, William Dauer, Jeffrey M. Hausdorff, and Ivo D. Dinov. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in parkinson's disease. *Scientific Reports*, 8(1):7129, 2018.

Xinzhou Ge, Yiling Elaine Chen, Dongyuan Song, MeiLu McDermott, Kyla Woyshner, Antigoni Manousopoulou, Ning Wang, Wei Li, Leo D. Wang, and Jingyi Jessica Li. Clipper: p-value-free fdr control on high-throughput data from two conditions. *Genome Biology*, 22(1):288, 2021.

Jaime Roquero Gimenez and James Zou. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In *Proceedings of the 22nd International*

Conference on Artificial Intelligence and Statistics, pages 2184–2192, Naha, Okinawa, Japan, 2019.

Jaime Roquero Gimenez, Amirata Ghorbani, and James Zou. Knockoffs for the mass: new feature importance statistics with false discovery guarantees. In *The 22nd international conference on artificial intelligence and statistics*, pages 2125–2133, Naha, Okinawa, Japan, 2019.

Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J. Garry. Drimpute: imputing dropout events in single cell rna sequencing data. *BMC Bioinformatics*, 19(1): 220, 2018.

Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, 2014.

Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. Imputing missing data for gene expression arrays. Technical report, Statistics Department, Stanford University, Stanford, CA, 1999.

Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

Zihuai He, Linxi Liu, Chen Wang, Yann Le Guen, Justin Lee, Stephanie Gogarten, Fred Lu, Stephen Montgomery, Hua Tang, Edwin K. Silverman, Michael H. Cho, Michael Greicius, and Iuliana Ionita-Laza. Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nature Communications*, 12(1):3152, 2021.

Zihuai He, Benjamin Chu, James Yang, Jiaqi Gu, Zhaomeng Chen, Linxi Liu, Tim Morrison, Michael E Belloy, Xinran Qi, Nima Hejazi, Maya Mathur, Yann Le Guen, Hua Tang, Trevor Hastie, Iuliana Ionita-Laza, Chiara Sabatti, and Emmanuel Candès. Beyond guilty by association at scale: searching for causal variants on the basis of genome-wide summary statistics. *bioRxiv*, 2024.

Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, Incorporated, Somerset, 2013.

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I. Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, 15(7):539–542, 2018.

Lucas Janson and Weijie Su. Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1):960–975, 2016.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(82):2869–2909, 2014a.

Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014b.

Jing Jiang, Cankun Wang, Ren Qi, Hongjun Fu, and Qin Ma. scread: A single-cell rna-seq database for alzheimer's disease. *iScience*, 23(11), 2020.

Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome Biology*, 23(1):31, 2022.

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks. In *International conference on learning representations*, 2018.

Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3):e694, 2022.

Arnav Kapur, Kshitij Marwah, and Gil Alterovitz. Gene expression prediction using low-rank matrix completion. *BMC Bioinformatics*, 17(1):243, 2016.

Zheng Tracy Ke, Jun S Liu, and Yucong Ma. Power of fdr control methods: The impact of ranking algorithm, tampered design, and symmetric statistic. *arXiv preprint arXiv:2010.08132*, 2020.

Peter V. Kharchenko, Lev Silberstein, and David T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. The technology and biology of single-cell rna sequencing. *Molecular Cell*, 58(4):610–620, 2015.

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E. Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M. Keizer, Indu Khatri, Szymon M. Kielbasa, Jan O. Korbel, Alexey M. Kozlov, Tzu-Hao Kuo, Boudewijn P.F. Lelieveldt, Ion I. Mandoiu, John C. Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J. Theis, Huan Yang, Alex Zelikovsky, Alice C. McHardy, Benjamin J. Raphael, Sohrab P. Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

Jun Li and Robert Tibshirani. Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data. *Statistical Methods in Medical Research*, 22(5): 519–536, 2013.

Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature Communications*, 9(1):997, 2018.

Jingbo Liu and Philippe Rigollet. Power analysis of knockoff filters for correlated designs. In *Advances in Neural Information Processing Systems*, pages 15446–15455, 2019.

Ying Liu and Cheng Zheng. Auto-encoding knockoff generator for fdr controlled variable selection. *arXiv preprint arXiv:1809.10765*, 2018.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413–468, 2014.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

Yang Lu, Yingying Fan, Jinchi Lv, and William Stafford Noble. Deeppink: reproducible feature selection in deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8676–8686, Montréal, Canada, 2018.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar. Mcimpute: Matrix completion based imputation for single cell rna-seq data. *Frontiers in Genetics*, 10, 2019.

Tuan-Binh Nguyen, Jerome-Alexis Chevalier, Bertrand Thirion, and Sylvain Arlot. Aggregation of multiple knockoffs. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7283–7293, Vienna, Austria, 2020.

Vasilis Ntranos, Lynn Yi, Páll Melsted, and Lior Pachter. A discriminative learning approach to differential expression analysis for single-cell rna-seq. *Nature Methods*, 16(2):163–166, 2019.

Zhimei Ren and Rina Foygel Barber. Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*, 2022.

Zhimei Ren, Yuting Wei, and Emmanuel J. Candès. Derandomizing knockoffs. *Journal of the American Statistical Association*, 118(542):948–958, 2023.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, 9(1):284, 2018.

Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 32, 2005.

Matteo Sesia, Chiara Sabatti, and Emmanuel J. Candès. Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1):1–18, 2019.

Matteo Sesia, Stephen Bates, Emmanuel Candès, Jonathan Marchini, and Chiara Sabatti. False discovery rate control in genome-wide association studies with population structure. *Proceedings of the National Academy of Sciences*, 118(40):e2105841118, 2021.

S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

Asher Spector and Lucas Janson. Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, 50(1):252–276, 2022.

Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

Weijie Su, Junyang Qian, and Linxi Liu. Communication-efficient false discovery rate control via knockoff aggregation. *arXiv preprint arXiv:1506.05446*, 2015.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, and M. Azim Surani. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Po-Yuan Tung, John D. Blischak, Chiaowen Joyce Hsiao, David A. Knowles, Jonathan E. Burnett, Jonathan K. Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, 7(1):39921, 2017.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

Asaf Weinstein, Rina Barber, and Emmanuel J. Candès. A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*, 2017.

Eric P Xing, Michael I Jordan, Richard M Karp, et al. Feature selection for high-dimensional genomic microarray data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 601–608, San Francisco, CA, USA, 2001.

Andrew C. Yang, Ryan T. Vest, Fabian Kern, Davis P. Lee, Maayan Agam, Christina A. Maat, Patricia M. Losada, Michelle B. Chen, Nicholas Schaum, Nathalie Khoury, Angus Toland, Kruti Calcuttawala, Heather Shin, Róbert Pálovics, Andrew Shin, Elizabeth Y. Wang, Jian Luo, David Gate, Walter J. Schulz-Schaeffer, Pauline Chu, Julie A. Siegenthaler, M. Windy McNerney, Andreas Keller, and Tony Wyss-Coray. A human brain vascular atlas reveals diverse mediators of alzheimer's risk. *Nature*, 603(7903):885–892, 2022.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Zifan Zhu, Yingying Fan, Yinfei Kong, Jinchi Lv, and Fengzhu Sun. Deeplink: Deep learning inference using knockoffs with applications to genomics. *Proceedings of the National Academy of Sciences*, 118(36):e2104683118, 2021.