False Discovery Rates in Biological Networks

Lu Yu
Department of Statistical Sciences
University of Toronto

Tobias Kaufmann
NORMENT, University of Oslo
Oslo University Hospital

Johannes Lederer
Department of Mathematics
Ruhr-University Bochum

Abstract

The increasing availability of data has generated unprecedented prospects for network analyses in many biological fields, such as neuroscience (e.g., brain networks), genomics (e.g., gene-gene interaction networks), and ecology (e.g., species interaction networks). A powerful statistical framework for estimating such networks is Gaussian graphical models, but standard estimators for the corresponding graphs are prone to large numbers of false discoveries. In this paper, we introduce a novel graph estimator based on knockoffs that imitate the partial correlation structures of unconnected nodes. We then show that this new estimator provides accurate control of the false discovery rate and yet large power.

1 Introduction

Biological processes can often be formulated as networks; examples include gene-gene regulation networks (Emmert-Streib et al., 2014; Hecker et al., 2009), functional brain networks (Bullmore and Sporns, 2009), and microbiome networks (Kurtz et al., 2015). A common statistical framework for such networks are Gaussian graphical models (Lauritzen, 1996). (Undirected) Gaussian graphical models describe the biological data as i.i.d. observations of a random vector $x := (x_1, \ldots, x_p)^\top$ that follows a multivariate normal distribution $N_p(0_p, \Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$ is a symmetric, positive definite matrix. The graph $G := (V, E)$ with node set $V := \{1, \ldots, p\}$ and edge set $E := \{(i, j) \in V \times V : i \neq j, \Sigma^{-1}_{ij} \neq 0\}$ then captures which pairs of the sample vector’s coordinates are dependent conditionally on all other coordinates: $x_i$ is conditionally independent of $x_j$ given all other coordinates of $x$ if and only if $(i, j) \in E$. For example, in modeling functional brain networks based on functional Magnetic Resonance Imaging (fMRI), $p$ is the number of brain regions under consideration, $x_i$ is the activity in the $i$th region, and the edge set $E$ denotes the directly connected pairs of regions.

A number of estimators for the edge set $E$ are known. Besides simplistic correlational approaches, popular estimators are neighborhood selection (Meinshausen and Bühlmann, 2006), which combines node-wise lasso estimates, and graphical lasso (Friedman et al., 2008; Yuan and Lin, 2007), which maximizes an $\ell_1$-penalized log-likelihood. These two estimators have been equipped with sharp prediction and estimation guarantees even for high-dimensional settings, where the number of samples is not much larger than the number of nodes $p$ (Ravikumar et al., 2011; Rothman et al., 2008; Zhuang and Lederer, 2018). In contrast to such prediction and estimation results, what is less well understood for high-dimensional Gaussian graphical models is inference.

Our objective is inference in terms of control over the false discovery rate (FDR), which is the expected proportion of falsely selected edges over all selected edges. Such control can make network estimation more reliable, which is particularly useful in biology as many biological networks seem to be hard to unravel—see (Zhang et al., 2018) for corresponding comments regarding brain imaging, for example. Formally, the FDR is defined as

$$
\text{FDR} := \mathbb{E}[\text{FDP}],
$$

where

$$
\text{FDP} := \frac{\#\{(i, j) : (i, j) \notin E \text{ and } (i, j) \in \hat{E}\}}{\#\{(i, j) : (i, j) \in \hat{E}\} \lor 1}
$$

is the false discovery proportion for an estimator that returns the edge set $\hat{E} \subset V \times V$, and $a \lor b := \max\{a, b\}$. We say that an estimator controls the FDR at level $q$ if $\text{FDR} \leq q$. In the language of hypothesis testing, FDR
control is the adjustment to multiple testing for the hypotheses $\mathcal{H}_{ij} : \Sigma^{-1}_{ij} = 0$ for $i \neq j$.

We establish an estimator based on knock-offs. In a regression-type setting, knock-offs are “fake predictors” that allow one to approximately count the number of falsely included variables (Barber and Candès, 2015; Candès et al., 2018; Dai and Barber, 2016). The knock-offs are supposed to maintain the original features’ correlation structure but to be only weakly correlated with the original features. Since the relevant predictors tend to have stronger association with the response than their knock-off counterparts, the number of falsely included variables can be approximated by comparing the estimated signals of the original predictors and their knock-off counterparts. In a graphical model setting, we introduce knock-offs as “fake edges.” Rather than maintaining correlation structures among the original nodes, they mimic partial correlations between separate, conditionally independent pairs of nodes. We then compare the signals of the sample partial correlations and their knock-off counterparts.

Our contributions can be summarized as follows.

1. We introduce a method for FDR control in graphical modeling, where inferential methods have been scarce: we compare the existing popular graph estimation methods and show their limitations on the FDR control.

2. We further support our method both mathematically and numerically: we establish theoretical guarantee for both approximate and exact FDR control in Theorems 3.1 and 3.2; and in Section 3.3, we demonstrate the proposed method achieves the FDR control and yields higher power than other popular graph estimation methods.

3. We apply the proposed method to three biological network data sets, and show in Section 4 that our method provides new insights into biological data.

We provide a free implementation that can be applied to networks within and beyond the exemplified domains on https://github.com/LedererLab/GGM-FDR.

**Related literature**  Drton and Perlman (2004) provides conservative simultaneous confidence intervals for the elements of the precision matrix $\Sigma^{-1}$ in Gaussian graphical models. van der Laan et al. (2004) studies the tail probability of the proportion of false positives via the family-wise error rate to obtain asymptotic FDR control in $n \to \infty$. Liu (2013) uses a multiple testing framework about conditional independence to obtain asymptotic FDR control in $n, p \to \infty$. Jankova and van de Geer (2015) establishes element-wise confidence intervals for $\Sigma^{-1}$.

**Outline of the paper**  The rest of this paper is organized as follows. In Section 2, we demonstrate that new methodology is indeed needed for FDR control in Gaussian graphical models. In Section 3, we introduce our approach and prove its effectiveness both mathematically and numerically. In Section 4, we apply our pipeline to three biological network data sets. In Section 5, we conclude with a discussion. All the proofs are deferred to the supplement.

**2 Motivation**

![Figure 1: FDR and power for GLASSO, MB(and), MB(or), CT, and PT as functions of the tuning parameters. None of the five methods provides a tuning parameter that leads to both small FDR and large power, and in any case, it is not clear how to calibrate the tuning parameters accordingly in practice.](https://example.com/figure1.png)
and moderately dense. Finally, 20 independent data sets with each one consisting of $n = 800$ independent samples from $\mathcal{N}_p(0, \Sigma)$ are generated.

Using again the **huge** package, the estimators are computed along a fine grid of tuning parameters. The estimators’ accuracy is evaluated in terms of FDR—see (1)—and in terms of power

$$\text{Power} := \frac{\# \left\{ (i, j) : (i, j) \in \mathcal{E} \text{ and } (i, j) \in \mathcal{E} \right\}}{\# \left\{ (i, j) : (i, j) \in \mathcal{E} \right\} \lor 1},$$

which is the proportion of the number of correctly estimated edges to the total number of edges.

Both FDR and power are averaged over the 20 data sets.

Figure 1 contains the FDR/power-curves along the tuning parameter paths. There is not necessarily a tuning parameter that leads to small FDR and large power simultaneously. And more importantly, FDR and power can be measured in simulations but not in practice; this means that even if there was a tuning parameter that leads to small FDR and large power, it would be unclear how to find it in practice. In particular, known calibration schemes such as cross-validation (Arlot and Celisse, 2010), AIC (Akaike, 1974), BIC (Schwartz, 1978), permutation (Sabourin et al., 2015), and AV (Chichignoud et al., 2016) are designed for different objectives and are, therefore, not suitable for this task. Taken together, standard estimators for Gaussian graphical models do not imply sensible FDR control.

### 3 Method

In this section, we introduce our strategy to FDR control and establish both mathematical and numerical support for its accuracy. A main ingredient of our strategy are knockoffs that imitate additional partial correlations. Accordingly, we refer to our method as “KO.”

#### 3.1 The KO Strategy

The KO strategy consists of three steps: First, we equip the sample partial correlations with knock-off counterparts. Second, we compare the sample partial correlations and their counterparts through corresponding test statistics. Third, we produce estimates based on these test statistics by defining a data-driven threshold.

The three mentioned steps now read in detail:

**Step 1: Constructing knock-offs.**

The starting points of our statistical analysis are the partial correlations. The partial correlations give us support for its accuracy. A main ingredient of our strategy are knockoffs that imitate additional partial correlations. Accordingly, we refer to our method as “KO.”

$$x_i \perp x_j | \mathbf{x}_{\mathcal{V}\setminus\{i,j\}} \iff \Sigma^{-1}_{ij} = 0 \iff \rho_{ij, \mathcal{V}\setminus\{i,j\}} = 0,$$

where $\rho_{ij, \mathcal{V}\setminus\{i,j\}}$ denotes the partial correlation between the variables $x_i$ and $x_j$ given the remaining $p-2$-dimensional vector $\mathbf{x}_{\mathcal{V}\setminus\{i,j\}}$.

We now use classical properties of sample correlations and sample partial correlations derived by Fisher (1915, 1921, 1924). Consider the data matrix $X = (x^1, \ldots, x^n)^\top \in \mathbb{R}^{n \times p}$, where $x^1, \ldots, x^n \in \mathbb{R}^p$ are independent and identically distributed samples from $\mathcal{N}_p(0, \Sigma)$ and assume that $n > p$. (The latter condition does not exclude high-dimensional settings in general: $n \approx p$ cannot be approached with classical inferential methods, and the number of parameters in graphical models is $p(p-1)/2$, which can be much larger than $n$ even for $p < n$.) Fisher (1915) derives the distribution of the sample correlation

$$C_{ij} := \frac{\sum_{l=1}^{n} (x^l_i) (x^l_j)}{\sqrt{\sum_{l=1}^{n} (x^l_i)^2} \sqrt{\sum_{l=1}^{n} (x^l_j)^2}},$$

of the coordinates $i$ and $j$; in particular, that paper yields that if the population correlation is zero, the statistic

$$\frac{C_{ij}}{\sqrt{(1 - (C_{ij})^2)/(n - 2)}}$$

follows a Student’s $t$-distribution with $n - 2$ degrees of freedom. Fisher (1924) then shows that the corresponding sample partial correlation, which we write as an entry

$$R_{ij} := -\frac{(\hat{\Sigma}^{-1})_{ij}}{\sqrt{(\hat{\Sigma}^{-1})_{ii} (\hat{\Sigma}^{-1})_{jj}}},$$

of a matrix $R \in \mathbb{R}^{p \times p}$ with $\hat{\Sigma} := X^\top X$ the sample covariance matrix of the data matrix $X$, has the same distribution as $C_{ij}$ but with an effective sample size of $n - (p-2)$, where $p - 2$ is the number of elements in $\mathcal{V}\setminus\{i,j\}$. We can, therefore, conclude that assuming the null-hypothesis $H_{(i,j)} : \Sigma_{ij}^{-1} = 0$, then the random variable

$$Z := \frac{R_{ij}}{\sqrt{(1 - R_{ij}^2)/(n - p)}}$$

follows a Student’s $t$-distribution with $n - 2 - (p-2) = n - p$ degrees of freedom.
Motivated by the above observations, we define the entries of $R^o$ through

$$R^o_{ij} := R^2_{ij} := \begin{cases} 1 & \text{if } i = j \\ \frac{Z_{ij}}{\sqrt{n-p+Z^2_{ij}}} & \text{if } i \neq j \end{cases}, \quad (4)$$

where the $Z_{ij}$'s $(i, j \in \{1, \ldots, p\})$ are sampled independently from the Student’s t-distribution with $n-p$ degrees of freedom. These are our knockoff versions of the sample partial correlations: each element of this matrix mimics sample partial correlations between two conditionally independent nodes. The diagonal elements of $R^o$ are set to 1 to equal the diagonal elements of $R$; the off-diagonal elements of $R^o$ are in $(-1, 1)$.

**Step 2: Establishing the Test Statistics.** We now construct the test statistics for the entries of the sample partial correlation matrix $R$ and its knock-off counterpart $R^o$. We first apply elementwise hard-thresholding, which can be written as penalized empirical risk minimization

$$\hat{R}(t) \in \arg\min_{A \in S} \{|R - A|^2 + t^2 |A|_0\}, \quad (5)$$

where $t > 0$ is the thresholding parameter and $S$ is the set of symmetric and invertible matrices in $\mathbb{R}^{p \times p}$.

(Our pipeline also applies to soft-thresholding, which corresponds to the $\ell_0$-term swapped with an $\ell_1$-term, and other estimators, but to avoid digression, we omit the details.) The knock-off version of that estimator is

$$\hat{R}^o(t) \in \arg\min_{A \in S} \{|R^o - A|^2 + t^2 |A|_0\}. \quad (6)$$

We now use those estimators to quantify the signal strengths. We define the test statistics matrix $\hat{T}$ via

$$\hat{T}_{ij} := \sup \left\{ t : \left(\hat{R}(t)\right)_{ij} \neq 0 \right\}, \quad (7)$$

which is the point on the tuning parameter path (ranging from $+\infty$ to 0) at which the sample partial correlation between $x_i$ and $x_j$ controlling for other variables first enters the model. The test statistic $\hat{T}_{ij}$ indeed tends to be large if $R_{ij}$ (and, therefore, its underlying population versions $\rho_{ij}, \forall \{(i,j)\}$) are large. Similarly, we can evaluate the signal strength of $R^o_{ij}$ via

$$\hat{T}^o_{ij} := \sup \left\{ t : \left(\hat{R}^o(t)\right)_{ij} \neq 0 \right\}. \quad (8)$$

Large values of $\hat{T}_{ij}$ provide evidence against $\mathcal{H}_{(i,j)} : \Sigma_{ij}^{-1} = 0$, while large values of $\hat{T}^o_{ij}$ provide evidence for $\mathcal{H}_{(i,j)} : \Sigma_{ij}^{-1} = 0$; thus, the larger $\hat{T}_{ij}$ in comparison to $\hat{T}^o_{ij}$, the more confidently we can reject $\mathcal{H}_{(i,j)} : \Sigma_{ij}^{-1} = 0$.

For a detailed assessment of the signal strengths, we construct the matrix-valued test statistics $\hat{W} \in \mathbb{R}^{n \times p}$ via

$$\hat{W}_{ij} := \hat{W}_{ji} := \begin{cases} (\hat{T}_{ij} \lor \hat{T}^o_{ij}) \cdot \text{sign}(\hat{T}_{ij} - \hat{T}^o_{ij}) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}. \quad (9)$$

The test matrix $\hat{W}$ depends on $R$ and $R^o$ through $\hat{T}_{ij}$ and $\hat{T}^o_{ij}$. A positive $\hat{W}_{ij}$ states that the edge $(i,j)$ enters the model before its knock-off counterpart; more generally, the larger $\hat{W}_{ij}$, the more evidence we have against the hypothesis $\mathcal{H}_{(i,j)} : \Sigma_{ij}^{-1} = 0$.

**Step 3: Defining a Data-dependent Threshold.**

According to the previous step, large $\hat{W}_{ij}$ provide evidence against $\mathcal{H}_{(i,j)} : \Sigma_{ij}^{-1} = 0$. In this step, we quantify this by defining a data-driven threshold $\hat{t}$ and selecting the edges $(i,j)$ with $\hat{W}_{ij} \geq \hat{t}$, which yields the estimated edge set $\hat{E} = \{(i,j) \in V \times V : \hat{W}_{ij} \geq \hat{t}\}$. Given a target FDR level $q$, the threshold is defined as

$$\hat{t} := \min \left\{ t \in \hat{W} : \frac{\#{\{(i,j) : \hat{W}_{ij} \leq -t\}}} {\#{\{(i,j) : \hat{W}_{ij} \geq t\}} \lor 1} \leq q \right\}, \quad (10)$$

where $\hat{W} := \{|\hat{W}_{ij}| : i, j \in \{1, \ldots, p\}\} \setminus \{0\}$. We set $\hat{t} := \infty$ if the minimum is taken over the empty set. The minimum is always attained as $\hat{W}$ is finite.

Generally, our thresholding scheme aims at bounding the FDR by bounding an “empirical version” of it. According to Lemma B.3 in the supplement, it holds for the statistics matrix $\hat{W}$ defined in (9), any edge set $E$ that satisfies $E = E' := \{(i, j) \in V \times V : i \neq j, x_i \notin x_j\}$, and any threshold $t \geq 0$ that

$$\#\{(i,j) : (i,j) \notin E', \hat{W}_{ij} \leq -t\} =_{d} \#\{(i,j) : (i,j) \notin E', \hat{W}_{ij} \geq t\},$$

where $=_{d}$ means equivalence in distribution. Using this equivalence and that an edge $(i,j)$ is selected if and only if $\hat{W}_{ij} \geq t$, we can approximately bound FDP($t$), which we define as the FDP for our pipeline with threshold $t$ as

$$\text{FDP}(t) = \frac{\#{\{(i,j) : (i,j) \notin E', \hat{W}_{ij} \geq t\}}} {\#{\{(i,j) : \hat{W}_{ij} \geq t\}} \lor 1}$$

$$= \frac{\#{\{(i,j) : (i,j) \notin E', \hat{W}_{ij} \geq t\}}} {\#{\{(i,j) : \hat{W}_{ij} \geq t\}} \lor 1} \approx \frac{\#{\{(i,j) : (i,j) \notin E', \hat{W}_{ij} \leq -t\}}} {\#{\{(i,j) : \hat{W}_{ij} \geq t\}} \lor 1}$$

$$\leq \frac{\#{\{(i,j) : \hat{W}_{ij} \leq -t\}}} {\#{\{(i,j) : \hat{W}_{ij} \geq t\}} \lor 1} := \text{FDP}(t).$$
We interpret $\hat{FDP}(t)$ as an estimate of the FDR. One can check readily that
\[
\hat{t} = \min \left\{ t \in \hat{W} : \hat{FDP}(t) \leq q \right\}
\]
(and set $\hat{t} := \infty$ if no such $t$ exists), which means that our data-driven threshold $\hat{t}$ controls an empirical version of the FDR.

We show in the next section that the above scheme provides approximate FDR control. If exact FDR control is required, one can modify the scheme similarly as in mimic Barber and Candès (2015) by thresholding the numerator, which can make the threshold slightly larger (see Section A in the supplement for some intuition). We call the pipeline of Section 3.1 with $\hat{t}$ replaced by $\hat{t}+$ the $KO+$ scheme. In practice, however, we would typically recommend the KO scheme, as it has higher statistical power.

### 3.2 Mathematical Support

We now support our method mathematically. We first state the following (all proofs are deferred to the supplementary materials):

**Theorem 3.1** (Approximate FDR control). For any target level $q \in [0,1]$, the KO scheme established in Section 3.1 satisfies
\[
E \left[ \frac{\# \{ (i,j) : (i,j) \notin \mathcal{E}' \text{ and } (i,j) \in \hat{E}' \} + q^{-1}}{\# \{ (i,j) : (i,j) \in \hat{E}' \}} \right] \leq q.
\]

This bound establishes an FDR-type guarantee. The left-hand side differs from the FDR in (1) and (2) in two aspects, though: First, it contains an additional $q^{-1}$ in the denominator. But this difference is negligible unless the number of selected edges is very small, and it can even be removed by applying a more conservative threshold (see supplementary materials). Second, it contains $\mathcal{E}'$ rather than $\mathcal{E}$, that is, it concerns correlations rather than partial correlations. Since $\mathcal{E}'$ can be considerably larger than $\mathcal{E}$, this means that the theorem cannot guarantee FDR control in general. But still, it can serve as a first mathematical witness for the potency of our approach.

And this potency is confirmed in simulations indeed. The simulation setup is the one of Section 2. In addition, the number of samples $n$ and the number of parameters $p$ is varied, and our KO method is evaluated on a fine grid of target FDR levels. Recall that the setup involves a band graph, where $\mathcal{E}' \gg \mathcal{E}$. Hence, in view of the above theory (which does not apply to such cases), good results in this setup would give a particularly strong argument for our method.

In addition, we can also guarantee exact FDR control for the KO+ scheme:

**Theorem 3.2** (Exact FDR Control). For any target level $q \in [0,1]$, and $\mathcal{E} = \mathcal{E}'$, the KO+ scheme satisfies
\[
\text{FDR} = E \left[ \frac{\# \{ (i,j) : (i,j) \notin \mathcal{E} \text{ and } (i,j) \in \hat{E}' \}}{\# \{ (i,j) : (i,j) \in \hat{E}' \} + 1} \right] \leq q.
\]
4 Real Data Analyses

We now demonstrate the utility of our proposed knock-off method in uncovering biological networks. We give three examples: brain connectivity networks, microbiological networks in the human gut, and abundance networks of amphibians. The target FDR level is set to 0.2 across all analyses.

4.1 Brain Connectivity Analysis

Functional Magnetic Resonance Imaging (fMRI) is a powerful tool to unveil the brain’s functional interdependence structures. The data at hand, described and analyzed in Bu and Lederer (2017), consists of resting-state fMRI acquired at the Department of Neurology at Beijing Hospital from April 2012 through December 2013. The data set comprises \( n = 210 \) samples of the average voxel intensities in \( p = 116 \) anatomical volumes in \( n_{\text{NC}} = 10 \) individuals with normal cognition. In line with earlier work (Horwitz et al., 1987; Huang et al., 2010), we restrict our focus to 42 anatomical volumes, further referred to as regions of interest (ROI). The 42 ROIs are located in the frontal lobe, parietal lobe, occipital lobe, and temporal lobe.

Since we have the data of \( n_{\text{NC}} = 10 \) subjects, we can complement our pipeline with the multiple FDR scheme introduced in (Xie and Lederer, 2019) with target FDR level \( 0.2 \times 0.5^k \) for the \( k \)-th individual, \( k \in \{1, \ldots, 10\} \). We then obtain the continuous graph estimates \( \hat{R}_{ij}(t) \) for each individual \( k \), which is denoted by \( \hat{R}_{ij}^k(t) \). Then, we calculate the scaled cumulative signal strengths as

\[
\sum_{k \in \text{group}} |\hat{R}_{ij}^k(t)| / \max_{l,m} \left\{ \sum_{k \in \text{group}} |\hat{R}_{lm}^k(t)| \right\}.
\]

The scaled cumulative signal strengths are displayed in Figure 6. The plot demonstrates that strong connections are predominately between the left and right counterparts of a given region, which is in line with earlier work on the functional network architecture of the brain (Honey et al., 2009).

4.2 Human Microbiome Analysis

We now apply the knock-off method to the human microbiome data set of the American Gut Project (http://humanfoodproject.com/americangut/). Our specific goal is to learn how the microbiome is associated with smoking. We use the processed data that were collected before December, 2018. We classify the individuals with smoking frequencies Daily, Occasionally (1-2 times/week), Regularly (3-5 times/week), and Rarely (a few times/month) as smokers and the ones with smoking frequency Never as non-smokers. This yields \( n_{\text{smoker}} = 1234 \) smokers and \( n_{\text{non-smoker}} = 15640 \) non-smokers. We incorporate the centered log-ratio transformed (Aitchison, 1982) abundances of the \( p = 32 \) phyla that appear in at least 5% of the individuals.

To reduce the influence of the imbalanced samples sizes, we again add the multiple FDR scheme of (Xie and Lederer, 2019) to our method. Specifically, we uniformly subsample \( n_{\text{sample}} = 1234 \) individuals from the non-smoker group 10 times. At each time \( k \in \{1, \ldots, 10\} \), we apply the knock-off method to the corresponding \( n_{\text{sample}} \times p \)-dimensional data set with target FDR level \( 0.2 \times 0.5^k \). Finally, we calculate the scaled cumulative signal strengths, as showed in Figure 7. The smoker group’s data is treated with the vanilla version of our scheme from Section 3.1.

We find strong evidence that the graphs of the smokers and non-smokers differ in their connectivities: the \( p \)-value of a corresponding Wilcoxon signed-rank test is \( \ll 10^{-10} \). In fact—see also the visualization in Figure 5—we find that there are much more interactions in the non-smokers’ guts, which is in agreement with findings in the literature (Biedermann et al., 2013; Savin et al., 2018; Stewart et al., 2018).

4.3 Atlantic Amphibians Abundance Analysis

We finally analyze abundance data from the Atlantic Forest Biome in South America (Vaccine et al., 2018). We specifically consider the \( p = 30 \) most abundant...
endemic (occurring uniquely in Atlantic Forest) and \( p = 30 \) most abundant non-endemic species of the order Anura. This ensures that the species appear in at least 0.9% of the observations.

The corresponding number of study sites for which species abundances are fully documented is \( n = 346 \). Again, we apply the centered log-ratio transformation to the data. We find strong evidence for differences in the connectivities of the graphs of the two groups: the p-value of a corresponding Wilcoxon signed-rank test is \( \ll 10^{-5} \). There are more interactions between the endemic species than between the non-endemic species, that is, abundances of endemic species are more interconnected among the different species. See also Figure 4, which visualizes the scaled connectivity estimates \( |\hat{R}_{ij}(t)| / \max_{l,m} \{|\hat{R}_{lm}(t)|\} \) from our pipeline. Since the total number of endemic and non-endemic species is comparable, we hypothesize that this difference is due to a higher level of adaptation of endemic species. This is in line with with Gorman et al. (2014), which indicates that endemic plants have an increased level of adaptation. However, to the best of our knowledge, our result is the first rigorous quantitative formulation of such a difference between endemic and non-endemic species.

5 Discussion

We have shown that our KO pipeline provides effective FDR control and that it can provide new insights into biological networks.

A topic for further research is the theory: We provide first theoretical insights in Section 3.3, but current theorems do yet not establish exact FDR control in general. However, our numerical results suggest that our theory can be sharpened accordingly. (In a paper that appeared after ours, Li and Maathuis (2019) were able to establish a general theory, but their method is different and computationally much more demanding.)

Another topic for further research are extensions to \( p > n \) along the lines of Candès et al. (2018). Our methodology applies very generally otherwise; in particular, it applies to arbitrary covariance matrices \( \Sigma \) and asymptotically even to non-Gaussian data.

In summary, its simplicity and convincing performance make our pipeline useful for a wide range of applications.

6 Acknowledgements

We sincerely thank Jinzhou Li for his insightful comments and the inspiring discussions.
Figure 5: Signal strengths for endemic species and non-endemic species in the Atlantic Forest Biome. The difference between the two plots in their numbers of gray cells indicates that there are more connections among endemic species than among non-endemic species.

Figure 6: Cumulative signal strengths for smoker group and cumulative signal strengths for non-smoker group. The graphs show there are more connections among the gut microbiome for non-smokers than for smokers.
Figure 7: Cumulative signal strength across $n_{NC} = 10$ individuals for connections among the 42 ROIs. The four red squares highlight the intra-lobe connections. The graph shows that strong connections are most common between regional counterparts in the left and right hemisphere.
Bibliography


