Supporting Information for “A direct approach to estimating false discovery rates conditional on covariates”

Simina M. Boca and Jeffrey T. Leek

June 19, 2017

1 Proofs of analytical results

Proof of Theorem 3

$$\begin{aligned} E[Y_i|\mathbf{X}_i=\mathbf{x}_i] &= Pr(P_i>\lambda|\mathbf{X}_i=\mathbf{x}_i)\\ &= Pr(P_i>\lambda|\theta_i=1,\mathbf{X}_i=\mathbf{x}_i)P(\theta_i=1|\mathbf{X}_i=\mathbf{x}_i)\\ &+ Pr(P_i>\lambda|\theta_i=0,\mathbf{X}_i=\mathbf{x}_i)P(\theta_i=0|\mathbf{X}_i=\mathbf{x}_i). \end{aligned}$$

Then, using the assumption that conditional on the null, the p-values do not depend on the covariates:

$$\begin{aligned} E[Y_i|\mathbf{X}_i=\mathbf{x}_i] &= Pr(P_i>\lambda|\theta_i=1)P(\theta_i=1|\mathbf{X}_i=\mathbf{x}_i)\\ &+ Pr(P_i>\lambda|\theta_i=0)P(\theta_i=0|\mathbf{X}_i=\mathbf{x}_i)\\ &= (1-\lambda)\pi_0(\mathbf{x}_i)+\{1-G(\lambda)\}\{1-\pi_0(\mathbf{x}_i)\}. \end{aligned}$$

Proof of Corollary 4

Applying the law of iterated expectations:

$$E[Y_i]=E[E[Y_i|\mathbf{X}_i]]=(1-\lambda)E[\pi_0(\mathbf{X}_i)]+\{1-G(\lambda)\}\{1-E[\pi_0(\mathbf{X}_i)]\}.$$

We complete the proof by using:

$$\begin{aligned} \pi_0 &= Pr(\theta_i=1)=\int Pr(\theta_i=1,\mathbf{X}_i=\mathbf{x})d\nu(\mathbf{x})\\ &= \int Pr(\theta_i=1|\mathbf{X}_i)dF_{\mathbf{X}_i}=E[Pr(\theta_i=1|\mathbf{X}_i)]=E[\pi_0(\mathbf{X}_i)], \end{aligned}$$

where ν is typically either the Lebesgue measure over a subset $\mathbb{R}$ or the counting measure over a subset of $\mathbb{Q}$, and $F_{\mathbf{X}_i}$ is the cumulative distribution function for $\mathbf{X}_i$. Here we are implicitly assuming some distribution for $\mathbf{X}_i$ as well. Everywhere else we are conditioning on $\mathbf{X}$.

Proof of Result 6

We prove this result by showing that:

$$E[(\hat{\pi}_0(\mathbf{x}_i)-\pi_0(\mathbf{x}_i))^2|\hat{\pi}_0(\mathbf{x}_i)>1]>E[(\hat{\pi}_0(\mathbf{x}_i)^C-\pi_0(\mathbf{x}_i))^2|\hat{\pi}_0(\mathbf{x}_i)>1] \quad (1)$$

and:

$$E[(\hat{\pi}_0(\mathbf{x}_i)-\pi_0(\mathbf{x}_i))^2|\hat{\pi}_0(\mathbf{x}_i)<0]>E[(\hat{\pi}_0^C(\mathbf{x}_i)-\pi_0(\mathbf{x}_i))^2|\hat{\pi}_0(\mathbf{x}_i)<0]. \quad (2)$$

Then, we can combine them as follows:

$$\begin{aligned} & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2] - E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2] = \\ = \; & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] - E[(\hat{\pi}_0(\mathbf{x}_i)^C - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] P(\hat{\pi}_0(\mathbf{x}_i) > 1) \\ + \; & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] - E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] P(\hat{\pi}_0(\mathbf{x}_i) < 0) \\ \geq \; & 0. \end{aligned}$$

In Eq. (1):

$$\begin{aligned} & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] - E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) > 1] = \\ = \; & E[(\hat{\pi}_0(\mathbf{x}_i) - 1)(\hat{\pi}_0(\mathbf{x}_i) + 1 - 2\pi_0(\mathbf{x}_i)) | \hat{\pi}_0(\mathbf{x}_i) > 1] > 0, \end{aligned}$$

because in this region $\hat{\pi}_0(\mathbf{x}_i) + 1 > 2 \geq 2\pi_0(\mathbf{x}_i)$.
In Eq. (2):

$$\begin{aligned} & E[(\hat{\pi}_0(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] - E[(\hat{\pi}_0^C(\mathbf{x}_i) - \pi_0(\mathbf{x}_i))^2 | \hat{\pi}_0(\mathbf{x}_i) < 0] = \\ = \; & E[(a - \hat{\pi}_0(\mathbf{x}_i))(2\pi_0(\mathbf{x}_i) - \hat{\pi}_0(\mathbf{x}_i) - 0) | \hat{\pi}_0(\mathbf{x}_i) < 0] > 0, \end{aligned}$$

because in this region $2\pi_0(\mathbf{x}_i) \geq 0 > \hat{\pi}_0(\mathbf{x}_i)$.

2 Functions $\pi_0(\mathbf{x}_i)$ used in simulation scenarios

Below, we refer to scenarios I-IV, as in Figure 3:
In scenarios I-IV, the values of x_1 are equally spaced between 0 and 1, with the number of points being equal to m, the number of features considered.

- Scenario I: $\pi_0(x_1) = 0.9$
- Scenario II: $\pi_0(x_1) = \pi_{01}(x_1) + \pi_{02}(x_1) + 0.12\pi_{03}(x_1)$, where:

$$\pi_{01}(x_1) = \begin{cases} 1 \text{ if } 0 \leq x_1 \leq 0.5 \\ -4/1.96(x_1 + 0.2)(x_1 - 1.2) \text{ if } 0.5 < x_1 < 0.7 \\ 4/1.96 \times 0.45 \text{ if } 0.7 \leq x_1 \leq 1, \end{cases} \quad \pi_{02}(x_1) = \begin{cases} 0 \text{ if } 0 \leq x_1 < 0.7 \\ -2.5(x - 0.7)^2 \text{ if } 0.7 \leq x_1 \leq 1 \end{cases}$$

$$\pi_{03}(x_1) = \begin{cases} 0 \text{ if } 0 \leq x_1 \leq 0.1 \\ -(x - 0.1)^2 \text{ if } 0.1 < x_1 < 0.7 \\ -0.36 \text{ if } 0.7 \leq x_1 \leq 1. \end{cases}$$

- Scenario III:

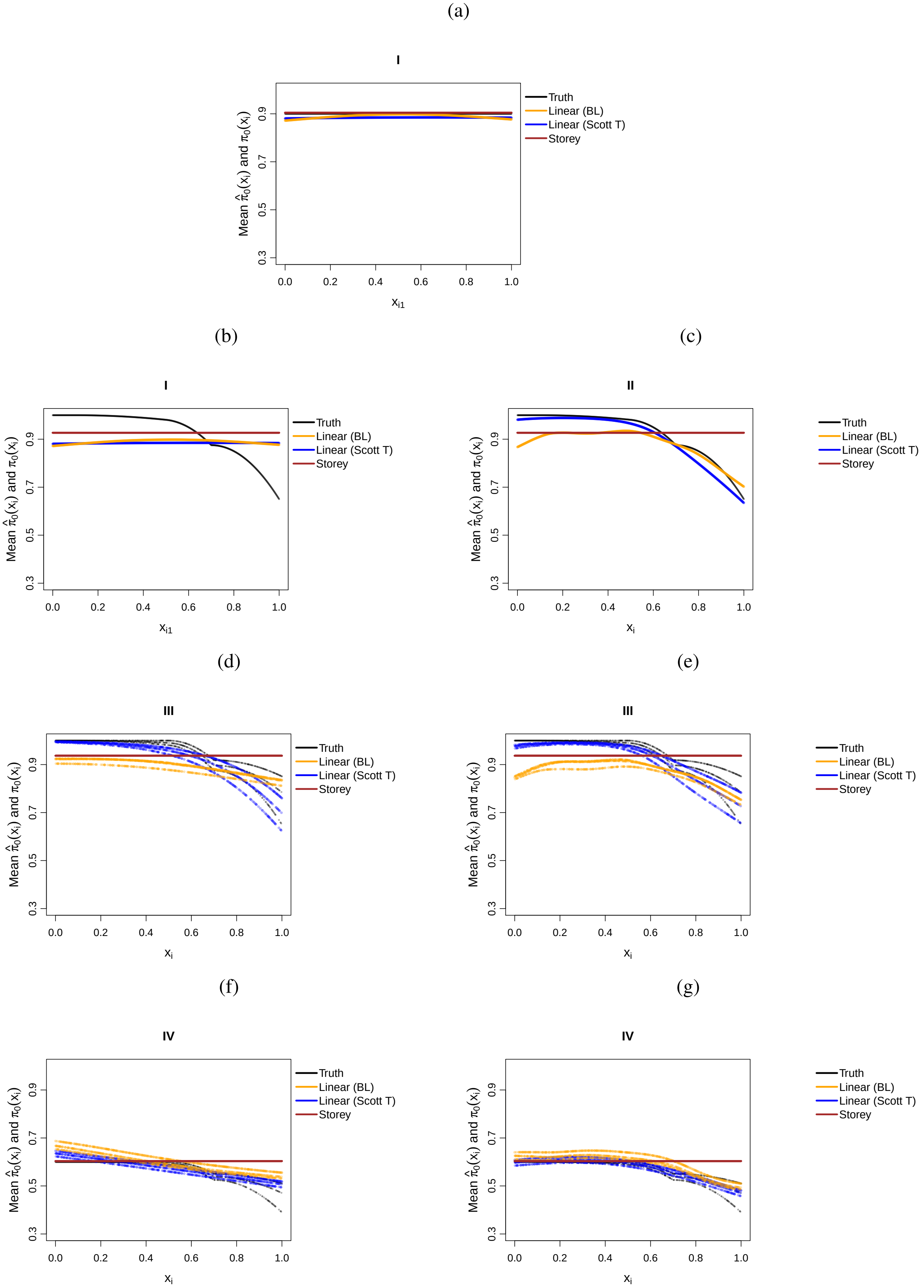
$$\pi_0(x_1, x_2) = \begin{cases} \pi_{01}(x_1) + \pi_{02}(x_1) + 0.12\pi_{03}(x_1) \text{ if } x_2 = 1 \\ \pi_{01}(x_1) + 0.5\pi_{02}(x_1) + 0.06\pi_{03}(x_1) \text{ if } x_2 = 2 \\ \pi_{01}(x_1) + 0.3\pi_{02}(x_1) \text{ if } x_2 = 3, \end{cases}$$

  where x_2 is defined by first randomly generating m points from Unif$(0, 0.5)$, then creating discrete categories by using the thresholds 0.127 and 0.302 and $\pi_{01}, \pi_{02}, \pi_{03}$ are defined as in Scenario II.

- Scenario IV: $\pi_0(x_1, x_2)$ is the same function as in scenario III multiplied by 0.6.

3 Supplementary figures

Figure S1: Simulation scenarios with m=1,000 features and normally-distributed independent test statistics (Table 3) showing the true function $\pi_0(\mathbf{x}_i)$ in black and the empirical means of $\hat{\pi}_0(\mathbf{x}_i)$, assuming different modelling approaches in the orange (for our approach, Boca-Leek = BL), blue (for the Scott approach with the theoretical null = Scott T), and brown for the Storey approach.



(a)

(b)

(c)

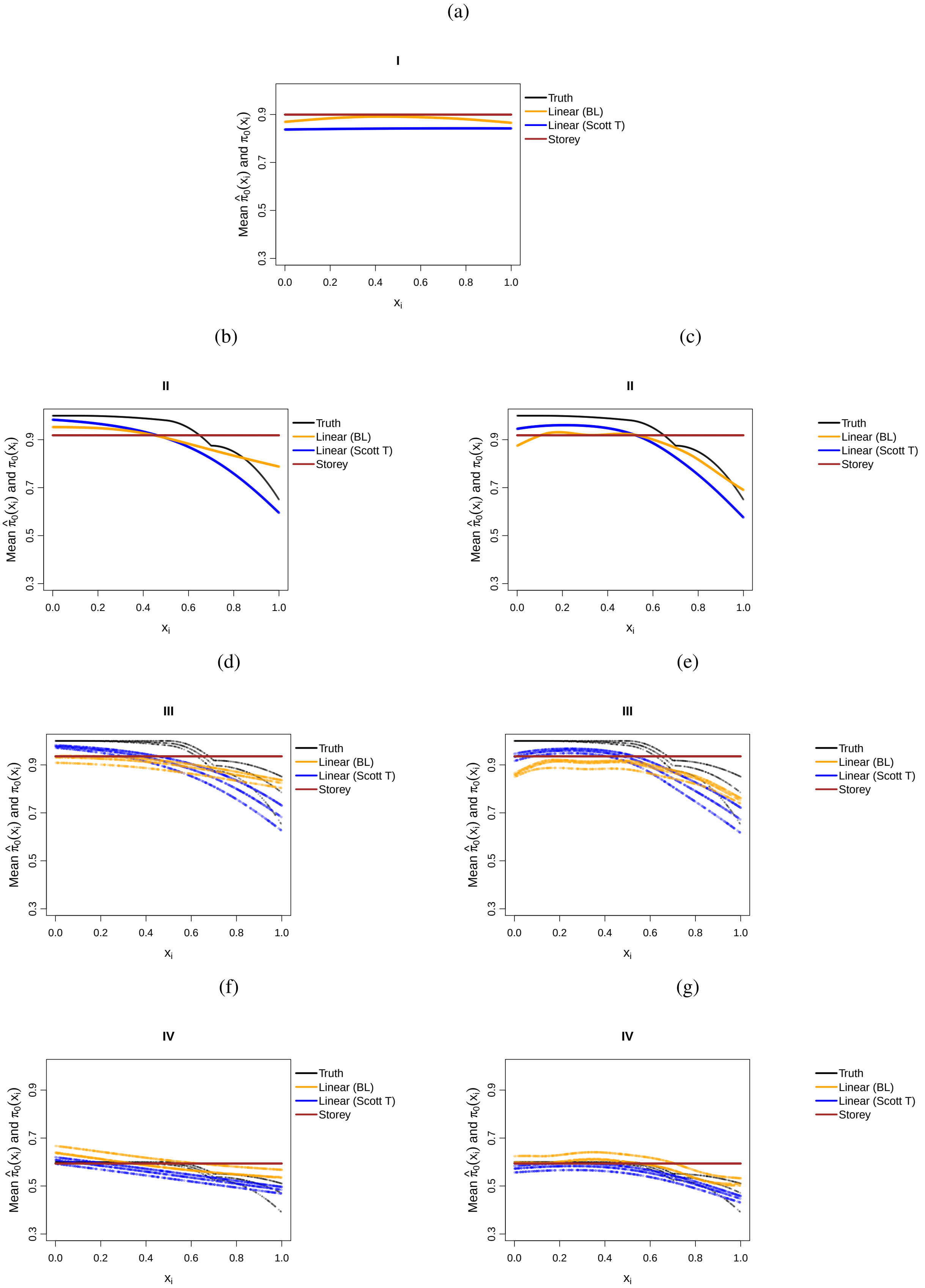
(d)

(e)

(f)

(g)

Figure S2: Simulation scenarios with m=1,000 features and t-distributed independent test statistics (Table 3) showing the true function $\pi_0(\mathbf{x}_i)$ in black and the empirical means of $\hat{\pi}_0(\mathbf{x}_i)$, assuming different modelling approaches in the orange (for our approach, Boca-Leek = BL), blue (for the Scott approach with the theoretical null = Scott T), and brown for the Storey approach.


(a)

(b)

(c)

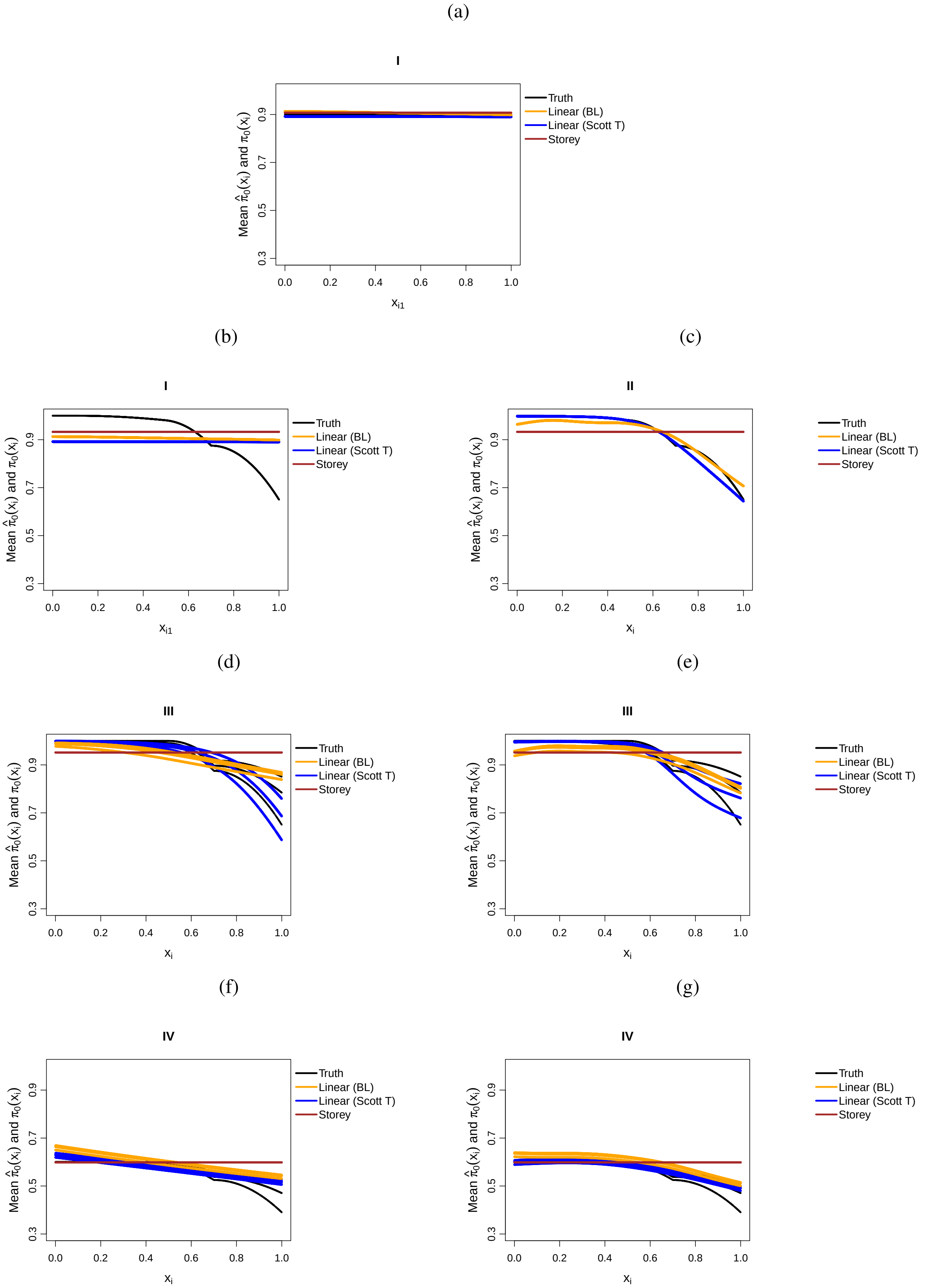
(d)

(e)

(f)

(g)

Figure S3: Simulation scenarios with m=10,000 features and normally-distributed independent test statistics (Table 4) showing the true function $\pi_0(\mathbf{x}_i)$ in black and the empirical means of $\hat{\pi}_0(\mathbf{x}_i)$, assuming different modelling approaches in the orange (for our approach, Boca-Leek = BL), blue (for the Scott approach with the theoretical null = Scott T), and brown for the Storey approach.


(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure S4: Simulation scenarios with m=10,000 features and t-distributed independent test statistics (Table 4) showing the true function $\pi_0(\mathbf{x}_i)$ in black and the empirical means of $\hat{\pi}_0(\mathbf{x}_i)$, assuming different modelling approaches in the orange (for our approach, Boca-Leek = BL), blue (for the Scott approach with the theoretical null = Scott T), and brown for the Storey approach.

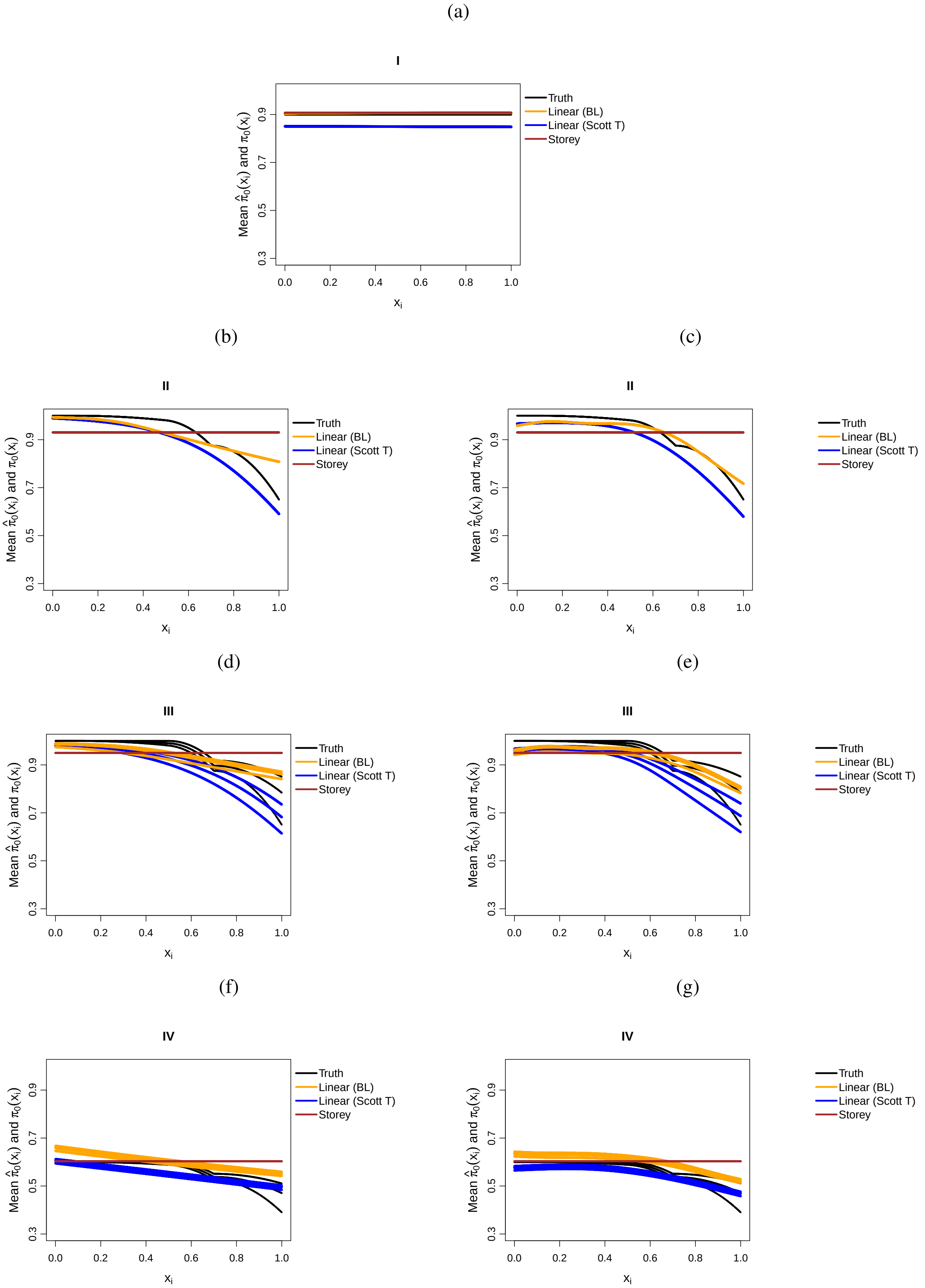
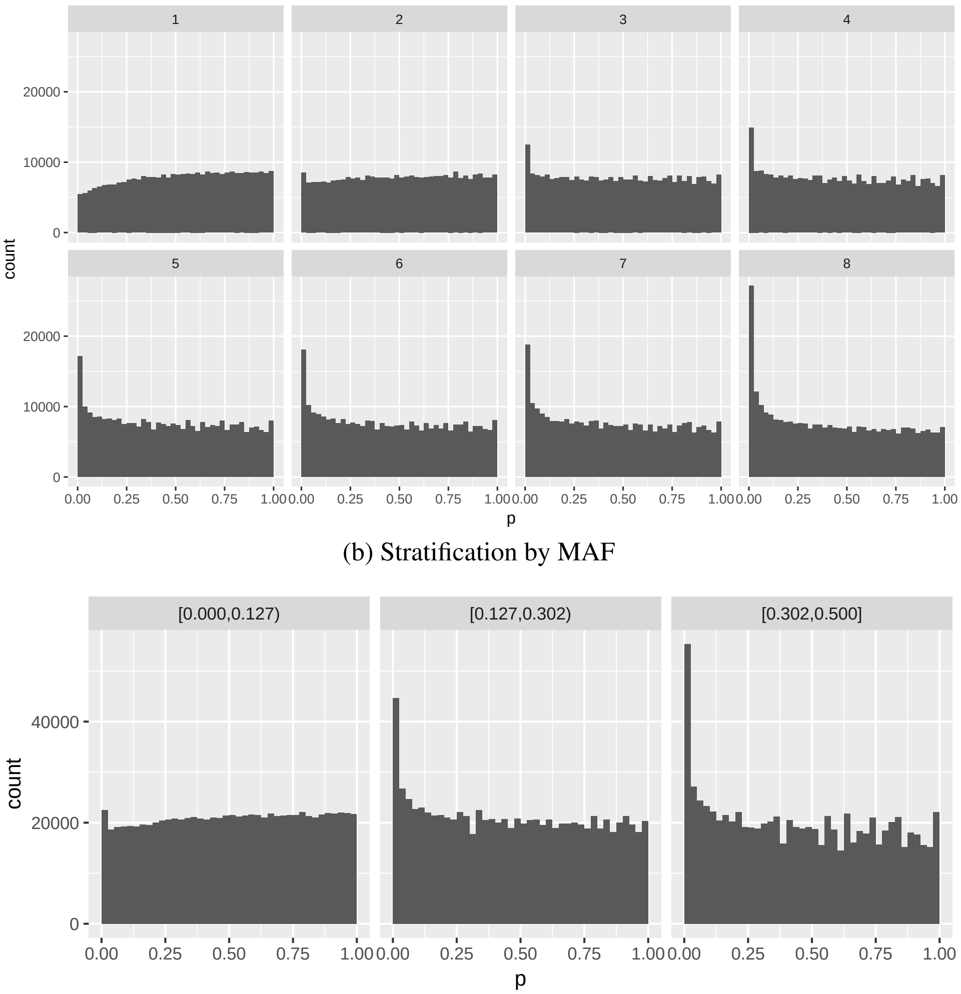


Figure S5: Diagnostic plots for assessing whether, in the BMI GWAS meta-analysis, the p-values and the covariates are conditionally independent under the null. Panel a) stratifies according to N, splitting up the dataset into 8 approximately equal datasets, panel b) uses the MAF stratification



(a) Stratification by N

(b) Stratification by MAF