Supporting Information for "A regression framework for the proportion of true null hypotheses"

Simina M. Boca and Jeffrey T. Leek

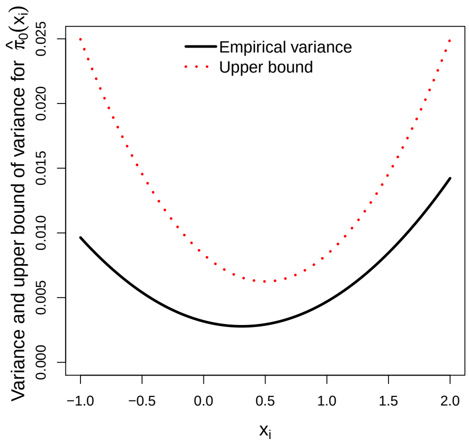
January 6, 2017

Table S1: Values of $max_{i=1}^{m} S_{ii}$ for the simulation scenarios in Figures 1 and S1, using the panel labels in Figure S1, for different values for the number of features, m, and $\lambda = 0.8$. Scenarios a) and b) fit the same model, therefore have the same upper bounds.

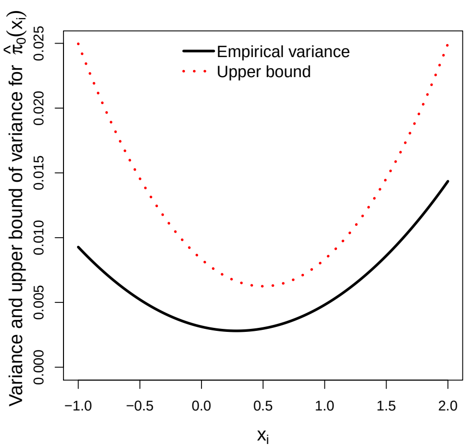
Scenario	m=10	m=100	m=1,000	m=10,000
a), b)	2.159	0.246	0.025	0.002
c)	4.697	0.726	0.076	0.008
d)	4.798	0.755	0.079	0.008
e)	6.250	3.540	0.491	0.051

Figure S1: The same simulation scenarios as in Figure 1, for $\lambda = 0.8$. The empirical variance of $\hat{\pi}_0(\mathbf{x}_i)$ is a continuous black line, the upper bound obtained in Lemma 7 is a dotted red line. Panels a) and b) fit linear regression; panel c) fits B-splines with 3 degrees of freedom (df); panel d) fits B-splines with 3 df for x_{i1}, true values for x_{i2}; panel e) fits B-splines with 20 df for x_{i1}, true values for x_{i2}.

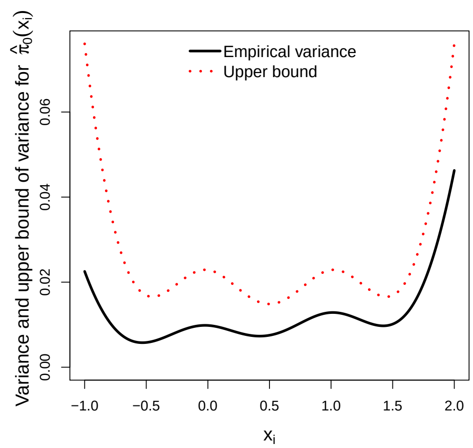
(a) $\pi_0(x_i) = x_i/4 + 1/2$



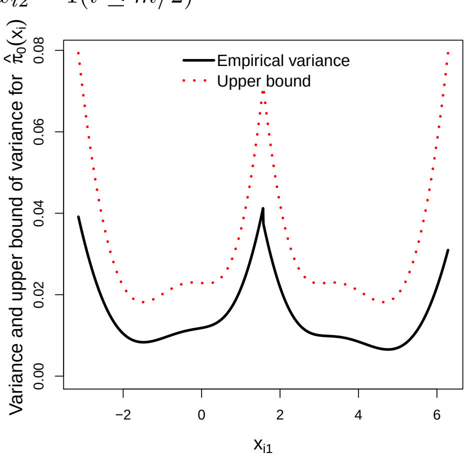
(b) $\pi_0(x_i) = \phi(x_i)$



(c) $\pi_0(x_i) = \phi(x_i)$



(d) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$, where $x_{i2} = 1(i \le m/2)$



(e) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$, where $x_{i2} = 1(i \le m/2)$

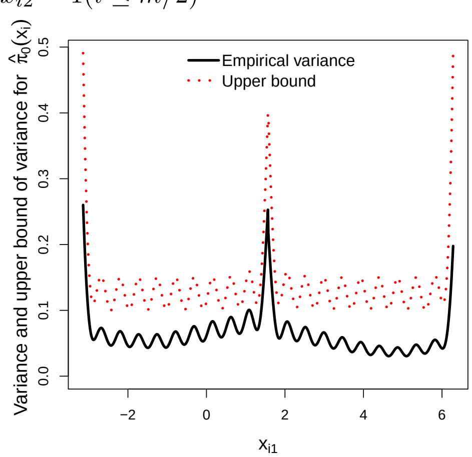
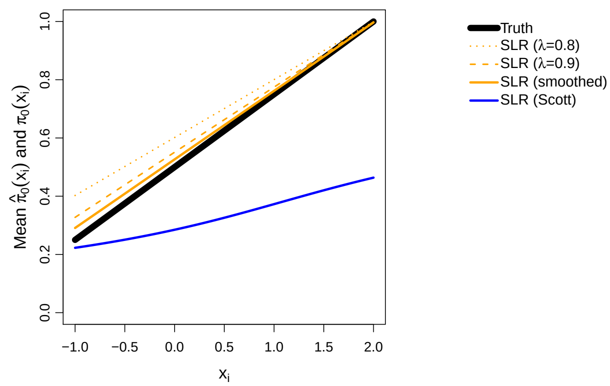
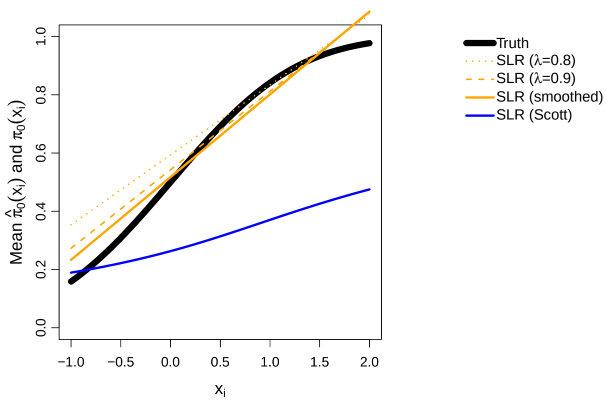


Figure S2: Different simulation scenarios using a block correlation structure for the latent variables which encode whether a particular feature is draw from the null distribution. The true function $\pi_0(\mathbf{x}_i)$ is plotted in a thick black line, while the empirical means of $\hat{\pi}_0(\mathbf{x}_i)$, assuming different modelling approaches are shown in the orange lines (for our approach) and in the blue lines (for the Scott approach). In panels b) and c) the same underlying truth is considered; this is also the case for panels d) and e). In d) and e), different terms are used in the regression for x_{i1}, while the true values are used for x_{i2}. SLR = simple (univariate) linear regression, df = degrees of freedom. No thresholding at 0 or 1 is considered for our approach.

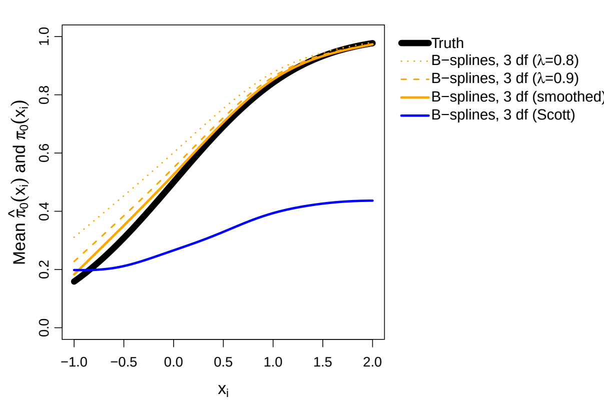
(a) $\pi_0(x_i) = x_i/4 + 1/2$



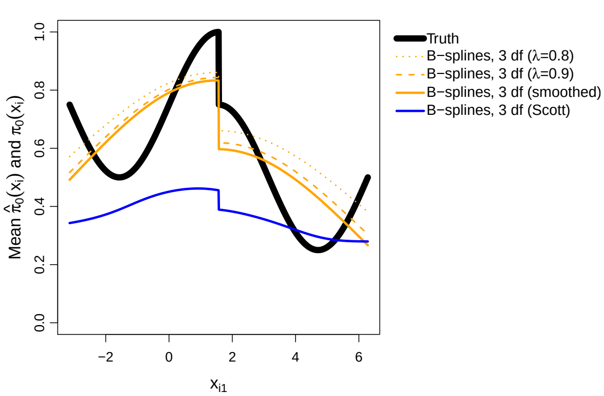
(b) $\pi_0(x_i) = \phi(x_i)$



(c) $\pi_0(x_i) = \phi(x_i)$



(d) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$, where $x_{i2} = 1(i \leq m/2)$.



(e) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$, where $x_{i2} = 1(i \leq m/2)$.

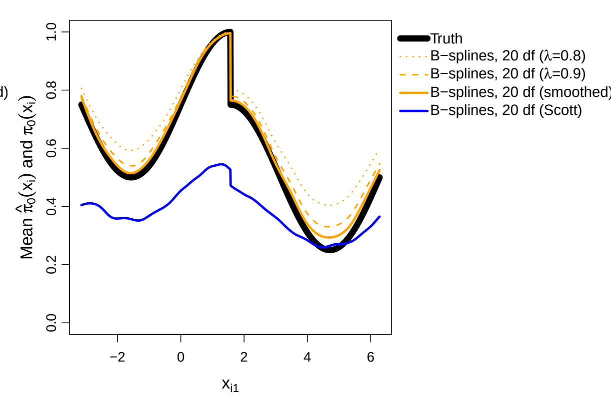
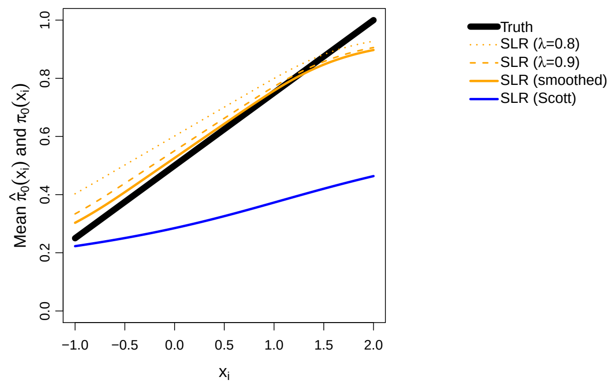
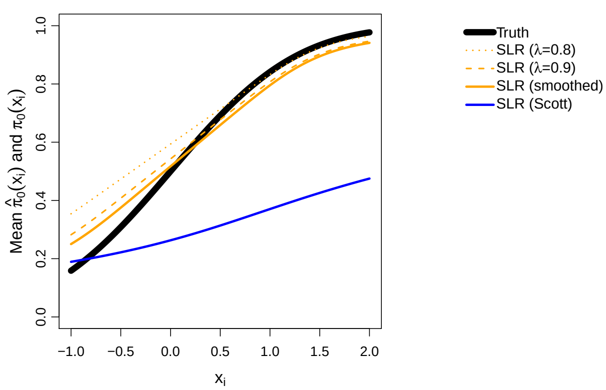


Figure S3: The same scenarios as in Figure S2 but considering thresholding at 0 and 1 in our approach, the Scott approach being the same as in Figure S2.

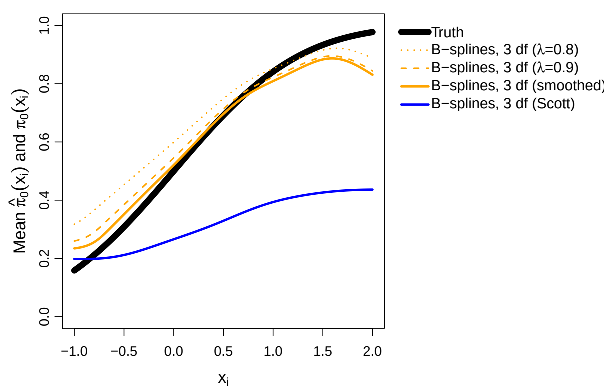
(a) $\pi_0(x_i) = x_i/4 + 1/2$



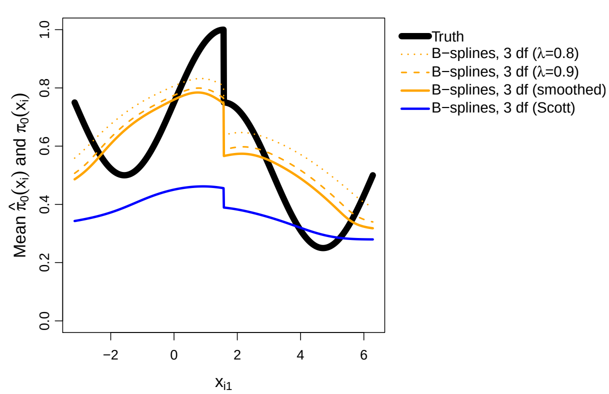
(b) $\pi_0(x_i) = \phi(x_i)$



(c) $\pi_0(x_i) = \phi(x_i)$



(d) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$, where $x_{i2} = 1(i \leq m/2)$.



(e) $\pi_0(\mathbf{x}_i) = \sin(x_{i1})/4 + x_{i2}/4 + 1/2$, where $x_{i2} = 1(i \leq m/2)$.