Supplementary Material for "Selection of nonlinear interactions by a forward stepwise algorithm: Application to identifying environmental chemical mixtures affecting health outcomes"

Naveen N. Narisetty[1], Bhramar Mukherjee[2], Yin-Hsiu Chen[2], Richard Gonzalez[3], and John D. Meeker[4]

[1]Department of Statistics, University of Illinois at Urbana-Champaign
[2]Department of Biostatistics, University of Michigan, Ann Arbor
[3]Department of Psychology, University of Michigan, Ann Arbor
[4]Department of Environmental Health Sciences, University of Michigan, Ann Arbor

In this supplementary material, we provide simulation results for the different simulation settings considered in Table 2 of the manuscript.

1 Simulation Results

We provide simulation results for the settings considered in the paper. We firs remind the simulation settings.

1.1 Simulation Settings

We consider the following simulation settings for comparison of different methods. We consider $n = 500$ observations and $p = 10/20$ covariates. The regression model is

$$y = \mu(x) + \epsilon,$$

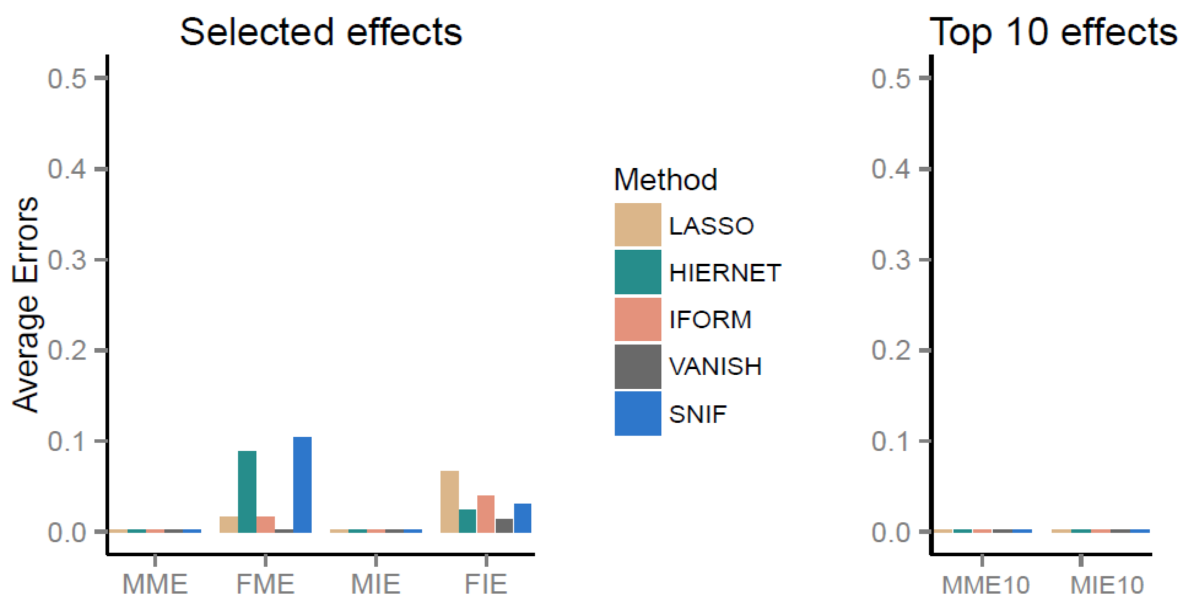
with $\epsilon \sim N(0, \sigma^2)$. Different settings for the conditional mean function $\mu(x)$ are considered. In Table 1, we provide the different choices considered for $\mu(x)$.

Table 1: Simulation settings: in Columns Main (and Inter), "L" indicates linear main (ineraction) effects, "NL" indicates nonlinear main (ineraction) effects, "No" indicates no ineraction effect. True Effects column gives indices of active main and interaction effects where "*" denotes nonlinearity.

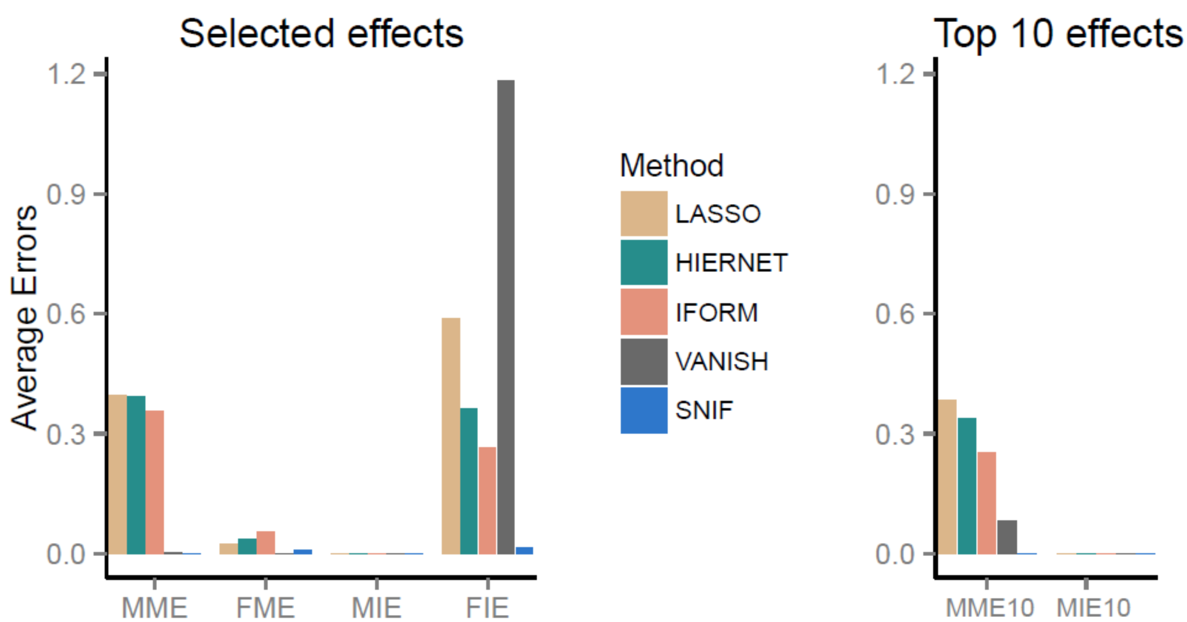
Main	Inter	Mean Function	True Effects
L	No	(a) $\mu_a(x) = 2 + \sum_{i=1}^{5} x_i$	1,2,3,4,5
NL	No	(b) $\mu_b(x) = 2 + 8(\lvert x_1\rvert - 1)^2 + 4\lvert\lvert x_2\rvert - 1\rvert + \sum_{i=3}^{5} x_i$ (c) $\mu_c(x) = 2 + (\lvert x_1\rvert \geq 1.5\&\lvert x_1\rvert \leq 2)) + \mathbb{1}(\lvert x_1\rvert \leq 0.5)$ $+2\ \mathbb{1}(0.5 \leq \lvert x_1\rvert \leq 1.5) + 4\lvert\lvert x_2\rvert - 1\rvert + \sum_{i=3}^{5} x_i$ (d) $\mu_d(x) = 2 + 2\lvert x_1\rvert\ \mathbb{1}(\lvert x_1\rvert < 1) + 2\ \mathbb{1}(\lvert x_1\rvert > 1)$ $+4\lvert\lvert x_2\rvert - 1\rvert + \sum_{i=3}^{5} x_i$	1^*,2^*,3,4,5
L	L	(e) $\mu_d(x) = 2 + \sum_{i=1}^{5} x_i + 6x_4x_5$	1,2,3,4,5,(4×5)
NL	L	(f) $\mu_f(x) = \mu_b(x) + 6x_4x_5$ (g) $\mu_g(x) = \mu_c(x) + 6x_4x_5$ (h) $\mu_h(x) = \mu_d(x) + 6x_4x_5$	1^*,2^*,3,4,5, (4×5)
NL	NL	(i) $\mu_i(x) = \mu_b(x) + 8\lvert x_1\rvert\lvert\lvert x_2\rvert - 1\rvert$ (j) $\mu_j(x) = \mu_c(x) + 8\lvert x_1\rvert\lvert\lvert x_2\rvert - 1\rvert$ (k) $\mu_k(x) = \mu_d(x) + 8\lvert x_1\rvert\lvert\lvert x_2\rvert - 1\rvert$	1^*,2^*,3,4,5, $(1^* \times 2^*)$
NL	NL	(l) $\mu_l(x) = \mu_i(x) + 8x_3\sqrt{\lvert x_2\rvert}$ (m) $\mu_m(x) = \mu_j(x) + 8x_3\sqrt{\lvert x_2\rvert}$ (n) $\mu_n(x) = \mu_k(x) + 8x_3\sqrt{\lvert x_2\rvert}$	1^*,2^*,3,4,5, $(1^* \times 2^*), (2^* \times 3^*)$
NL	NL	(o) $\mu_o(x) = 2 +\ \mathbb{1}(1.5 \leq \lvert x_1\rvert \leq 2) +\ \mathbb{1}(\lvert x_1\rvert \leq 0.5)$ $+2\ \mathbb{1}(0.5 \leq \lvert x_1\rvert \leq 1.5) + \sum_{i=3}^{5} x_i + 8\lvert x_1\rvert\lvert\lvert x_2\rvert - 1\rvert$ (p) $\mu_p(x) = 2 +\ \mathbb{1}(1.5 \leq \lvert x_1\rvert \leq 2) +\ \mathbb{1}(\lvert x_1\rvert \leq 0.5)$ $+2\ \mathbb{1}(0.5 \leq \lvert x_1\rvert \leq 1.5) + \sum_{i=3}^{5} x_i + 8\lvert x_1\rvert\lvert\lvert x_2\rvert - 1\rvert$ (q) $\mu_q(x) = 2 + 2\lvert x_1\rvert\ \mathbb{1}(\lvert x_1\rvert < 1) + 2\ \mathbb{1}(\lvert x_1\rvert > 1)$ $+\sum_{i=3}^{5} x_i + 8\lvert x_1\rvert\lvert\lvert x_2\rvert - 1\rvert$	1^*,3,4,5,(1 × 2)

1.2 Results

We present the results for the different mean structures given by rows (a-q) given in Table 1 for $p = 10$ or $p = 20$ and noise standard deviation $\sigma = 1$. For screening, we present the results based on the top p effects. The conclusions from these additional simulation results also demonstrate that SNIF has very competitive performance across simulation settings.

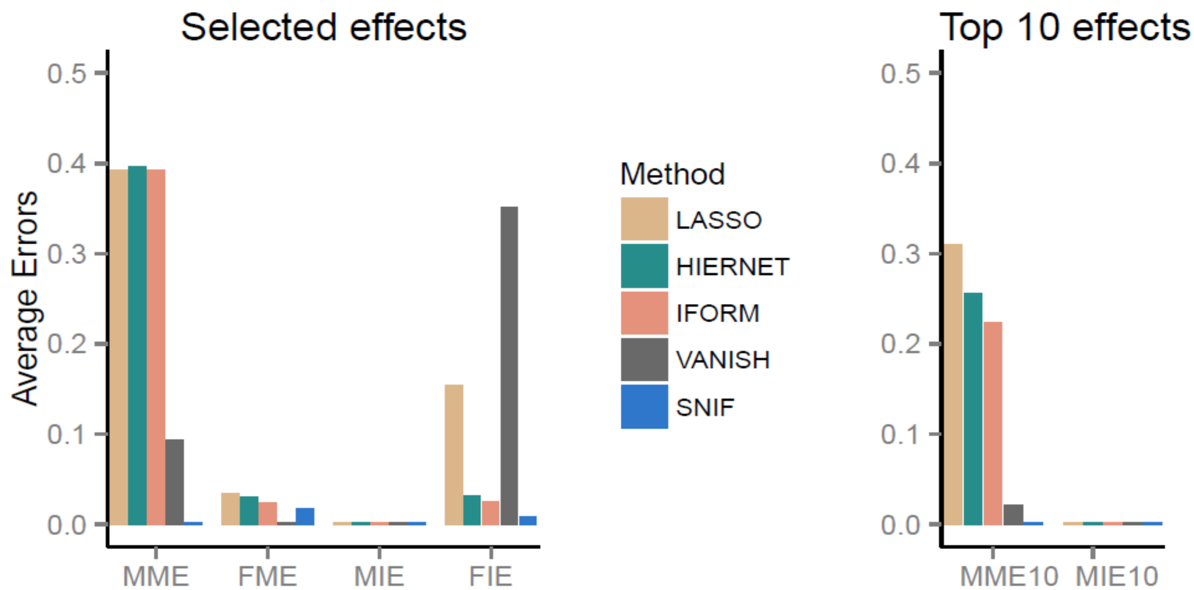


(a)

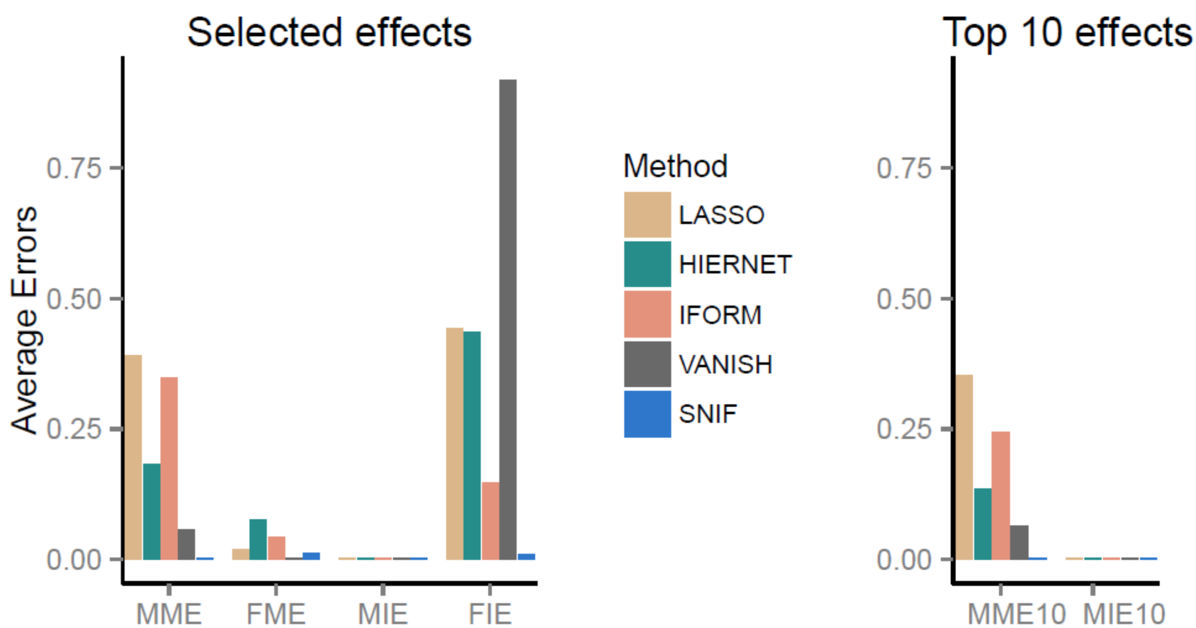


(b)

Models (a) - (b) with $p = 10$ covariates

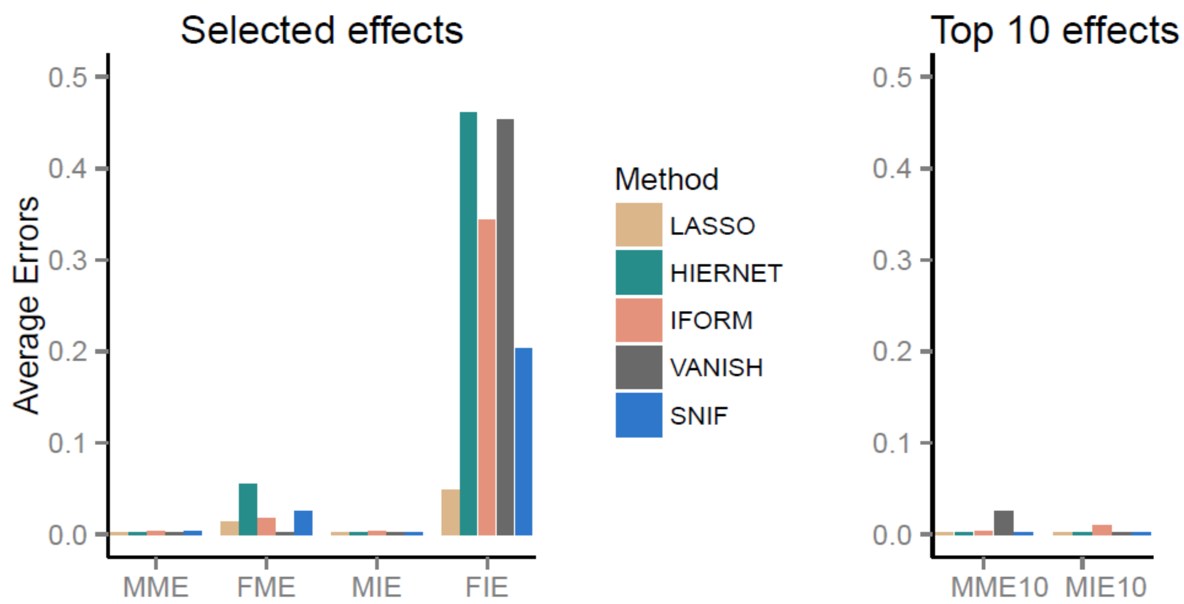


(c)

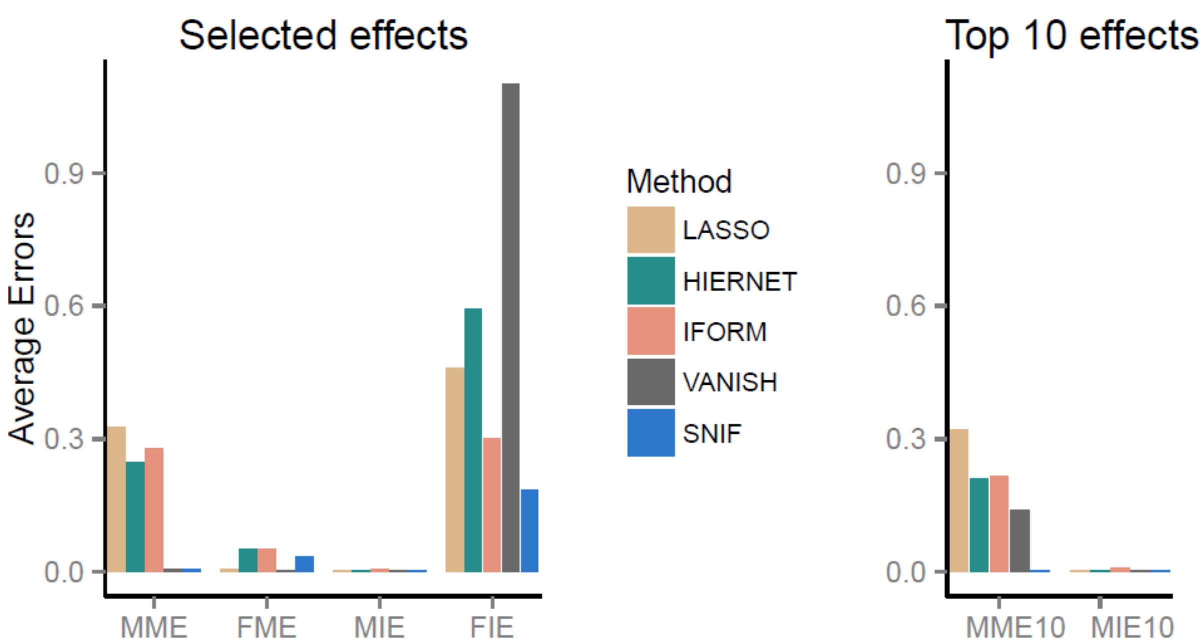


(d)

Models (c) - (d) with $p = 10$ covariates

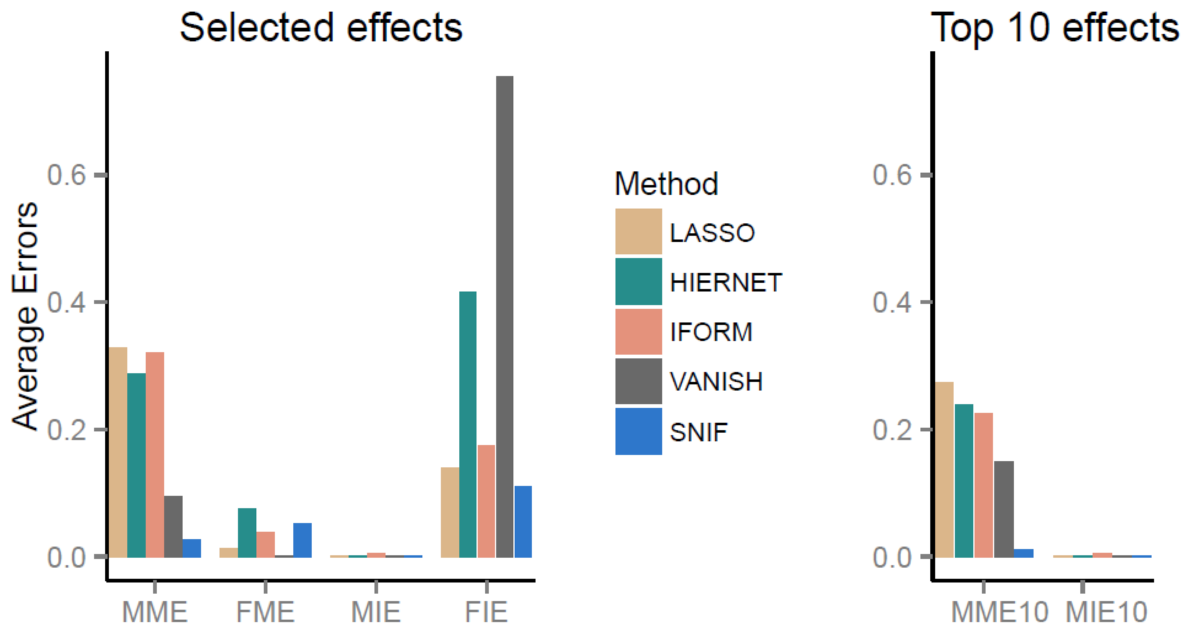


(e)

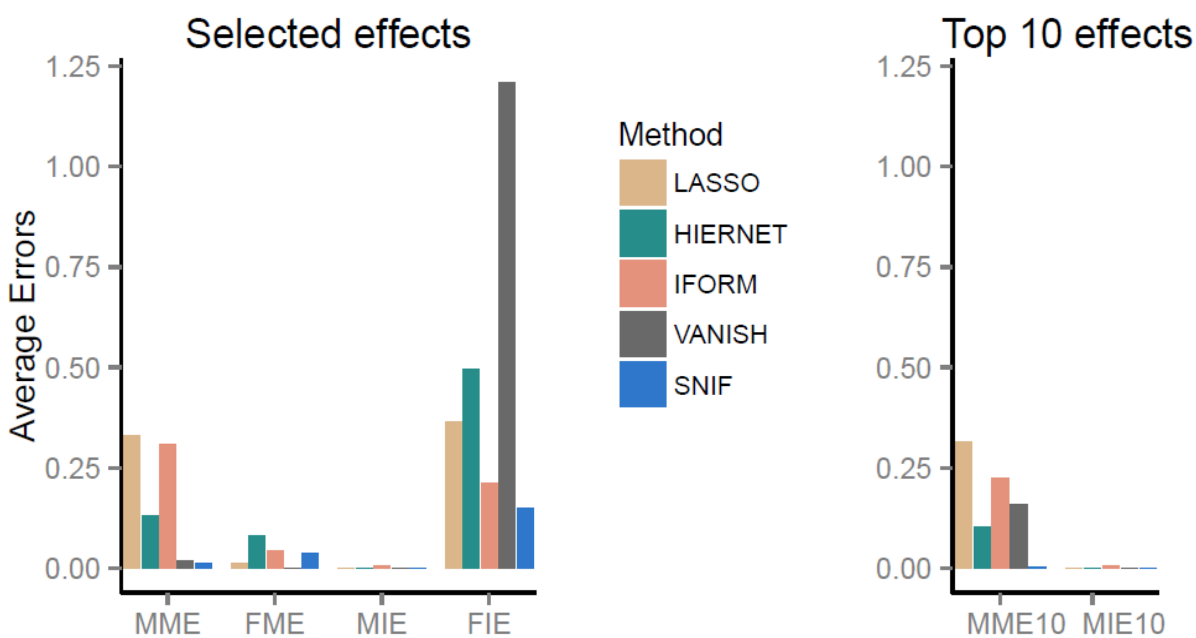


(f)

Models (e) - (f) with $p = 10$ covariates

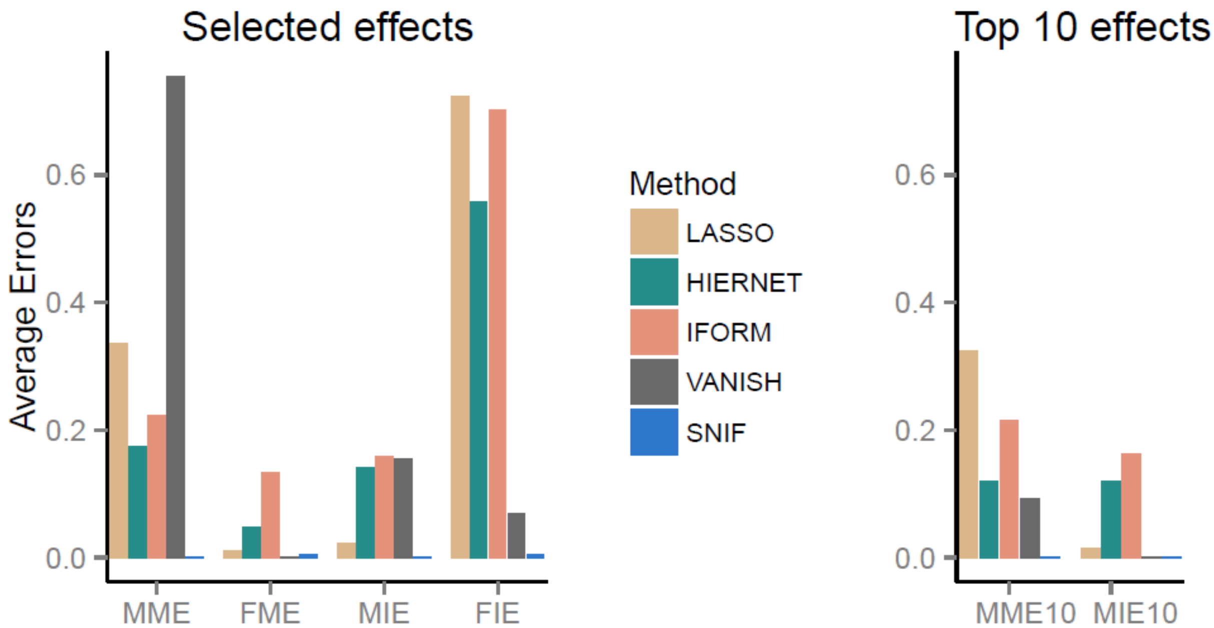


(g)

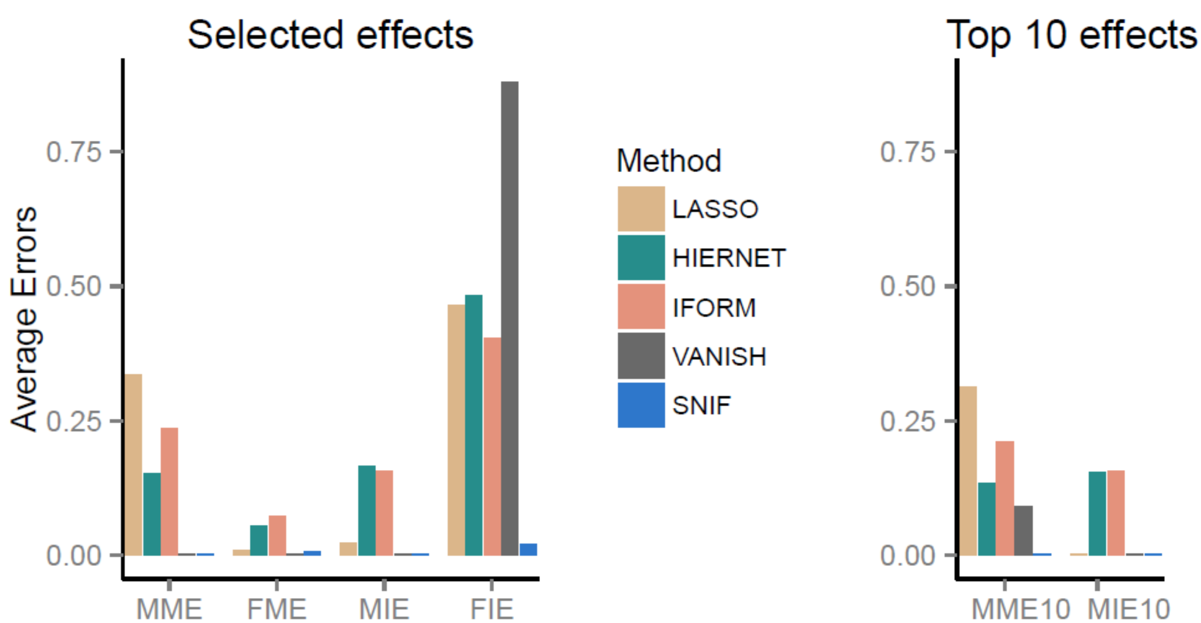


(h)

Models (g) - (h) with $p = 10$ covariates

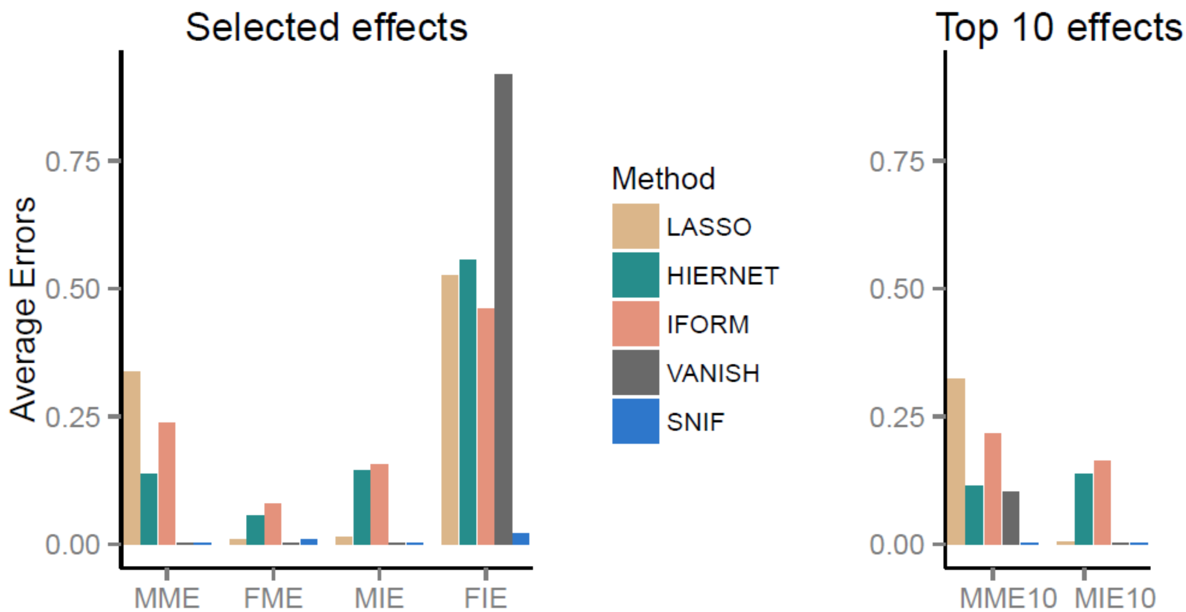


(i)

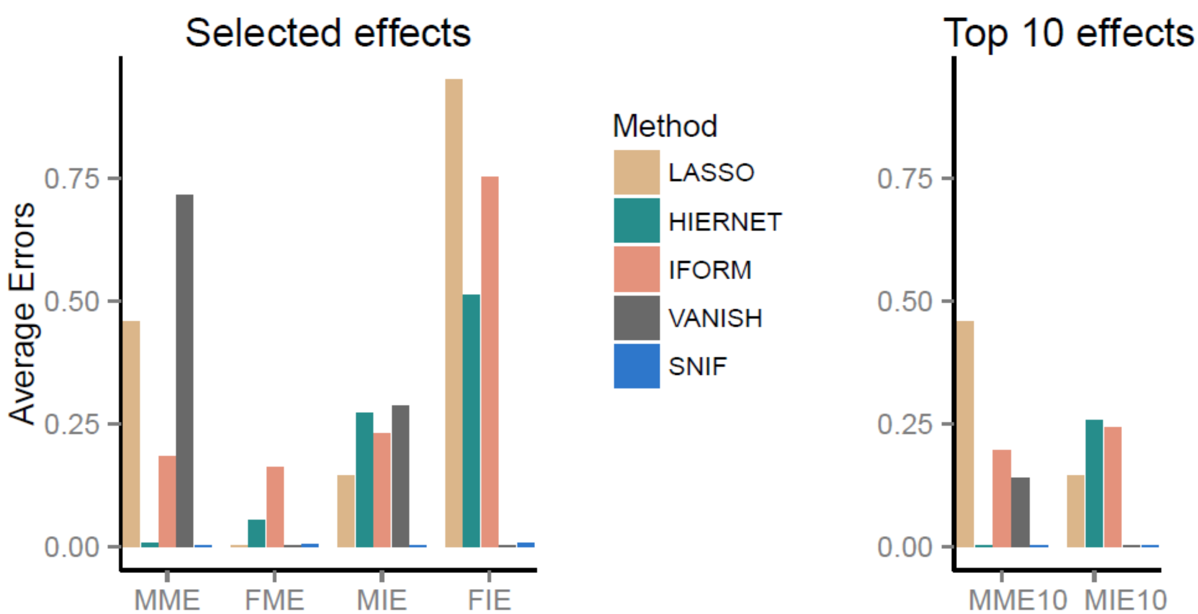


(j)

Models (i) - (j) with $p = 10$ covariates

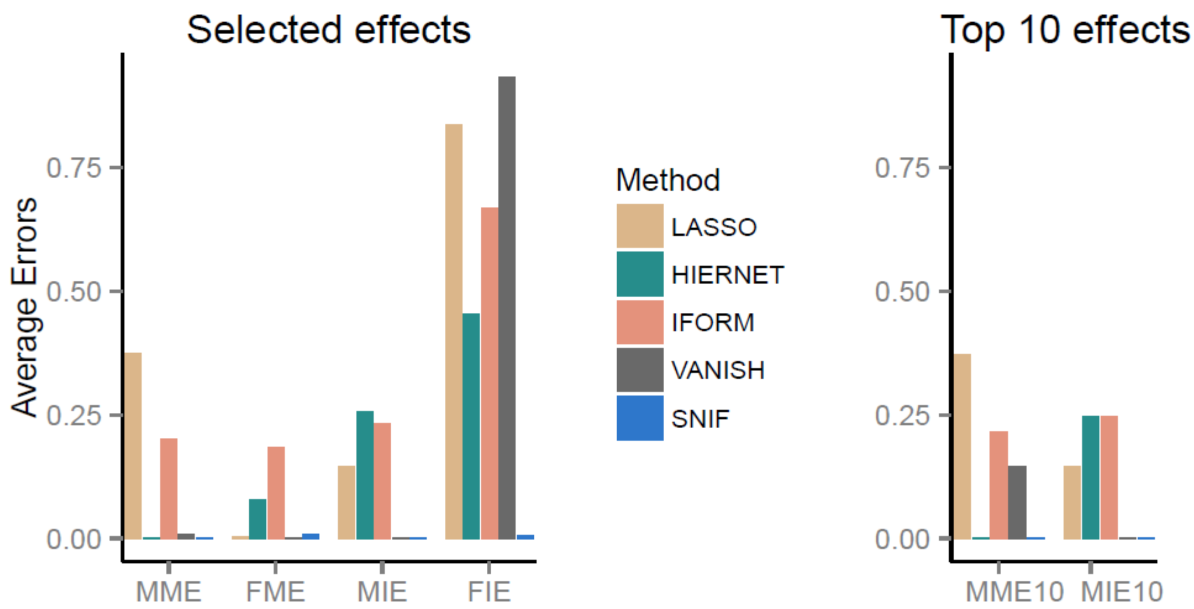


(k)

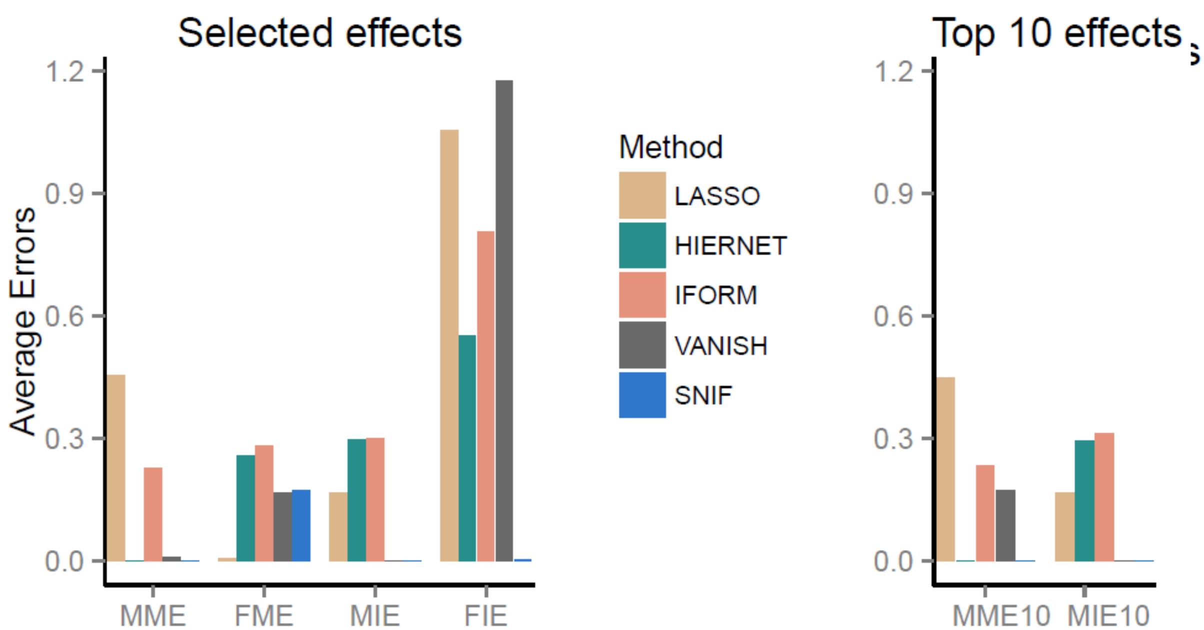


(l)

Models (k) - (l) with $p = 10$ covariates

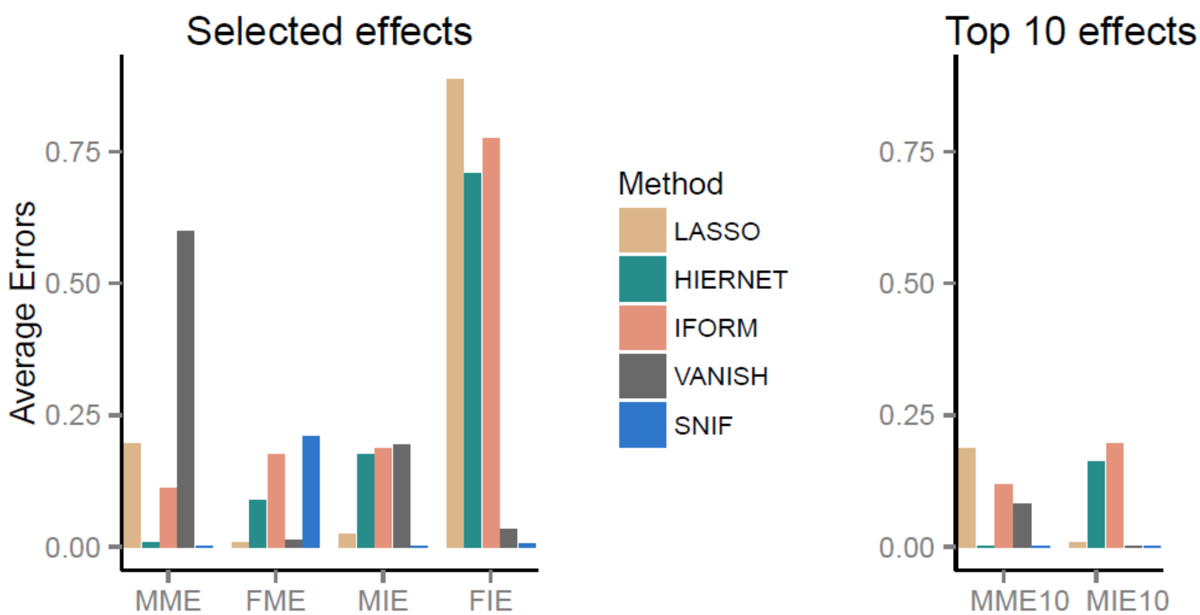


(m)

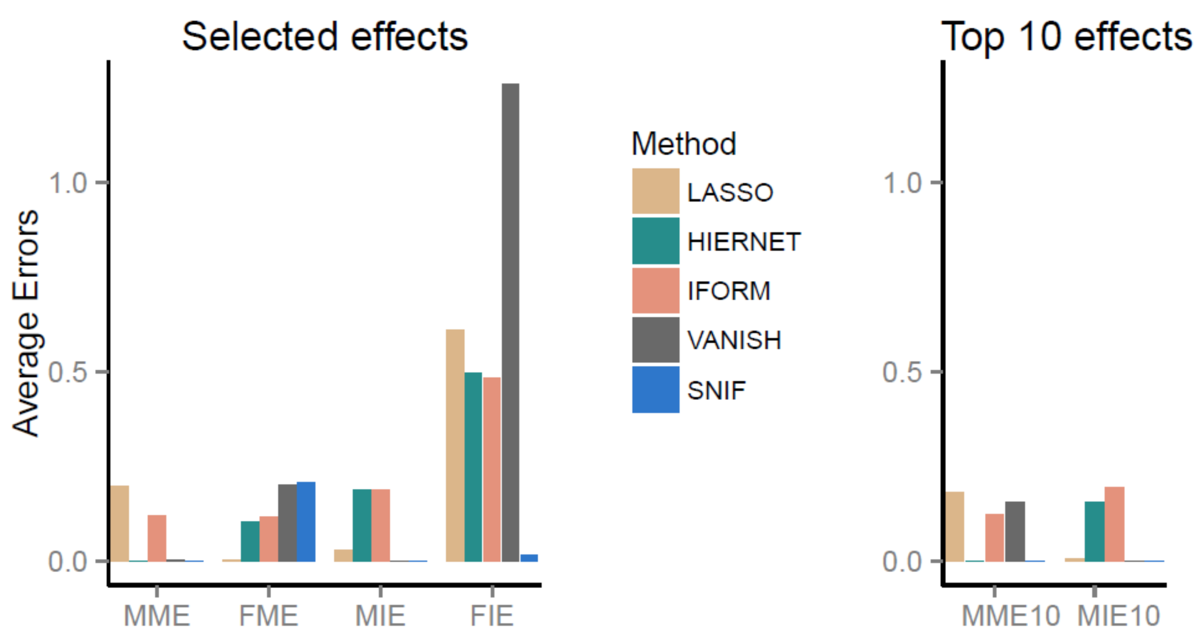


(n)

Models (m) - (n) with $p = 10$ covariates

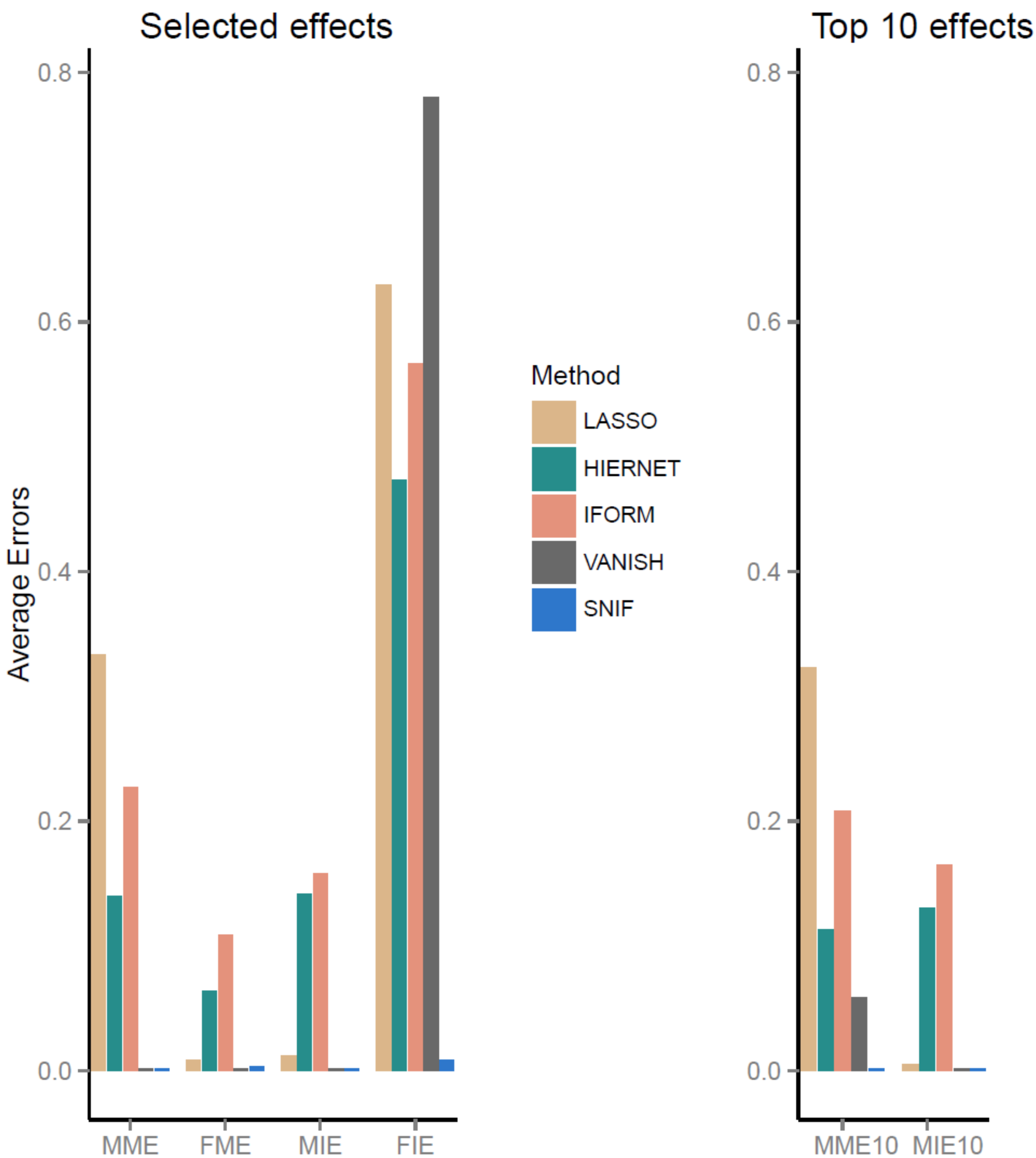


(o)



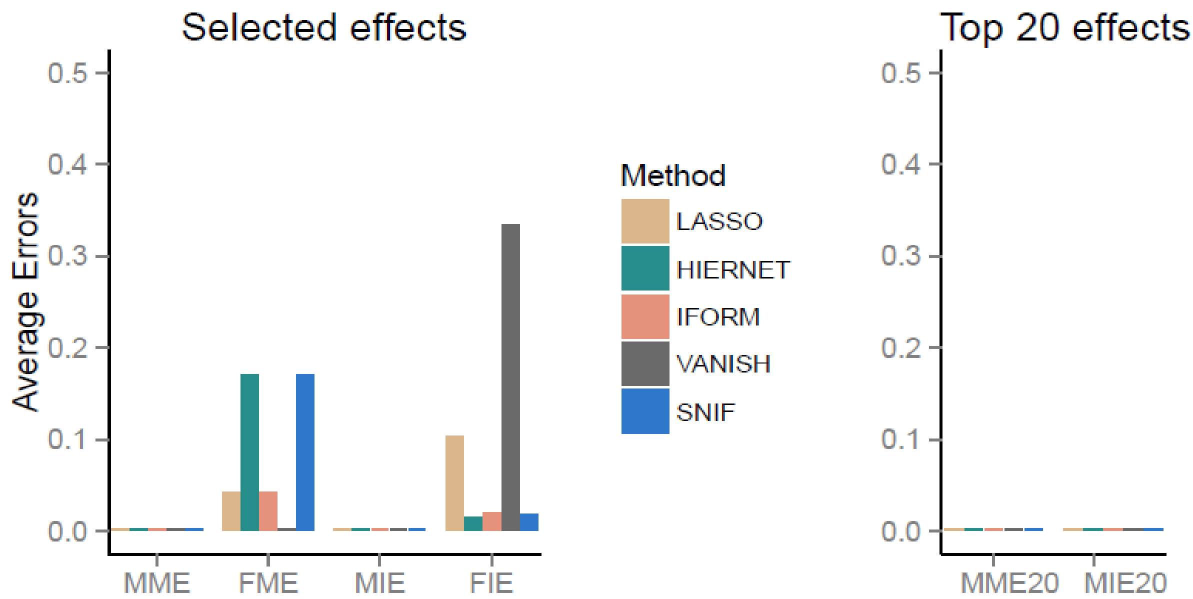
(p)

Models (o) - (p) with $p = 10$ covariates

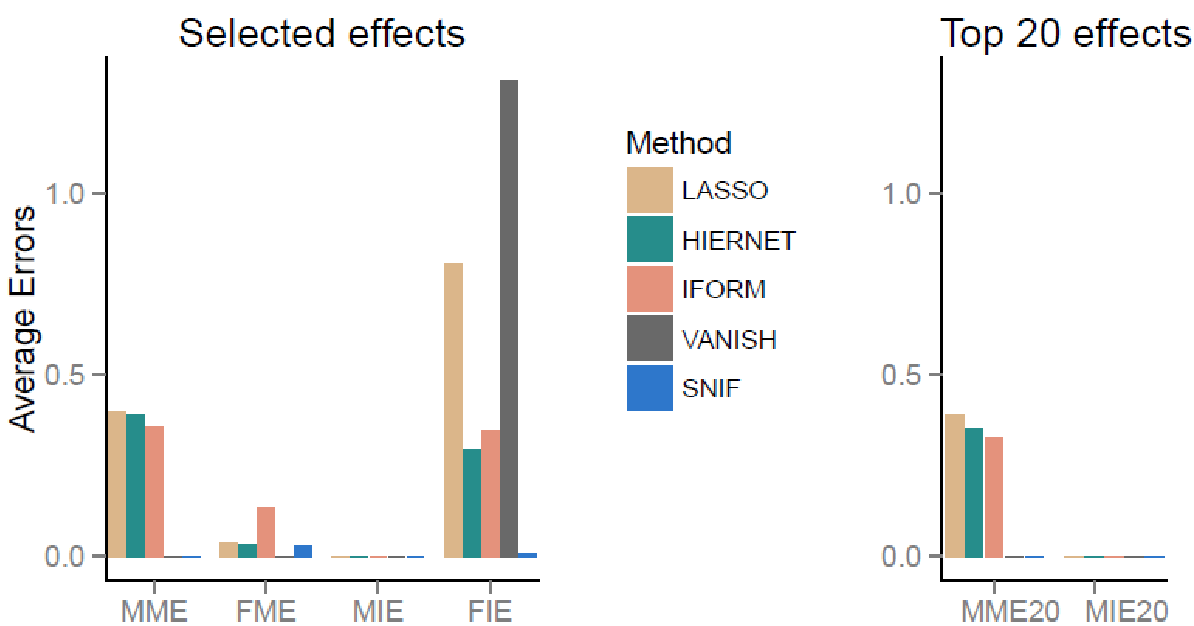


(q)

Model (q) with $p = 10$ covariates

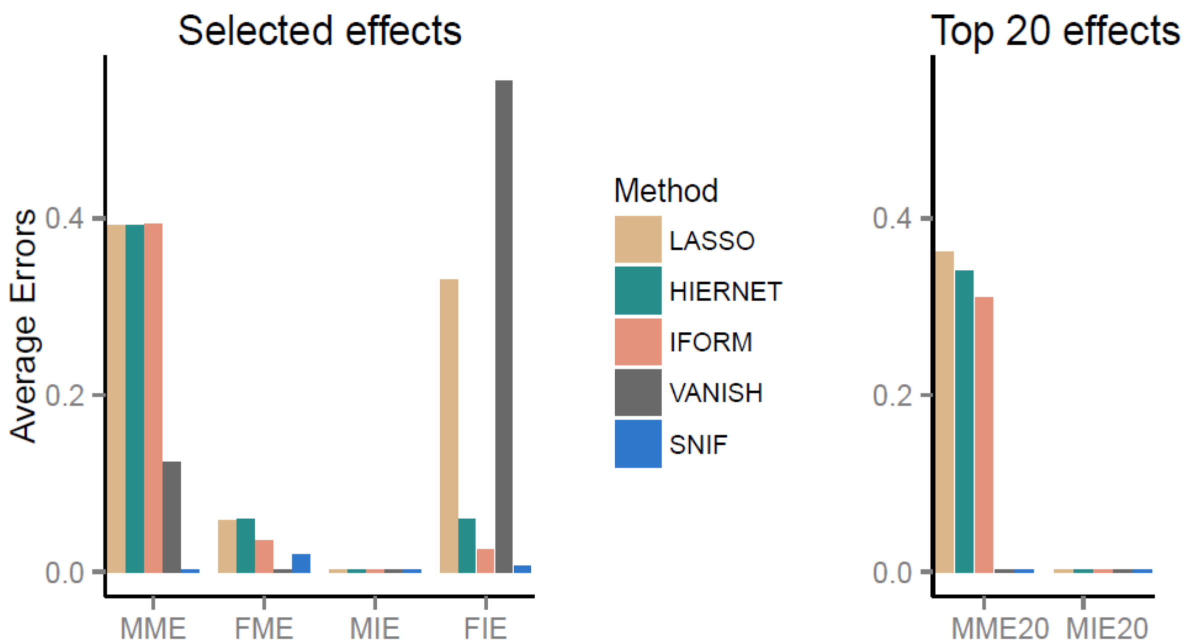


(a)

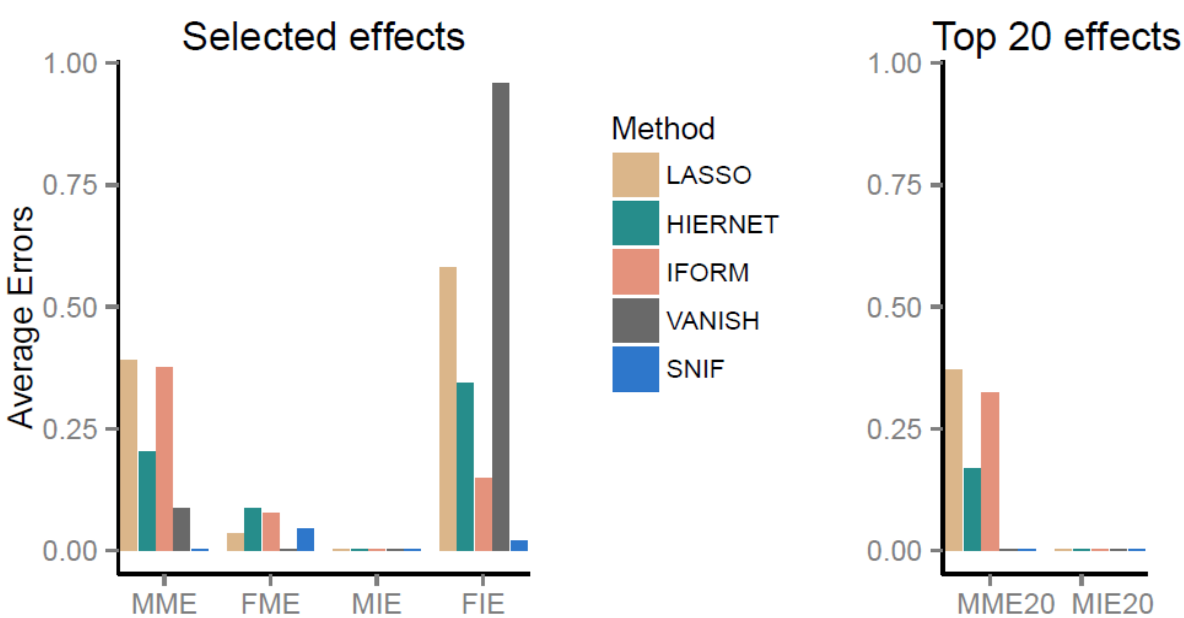


(b)

Models (a) - (b) with $p = 20$ covariates

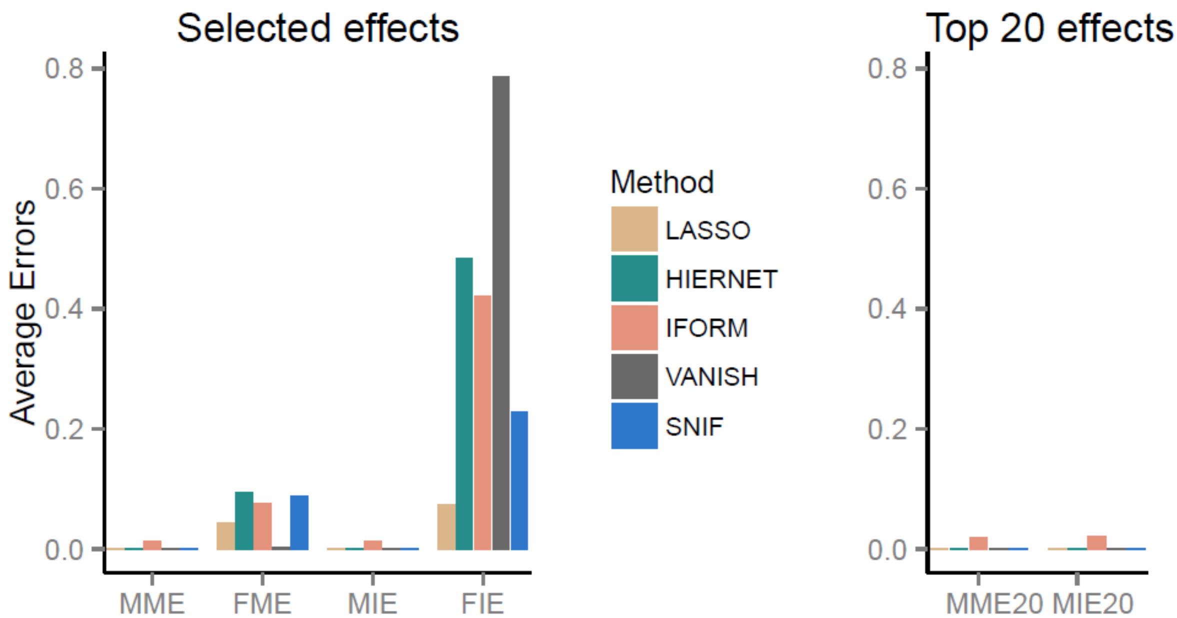


(c)

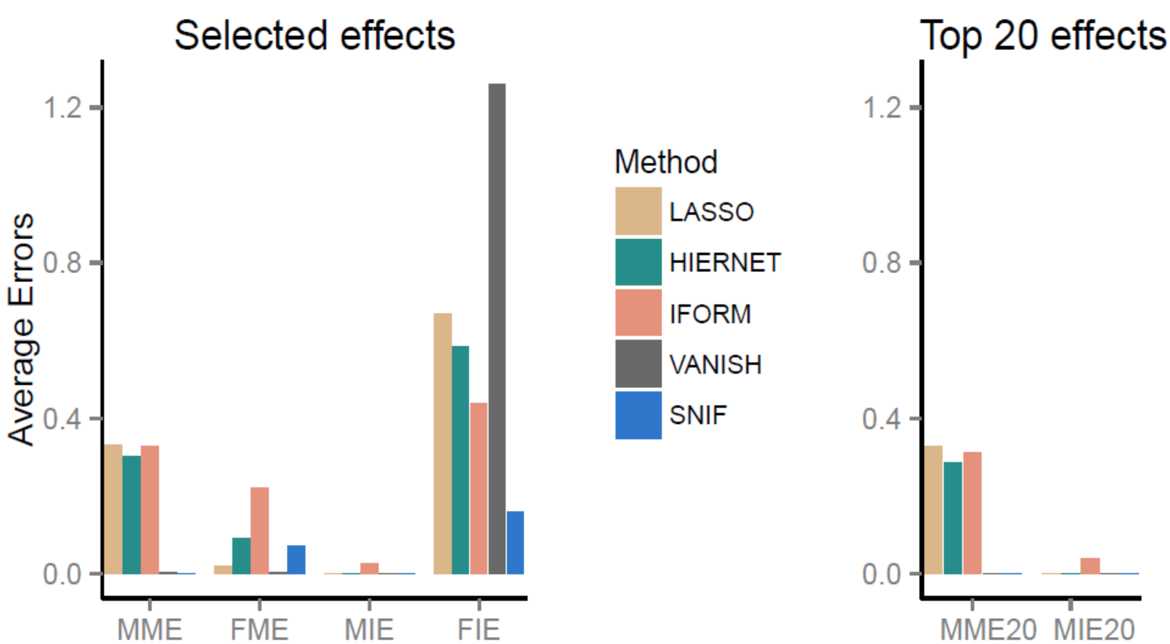


(d)

Models (c) - (d) with $p = 20$ covariates

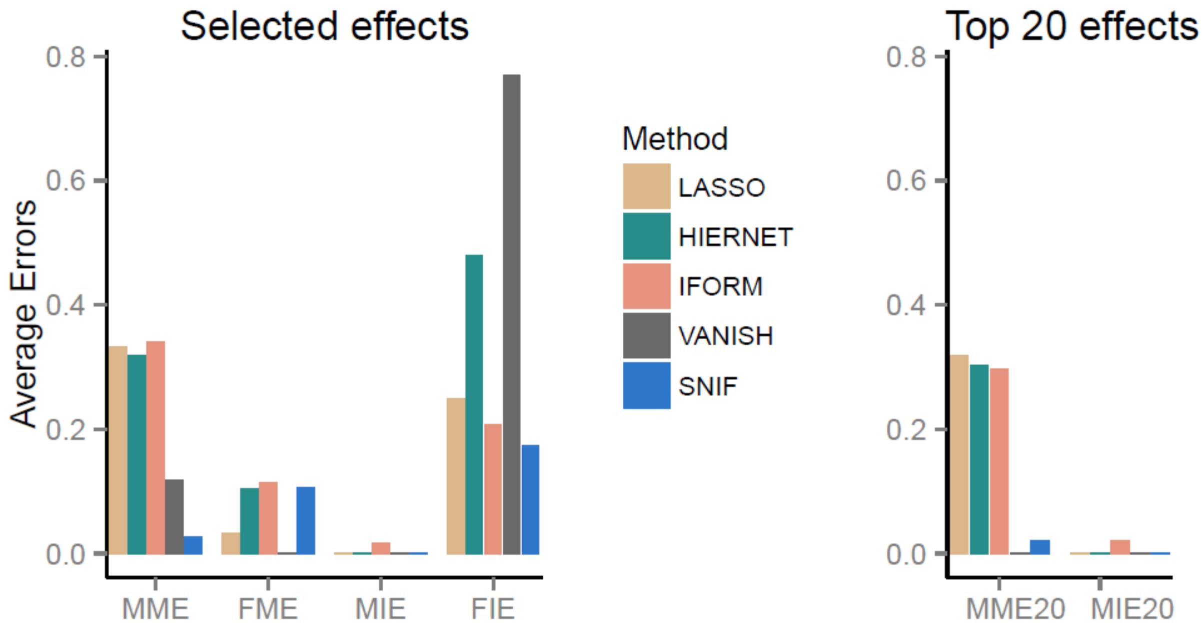


(e)

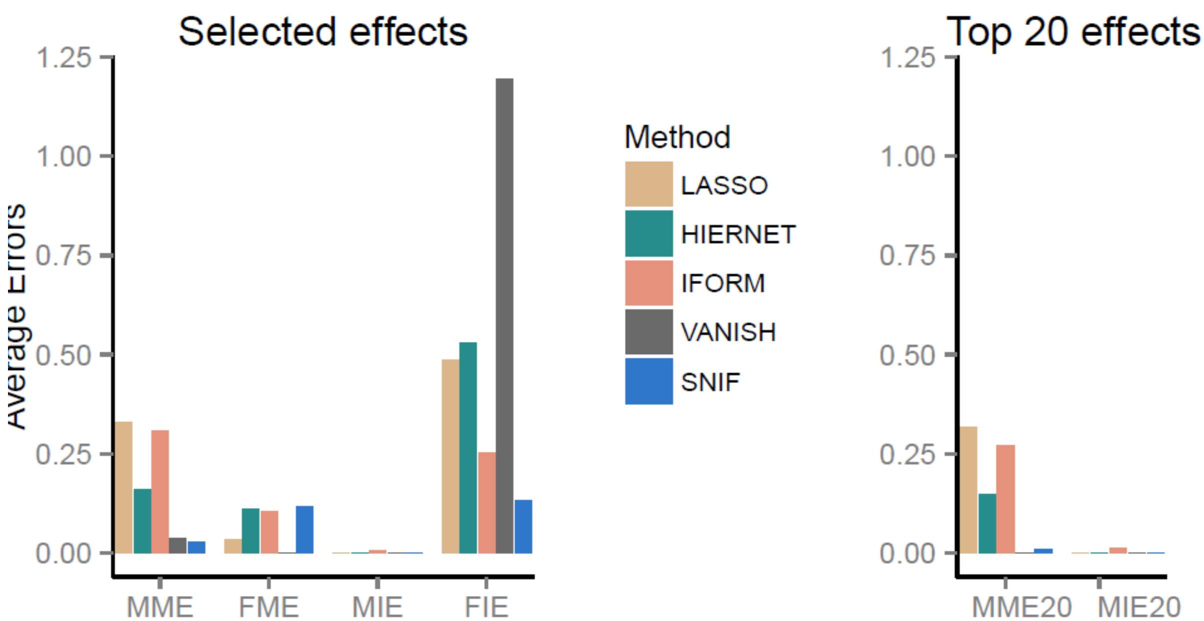


(f)

Models (e) - (f) with $p = 20$ covariates

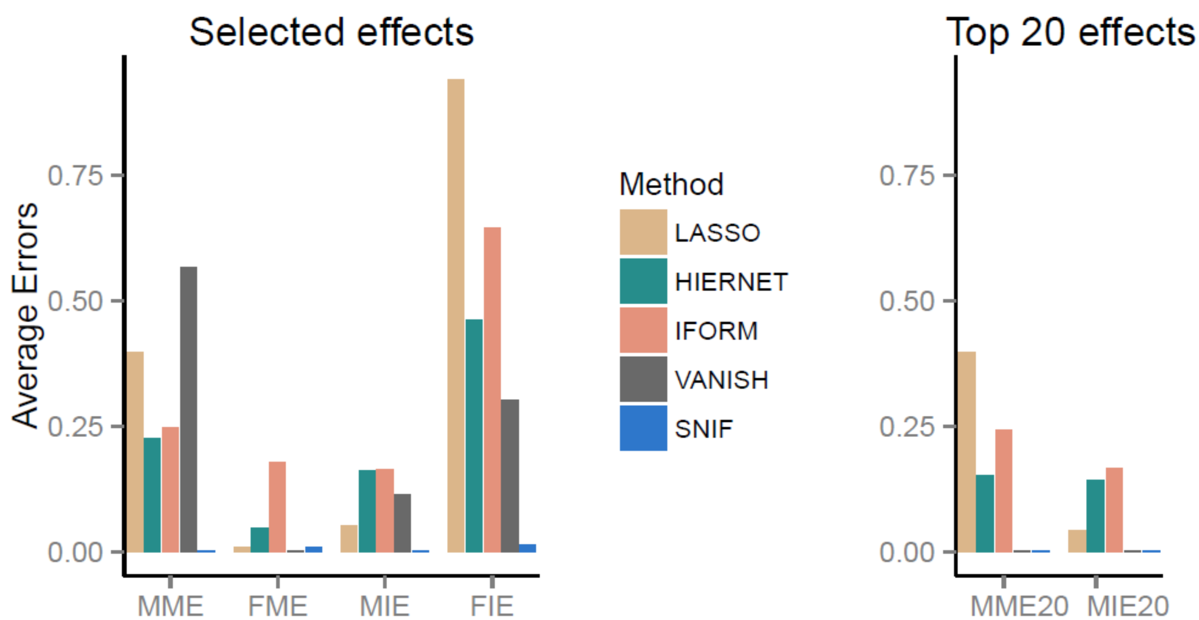


(g)

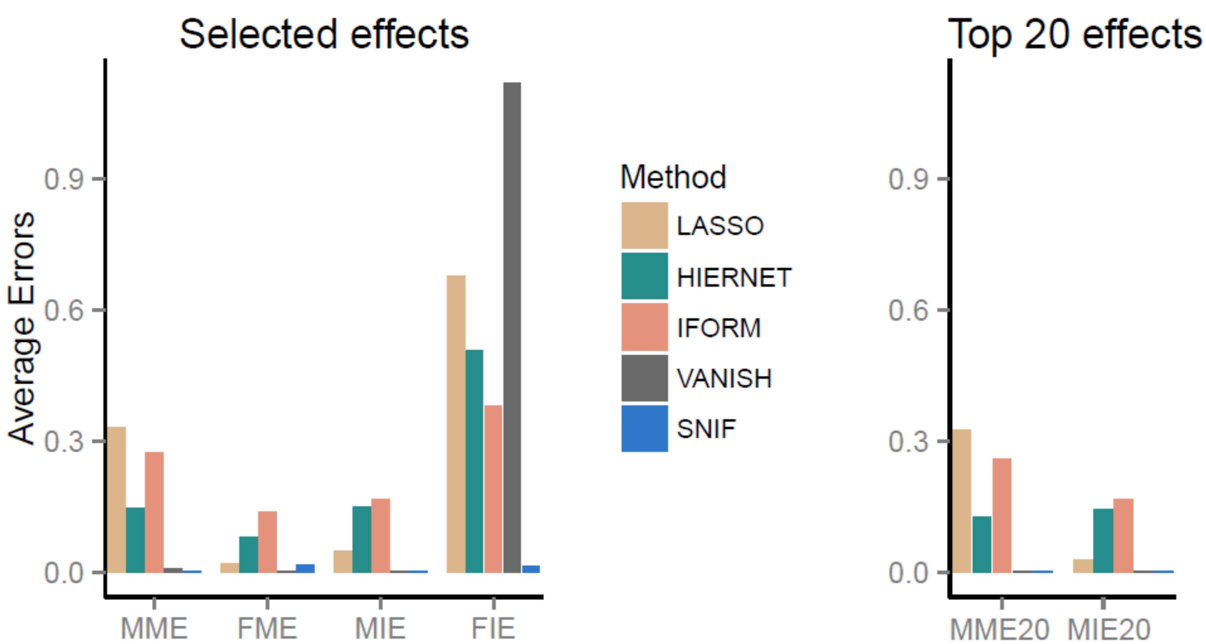


(h)

Models (g) - (h) with $p = 20$ covariates

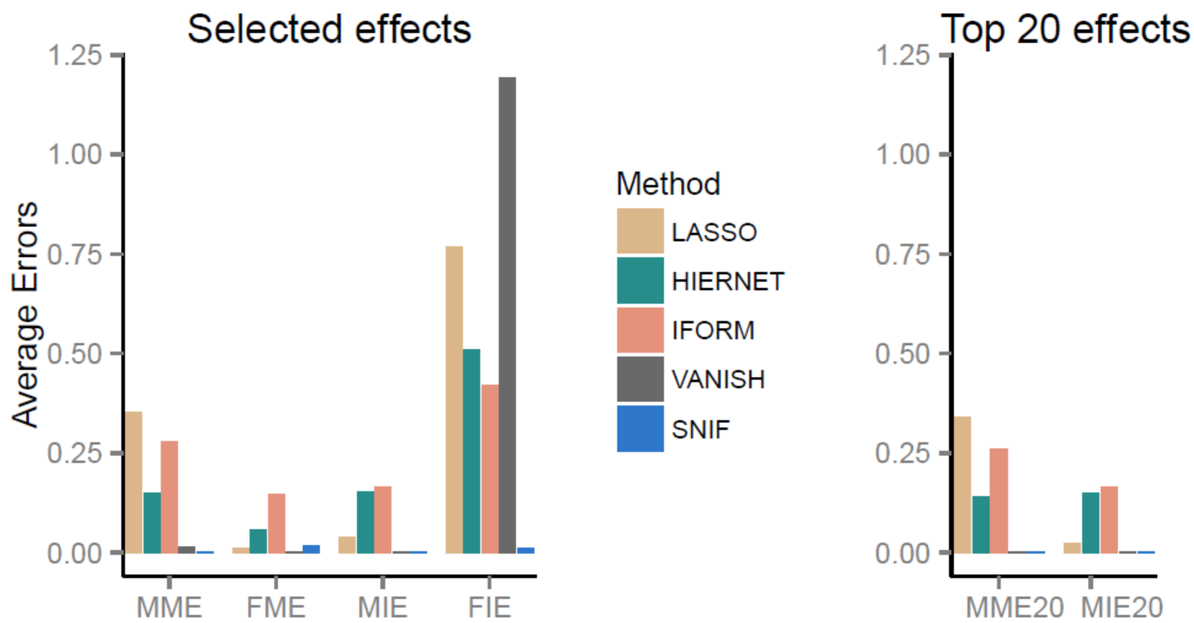


(i)

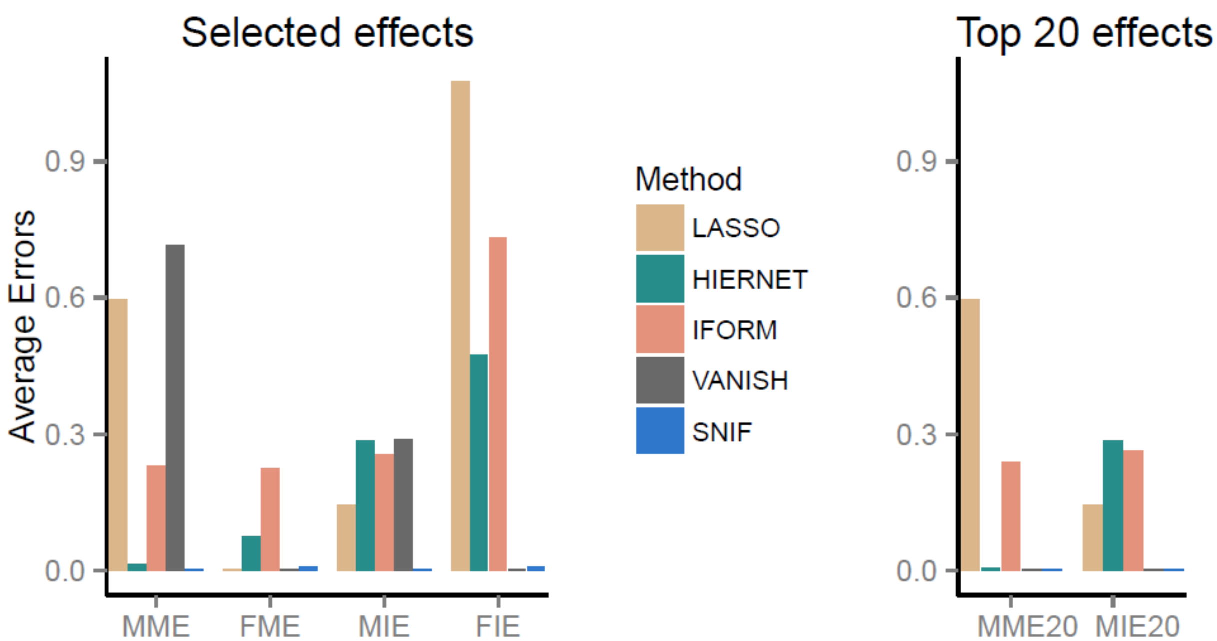


(j)

Models (i) - (j) with $p = 20$ covariates

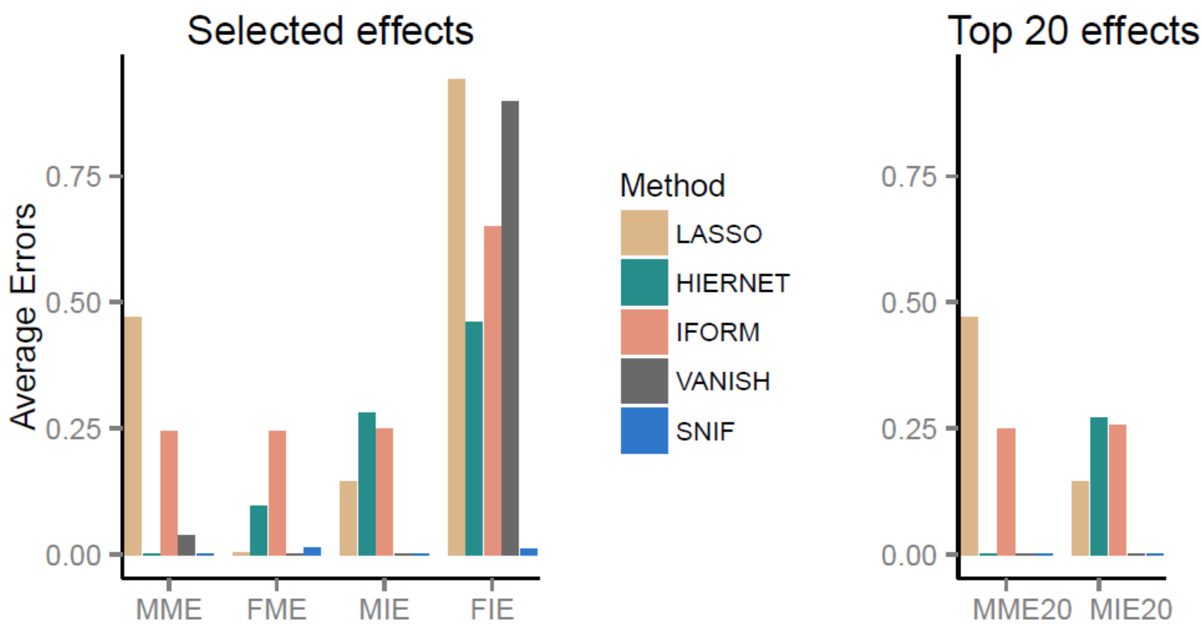


(k)

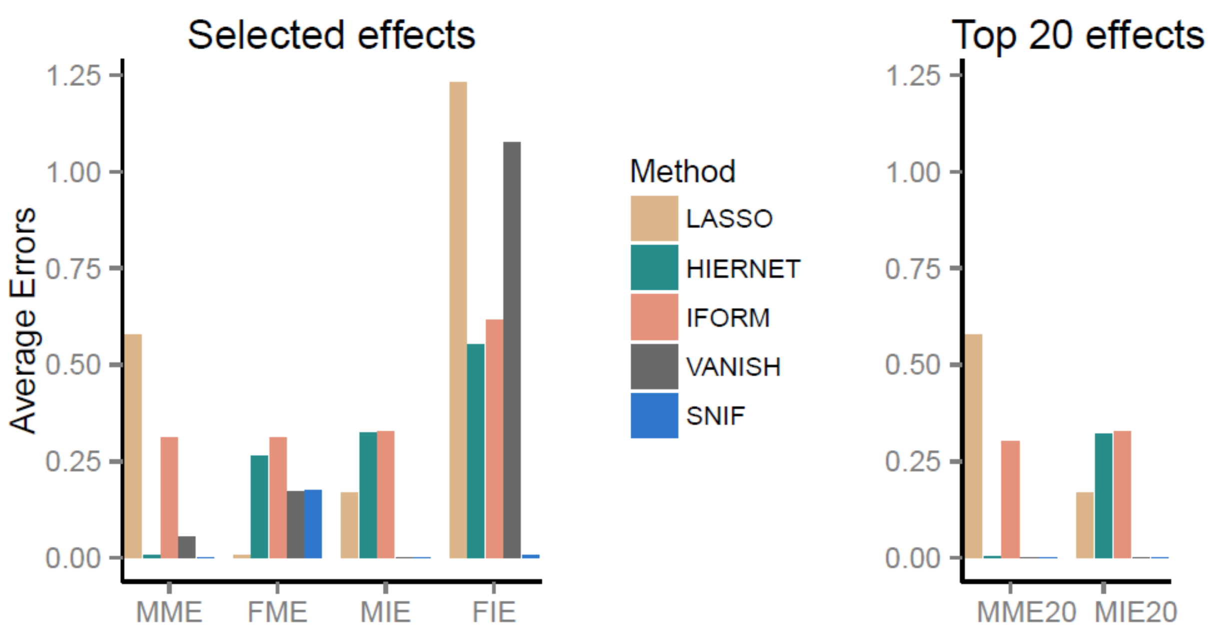


(l)

Models (k) - (l) with $p = 20$ covariates

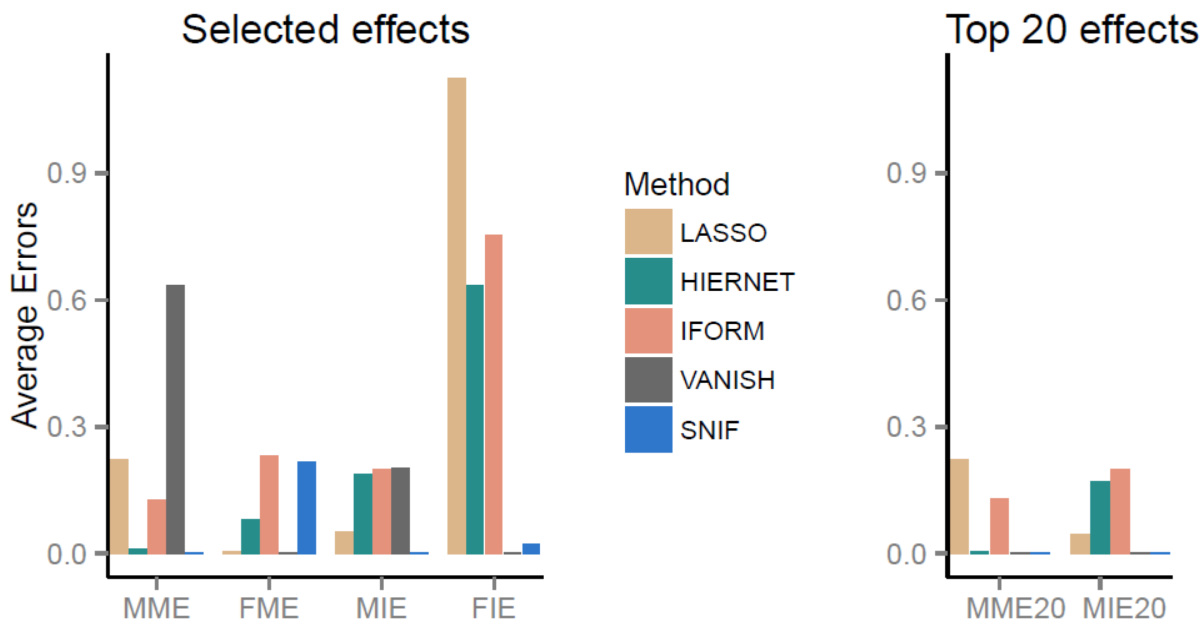


(m)

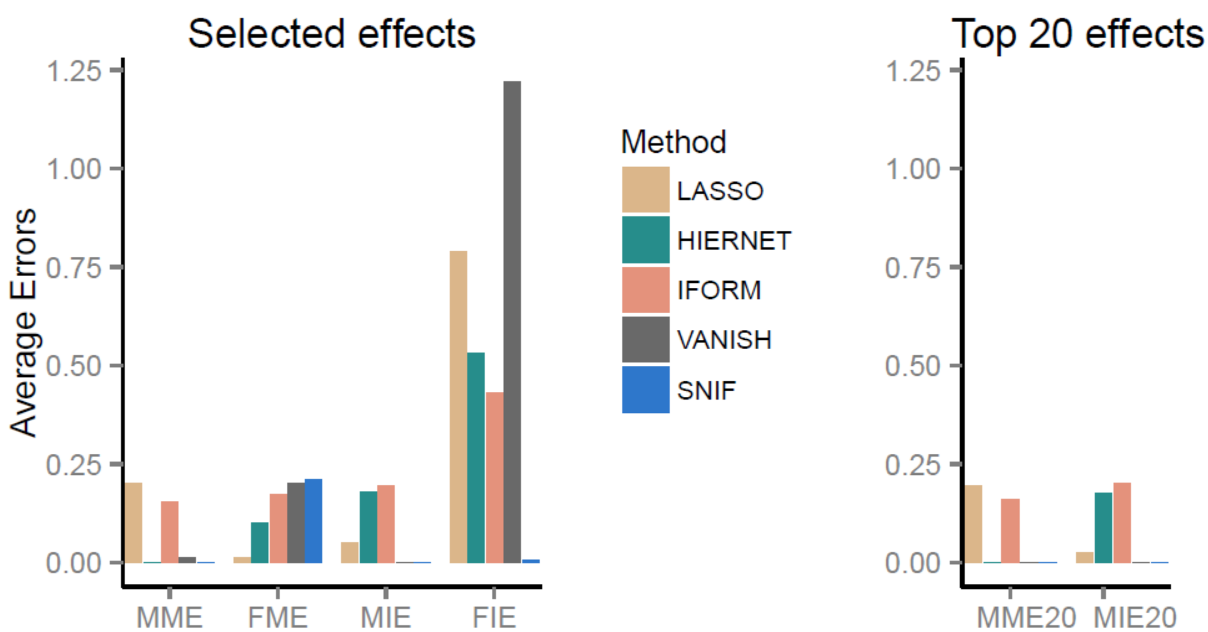


(n)

Models (m) - (n) with $p = 20$ covariates

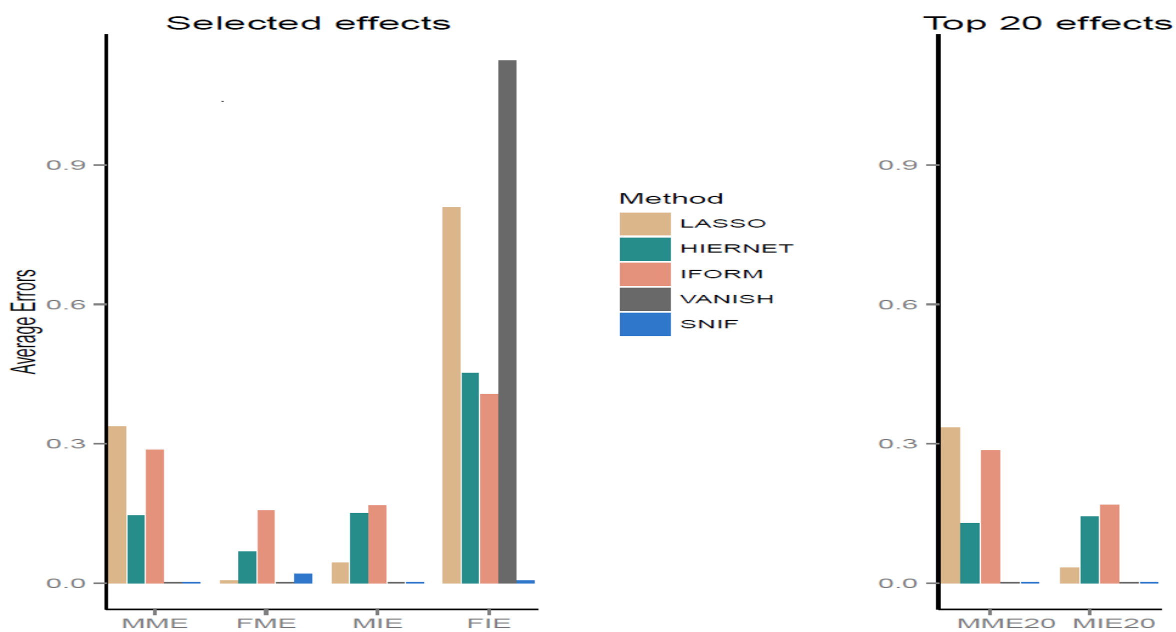


(o)



(p)

Models (o) - (p) with $p = 20$ covariates



(q)

Model (q) with $p = 20$ covariates