Global Drug Development

VS

Examining horseshoe prior and knockoffs for variable selection problems in drug development

David Ohlssen, Head of Advanced Exploratory Analytics
Matthias Kormaksson & Kostas Sechidis (Advanced Exploratory Analytics)
September 11th , 2020

NOVARTIS | Reimagining Medicine

Acknowledgements

- Ryan Murphy (summer intern at Novartis)
- Sebastian Weber

Agenda

- Problem background and motivation
- Traditional approaches
- Horseshoe prior method
  - Implementation using Stan and the brms package for a wide variety of models
- Knockoff method
  - Simulation study
  - Comparison with Horseshoe
- Conclusions

A bit about me

Areas of Expertise Bayesian statistics and modeling, data science, Meta-analysis, Model diagnostics

Projects / Deliverables

- Currently Head of Advanced Exploratory Analytics group
- Focused on building and leading quantitative innovation teams to impactful drug development contributions
- Consultancy on a wide range of quantitative problems in drug development
- Digging into methodology to see if it makes sense in practice
- Current/ Recent projects include: Complex innovative trial design for a pediatrics program; Oxford BDI collaboration (knockoff methodology)

Work History

- Novartis (2007-Present)
- 2017 Head of Advanced Exploratory Analytics group
- 2015-2017 of Advanced Exploratory Analytics team lead
- 2007-2015 Statistical Methodologist- Biometrical Fellow
- Research Fellow MRC Biostatistics unit (2003-2006)

Education History

- PhD Biostatistics, University of Cambridge 2000-2003
- M.Sc. Medical Statistics, University of Leicester 1999-2000
- B.Sc. in MORSE, University of Warwick 1996-1999

Framing the problem

- A successful program will often lead to a substantial amount of clinical trial data that can be synthesized and modeled to answer questions beyond the intended primary purpose of any one study.
- For example, the larger pool of data could be used to examine a greater level of understanding of factors affecting either treatment effect heterogeneity or long term prognosis of a patient.
- The search for such prognostic or predictive factors may be framed as a variable selection problem and mathematically, this could be expressed as understanding the conditional distribution of an outcome Y given a set of covariates X

Mathematical frame

X_1	X_2	...	X_p	Y
-0.300	0.416	...	-0.328	1.128
-0.310	-0.568	...	-0.396	-0.725
-0.876	-1.689	...	-2.554	-0.107
0.308	0.804	...	-0.515	0.791
-0.038	0.425	...	-1.015	0.233
0.931	-1.041	...	0.818	-0.350
-1.402	0.472	...	-0.208	-0.849
0.215	-0.513	...	1.822	-0.386
⋮	⋮	⋮	⋮	⋮
0.931	-1.041	...	0.818	-0.350

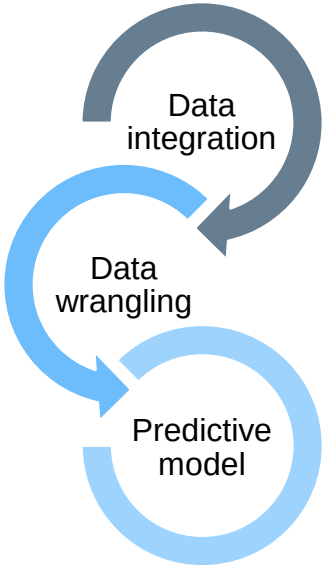
A variable is of interest if:
$p(Y|X\_j, \text{other\_variables}) \neq p(Y|\text{other\_variables})$

Formally, we want to test: $Y \perp X_j|X_{-j}$?

Analytical challenges at Novartis

Typical drug development program

~22,000 PYS exposure in 50 trials



eCRF data

- Age, Sex, Race
- Smoking History
- Medical History
- Disease History
- CV History
- Prior Biologics
- Con-Meds
- Adverse Events
- Key outcome scores

Other data

- Genomic
- Sensors
- Images

PRO data

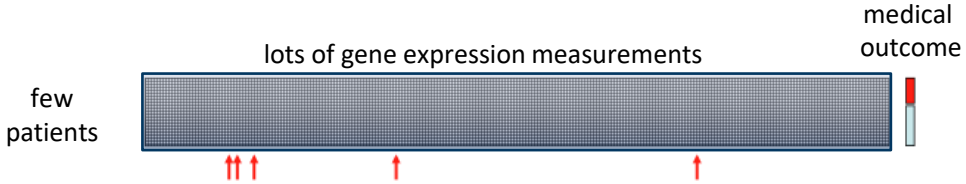
- Diary
- QoL

Lab & biomarker

- Chemistry
- CRP, ESR, Lipids

Feature selection

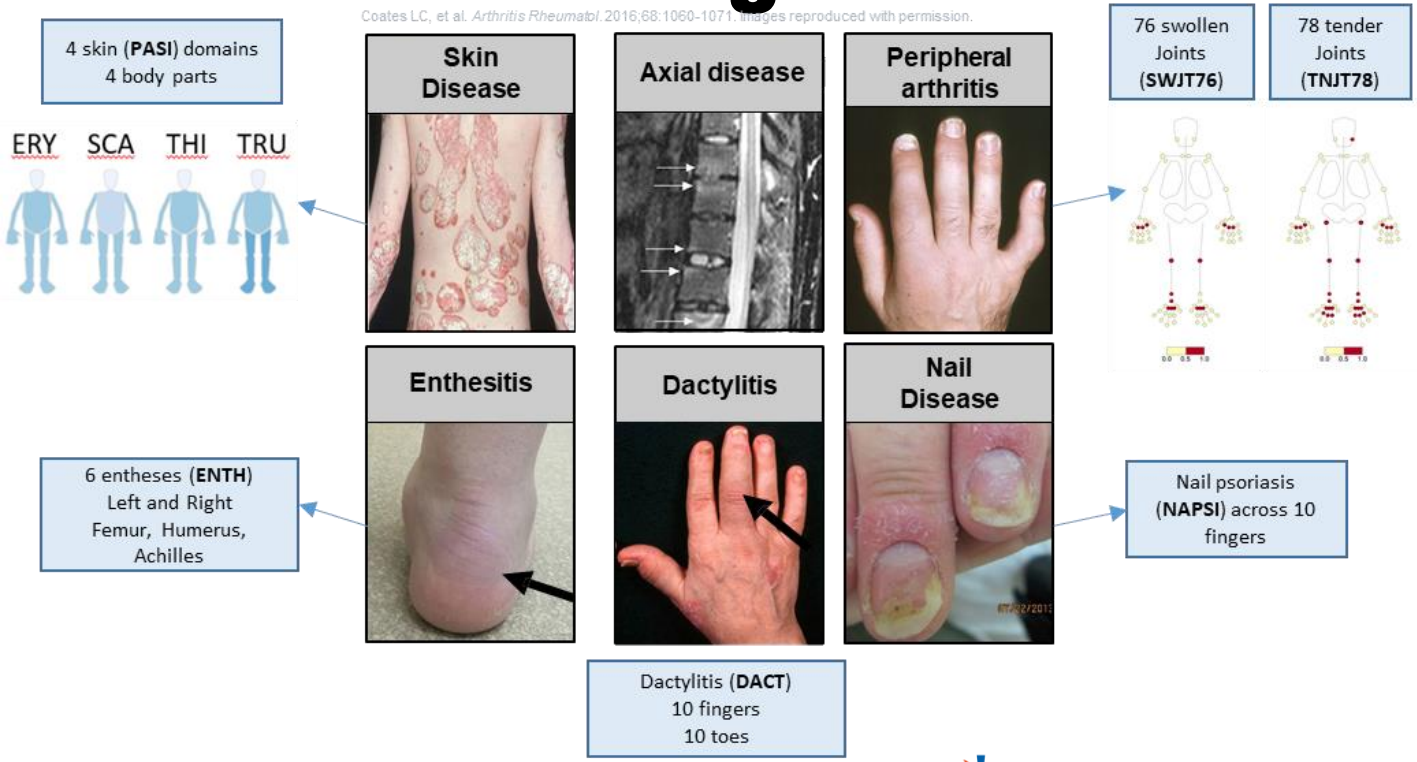
- One response *Y*, e.g. disease status, disease progression
- Thousands of variables *X*: genotype information, censored ...



Only a <u>subset</u> of features actually influences the outcome
Important in healthcare, *i.e. identify prognostic biomarkers*

NOVARTIS | Reimagining Medicine

Richness of clinical data: Example from Psoriatic Arthritis Program



Standard solutions: Independent Screening

- Test for an association between two variables at a time (e.g, a clinical outcome & Baseline characteristic)
- Repeat this for all possible characteristics, getting a large number of p-values
- Choose p-value threshold controlling False Discovery Rate (FDR) – e.g. Benjamini-Hochberg (BH)

- Appealing due to simplicity
- Does not directly address the question of interest (what is directly connected to the outcome)
- FDR control does not hold if the baseline factors are correlated
- Extensions that account for this tend to be on the conservative (Benjamini and Yekutieli, 2001)

Standard solutions: Regression model

- Perhaps the most widely utilized approach to assessing conditional probability relationships, particularly in epidemiological
- As all models are wrong, at best with careful use of model diagnostics and an iterative cycle of model criticism and elaboration, a model that is good approximation could be formed.
- Harrell (2001) provides general problem solving strategies for model selection
- Results from fitting such models are often interpreted in terms of conditional relationships (by using confidence intervals or p-values associated with model parameters)
- The interpretation is rather unclear due to the need to account for multiple comparisons.
- strategies are difficult to automate and focus on inference as opposed to variable selection and false discovery rate control.

Standard solution: Statistical Learning

- Statistical learning tends to focus on finding a good model for prediction
- Regularized regression (lasso, Tibshirani (1996) or elastic net, Zou and Hastie(2005)) provide final set of selected variables
- These methods have gained widespread popularity since their introduction to the literature, in particular due to the fact that they can handle a large number of explanatory variables.

- Does not directly focus on the problem of interest
- Recent work (see e.g. Su et al. (2017)) has observed that lasso has problems in selecting the proper model in practical applications, and that false discoveries may appear very early on the lasso path.

Define False Discovery Rate

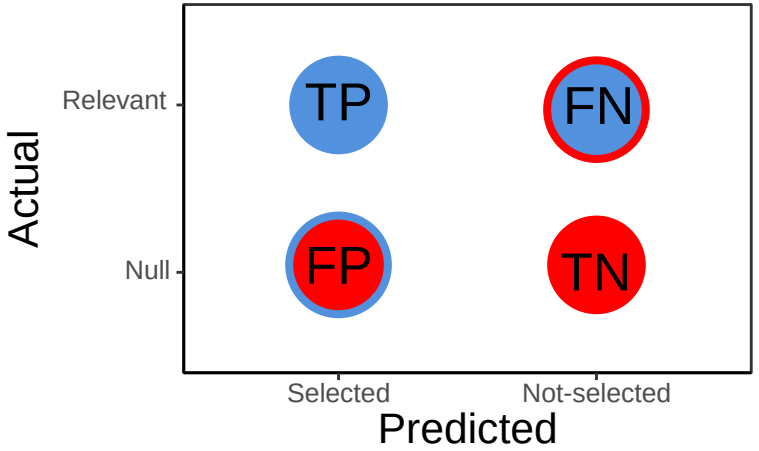
$\mathcal{H}_0$: the set of null features
$\hat{\mathcal{S}}$: the set of selected features

False Discovery Proportion:

$$\mathrm{FDP} = \frac{\left\|\hat{\mathcal{S}} \cap \mathcal{H}_0\right\|}{\left\|\hat{\mathcal{S}}\right\|} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}$$

False Discovery Rate:

$$\mathrm{FDR} \coloneqq \mathbb{E}\left[\,\mathrm{FDP}\right]$$



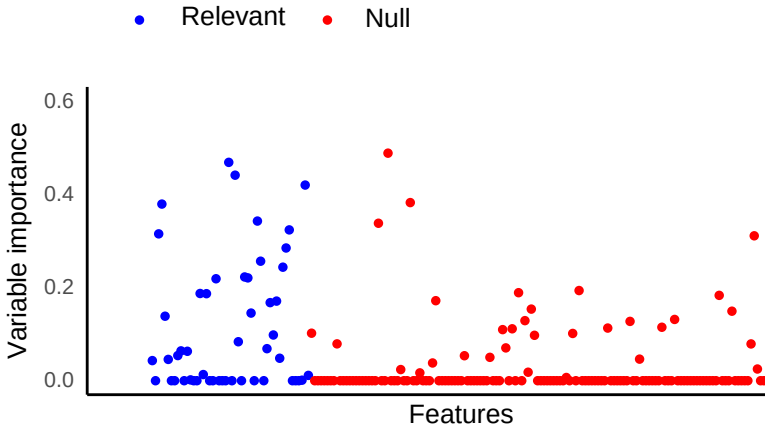
Statistical learning example …

n = 500 patients
d = 200 features

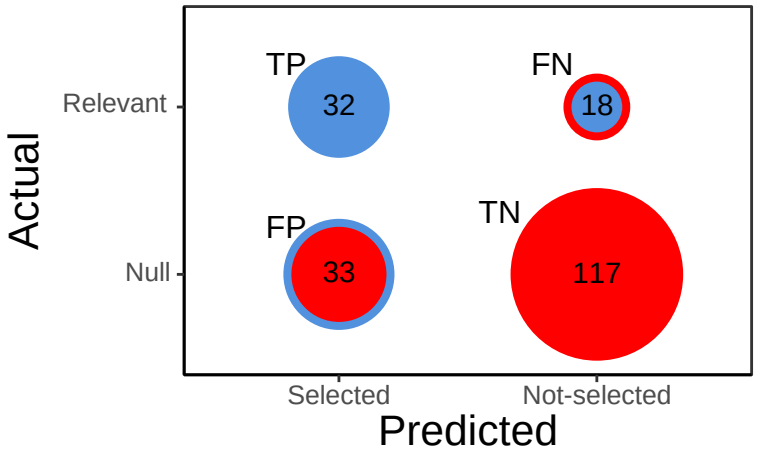
$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

$$Y = a(X_1 + \cdots + X_{50}) + \epsilon$$

Relevant



False Discovery Proportion = $\frac{\text{FP}}{\text{TP+FP}} = \frac{33}{32+33} = 0.51$



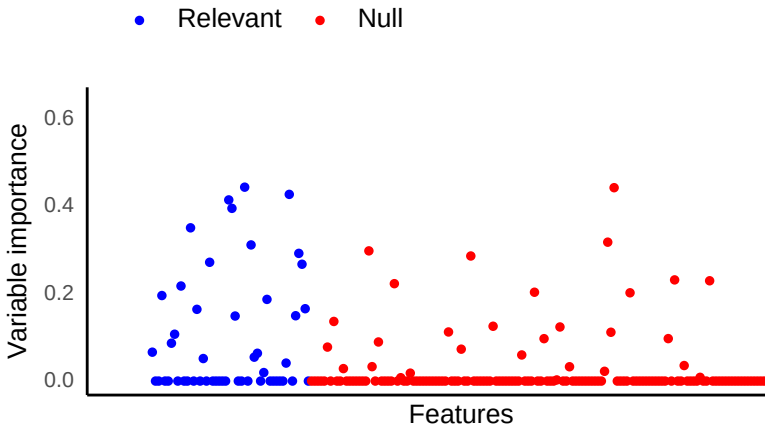
Statistical learning example …

n = 500 patients
d = 200 features

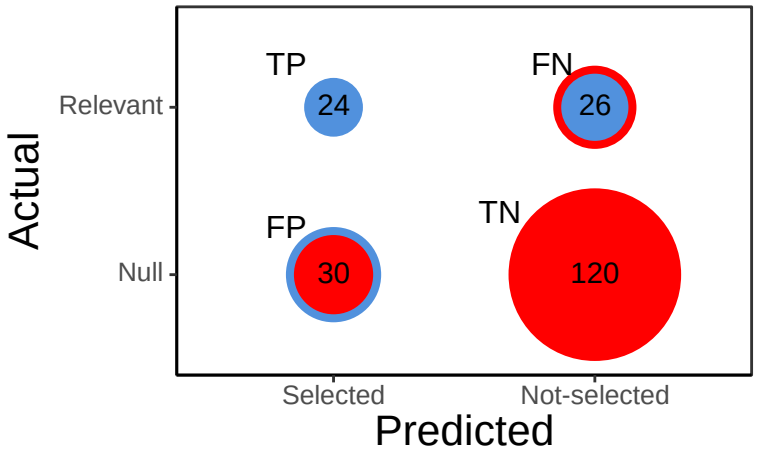
$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

$$Y = a(X_1 + \cdots + X_{50}) + \epsilon$$

Relevant



False Discovery Proportion $= \frac{\text{FP}}{\text{TP+FP}} = \frac{30}{24+30} = 0.56$



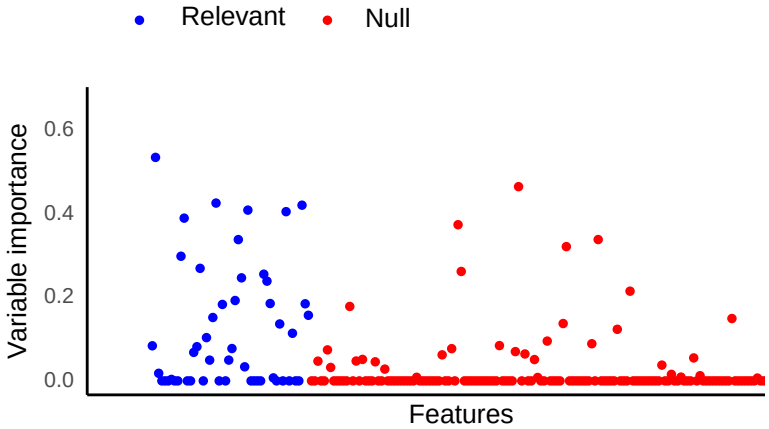
Statistical learning example …

n = 500 people
d = 200 features

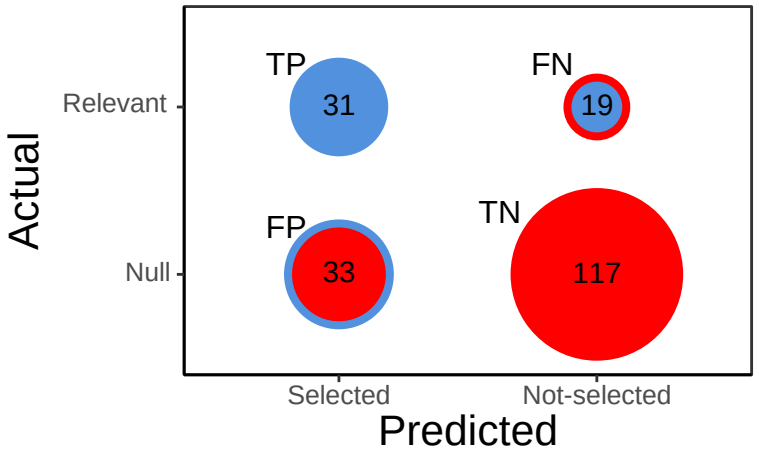
$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

$$Y = a(X_1 + \cdots + X_{50}) + \epsilon$$

Relevant



False Discovery Proportion $= \frac{\text{FP}}{\text{TP+FP}} = \frac{33}{31+33} = 0.52$



Conclusions on standard solutions

- The review underlines that the approaches commonly used in practice all have limitations
- Approaches with clear statistical characteristics or an understanding of uncertainty are not easily available
- Natural to look at the mathematical statistics literature for developments that have the potential to move into practice:
- Bayesian shrinkage priors and particularly the horseshoe prior (Carvalho, Polson and Scott ;2009, Polson and Scott; 2011, Piironen and Vehtari ; 2017) provide an approach accounting for uncertainty
- A further development with FDR control is the so called “knockoff” approach (Barber and Candes, 2015; Candes et al., 2018; Romano et al., 2019).

Horseshoe prior

Horseshoe prior: Carvalho, Polson and Scott (2009)

$$y_i = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

$$\beta_j|\lambda_j, \tau \sim \mathrm{N}(0, \tau^2\lambda_j^2)$$

$$\lambda_j \sim \mathrm{C}^+(0, 1)$$

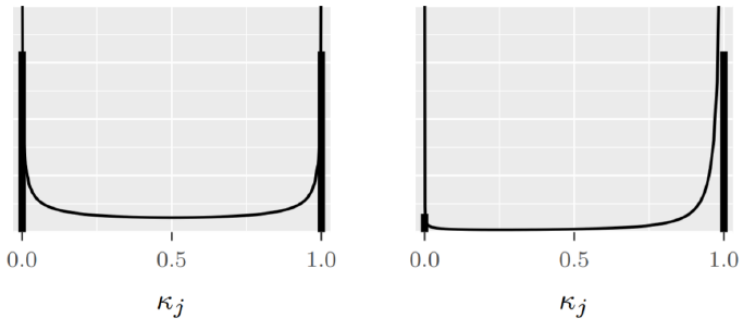
$\lambda_j^2 \sim$ Bernoulli	for spike-and-slab
$\lambda_j^2 \sim$ Exponential	for Laplace
$\lambda_j \sim$ Half-Cauchy	for Horseshoe

- τ is the global shrinkage parameter
- λ_j are the local shrinkage parameters
- With normalized covariates, the posterior mean of each regression coefficient is shrunk from the maximum likelihood solution by a factor κ_j

$$\overline{\beta}_j = (1 - \kappa_j)\hat{\beta}_{j,\mathrm{ML}} \qquad \kappa_j = \frac{1}{1+n\sigma^{-2}\tau^2\lambda_j^2}$$

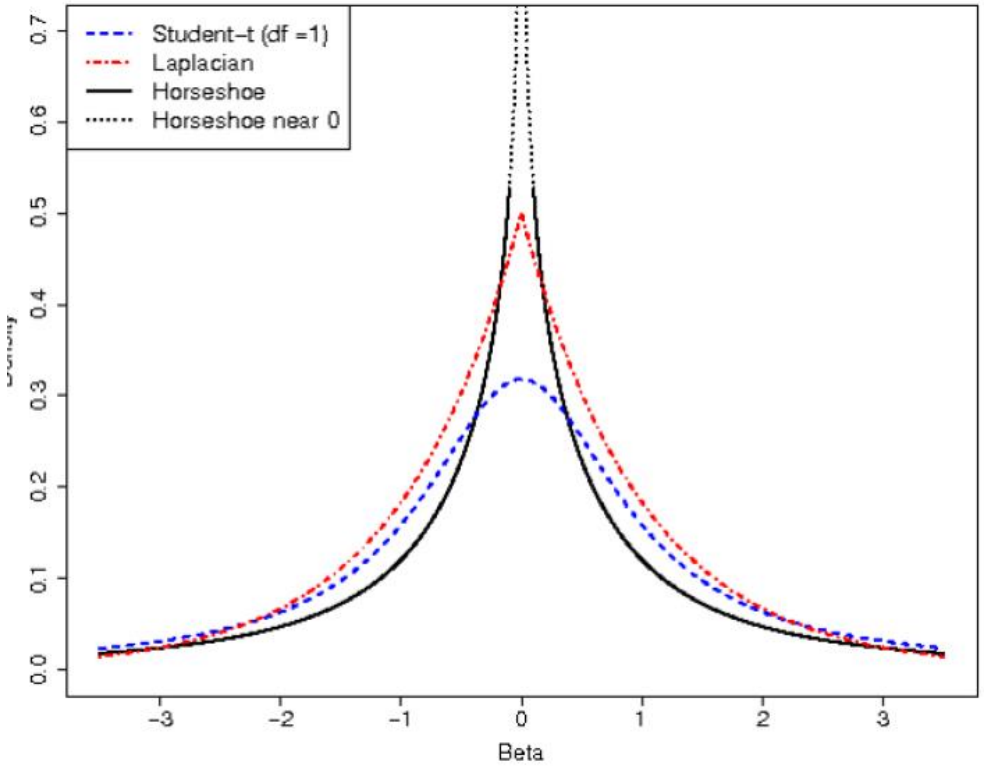
The name Horseshoe

- The name horseshoe comes from the distribution that is imposed on the shrinkage factors κ_j



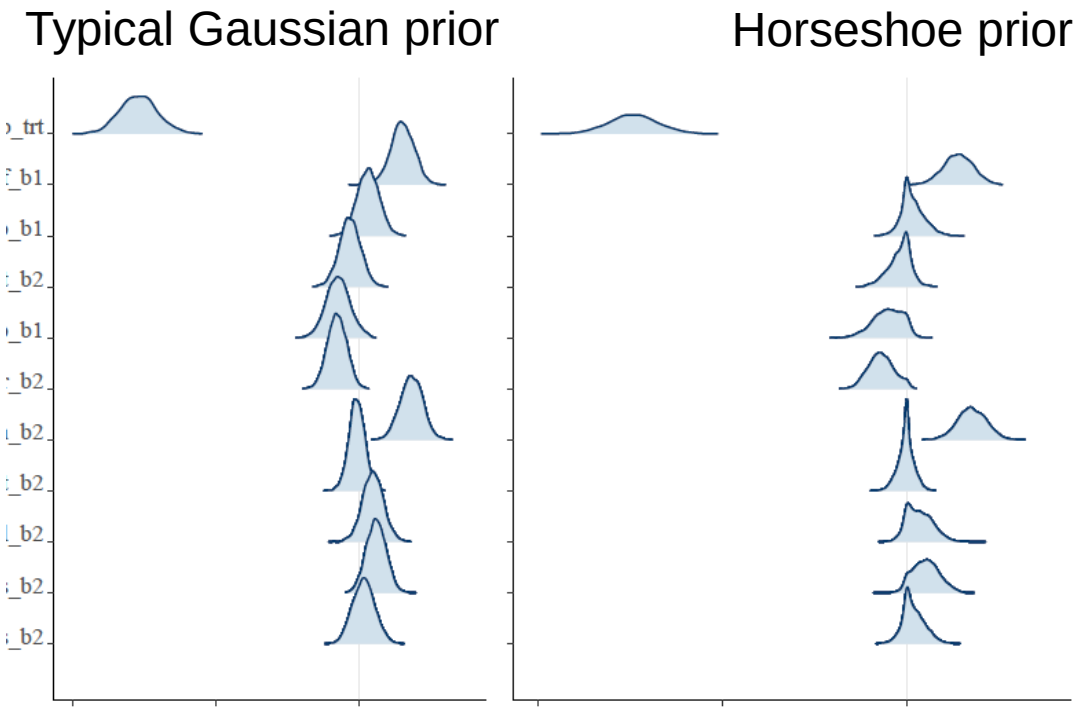
$\tau = 1$ $\tau = 0.1$

Visualizing types of shrinkage priors



http://proceedings.mlr.press/v5/carvalho09a/carvalho09a.pdf
9 j

Typical results of horseshoe shrinkage



Controlling the amount of shrinkage in the Horseshoe prior

- τ drives the amount of shrinkage
- The amount of shrinkage is typically unknown – τ is given a wide hyper prior
- The original horseshoe used a standard half Cauchy distribution
- Polson and Scott scaled this by the error variance
- Neither directly allow prior knowledge on the degree of sparsity to be incorporated
- However, Piironen and Vehtari (2017) developed a method to incorporate prior information about the number of non zero coefficients
- Extends to the GLM setting

Illustration with test data

- 1000 hypothetical patients from a clinical trial
- The endpoint tells you whether the patient has responded (y=TRUE) or not (y=FALSE) to a hypothetical drug.
- Also simulated data sets with count or survival outcomes (piecewise exponential model)
- For each patient we also have baseline covariates (age, sex, and additional 15 binary covariates, which you can think of as simplified genomic markers, e.g. SNPs).

Software and implementation

- Stan uses Hamilton Monte Carlo to implement Bayesian inference
- brms implements Stan models all standard biostatistics models (glms, survival mixed models) using code that mimics standard R models but also allows full prior specification
- ShinyStan for checking convergence and summary statistics

Fitting Horseshoe models using brms

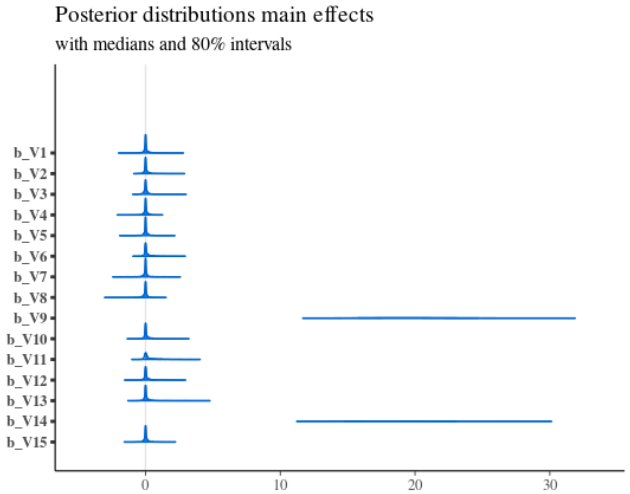
```
m_hs <- brm(as.numeric(y)~age + (V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11 +
                                 V12 + V13 + V14 + V15)^2,  data = dat,family=bernoulli(),
            prior = c(prior(normal(0, 1), class = "Intercept"),
                      # Prior guess of 20% of the terms are non-zero
                      prior(horseshoe(par_ratio =0.2), class = "b")),
            iter = 1000,  # just to save time
            chains = 2L,
            cores = 2L,
            # Need higher adapt_delta
            control = list(adapt_delta = .99),
            seed = 2217)
```

Shiny Stan for model diagnostics

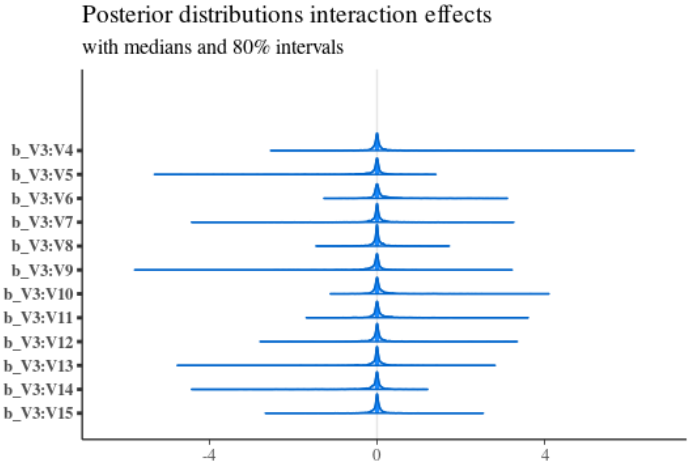

```
# summary statistics
m_hs
launch_shinystan(m_hs)
# posterior sample
```

Summarizing the posterior

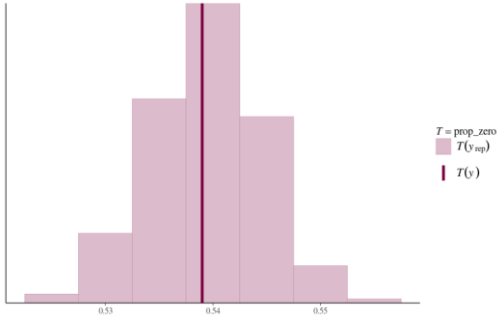
```
# posterior plots
plot_title <- ggtitle("Posterior distributions main effects",
                      "with medians and 80% intervals")
mcmc_areas(posterior,
           pars = paste("b_V",1:15,sep=""),
           prob = 0.8) + plot title
```



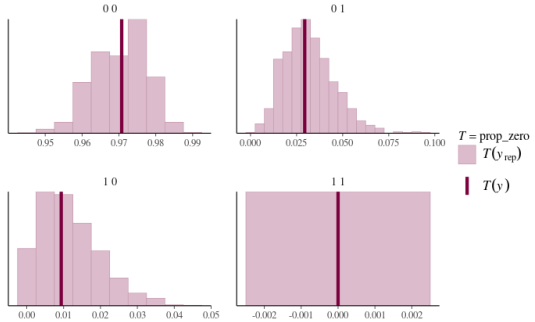

```
mcmc_areas(posterior,
           pars = colnames(posterior)[-c(1:17,123,124)][28:39],
           prob = 0.8) + plot_title
```



Model diagnostics: Posterior predictive checks




```
#posterior predictive checks
yrep<-posterior_predict(m_hs)
# 10 random histograms
ppc_hist(as.numeric(dat$y), yrep[1:10, ])
prop_zero <- function(x) mean(x == 0)
# overall posterior predictive check
prop_zero(dat$y) # check proportion of zeros in y
ppc_stat(as.numeric(dat$y),yrep, stat = "prop_zero", binwidth = 0.005)
```




```
ppc_stat_grouped(as.numeric(dat$y),
yrep,group=paste(dat$V9,dat$V14), stat = "prop_zero", binwidth = 0.005)
```

Negative binomial regression with spline

```
m_hs_count <- brm(bf(ycount~s(age,k=5) + (sex+V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 +
V10 + V11 +
                          V12 + V13 + V14 + V15)^2),  data = dat,family=negbinomial,
          prior = c(prior(normal(0, 1), class = "Intercept"),
                    # Prior guess of 20% of the terms are non-zero
                    prior(horseshoe(par_ratio =0.2), class = "b")),
          iter = 1000,  # just to save time
          chains = 2L,
          cores = 2L,
          # Need higher adapt_delta
          control = list(adapt_delta = .99),
          seed = 2217)
```

Proportional odds model

```
# A model with all main and interaction effects
m_hs_ord <- brm(bf(yord2~s(age,k=5) + (sex+V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11 +
                                       V12 + V13 + V14 + V15)^2), data = dat,family=cumulative("logit"),
                prior = c(prior(normal(0, 1), class = "Intercept"),
                          # Prior guess of 20% of the terms are non-zero
                          prior(horseshoe(par_ratio =0.2), class = "b")),
                iter = 1000,  # just to save time
                chains = 2L,
                cores = 2L,
                # Need higher adapt_delta
                control = list(adapt_delta = .99),
                seed = 2217)
```

Survival data convert to piecewise exponential set-up

L	M	N	O	P	Q	T	U	V
V10	V11	V12	V13	V14	V15	id	Event	Time
1	0	0	1	0	0	1	1	243
0	0	1	0	0	0	2	0	130
0	0	0	1	0	0	3	0	110
0	0	0	0	1	0	4	1	189
1	1	0	0	0	1	5	0	274
1	0	0	0	1	0	6	1	174
0	0	0	0	1	0	7	1	256
0	0	0	0	0	0	8	0	229
0	0	1	0	0	0	9	0	297
0	0	1	0	0	0	10	0	273
0	0	1	0	1	0	11	0	199
1	0	0	0	1	0	12	0	247
0	0	0	0	1	0	13	1	157
1	1	0	0	0	1	14	0	229
1	1	0	0	0	1	15	0	238
1	1	0	0	0	1	16	0	190
0	0	1	0	0	0	17	0	181
0	0	0	0	0	0	18	0	141
0	1	1	0	0	1	19	0	266

```
R-3.6.1> dat <- read.csv("datsurv.csv", header=TRUE)
R-3.6.1> dat<-select(dat,-c(y,ycount))
R-3.6.1> dat$Time<-ifelse(dat$Time==0,0.1,dat$Time)
R-3.6.1>
R-3.6.1> ped <- as_ped(Surv(Time,Event) ~ ., data = dat, cut = seq(0, 365, (365/8)),
+                id = "id")
R-3.6.1> ped
   id  tstart    tend         interval     offset ped_status age sex V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
1   1   0.000  45.625       (0,45.625]  3.8204558          0  28   0  0  1  0  1  1  1  0  0  1   1
2   1  45.625  91.250   (45.625,91.25]  3.8204558          0  28   0  0  1  0  1  1  1  0  0  1   1
3   1  91.250 136.875  (91.25,136.875]  3.8204558          0  28   0  0  1  0  1  1  1  0  0  1   1
4   1 136.875 182.500 (136.875,182.5]  3.8204558          0  28   0  0  1  0  1  1  1  0  0  1   1
5   1 182.500 228.125 (182.5,228.125]  3.8204558          0  28   0  0  1  0  1  1  1  0  0  1   1
6   1 228.125 273.750 (228.125,273.75]  2.6996820          1  28   0  0  1  0  1  1  1  0  0  1   1
7   2   0.000  45.625       (0,45.625]  3.8204558          0  29   0  1  1  0  0  0  1  0  0  0   0
8   2  45.625  91.250   (45.625,91.25]  3.8204558          0  29   0  1  1  0  0  0  1  0  0  0   0
9   2  91.250 136.875  (91.25,136.875]  3.6571308          0  29   0  1  1  0  0  0  1  0  0  0   0
```

```
# A model with all main and interaction effects
# piecewise exponential
# need to add prior for intervals so it works properly currently HS prior on them
n_hs_surv <- brm(bf(ped_status~offset(offset)+interval+s(age,k=5) + (sex+V1 + V2 + V3 + V4
                                    V12 + V13 + V14 + V15)^2), data = ped,family=poisson(),
            prior = c(prior(normal(0, 1), class = "Intercept"),
                      # Prior guess of 20% of the terms are non-zero
                      prior(horseshoe(par_ratio =0.2), class = "b")),
            iter = 1000,  # just to save time
            chains = 2L,
            cores = 2L,
            # Need higher adapt_delta
            control = list(adapt_delta = .99),
            seed = 2217)
```

Discussion points on horseshoe priors

- Produces a a final model with uncertainty surrounding all parameters
- Easy to implement for most standard biostatistics models
- Does not directly produce a final list of selected variables
  - No parameter is zero with probability 1
  - Predictive projection is an extension that looks into this
- Stan and brms work well for drug development size data sets (e.g. large pools)
  - Some limitation when conducting simulation studies
  - Literature on scalable computation using MCMC - e.g, Johndrow et al arXiv:1705.00841

Knockoffs

Purpose of Knockoff procedures

- Variable selection procedure, controls False Discovery Rate (FDR)
  - If we collect data on p independent variables and model

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

    only a subset of variables exist in true model
  - Goal is to “discover”, or select, variables that belong

$$\text{FDR} = \mathbb{E}\left(\frac{\#\text{ Incorrect "discoveries"}}{\#\text{Discoveries made}}\right)$$

$$= \mathbb{E}\left(\frac{\#\text{ Covariates we select that are not in true model}}{\#\text{ Covariates selected}}\right)$$

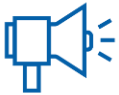
Application of Knockoff procedures

- Discovering clinically meaningful predictors

$$Disease\_Activity = \beta_1 AGE + \beta_2 BMI + \cdots$$

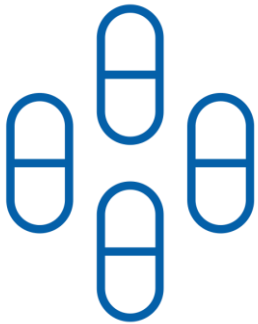
True discoveries
✓ Scientific insight
✓ Treatments

False discoveries
× Wasted time
× Wasted effort



- Maximize discoveries while keeping false discoveries manageable
- Desire: a procedure with high statistical power that controls FDR at specified level.

Intuition: knockoff methodology



Effective drugs improve outcomes compared to placebo
→ Placebos look like medicine, but should have no relationship with outcome

Select variables that are “more significant” than their knockoff variable
→ Knockoff variables look like real variables but have no relationship with outcome (given the real variables)

Knockoffs: In Detail

* Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055-2085.
* Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). 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.

Variable selection algorithm

1. Generate valid knockoffs

y_1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$\tilde{x}_{1,1}$	$\tilde{x}_{1,2}$	$\tilde{x}_{1,3}$
y_2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$\tilde{x}_{2,1}$	$\tilde{x}_{2,2}$	$\tilde{x}_{2,3}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮

2. a) Fit model to real and knockoff variables

$$y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \tilde{\beta}_1 \tilde{x}_{i,1} + \cdots + \tilde{\beta}_p \tilde{x}_{i,p} + e_i$$

2 b) Compare importance of real variable and its knockoff in model

$$W_j = |\hat{\beta}_j| - |\hat{\tilde{\beta}}_j|$$

3) Select variable X_j if W_j is large in a way that controls false discovery rate

1st step: Construct knockoffs

- Two properties
  1. **Exchangeability**: The original and the knockoff variables are pairwise exchangeable, in other words, the joint distribution of $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ is invariant under any swapping of X_j and $\tilde{X}_j$.

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \overset{\mathrm{d}}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$
$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \overset{\mathrm{d}}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

  2. **Conditional independence**: Given the original features $\boldsymbol{X}$ the knockoffs are conditionally independent with the target: $\tilde{\boldsymbol{X}} \perp\!\!\!\perp Y \,|\, \boldsymbol{X}$

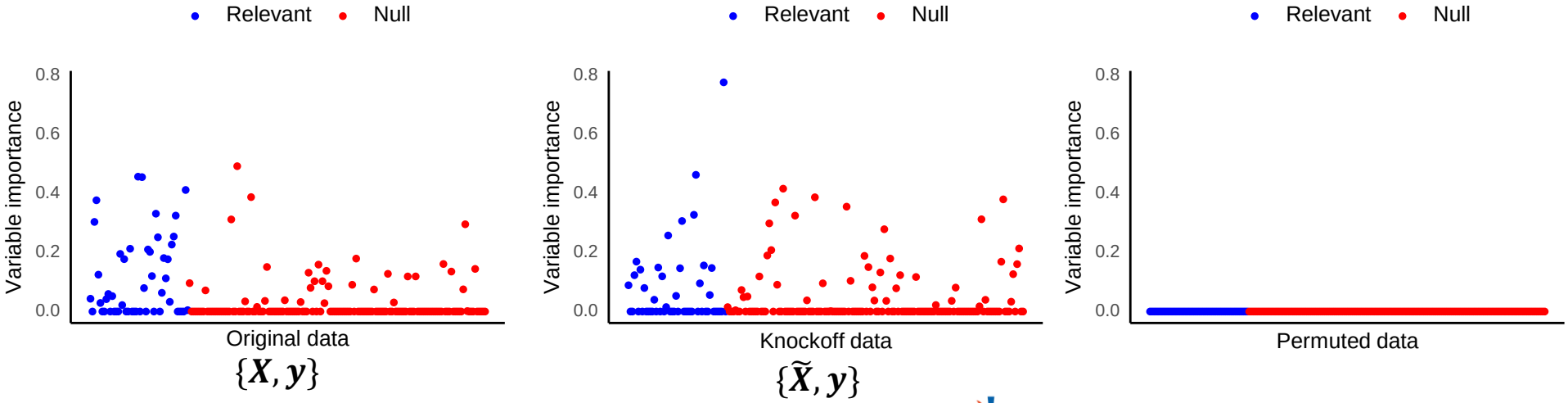
- Active area of research:
  - **Gaussianity assumption (Candes et al. 2016)**
  - Hidden-markov-model (Sesia et al. 2019)
  - General Graphical models(Bates et al. 2020)
  - Deep learning (Romano et al. 2019)
  - KnockoffGAN (Jordan et al. 2019)

1st step: Construct knockoffs (Candes et al. 2016)

- $\boldsymbol{X} \sim \boldsymbol{N}(\mu, \boldsymbol{\Sigma})$ and $\tilde{\boldsymbol{X}} \sim \boldsymbol{N}(\mu, \boldsymbol{\Sigma})$
- $(\boldsymbol{X}, \tilde{\boldsymbol{X}}) \sim \boldsymbol{N}\left(\begin{bmatrix}\mu \\ \mu\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{\mathbf{s}\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{\mathbf{s}\} & \boldsymbol{\Sigma}\end{bmatrix}\right)$
- Sample $\tilde{\boldsymbol{X}}$ from $\tilde{\boldsymbol{X}}|\boldsymbol{X}$

$(X_1, X_2, \tilde{X}_1, \tilde{X}_2)$

Y	X_1	X_2	...	X_p	$\tilde{X}_1$	$\tilde{X}_2$	...	$\tilde{X}_p$
1.128	-0.300	0.416	...	-0.328	-0.310	-0.568	...	-0.396
-0.725	-0.310	-0.568	...	-0.396	0.215	-0.513	...	1.822
-0.107	-0.876	-1.689	...	-2.554	0.931	-1.041	...	0.818
0.791	0.308	0.804	...	-0.515	-0.876	-1.689	...	-2.554
0.233	-0.038	0.425	...	-1.015	-0.300	0.416	...	-0.328
-0.350	0.931	-1.041	...	0.818	1.402	0.472	...	-0.208
-0.849	-1.402	0.472	...	-0.208	-0.038	0.425	...	-1.015
-0.386	0.215	-0.513	...	1.822	0.931	-1.041	...	0.818
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-0.350	0.931	-1.041	...	0.818	0.308	0.804	...	-0.515



2nd step: Calculate a knockoff statistic

- Step 2a: Use dataset $\{[\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y}\}$ to derive importance scores for original and knockoff features:

$$Q \coloneqq (Z_1, Z_2, \ldots, Z_p, \tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_p)$$

- Step 2b: We construct the knockoff statistic W_j as $W_j = f(Z_j, \tilde{Z}_j)$ for some function f

LASSO regression

$$\hat{\mathbf{b}}(\lambda) = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\, \mathbf{b} \right\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}$$

$$Z_j = |\hat{b}_j(\lambda)|, \quad \tilde{Z}_j = |\hat{b}_{j+p}(\lambda)|$$

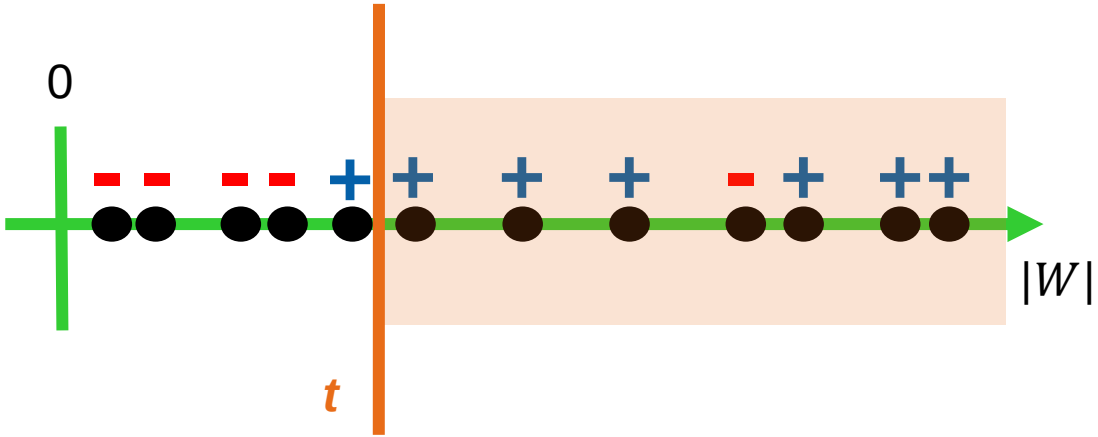
$$W_j^{\mathrm{LCD}} = |\hat{b}_j(\lambda)| - |\hat{b}_{j+p}(\lambda)|$$

Random forests

Z_j can be an importance score
Gini importance, permutation importance etc

$$W_j^{\mathrm{RF}} = |Z_j| - |\tilde{Z}_j|$$

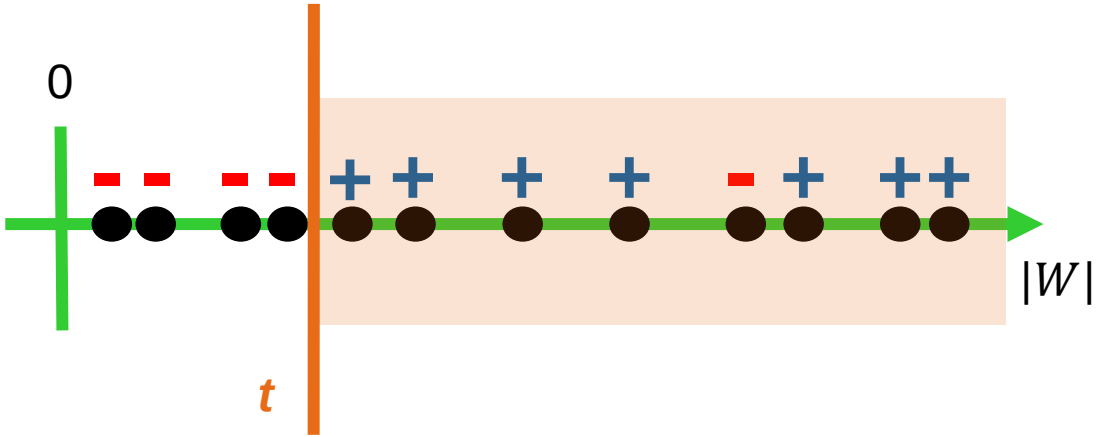
3rd step: Calculate threshold for controlled FDR



FDR = 0.30

$$\widehat{\mathrm{FDP}}(t) = \frac{1 + \left|\{j : W_j \leq -t\}\right|}{\left|\{j : W_j \geq t\}\right|} = 0.33$$

3rd step: Calculate threshold for controlled FDR



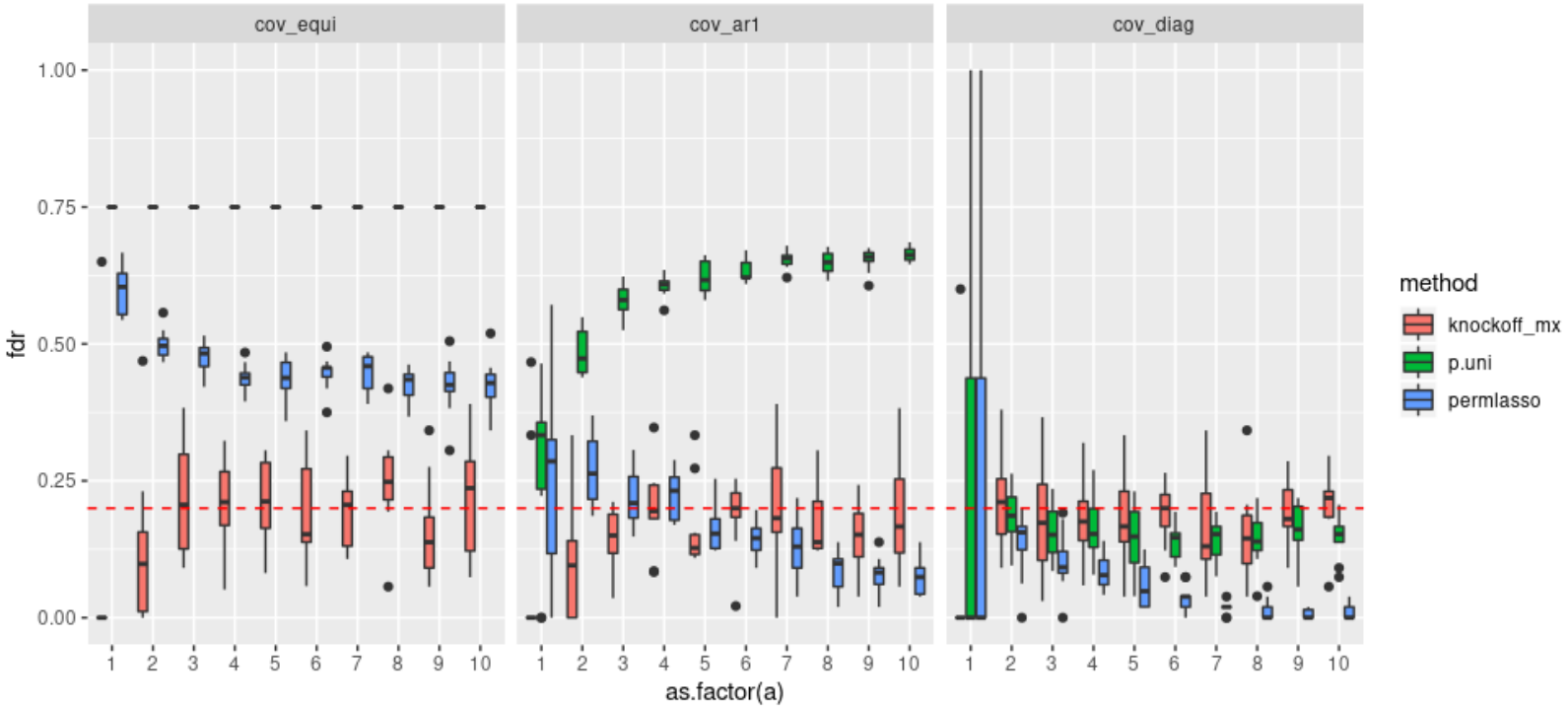
FDR = 0.30

$$\widehat{\mathrm{FDP}}(t) = \frac{1 + \left|\{j : W_j \leq -t\}\right|}{\left|\{j : W_j \geq t\}\right|} = 0.28$$

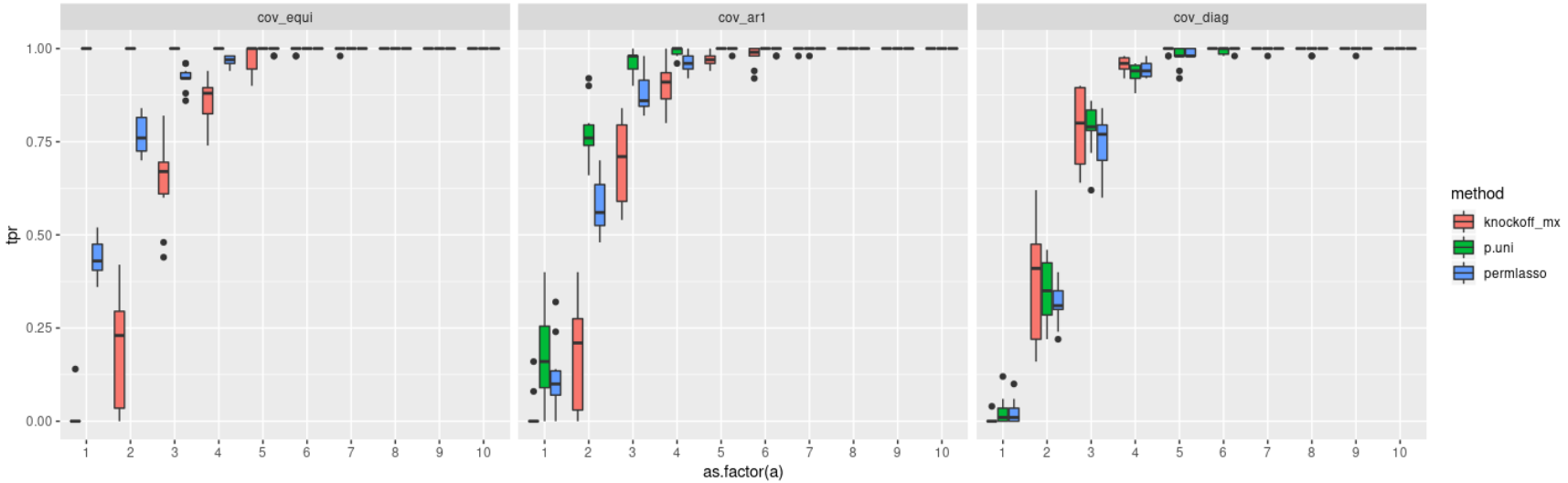
Simulation study

- We generated synthetic data sets of a mimicking the size and structure of a clinical trial pool for a marketed drug
- In all cases, we sampled n = 2000 covariate-response pairs (X, Y), where
  - Predictor X is continuous (p=200)
  - Response Y depends on only 50 of the *200* measurements in X, with strength of association or amplitude a (varied between 1 and 6)
- Different correlation structures are applied to X
  - Equi-correlated covariates ρ = 0.5
  - AR(1) covariates, ρ = 0.5
  - Independent
- Compare the knockoff approach with independent screening using the BH FDR and the permutation lasso

False discovery proportion



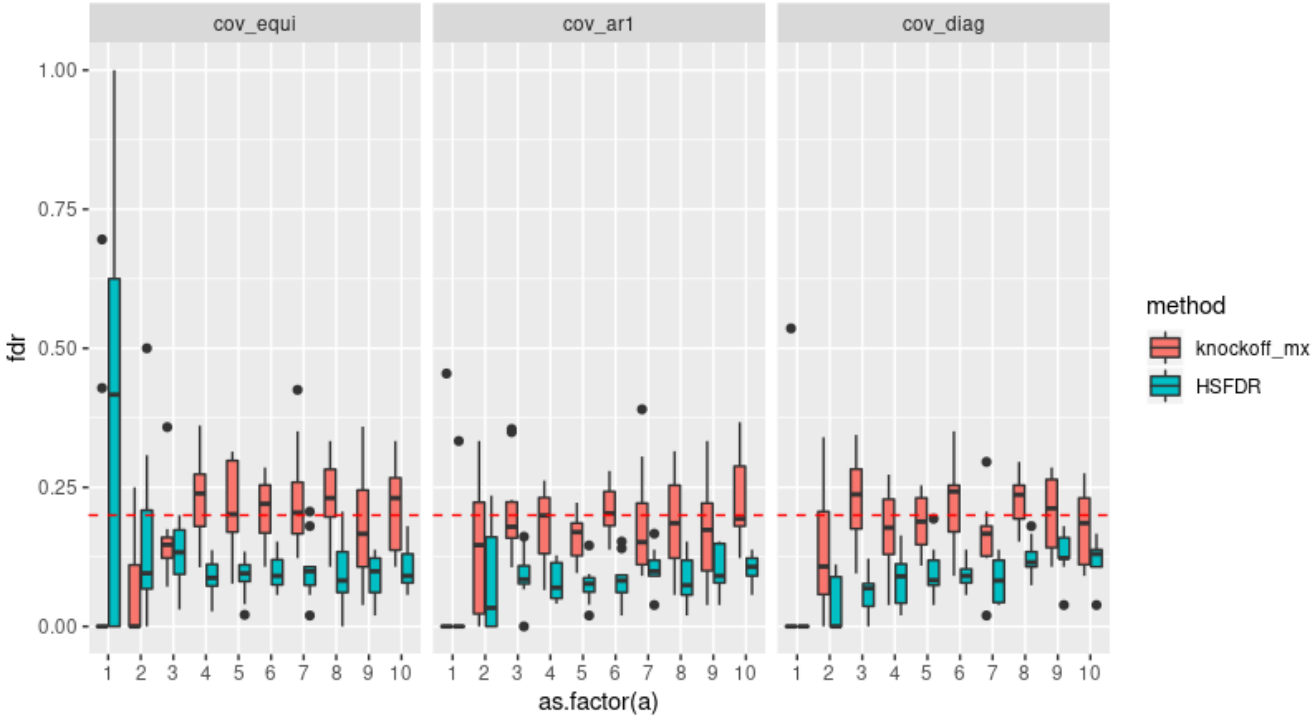
True discovery proportion (power)



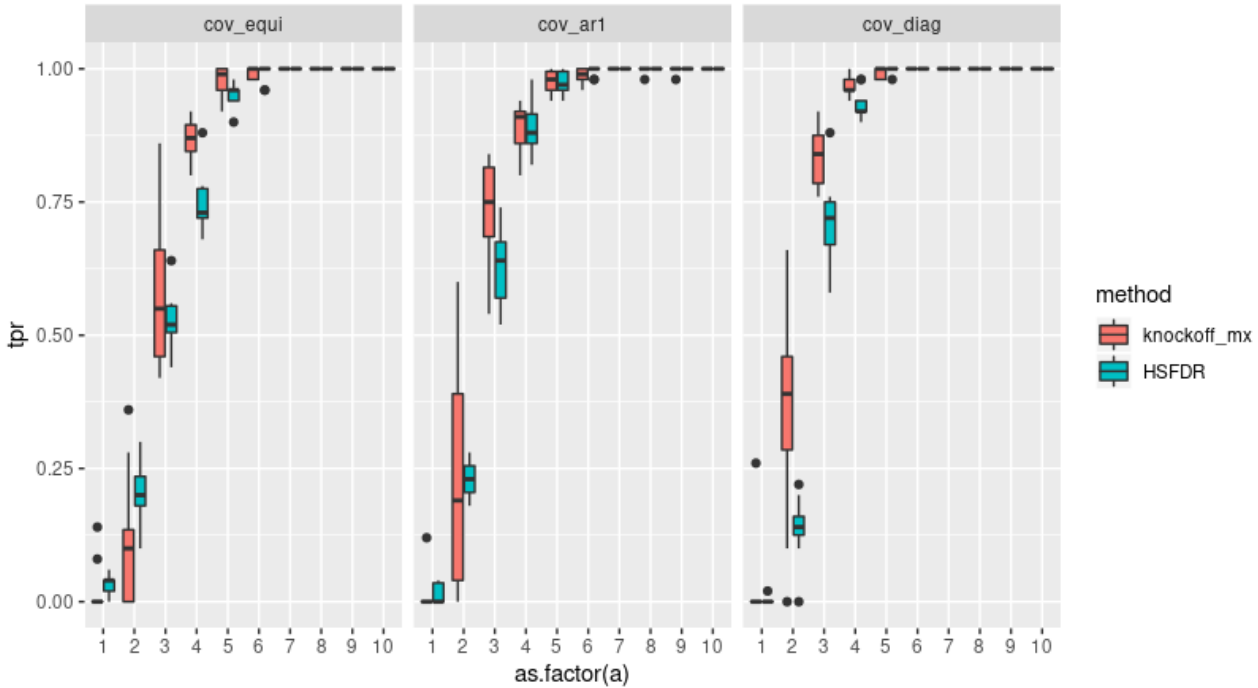
Knockoff v Horseshoe

- It is difficult to directly compare as they answer slightly different questions
  - Remember the HS does not directly produce a final list of selected variables
- Need to define a way to do this
  - Based on links between the HS and other Bayesian shrinkage methods (spike and slab type priors), some implementations do offer an explicit probability of a covariate being 0 (the monomvn R package)
  - However, this probability can only be thought of as a local false discovery rate so some clever post processing is required
  - Therefore, we create something akin to a two sided p-value from the posterior probability of a covariate being greater than or less than 0 and post-process similar to the BH FDR

False discovery proportion



True discovery proportion (power)



Discussion and extensions

- We only covered the model X-knockoff approach that is suitable for continuous covariates
- The knockoff methodology is based on simulating a single knockoff
  - Remember FDR is the expectation of the FDP
- Therefore, different conclusions could be reached if the analysis is re-run
- Currently working on an extensions with the Oxford big data institute:
  - Fully cover a mixture of discrete and continuous covariates
  - Using multiple knockoffs
- A further extension to predictive markers (markers affecting treatment effect) is also being developed

Thank you and questions

Knockoff references

- Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055-2085.
- Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). 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.

Horseshoe prior references

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009). Handling sparsity via the horseshoe. In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (D. van Dyk and M. Welling, eds.). Proceedings of Machine Learning Research 5 73–80. PMLR. MR2650751

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. Biometrika 97 465–480. MR3036256

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. In Bayesian statistics 9 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 501–538. Oxford University Press, Oxford. MR3204017

Piironen, J. and Vehtari, A. (2017) Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics, 11(2):5018–5051