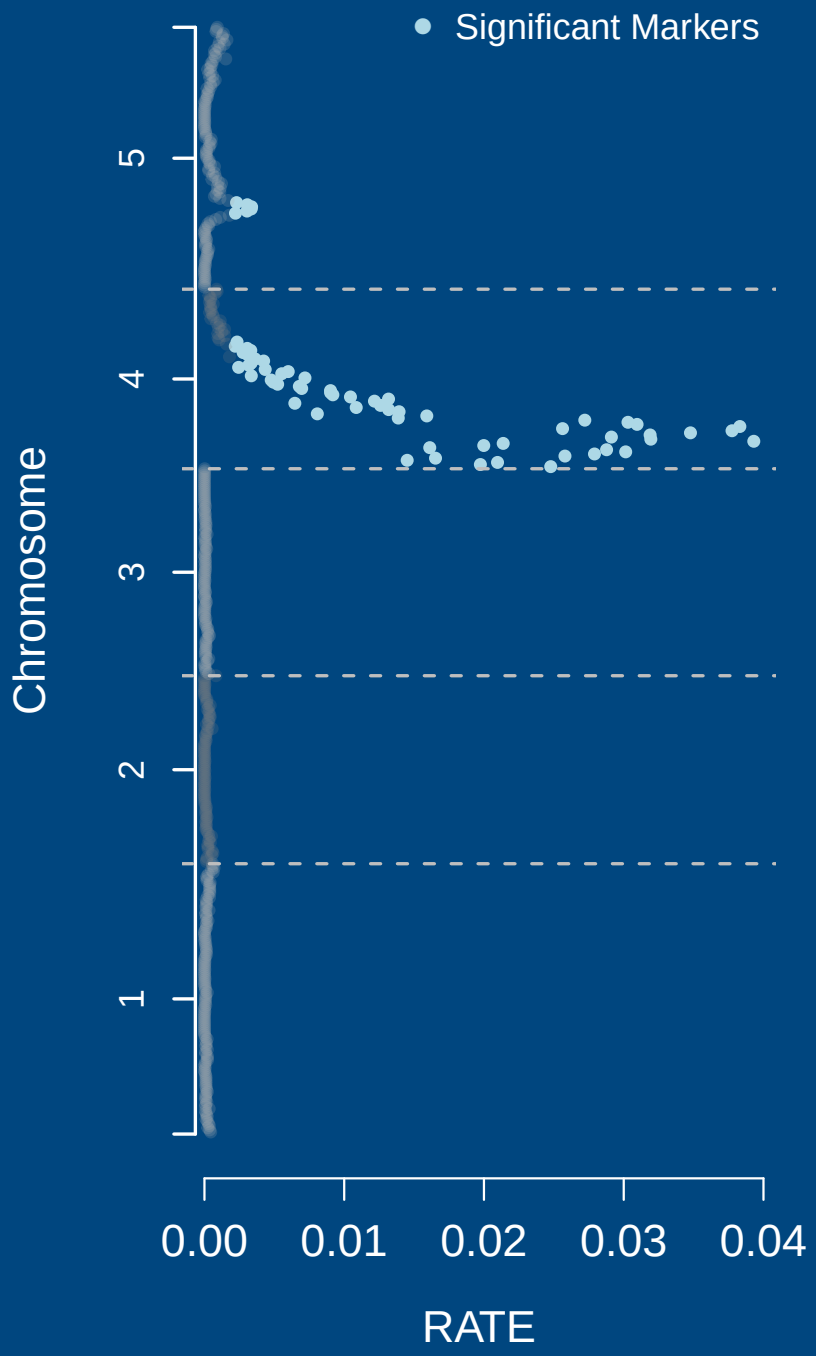


Predictor Variable Prioritization in Nonlinear Models with RATE

Seth Flaxman

25 April 2018

Explain AI@Imperial Workshop

Imperial College London

Interpretation

Regression:

$$Y = \boldsymbol{X}\beta + \epsilon$$

$$f(\boldsymbol{x}) = E[y|\boldsymbol{x}]$$

- Which predictor variables are important? (variable importance)
- Does the model fit the data? Better or worse than another model? (model criticism / selection)
- Can we exclude some of the variables? (variable selection)
- What is the relationship between a given predictor variable and the output? (effect size)
- What statistical guarantees (confidence intervals) or probabilities (posterior uncertainty intervals) can we attach to the above answers?

Explanation

Source: Goodman and Flaxman, "European Union Regulations on Algorithmic Decision Making and a 'Right to Explanation' ", *AI Magazine*, 2017

Explanation

- What data about me does the algorithm use?
- Why was I shown this ad?
- Why was my insurance application declined?
- Why was I given this offer (features, down payment, overall price, financing) for a car?

Putting it together: inference

inference: drawing conclusions from data

[1] https://www.youtube.com/watch?v=Qi1Yry33TQE and http://www.argmin.net/2017/12/05/kitchen-sinks/

Putting it together: inference

inference: drawing conclusions from data

- This is already hard with classical statistics (cf. scientific replicability crisis)!

[1] https://www.youtube.com/watch?v=Qi1Yry33TQE and http://www.argmin.net/2017/12/05/kitchen-sinks/

Putting it together: inference

inference: drawing conclusions from data

- This is already hard with classical statistics (cf. scientific replicability crisis)!
- What about with machine learning?

[1] https://www.youtube.com/watch?v=Qi1Yry33TQE and http://www.argmin.net/2017/12/05/kitchen-sinks/

Putting it together: inference

inference: drawing conclusions from data

- This is already hard with classical statistics (cf. scientific replicability crisis)!
- What about with machine learning?

Rahimi and Recht, NIPS 2017[1]:

> *Machine learning has become alchemy.*
> *Alchemy worked.*
> *If you're building photo sharing systems, alchemy is ok.*
> *We're building systems that govern healthcare, and mediate our civic dialogue.*
> ⚠ *We influence elections.* ⚠

[1] https://www.youtube.com/watch?v=Qi1Yry33TQE and http://www.argmin.net/2017/12/05/kitchen-sinks/

Disclaimer: causality and representativeness

- Without extra assumptions about data collection or underlying mechanisms we are explaining / interpreting statistical associations (i.e. joint or conditional distributions, posterior predictive distributions, predictions, etc.) rather than cause and effect.

Disclaimer: causality and representativeness

- Without extra assumptions about data collection or underlying mechanisms we are explaining / interpreting statistical associations (i.e. joint or conditional distributions, posterior predictive distributions, predictions, etc.) rather than cause and effect. But scientifically relevant inferences can be evaluated with other evidence!

Disclaimer: causality and representativeness

- Without extra assumptions about data collection or underlying mechanisms we are explaining / interpreting statistical associations (i.e. joint or conditional distributions, posterior predictive distributions, predictions, etc.) rather than cause and effect. But scientifically relevant inferences can be evaluated with other evidence!
- Results may not generalize to different settings or different populations.

Disclaimer: causality and representativeness

- Without extra assumptions about data collection or underlying mechanisms we are explaining / interpreting statistical associations (i.e. joint or conditional distributions, posterior predictive distributions, predictions, etc.) rather than cause and effect. But scientifically relevant inferences can be evaluated with other evidence!
- Results may not generalize to different settings or different populations. But inferences can be compared across settings more meaningfully than predictive models.

From linear methods to nonlinear methods

$$Y \sim X_1 + X_2 + X_3$$

From linear methods to nonlinear methods

$$Y \sim X_1 + X_2 + X_3$$
$$Y \sim X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3$$

From linear methods to nonlinear methods

$$Y \sim X_1 + X_2 + X_3$$
$$Y \sim X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3$$
$$Y \sim f_1(X_1) + f_2(X_2) + f_3(X_3)$$

From linear methods to nonlinear methods

$$
\begin{aligned}
Y &\sim X_1 + X_2 + X_3 \\
Y &\sim X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3 \\
Y &\sim f_1(X_1) + f_2(X_2) + f_3(X_3) \\
&\vdots \\
Y &\sim f(X_1, X_2, X_3)
\end{aligned}
$$

From linear methods to nonlinear methods

$$
\begin{aligned}
Y &\sim X_1 + X_2 + X_3 \\
Y &\sim X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3 \\
Y &\sim f_1(X_1) + f_2(X_2) + f_3(X_3) \\
&\vdots \\
Y &\sim f(X_1, X_2, X_3)
\end{aligned}
$$

Choices for f: neural network, Gaussian process, decision tree, boosting, etc: methods that succeed by considering many and higher order interactions!

From linear methods to nonlinear methods

$$
\begin{aligned}
Y &\sim X_1 + X_2 + X_3 \\
Y &\sim X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3 \\
Y &\sim f_1(X_1) + f_2(X_2) + f_3(X_3) \\
&\vdots \\
Y &\sim f(X_1, X_2, X_3)
\end{aligned}
$$

Choices for f: neural network, Gaussian process, decision tree, boosting, etc: methods that succeed by considering many and higher order interactions!

Our central question: rank the importance of the predictor variables, not just marginally but taking into account these interactions.

Effect sizes in linear methods

Fit a linear model:

$$Y = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

The regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are effect sizes.

Effect sizes in linear methods

Fit a linear model:

$$Y = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

The regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are effect sizes.

Regression coefficients in an ordinary least squares setup, where $\boldsymbol{X}$ is the design matrix:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top y$$

Effect sizes in linear methods

Fit a linear model:

$$Y = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

The regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are effect sizes.
Regression coefficients in an ordinary least squares setup, where $\boldsymbol{X}$ is the design matrix:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top y$$

Define the projection operator:

$$\widehat{\boldsymbol{\beta}} = \text{Proj}(\boldsymbol{X}, y) := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top y$$

Effect sizes in linear methods

Fit a linear model:

$$Y = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

The regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are effect sizes.
Regression coefficients in an ordinary least squares setup, where $\boldsymbol{X}$ is the design matrix:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top y$$

Define the projection operator:

$$\widehat{\boldsymbol{\beta}} = \text{Proj}(\boldsymbol{X}, y) := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top y$$

What information does this vector contain? For standardised variables X_1 and Y:

$$\hat{\beta}_1 = \text{Corr}(X_1, Y)$$

$$\hat{\beta}_2 = \text{Corr}(X_2, Y)$$

Effect size "analogue"

Generic non-linear model:

$$Y \sim f(\boldsymbol{X})$$

Once we learn a function $\hat{f}$ we can calculate our predictions

$$\hat{Y} = \hat{f}(\boldsymbol{X})$$

Basic idea: $\widetilde{\beta} := \text{Proj}(\boldsymbol{X}, \hat{Y})$ is still a sensible quantity of interest.

$$\tilde{\beta}_1 = \text{Corr}(X_1, \hat{Y})$$
$$\tilde{\beta}_2 = \text{Corr}(X_2, \hat{Y})$$

Cf. model compression [Bucilă, Caruana, Niculescu-Mizil KDD 2006]

Gaussian processes

- Bayesian framework for specifying a prior over functions:

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}))$$

Gaussian processes

- Bayesian framework for specifying a prior over functions:

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}))$$

- Regression:

$$y_i = f(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Gaussian processes

- Bayesian framework for specifying a prior over functions:

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}))$$

- Regression:

$$y_i = f(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Conjugate model:

$$[f(\boldsymbol{x}_1) \dots f(\boldsymbol{x}_n)]^\top \sim \mathcal{N}(\boldsymbol{\mu}, K)$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \dots, n$$

Gaussian processes

- Bayesian framework for specifying a prior over functions:

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}))$$

- Regression:

$$y_i = f(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Conjugate model:

$$[f(\boldsymbol{x}_1) \dots f(\boldsymbol{x}_n)]^\top \sim \mathcal{N}(\boldsymbol{\mu}, K)$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \dots, n$$

Gaussian processes

- Bayesian framework for specifying a prior over functions:

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}))$$

- Regression:

$$y_i = f(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Conjugate model:

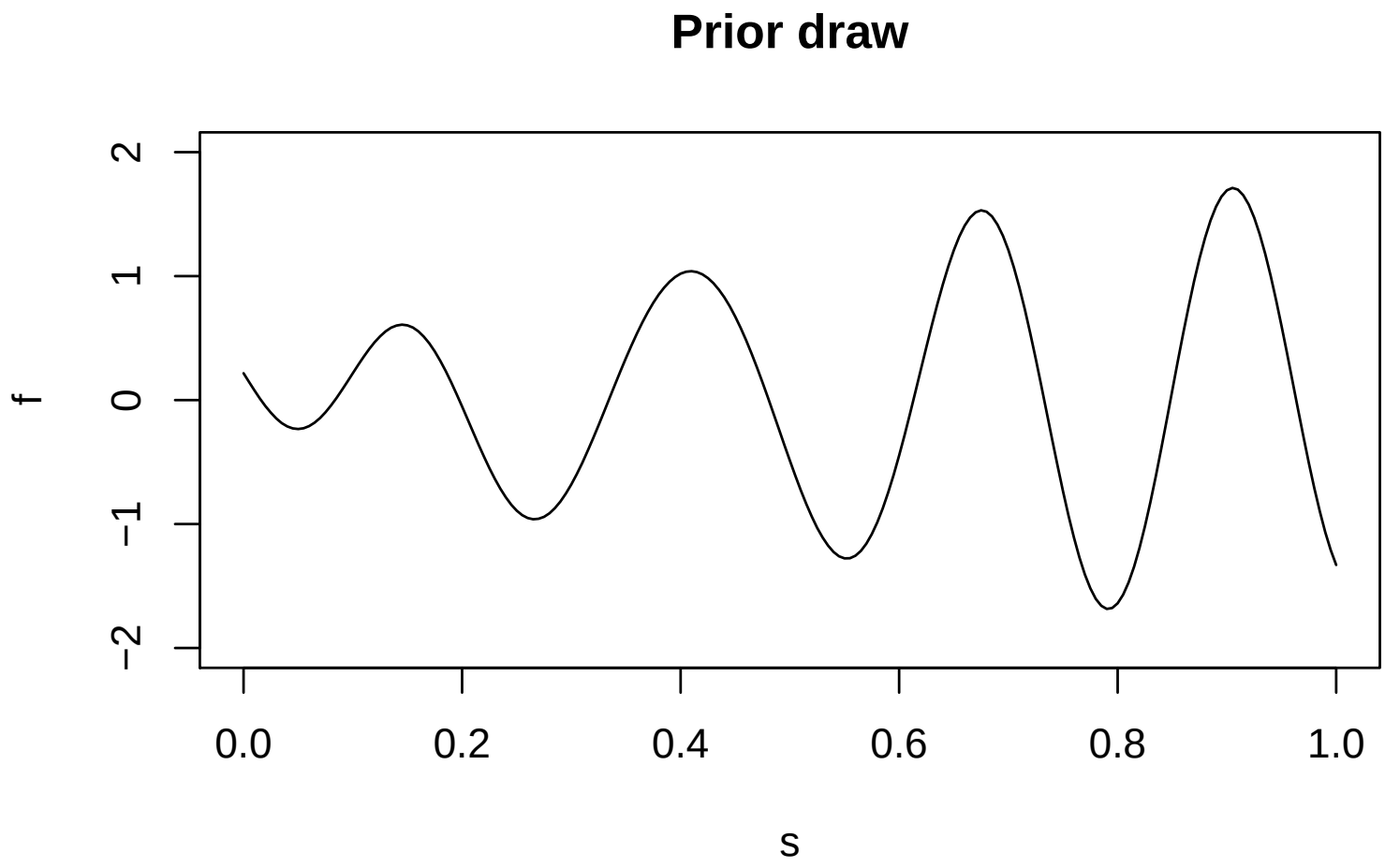
$$[f(\boldsymbol{x}_1) \dots f(\boldsymbol{x}_n)]^\top \sim \mathcal{N}(\boldsymbol{\mu}, K)$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \dots, n$$

- Closed form posterior!

$$f(\boldsymbol{x}) | Y \sim \mathcal{N}\left(K(K + \sigma^2 I)^{-1} Y, K - K(K + \sigma^2 I)^{-1} K^\top\right)$$

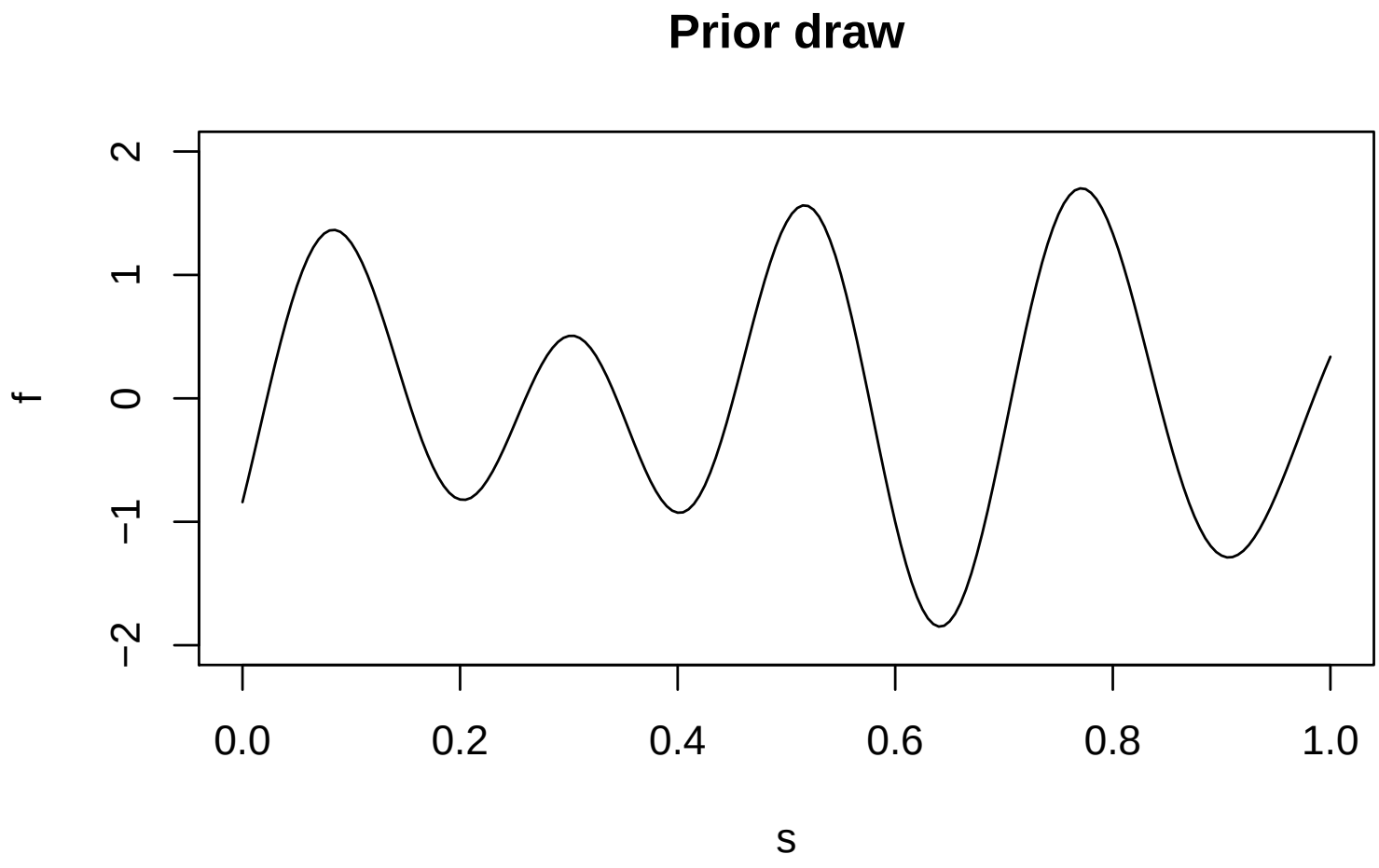
GP illustration

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \ldots, n$$



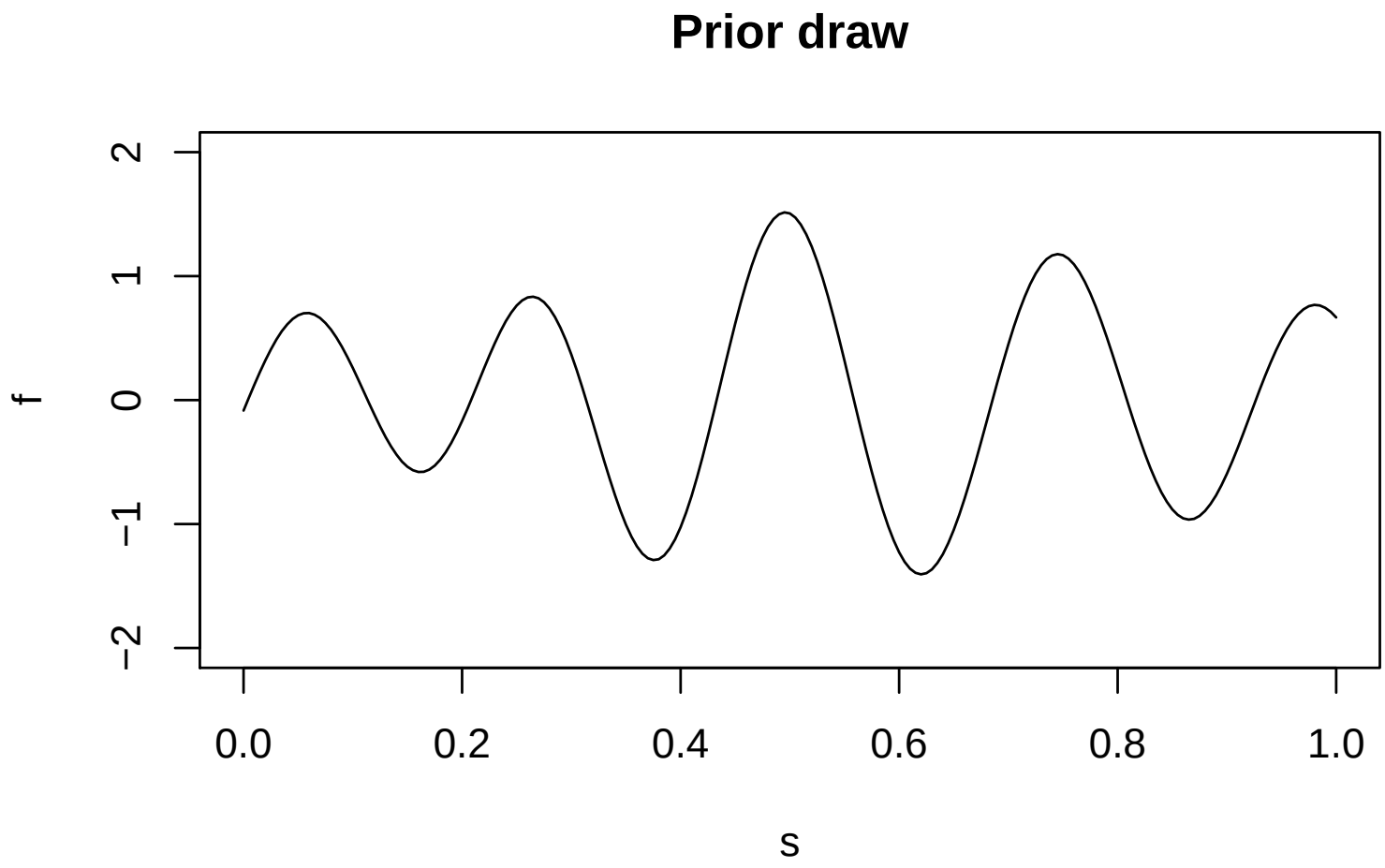
GP illustration

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \ldots, n$$



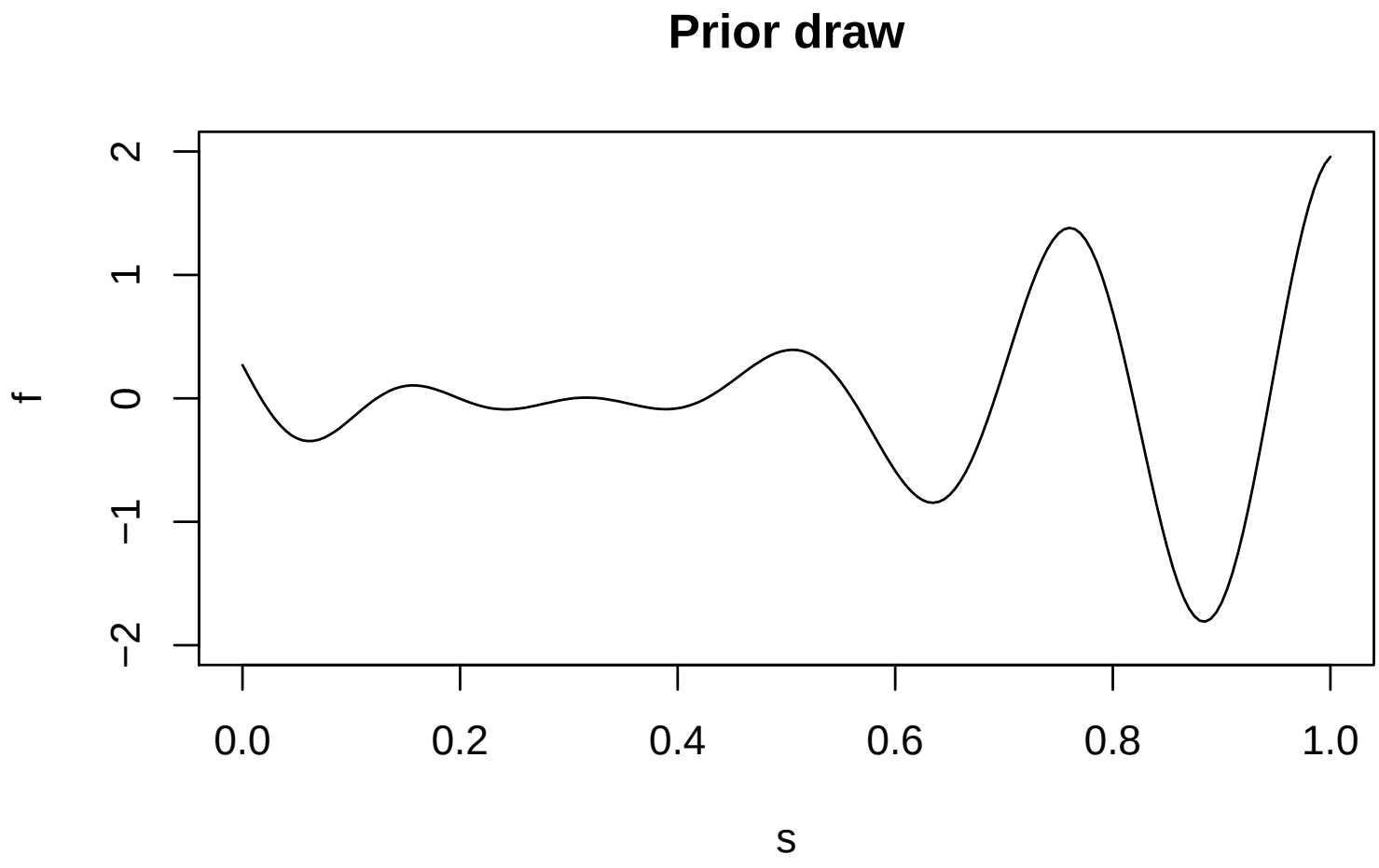
GP illustration

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \ldots, n$$



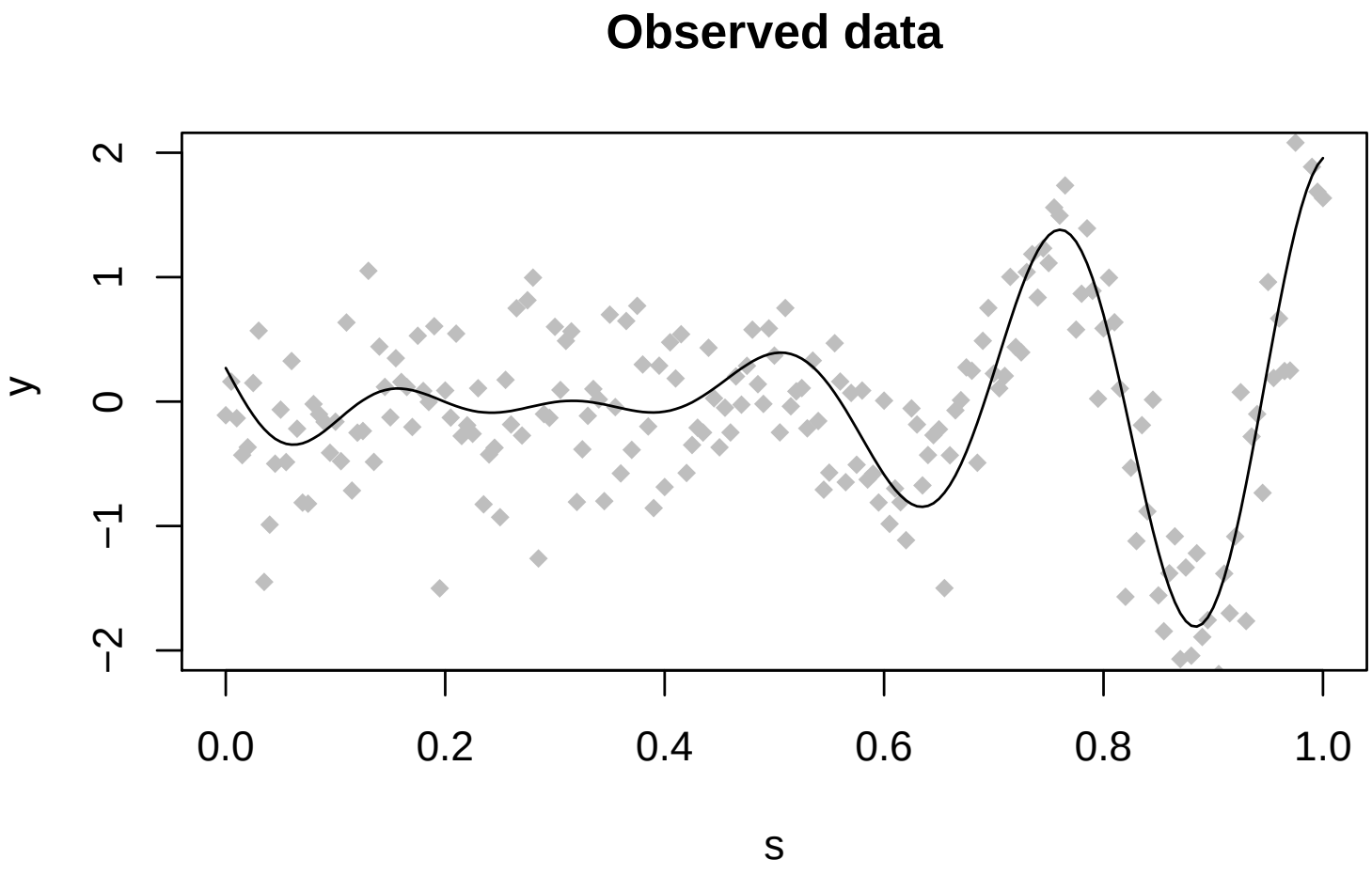
GP illustration

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \ldots, n$$



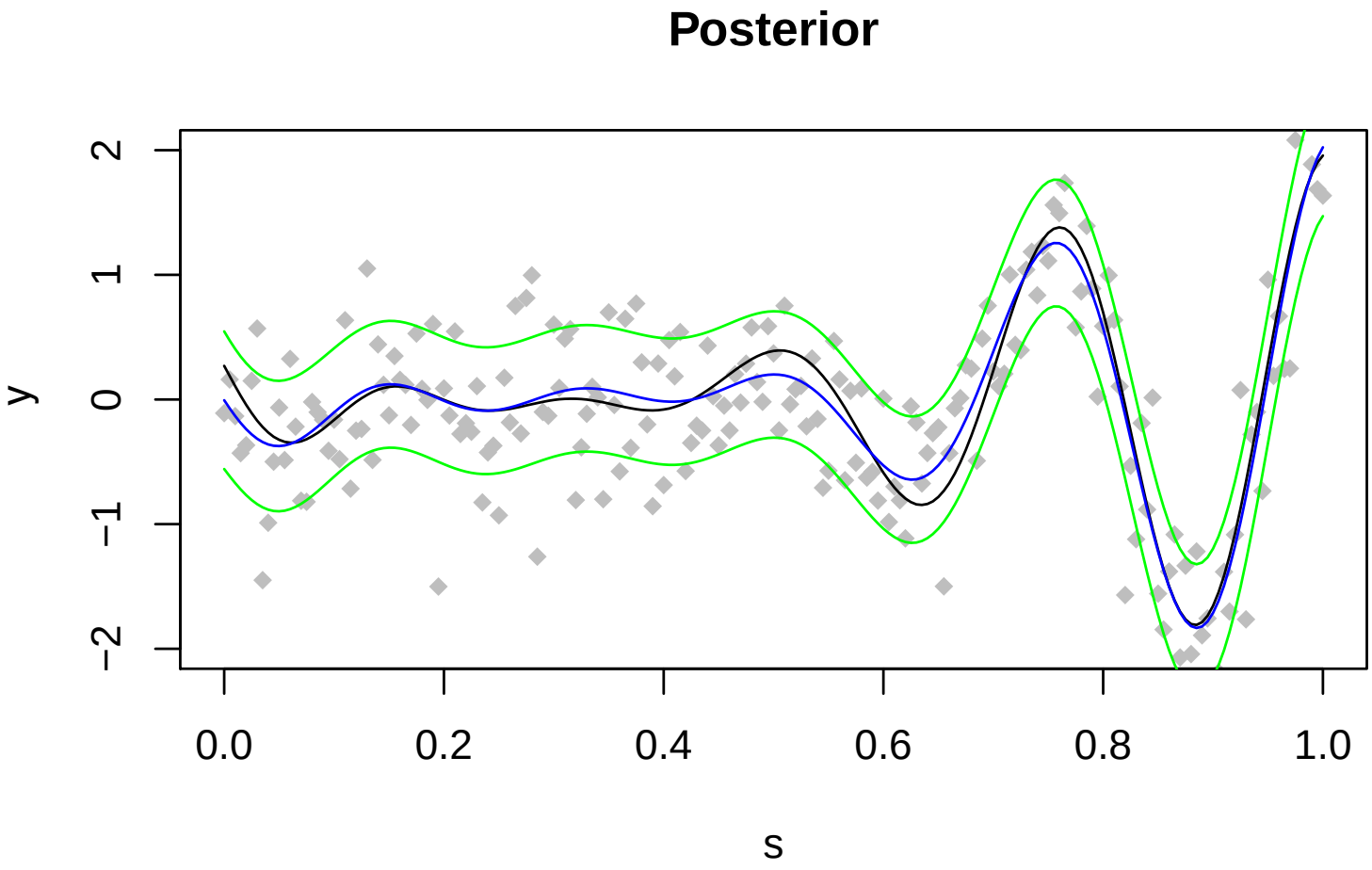
GP illustration

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \ldots, n$$



GP illustration

$$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$
$$y_i | f(\boldsymbol{x}_i) \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2), \quad i = 1, \ldots, n$$



Gaussian processes: the "weight space" view

Given K we can consider a decomposition such that $L^\top L = K$. Then

$$f \sim \mathcal{N}(0, K) \iff \gamma \sim \mathcal{N}(0, I), f := L^\top \gamma$$

This resembles the original linear regression problem, but with a new set of covariates. L is $n \times n$, so there are as many parameters as observations (thus "non-parametric"). We need to find a set of coefficients γ.

Now we return to the effect size analogue[2]:

$$\widetilde{\beta} = \text{Proj}(\boldsymbol{X}, \hat{y}) = \text{Proj}(\boldsymbol{X}, f) = \text{Prof}(\boldsymbol{X}, L^\top \gamma) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top L^\top \gamma$$

Notice:

$$f \approx \boldsymbol{X} \widetilde{\beta}$$

Just the starting point—while it can capture interaction effects, it's only assessing marginal importance.

[2] also see Crawford et al., 2017

Methodological contribution

Assume we have a posterior distribution over the function f. This induces a posterior distribution over $\widetilde{\beta}$. (Assume it's multivariate normal.)

Recall our goal: rank variables in terms of their importance in interaction with other variables.

A linear model assumes no interaction, while a GP with appropriate kernel or a fully connected neural network assumes all orders of interaction.

Measuring importance through centrality

Consider the posterior distribution $p(\widetilde{\boldsymbol{\beta}})$ and a particular variable of interest j with effect size analogue $\tilde{\beta}_j$. Denote the remaining effect size analogues $\widetilde{\boldsymbol{\beta}}_{-j}$. We are interested in the relationship betwen variable j and the rest of the variables, so we consider two distributions:

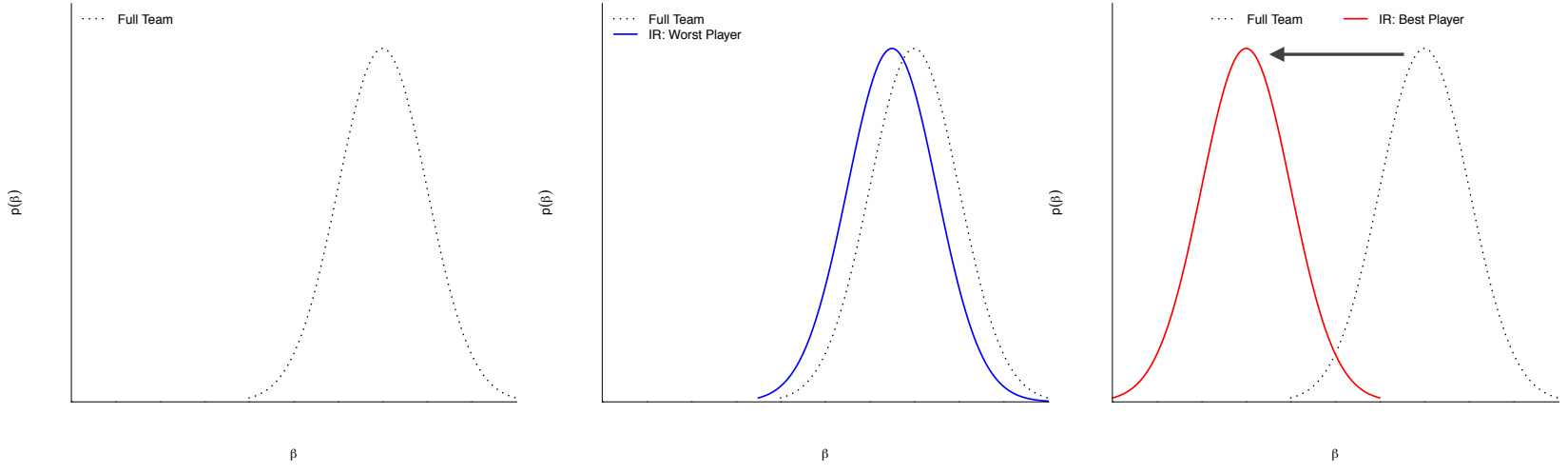
$$p(\widetilde{\boldsymbol{\beta}}_{-j}) \text{ and } p(\widetilde{\boldsymbol{\beta}}_{-j}|\tilde{\beta}_j)$$

One extreme: if the effect of the other variables are totally independent of the effect of $\tilde{\beta}_j$, then:

$$p(\widetilde{\boldsymbol{\beta}}_{-j}) = p(\widetilde{\boldsymbol{\beta}}_{-j}|\tilde{\beta}_j)$$

Another extreme: $\tilde{\beta}_j$ interacts with all other variables. Then $p(\widetilde{\boldsymbol{\beta}}_{-j})$ and $p(\widetilde{\boldsymbol{\beta}}_{-j}|\tilde{\beta}_j)$ are very far apart.

Illustration: Ranking Influential Players



[Source: Lorin Crawford]

Kullback-Leibler Divergence

We use the KLD to quantify the distance between the distributions.

$$\mathrm{KLD}(p(\boldsymbol{\beta}_{-j}) \| p(\boldsymbol{\beta}_{-j} | \beta_j)) = \int_{\widetilde{\boldsymbol{\beta}}_{-j}} \log \left(\frac{p(\widetilde{\boldsymbol{\beta}}_{-j})}{p(\widetilde{\boldsymbol{\beta}}_{-j} \mid \widetilde{\beta}_j)} \right) p(\widetilde{\boldsymbol{\beta}}_{-j}) \, \mathrm{d}\widetilde{\boldsymbol{\beta}}_{-j}.$$

Kullback-Leibler Divergence

We use the KLD to quantify the distance between the distributions.

$$\mathrm{KLD}(p(\boldsymbol{\beta}_{-j}) \| p(\boldsymbol{\beta}_{-j} | \beta_j)) = \int_{\widetilde{\boldsymbol{\beta}}_{-j}} \log \left(\frac{p(\widetilde{\boldsymbol{\beta}}_{-j})}{p(\widetilde{\boldsymbol{\beta}}_{-j} \mid \widetilde{\beta}_j)} \right) p(\widetilde{\boldsymbol{\beta}}_{-j}) \, \mathrm{d}\widetilde{\boldsymbol{\beta}}_{-j}.$$

This is a convenient choice because we will be able to derive it in closed form as the posterior distributions we consider will be multivariate normal.

Kullback-Leibler Divergence

We use the KLD to quantify the distance between the distributions.

$$\mathrm{KLD}(p(\boldsymbol{\beta}_{-j}) \| p(\boldsymbol{\beta}_{-j} | \beta_j)) = \int_{\widetilde{\boldsymbol{\beta}}_{-j}} \log \left(\frac{p(\widetilde{\boldsymbol{\beta}}_{-j})}{p(\widetilde{\boldsymbol{\beta}}_{-j} \mid \widetilde{\beta}_j)} \right) p(\widetilde{\boldsymbol{\beta}}_{-j}) \, \mathrm{d}\widetilde{\boldsymbol{\beta}}_{-j}.$$

This is a convenient choice because we will be able to derive it in closed form as the posterior distributions we consider will be multivariate normal.

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_j \\ \boldsymbol{\mu}_{-j} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_j & \boldsymbol{\sigma}_{-j}^{\mathsf{T}} \\ \boldsymbol{\sigma}_{-j} & \boldsymbol{\Sigma}_{-j} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_j & \boldsymbol{\lambda}_{-j}^{\mathsf{T}} \\ \boldsymbol{\lambda}_{-j} & \boldsymbol{\Lambda}_{-j} \end{pmatrix}$$

Kullback-Leibler Divergence

We use the KLD to quantify the distance between the distributions.

$$\text{KLD}(p(\boldsymbol{\beta}_{-j}) \| p(\boldsymbol{\beta}_{-j}|\beta_j)) = \int_{\widetilde{\boldsymbol{\beta}}_{-j}} \log\left(\frac{p(\widetilde{\boldsymbol{\beta}}_{-j})}{p(\widetilde{\boldsymbol{\beta}}_{-j} \mid \widetilde{\beta}_j)}\right) p(\widetilde{\boldsymbol{\beta}}_{-j})\, \mathrm{d}\widetilde{\boldsymbol{\beta}}_{-j}.$$

This is a convenient choice because we will be able to derive it in closed form as the posterior distributions we consider will be multivariate normal.

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_j \\ \boldsymbol{\mu}_{-j} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_j & \boldsymbol{\sigma}_{-j}^{\mathsf{T}} \\ \boldsymbol{\sigma}_{-j} & \boldsymbol{\Sigma}_{-j} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_j & \boldsymbol{\lambda}_{-j}^{\mathsf{T}} \\ \boldsymbol{\lambda}_{-j} & \boldsymbol{\Lambda}_{-j} \end{pmatrix}$$

$$\text{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + tr(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$$

RATE

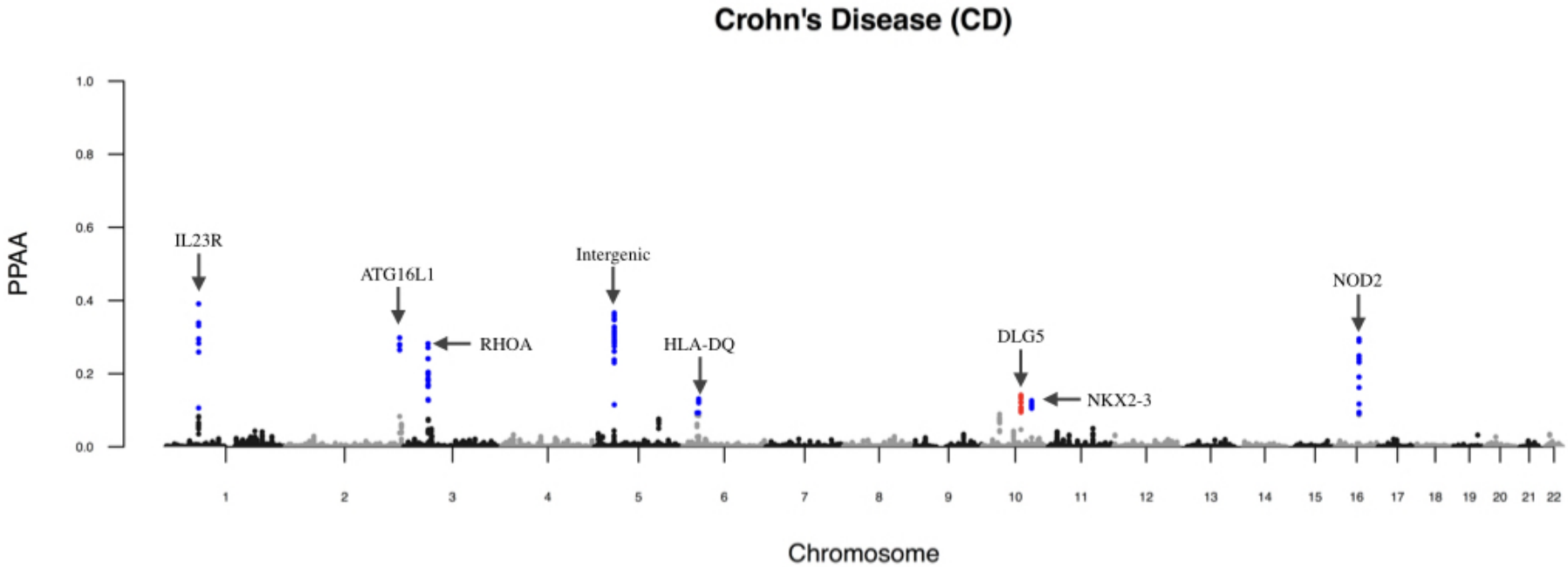
"RelATive cEntrality" (RATE):

$$\text{RATE}(\widetilde{\beta}_j) = \frac{\text{KLD}(\widetilde{\beta}_j)}{\sum_{\ell=1}^{p} \text{KLD}(\widetilde{\beta}_\ell)}$$

$$\sum_{j=1}^{p} \text{RATE}(\widetilde{\beta}_j) = 1$$

Biology application

- Goal: genetic association mapping
- Example: Genome Wide Association Study (GWAS)
- The Wellcome Trust Case Control Consortium (WTCCC)
- $n = 14{,}000$, 7 diseases, and 3,000 controls



blue: previously identified loci; red: potentially novel loci. [Source: Lorin Crawford]

Synthetic experimental setup

Data generating models are as follows:

(i) Standard model: $\mathsf{y} = \mathsf{X}\boldsymbol{\beta} + \mathsf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$,

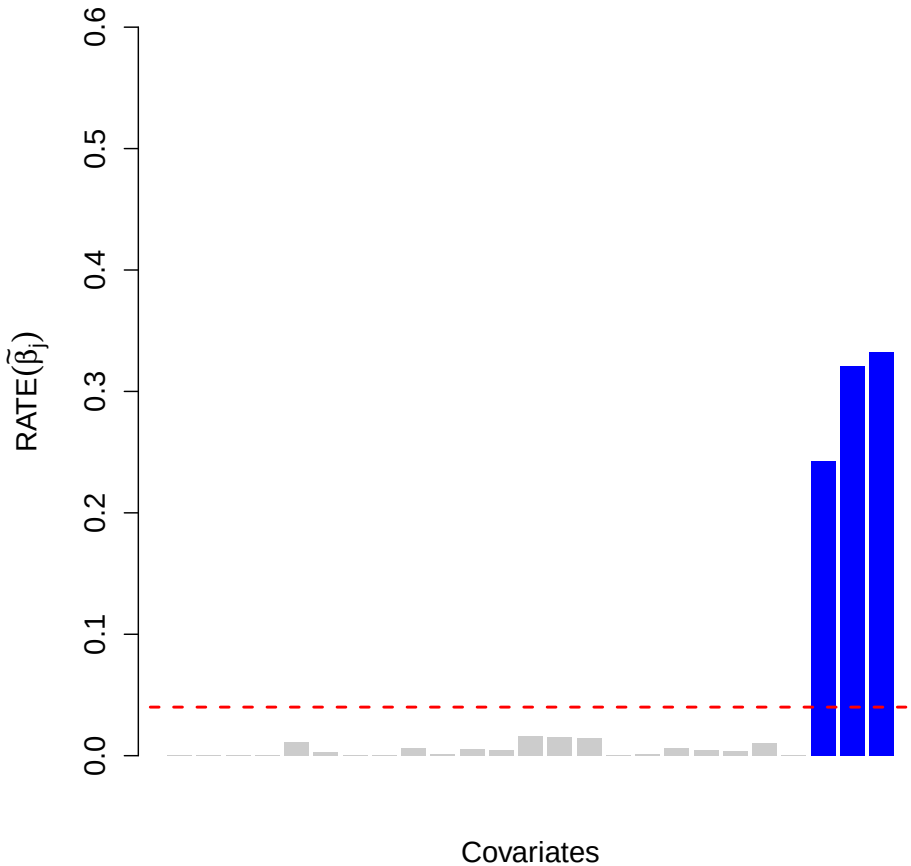
(ii) Population stratification model: $\mathsf{y} = \mathsf{X}\boldsymbol{\beta} + \mathsf{W}\boldsymbol{\gamma} + \mathsf{Z}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}$.

- p predictors total.
- Random subset j^* are truly associated ("causal") variables with $\beta_{j^*} \sim \mathcal{N}(0, 1)$.
- The rest have $\beta_j \sim \mathcal{N}(0, 0.001)$.
- Main effects are in X. Interaction effects (subset of the j^*) are in W. (Vary percent explained by each.)
- $\mathsf{Z}\boldsymbol{\varphi}$ is structured noise, to mimic population structure. Explains 10% of variance.
- We fit models using a variety of linear methods, and also GP regression + RATE.

Illustration

$n = 500$, $p = 25$ predictors. Predictors $\{23, 24, 25\}$ have additive and interaction effects.

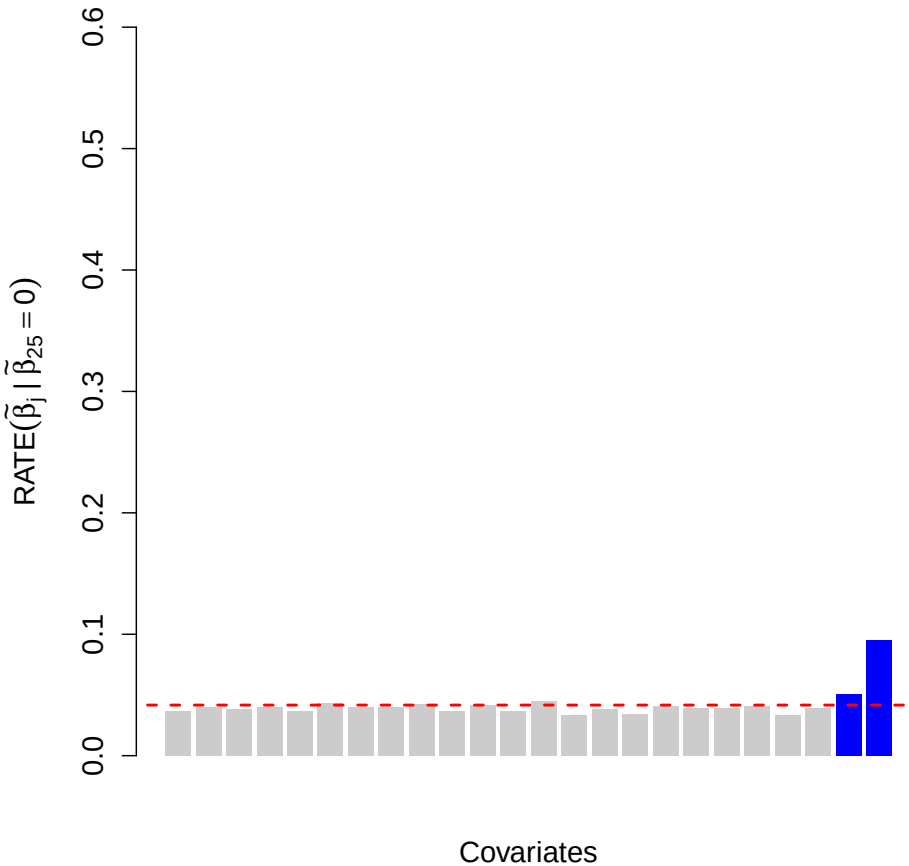
$X\beta + W\gamma$ explain 60% of variation, half from additive and half from interaction.



Illustration

$n = 500$, $p = 25$ predictors. Predictors $\{23, 24, 25\}$ have additive and interaction effects.

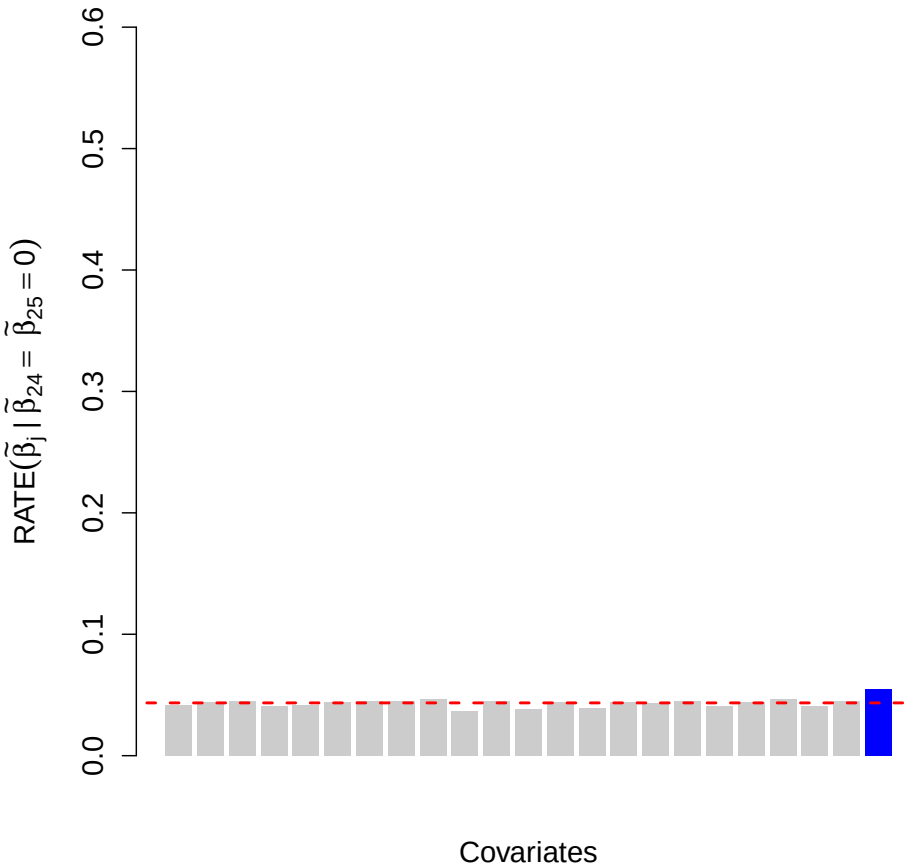
$X\beta + W\gamma$ explain 60% of variation, half from additive and half from interaction.



Illustration

$n = 500$, $p = 25$ predictors. Predictors $\{23, 24, 25\}$ have additive and interaction effects.

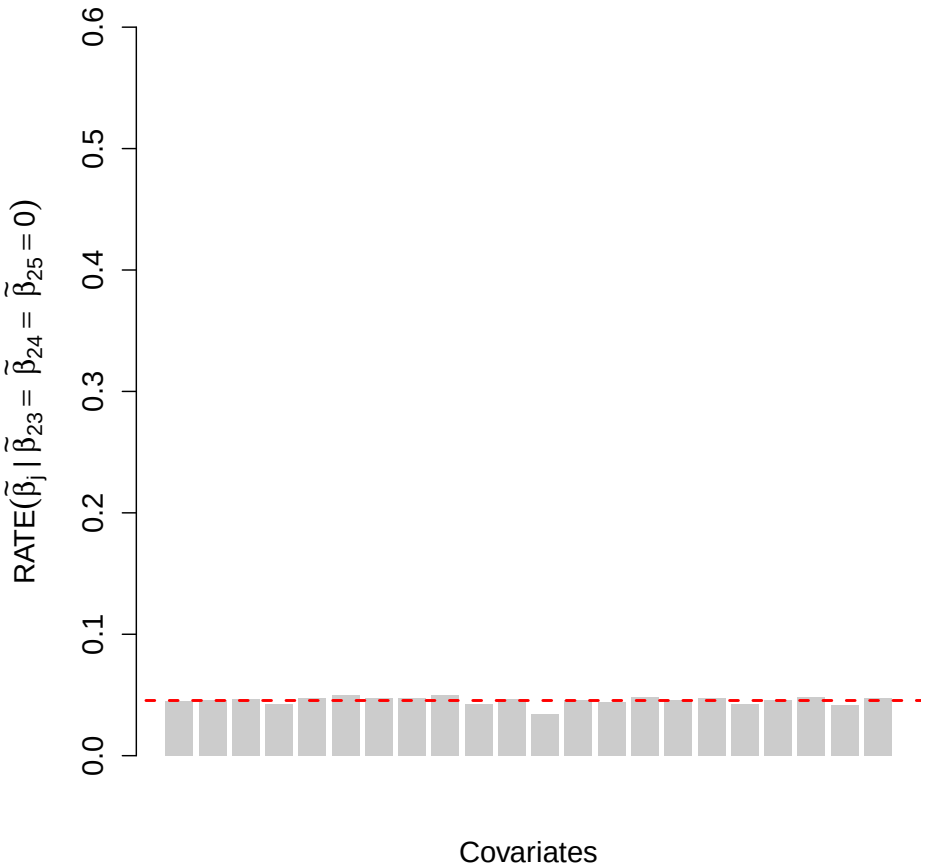
$X\beta + W\gamma$ explain 60% of variation, half from additive and half from interaction.



Illustration

$n = 500$, $p = 25$ predictors. Predictors $\{23, 24, 25\}$ have additive and interaction effects.

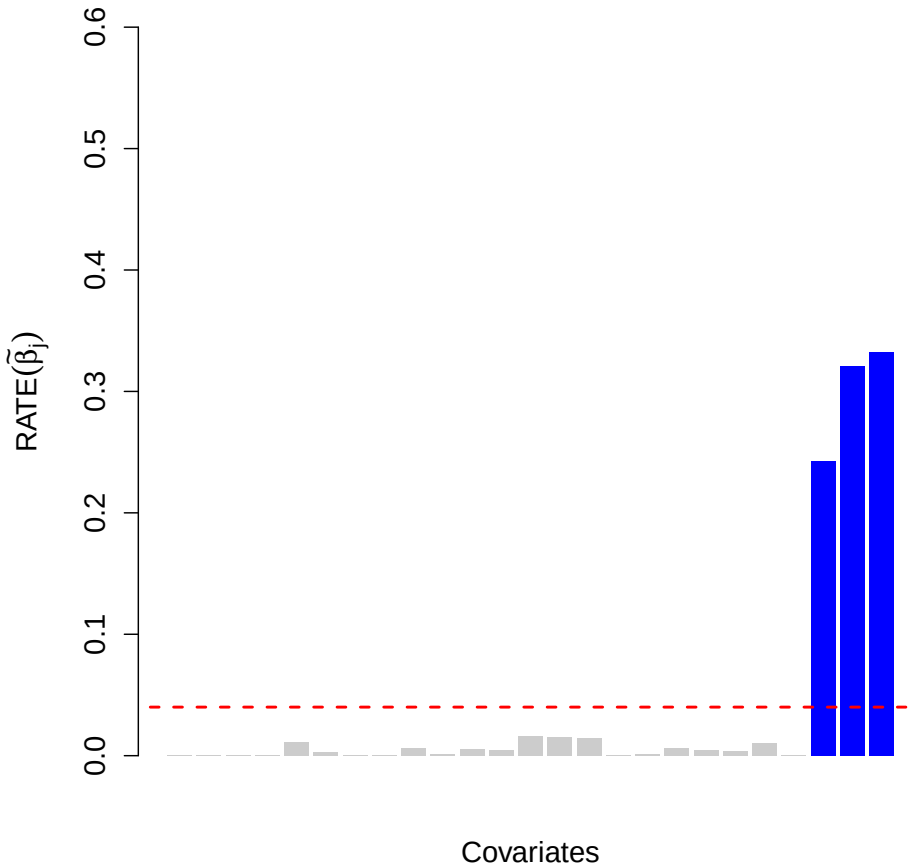
$X\beta + W\gamma$ explain 60% of variation, half from additive and half from interaction.



Illustration

$n = 500$, $p = 25$ predictors. Predictors $\{23, 24, 25\}$ have additive and interaction effects.

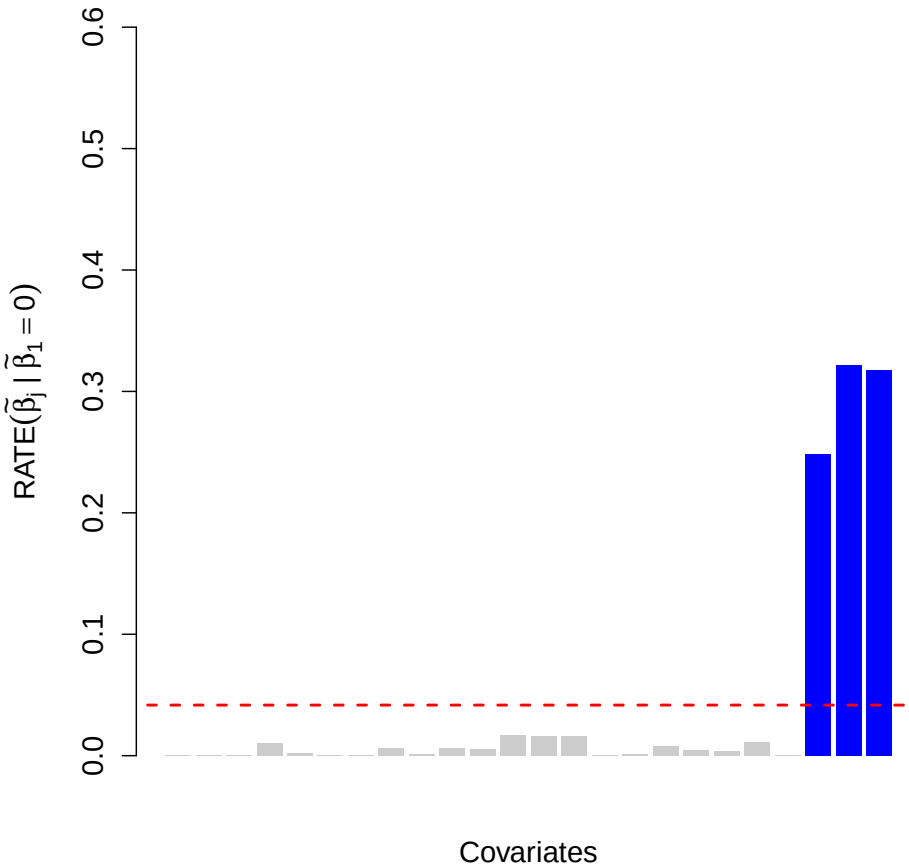
$X\beta + W\gamma$ explain 60% of variation, half from additive and half from interaction.



Illustration

$n = 500$, $p = 25$ predictors. Predictors $\{23, 24, 25\}$ have additive and interaction effects.

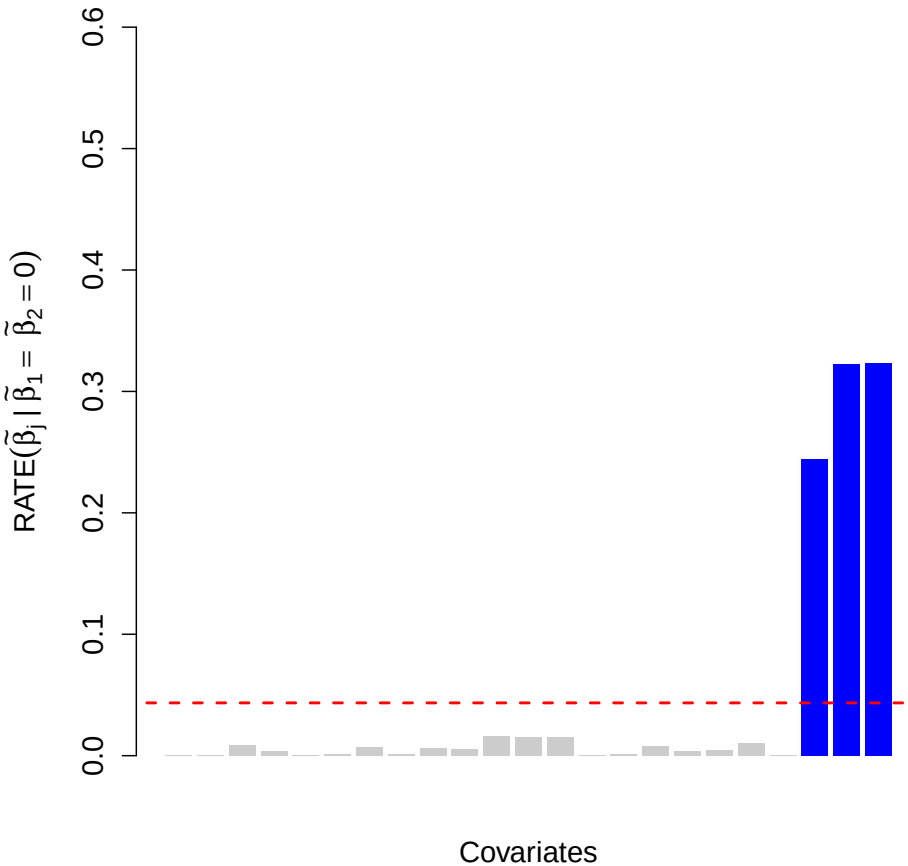
$X\beta + W\gamma$ explain 60% of variation, half from additive and half from interaction.



Illustration

$n = 500$, $p = 25$ predictors. Predictors $\{23, 24, 25\}$ have additive and interaction effects.

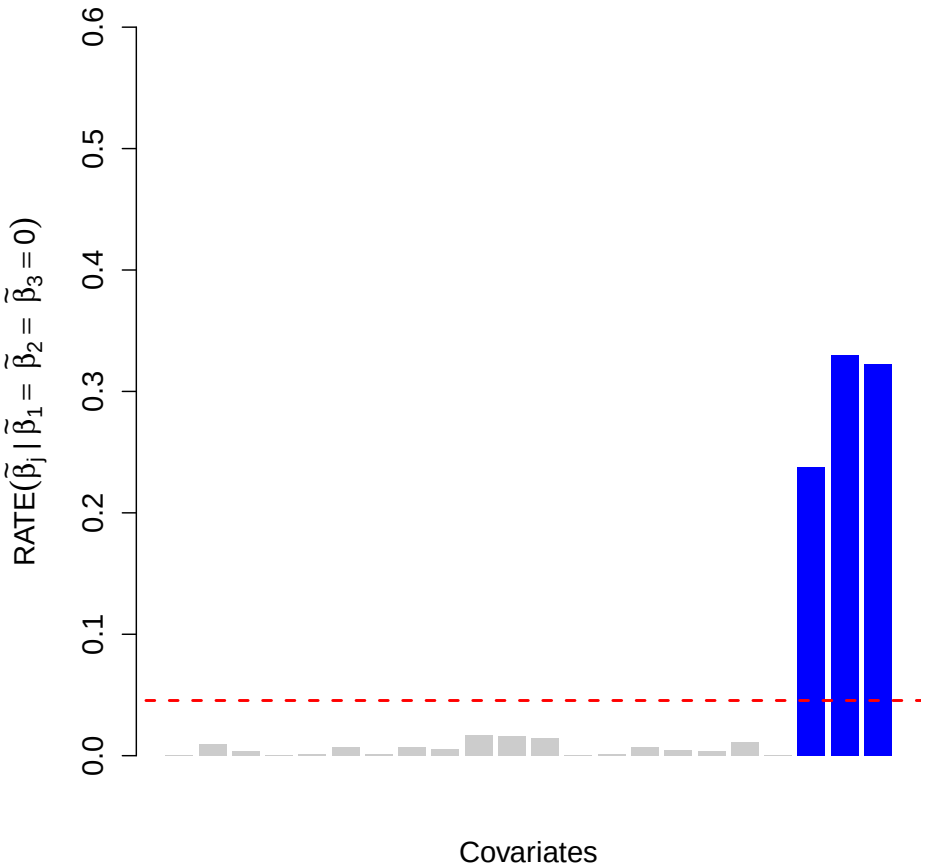
$X\beta + W\gamma$ explain 60% of variation, half from additive and half from interaction.



Illustration

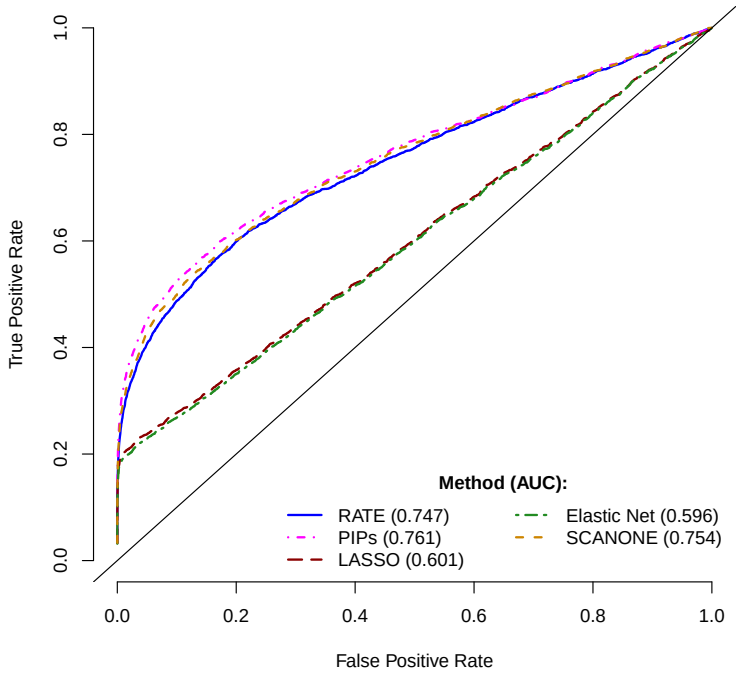
$n = 500$, $p = 25$ predictors. Predictors $\{23, 24, 25\}$ have additive and interaction effects.

$X\beta + W\gamma$ explain 60% of variation, half from additive and half from interaction.

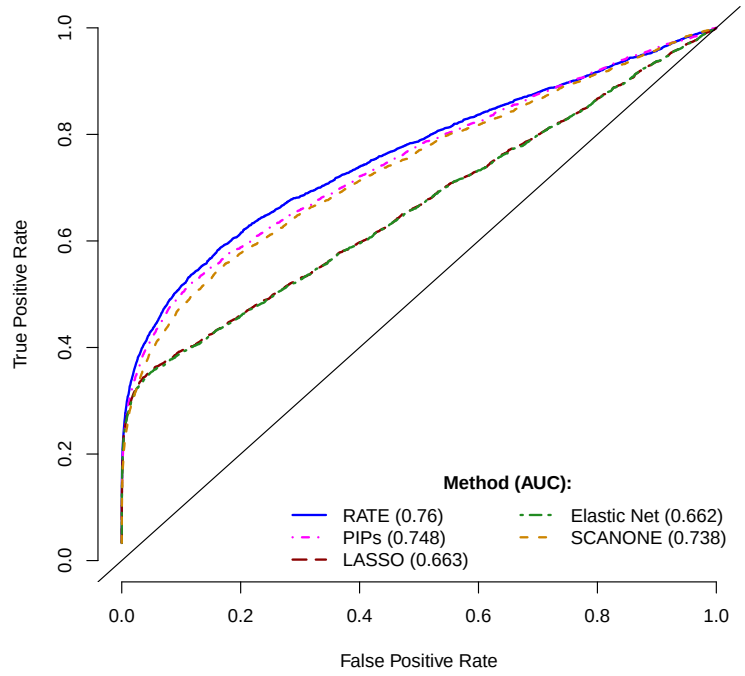


Power analysis

$n = 500$, $p = 2500$ predictors. 10 predictors have additive effects, 20 predictors have additive and interaction effects.



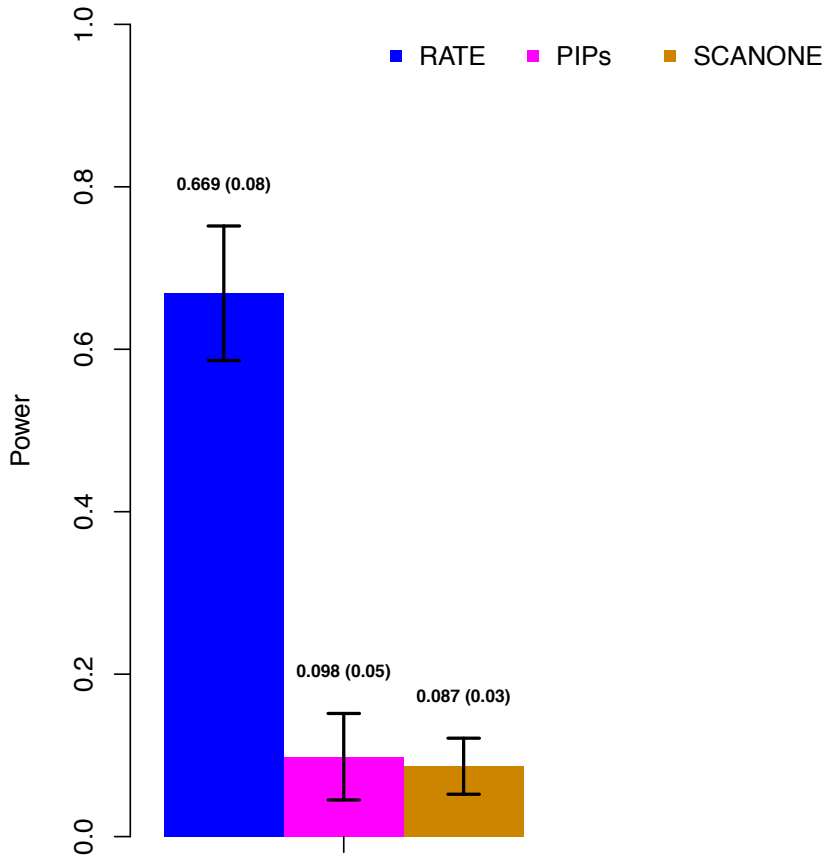
(a) Standard Model



(b) Population Stratification Model

Power analysis

Compare power for RATE $> 1/p$,
posterior inclusion probability > 0.5,
multiple testing corrected SCANONE method $P < 2 \times 10^{-5}$.



Real data

- Phenotypes of *Arabidopsis thaliana*
- Versailles Arabidopsis Stock Center `publiclines.versailles.inra.fr/page/33`
- Used in previous studies for similar methods [Demetrashvili et al 2013]
- $n = 403, p = 1028$.
- Many genotypes of perfect correlation $r^2 > 0.99$ so final dataset $p = 524$ covariates.
- Phenotypes are six biochemical content measurements: allyl, Indol-3-ylmethyl (I3M), 4-methoxy-indol-3-ylmethyl (MO4I3M), 4-methylsulfinylbutyl (MSO4), 8-methylthiooctyl (MT8), and 3-hydroxypropyl (OHP3)

Real data

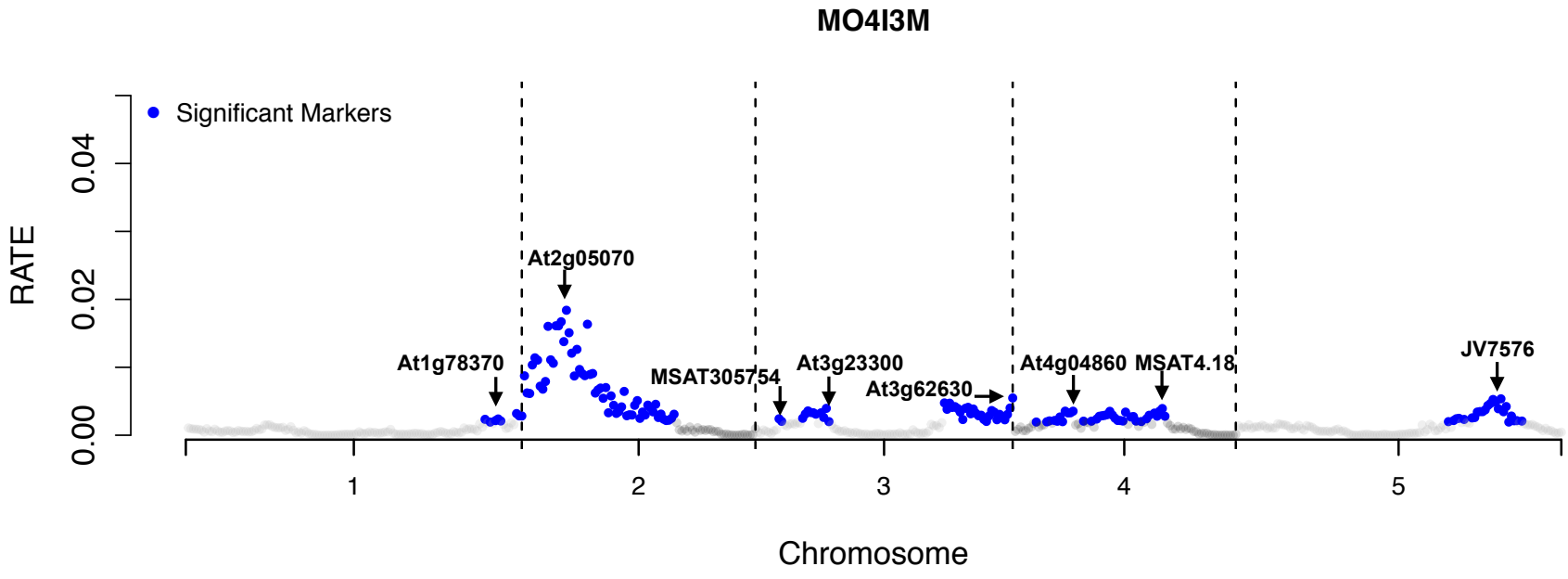


Figure: RATE

Real data

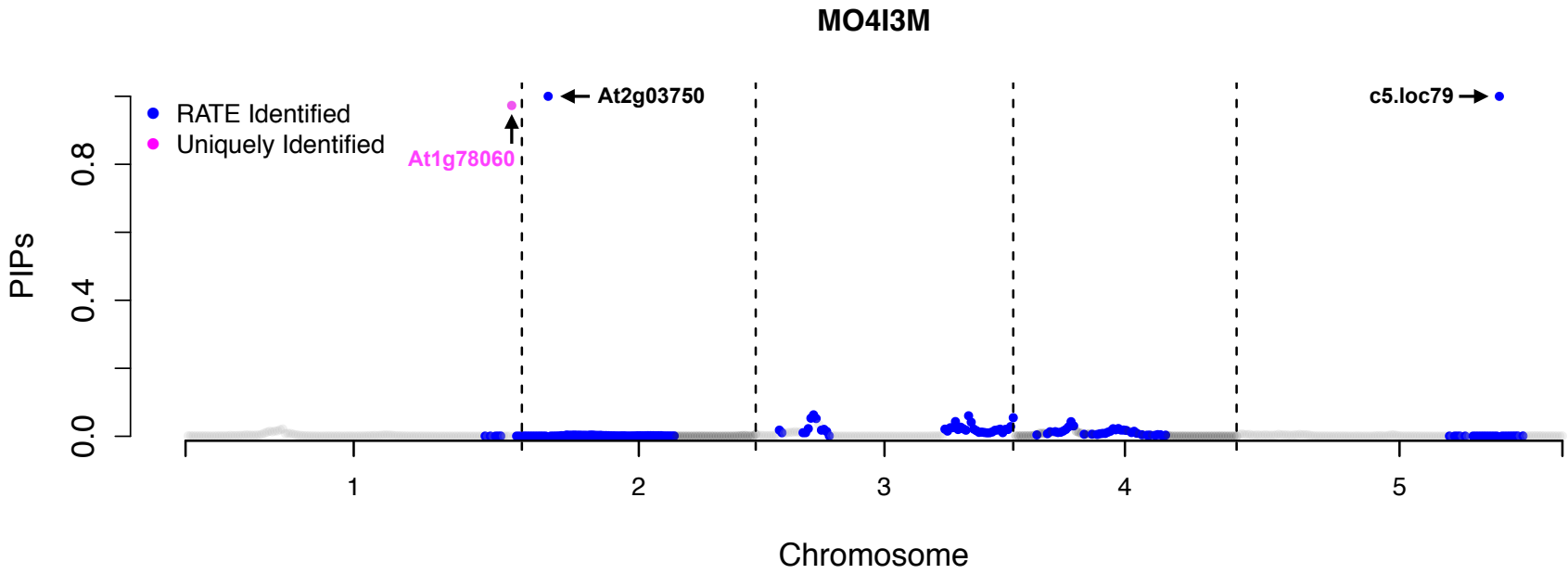


Figure: Bayesian spike and slab prior

Real data

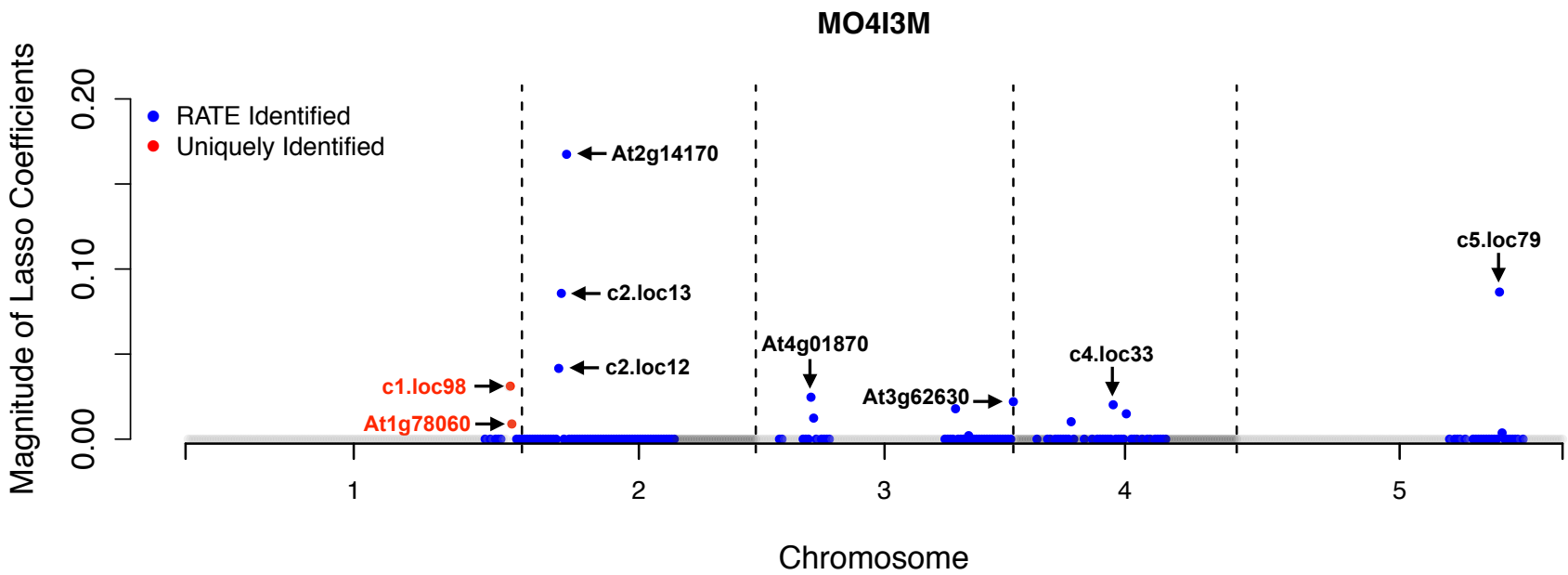


Figure: Lasso

Real data

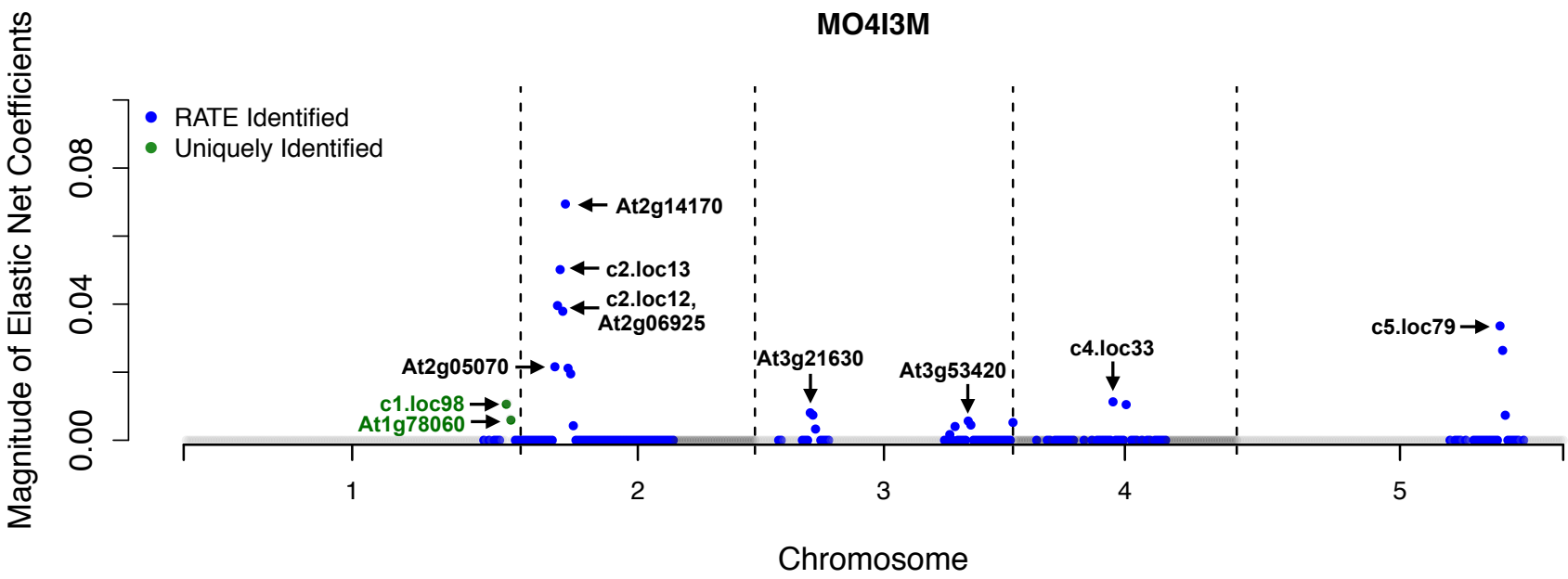


Figure: Elastic net

Real data

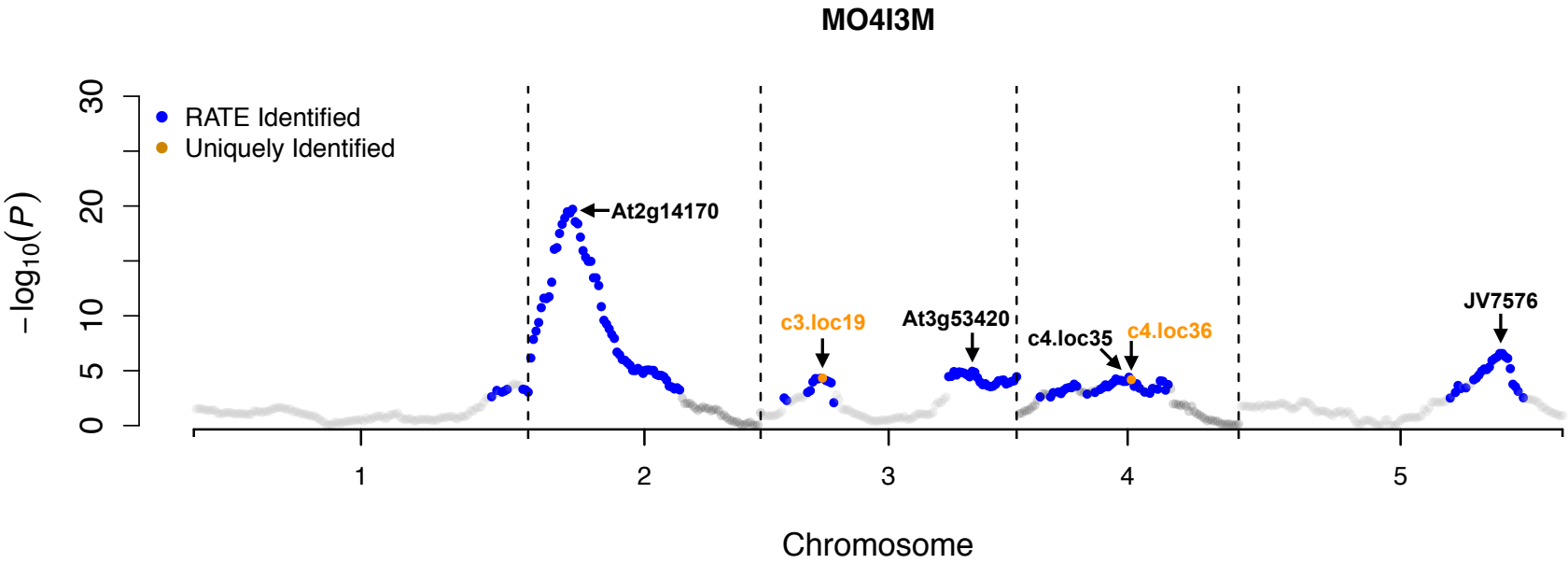


Figure: SCANONE

Takeaways from real study

- RATE consistently identifies genomic locations corresponding to known members of biosynthetic pathways in *Arabidopsis thaliana* and validated findings from previous experimentally based studies
- Same general regions also identified by other methods, but real differences amongst them
- Sparsity inducing methods not always appropriate approaches for mapping studies because we know that the true causal variants are likely not any of the observed marker—even with 10 million SNPs, many un-tagged variants, and groups of nearby variants that are almost perfectly correlated. Real goal is not to find the best single marker but to identify a region (or subnetwork) of the genome that contains the true variant.

Conclusion

- Draft of Crawford, Flaxman, Runcie, and West [2018] on arXiv:1801.07318.
- Replication code is on GitHub `https://github.com/lorinanthony/RATE`
- Next steps: scalability, deep neural networks, application to frequentist settings, connections with likelihood ratio test and Bayes Factor
- Explainability in Machine Learning Challenge has launched (partners: Imperial, FICO, Google, Berkeley, MIT) `http://explainable.ml/`

References

- L. Crawford, K.C. Wood, X. Zhou, S. Mukherjee (2017) "Bayesian approximate kernel regression with variable selection." *JASA*.
- B. Goodman and S. Flaxman (2017) "European Union Regulations on Algorithmic Decision Making and a 'Right to Explanation' ", *AI Magazine*. [arXiv:1606.08813]
- L. Crawford, S. Flaxman, D. Runcie, M. West (2018) "Predictor Variable Prioritization in Nonlinear Models: A Genetic Association Case Study." Draft on arXiv:1801.07318.