PhD oral examination

Spencer Woody

Department of Statistics and Data Science
The University of Texas at Austin

November 13, 2018

Acknowledgements

Committee members

- Professor James Scott[1 2] (advisor)
- Professor Carlos Carvalho[1 2]
- Professor Jared Murray[2]
- Professor Cory Zigler[1 3]

[1]Department of Statistics and Data Science
[2]Department of Information, Risk, and Operations Management
[3]Department of Women's Health, Dell Medical School

Outline

I. Bayes-optimal post-selection inference*
   (a) Sparse means
   (b) Spatial hotspot detection

II. Future directions: posterior summarization†

*Joint work with Prof. Scott

†Joint work with Profs. Carvalho and Murray

Bayesian models for sparsity

A large body of Bayesian work has focused on finding sparse signals, canonically in the Gaussian means model:

$$(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2)$$

Bayesian models for sparsity

A large body of Bayesian work has focused on finding sparse signals, canonically in the Gaussian means model:

$$(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2)$$

Two common Bayesian approaches:

- Two-groups model: $\theta_i \sim p \cdot \pi(\theta) + (1 - p) \cdot \delta_0$
- Continuous shrinkage priors: horseshoe, Laplace, etc.

Goal: *Quantify uncertainty for the "interesting" θ_i once we've found them, while adjusting for selection.*

Do Bayesians even need to worry about this?

Scenario 1: Imagine a genomics lab cherry-picking results from a study:

1. Draw $\theta_i \sim \pi(\theta)$ for gene $i = 1, \ldots, N$.
2. Observe data $y_i \sim \mathcal{N}(\theta_i, \sigma^2)$.
3. Select the (y_i, θ_i) pairs where $y_i \in S$ (e.g. $|y| > 2$).

Do Bayesians even need to worry about this?

Scenario 1: Imagine a genomics lab cherry-picking results from a study:

1. Draw $\theta_i \sim \pi(\theta)$ for gene $i = 1, \ldots, N$.
2. Observe data $y_i \sim \mathcal{N}(\theta_i, \sigma^2)$.
3. Select the (y_i, θ_i) pairs where $y_i \in S$ (e.g. $|y| > 2$).

The joint distribution of (θ, y) under selection is

$$p_S(\theta, y) = \frac{\pi(\theta) \cdot p(y \mid \theta)}{\Pr(y \in S)} \cdot \mathbf{1}(y \in S)\,.$$

So for any y that falls in S,

$$p_S(\theta \mid y) \propto \pi(\theta) \cdot p(y \mid \theta)\,.$$

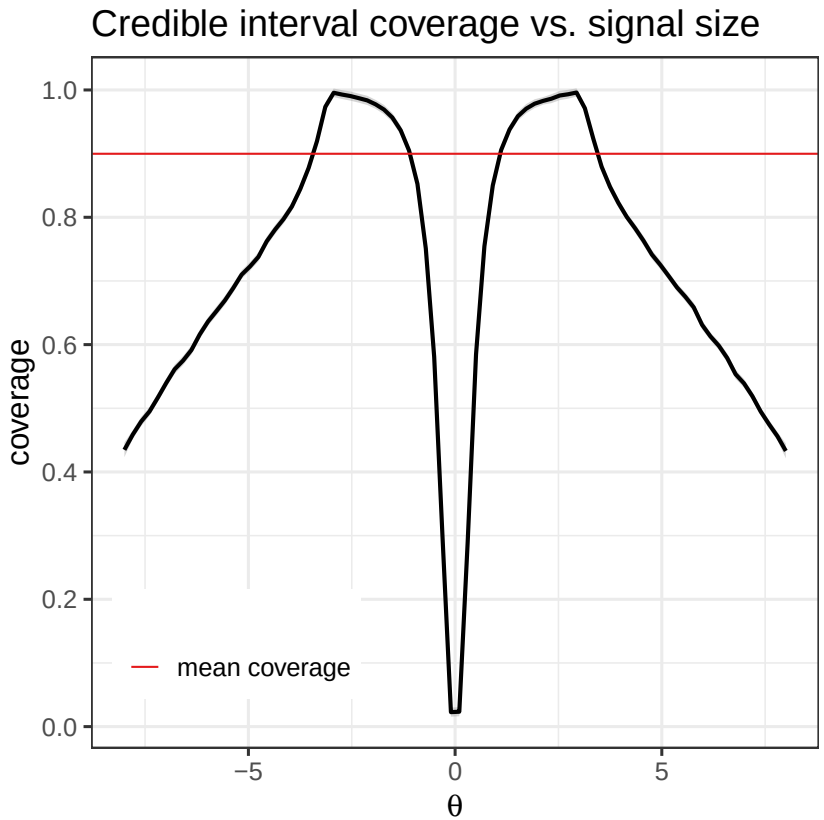
The posterior is unaffected by this form of selection. (Yekutieli, 2012)

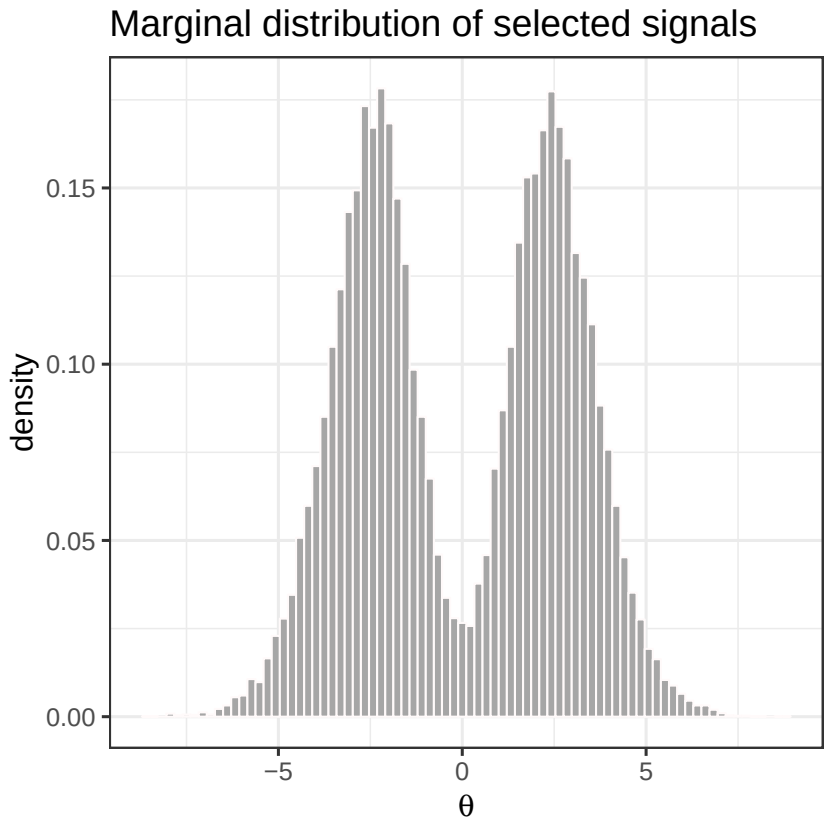
Do Bayesians even need to worry about this?

If you care about coverage, there is still a problem

Suppose $\theta_i \sim \mathcal{N}(0, 2^2),\ y_i \sim \mathcal{N}(\theta_i, 1).$

Select only θ_i where $|y_i| > 2$





Not all selection mechanisms are the same.

Scenario 2: Imagine a scientific field with many open questions, where journals publish results only if $y \in S$. For each question $(k = 1, 2, \ldots)$:

1. Draw $\theta^{(k)} \sim \pi(\theta)$.
2. Many labs $(i = 1, 2, \ldots)$ observe $y_i^{(k)} \sim \mathcal{N}(\theta^{(k)}, \sigma^2)$.
3. The first lab that observes $y_i^{(k)} \in S$ publishes its results.

Not all selection mechanisms are the same.

Scenario 2: Imagine a scientific field with many open questions, where journals publish results only if $y \in S$. For each question $(k = 1, 2, \ldots)$:

1. Draw $\theta^{(k)} \sim \pi(\theta)$.
2. Many labs $(i = 1, 2, \ldots)$ observe $y_i^{(k)} \sim \mathcal{N}(\theta^{(k)}, \sigma^2)$.
3. The first lab that observes $y_i^{(k)} \in S$ publishes its results.

Now the joint distribution of (θ, y) under selection is

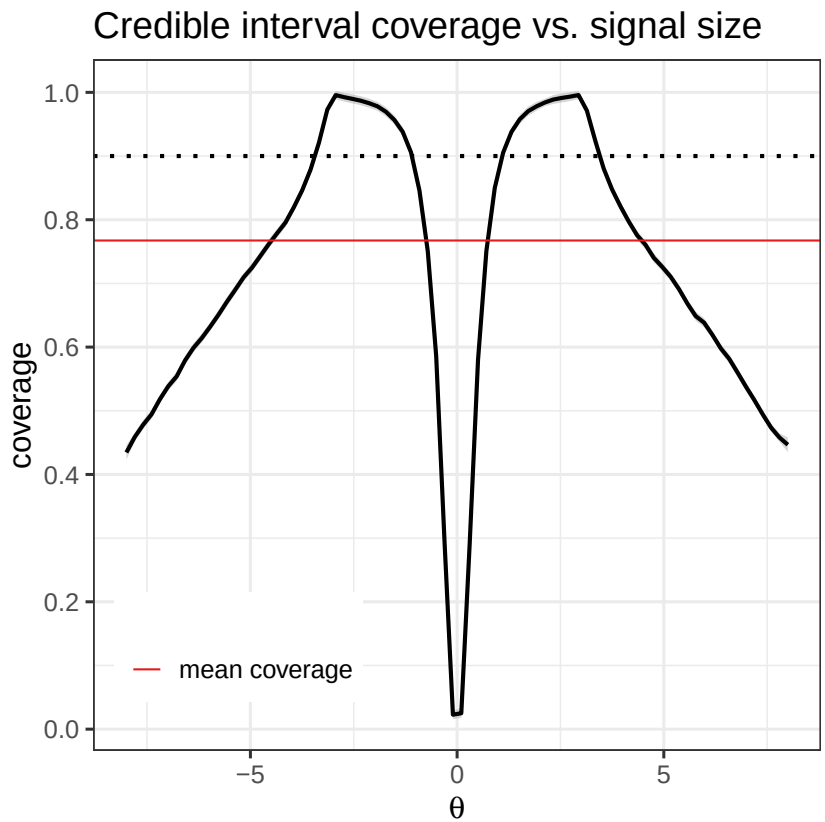
$$p_S(\theta, y) = \frac{\pi(\theta) \cdot p(y \mid \theta)}{\Pr(y \in S \mid \theta)} \cdot \mathbf{1}(y \in S)\,.$$

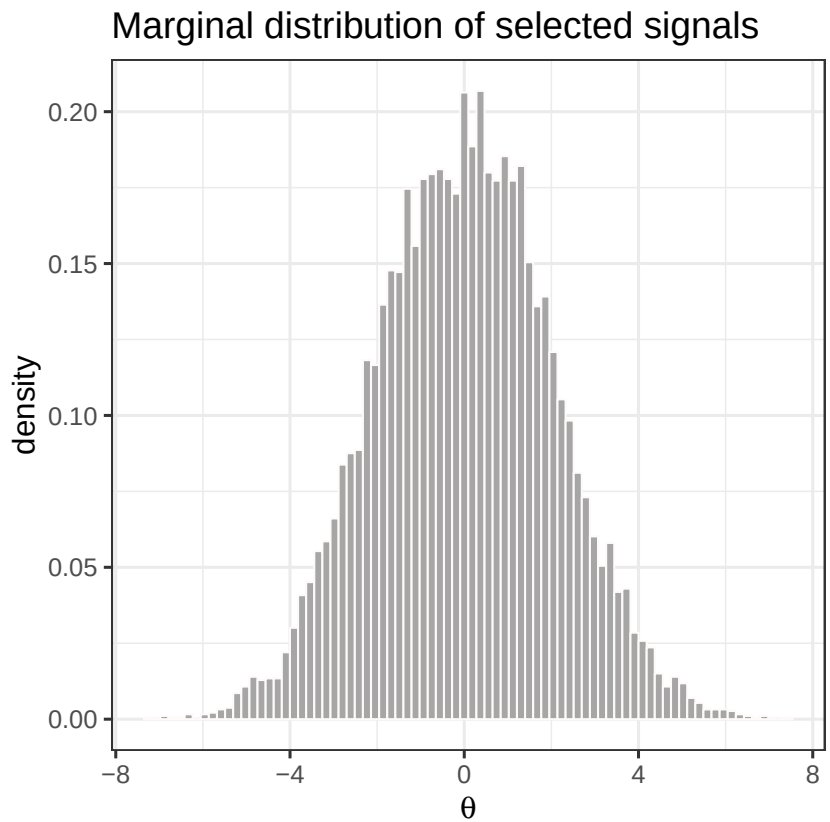
The posterior *is* affected by this form of selection (Yekutieli, 2012).

Not all selection mechanisms are the same

Unadjusted posterior credible intervals do not maintain nominal coverage on average

Suppose $\theta_i \sim \mathcal{N}(0, 2^2),\ y_i \sim \mathcal{N}(\theta_i, 1) \cdot \mathbf{1}(|y_i| > 2)$





Frequentist approaches to POSI

Most work from Jonathan Taylor and collaborators. For example:

- Fithian, Sun, and Taylor (2014): a data-splitting-like approach
- Reid, Taylor, and Tibshirani (2014): normal means
- Lee, Sun, Sun, and Taylor (2016): regression with the lasso
- Benjamini, Taylor, and Irizarry (2016): spatial hotspots

But:

- This line of research is still at the “pre-James–Stein” stage.
- The resulting intervals are often needlessly wide, and there is a lot of room for efficiency improvement.

Our contribution

There is an unmet need for inferential approaches that:

- correctly adjust for selection, of whatever form.
- incorporate the efficiency benefits of “information borrowing” from a prior distribution.
- maintain exact frequentist coverage, uniformly across the parameter space, *even if the prior is wrong*.

Our contribution

There is an unmet need for inferential approaches that:

- correctly adjust for selection, of whatever form.
- incorporate the efficiency benefits of "information borrowing" from a prior distribution.
- maintain exact frequentist coverage, uniformly across the parameter space, *even if the prior is wrong*.

Our approach does all three.

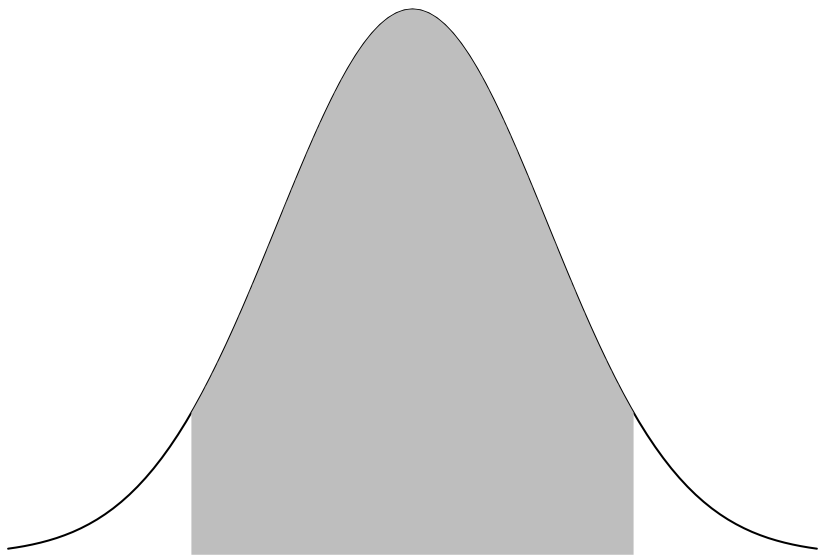
We will show:

- how the basic frequentist approach works.
- where the prior comes in (following Yu and Hoff, 2018).
- how our method performs.

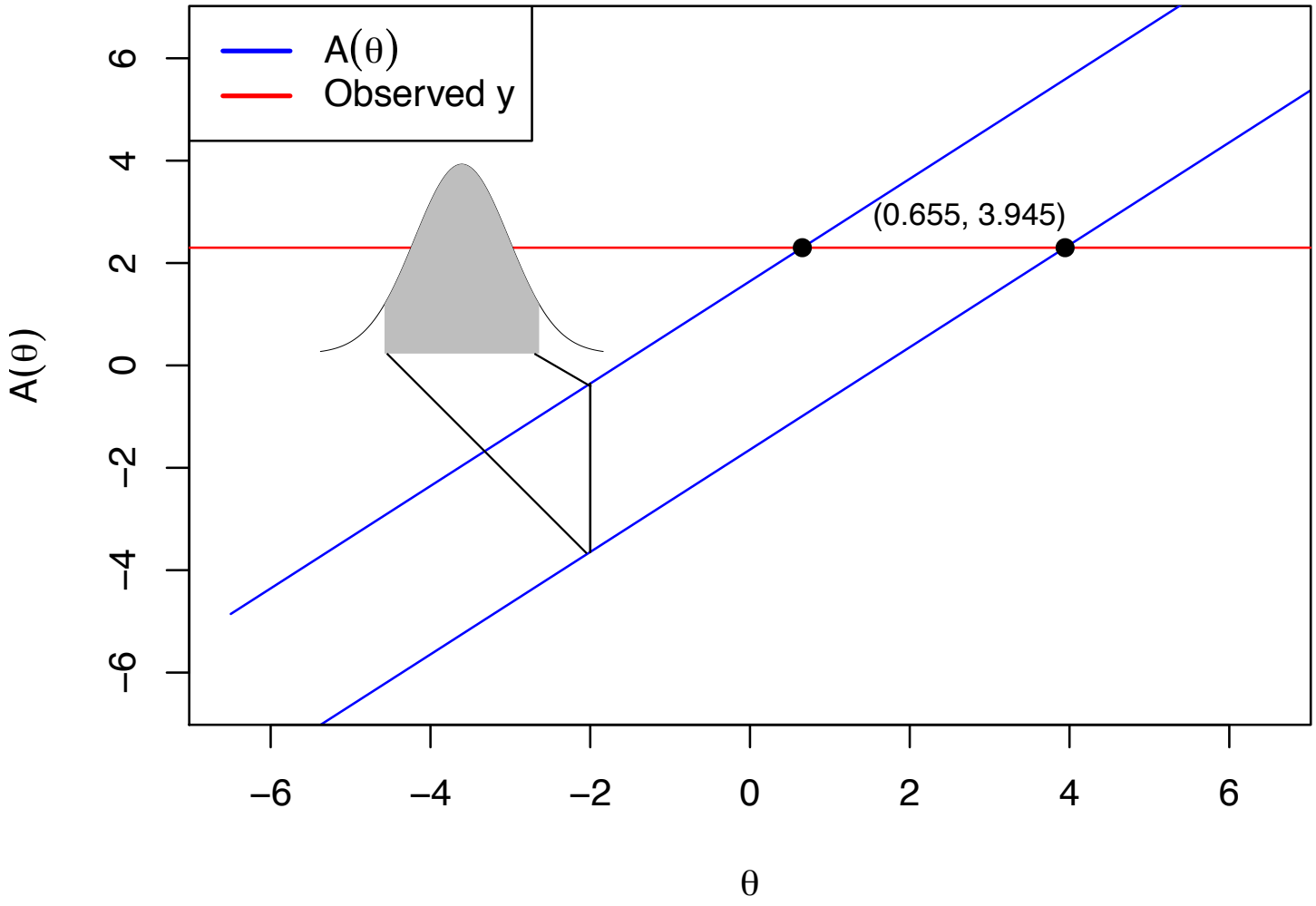
Post-selection inference by inverting a test

Recall that we can build a confidence set by inverting a test:

- Let $F(y \mid \theta)$ be the CDF of the sampling distribution for $(y \mid \theta)$.
- Now construct a size-α test of $H_0 : \theta = \theta_0$, with acceptance region $A(\theta_0) = (F^{-1}(\alpha/2 \mid \theta_0),\ F^{-1}(1 - \alpha/2 \mid \theta_0))$



Post-selection inference by inverting a test



Observe y, report $C(y) = \{\theta : y \in A(\theta)\}$, yielding the universally most accurate unbiased (UMAU) confidence set

Post-selection inference by inverting a test

We can do the same under the selection-adjusted model, using the *selection-adjusted likelihood*

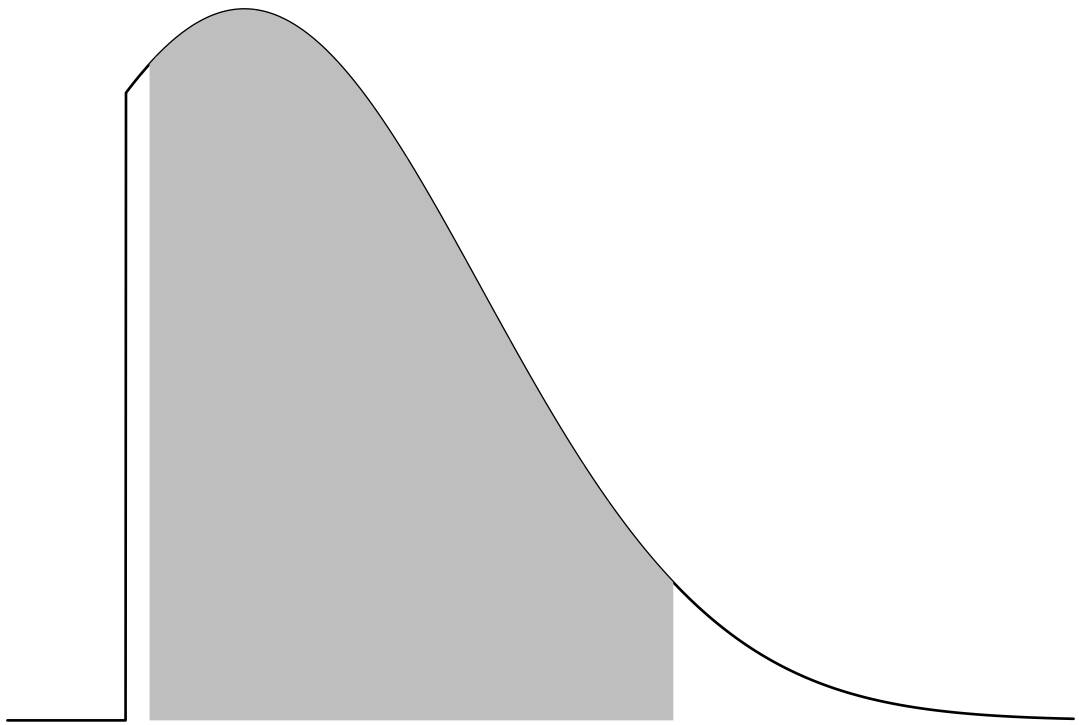
$$f_S(y \mid \theta) = \frac{f(y \mid \theta)}{\Pr(y \in S \mid \theta)} \cdot \mathbf{1}(y \in S)$$

Let $F_S(y \mid \theta)$ be the CDF for this sa-likelihood.

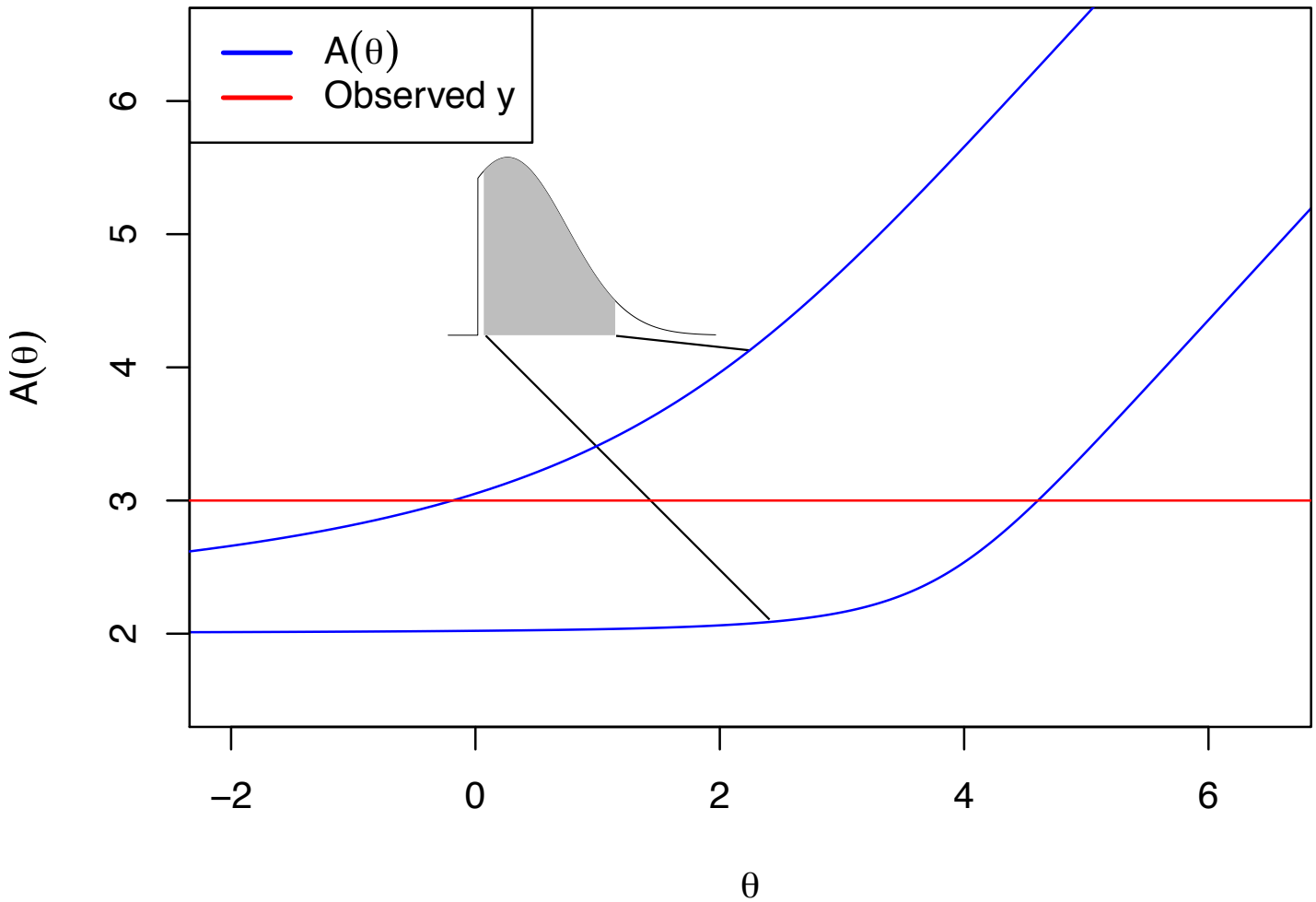
Post-selection inference by inverting a test

Below:

- Data is $(y \mid \theta) \sim \mathcal{N}_S(\theta, 1)$,
- The selection region is $S = (2, \infty)$.
- $A(\theta_0) = (F_S^{-1}(\alpha/2 \mid \theta_0), F_S^{-1}(1 - \alpha/2 \mid \theta_0))$.



Post-selection inference by inverting a test



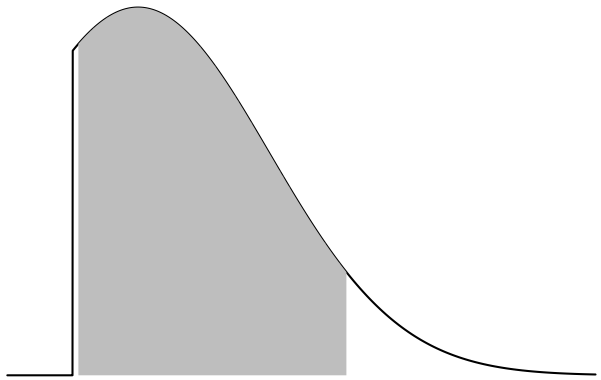
Observe $y \in S$, report $C(y) = \{\theta : y \in A(\theta)\}$. (Fithian et al., 2014)

Post-selection inference by inverting a test

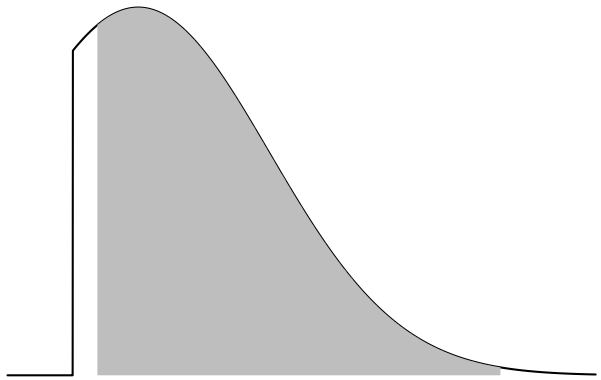
But $A(\theta)$ need not use equal tail areas. Define

$$A_w(\theta) = (l_w(\theta), u_w(\theta))$$
$$l_w(\theta) = F_S^{-1}(w\alpha \mid \theta)$$
$$u_w(\theta) = F_S^{-1}(w\alpha + 1 - \alpha \mid \theta)$$

Left tail = 0.2α, right tail = 0.8α



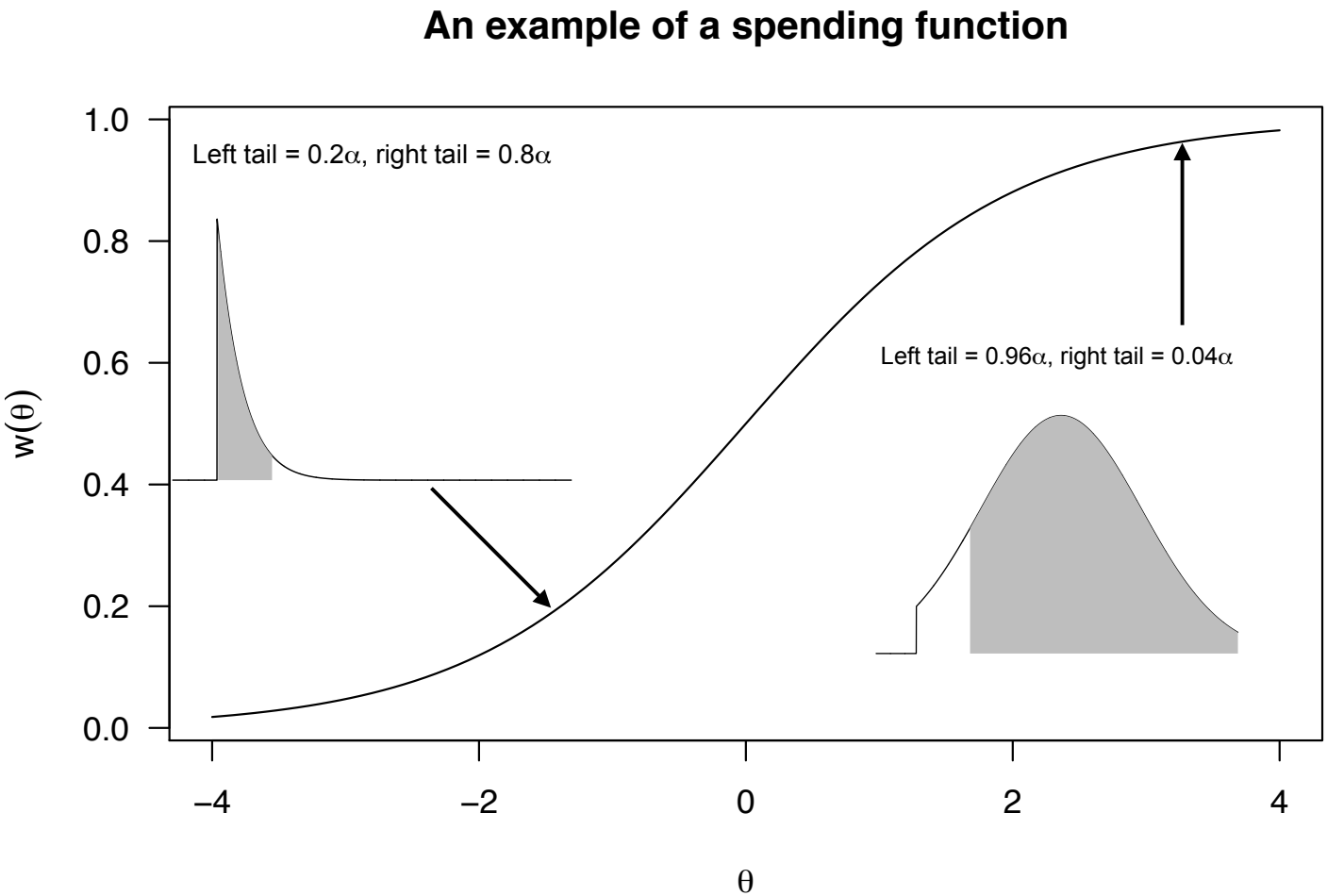
Left tail = 0.96α, right tail = 0.04α



$C_w(y) := \{\theta : y \in A_w(\theta)\}$ will retain $1 - \alpha$ coverage

Post-selection inference by inverting a test

We can even spend α in a different way for every θ.
This is controlled by a spending function $w(\theta)$:



Bayes-optimal post-selection inference

Any $w(\theta)$ ensures exact coverage, but which is the best?

Key idea (Pratt, 1963; Yu and Hoff, 2018): choose $w(\theta)$ to minimize expected size of the confidence sets under a prior $\pi(\theta)$.

Define the (frequentist) risk of a confidence set as its expected size, for fixed θ and random $(y \mid \theta)$:

$$R(\theta; w) = \int \int \mathbf{1}(y \in A_w(\tilde{\theta})) f(y|\theta) \mathrm{d}\tilde{\theta} \mathrm{d}y \,,$$

recalling that $A_w(\theta)$ is determined by $w(\theta)$.

Bayes-optimal post-selection inference

Now suppose that $\theta \sim \pi(\theta)$. Our decision variable is the spending function $w(\theta)$, and the Bayes loss is a functional of w:

$$L(\pi, w(\theta)) = \int R(\theta; w(\theta))\ \pi(\theta) \mathrm{d}\theta$$

FAB principle ("frequentist assisted by Bayes," per YH18):
For a given prior, choose $w(\theta)$ to minimize the Bayes loss.

This is a variational optimization problem that can be solved:

- analytically in simple settings.
- rapidly by brute force (i.e. pointwise in θ) everywhere else.

Requires sa-likelihood $f_S(y \mid \theta)$ and sa-marginal $m_S(y)$

Bayes-optimal post-selection inference

Yu and Hoff (2018) compute the optimal spending function under the one-way ANOVA model:

$$(y_{ij} \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2), \quad \theta_i \sim \mathcal{N}(\mu, \gamma^2)\,.$$

Our setting involves two very different assumptions.

- Sparsity: most θ_i are zero or very small.
- Post-selection inference: the same data is used to identify the interesting signals and to compute their confidence sets.

saFAB (selection-adjusted FAB) procedure

(1) Derive $f_S(y \mid \theta)$ and $m_S(y)$ using $\pi(\theta)$ and selection rule S
(2) Construct spending function $w(\theta)$ using $f_S(y \mid \theta)$ & $m_S(y)$
(3) Construct the family of biased tests from $w(\theta)$
(4) Invert this family of biased tests for observed y

saFAB (selection-adjusted FAB) procedure

(1) Derive $f_S(y \mid \theta)$ and $m_S(y)$ using $\pi(\theta)$ and selection rule S
(2) Construct spending function $w(\theta)$ using $f_S(y \mid \theta)$ & $m_S(y)$
(3) Construct the family of biased tests from $w(\theta)$
(4) Invert this family of biased tests for observed y

NB: Using $w_{\text{UMAU}}(\theta) \equiv 1/2$ returns the UMAU confidence sets (existing frequentist approach) (Fithian et al., 2014; Reid et al., 2014)

A toy example

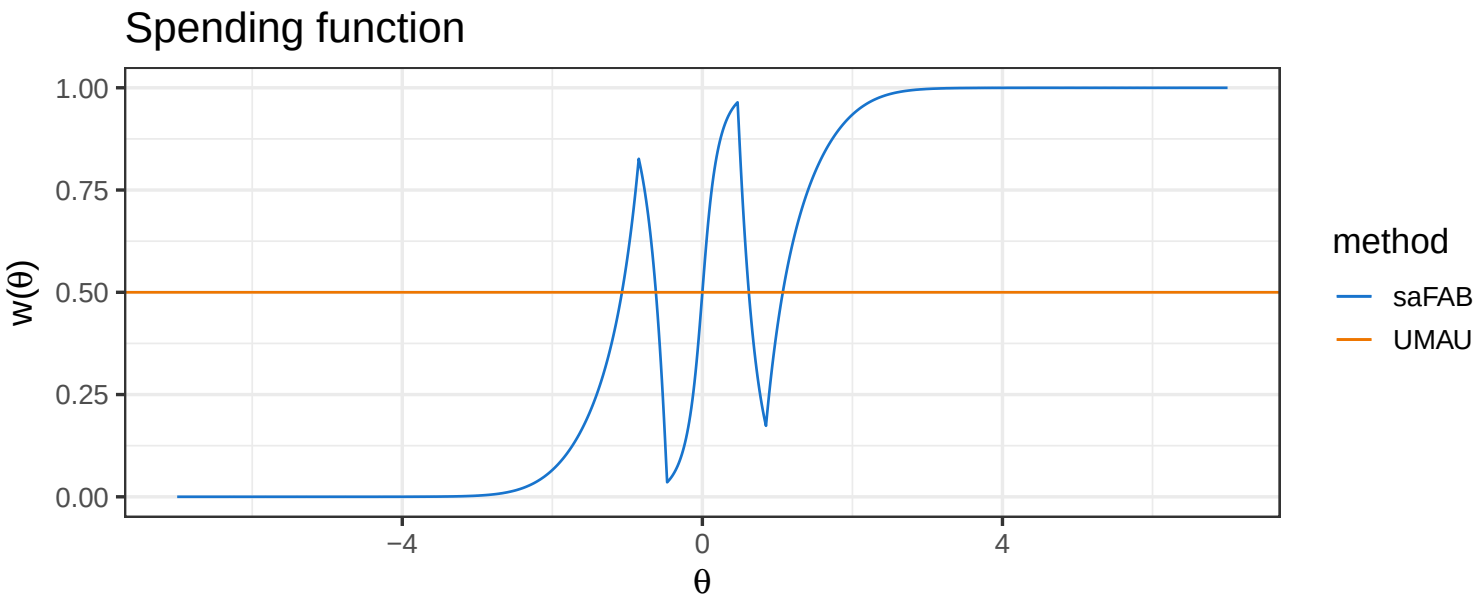
Model:

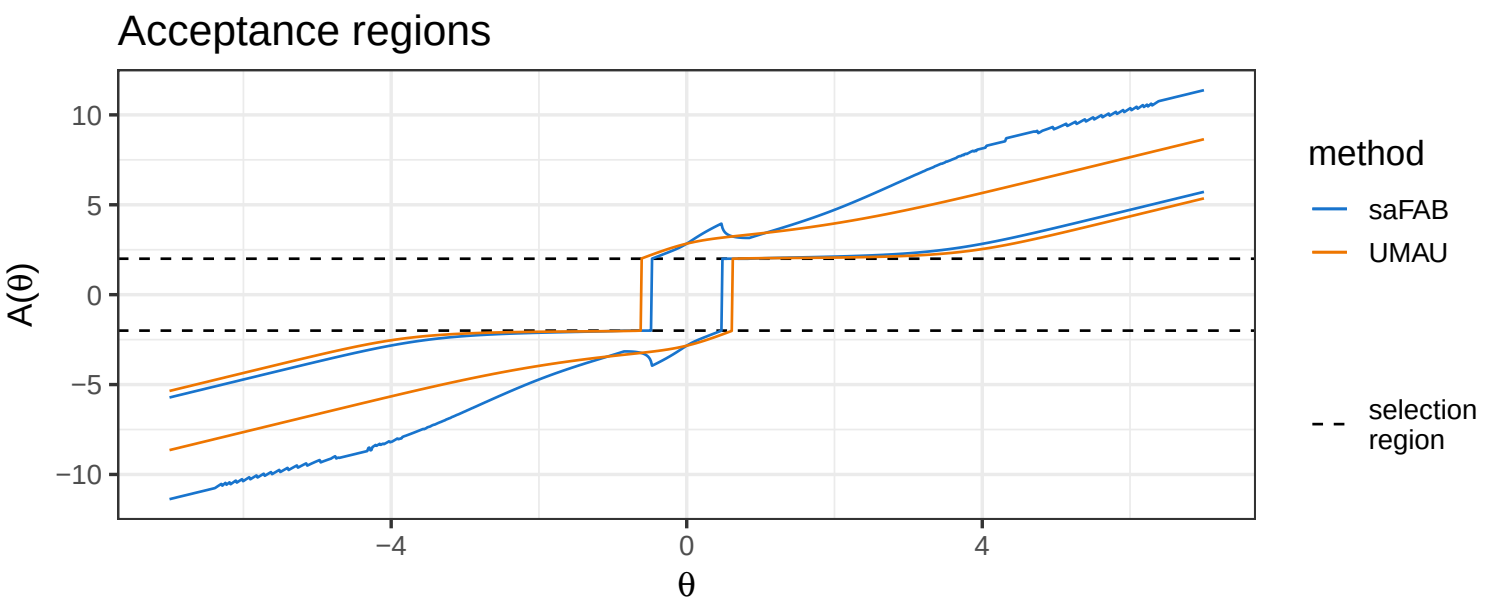
$$\theta_i \sim p \cdot \mathcal{N}(0, \tau^2) + (1 - p) \cdot \delta_0$$
$$(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2)$$

Set parameters to known values $\sigma^2 = 1, p = 0.2, \tau^2 = 3$.

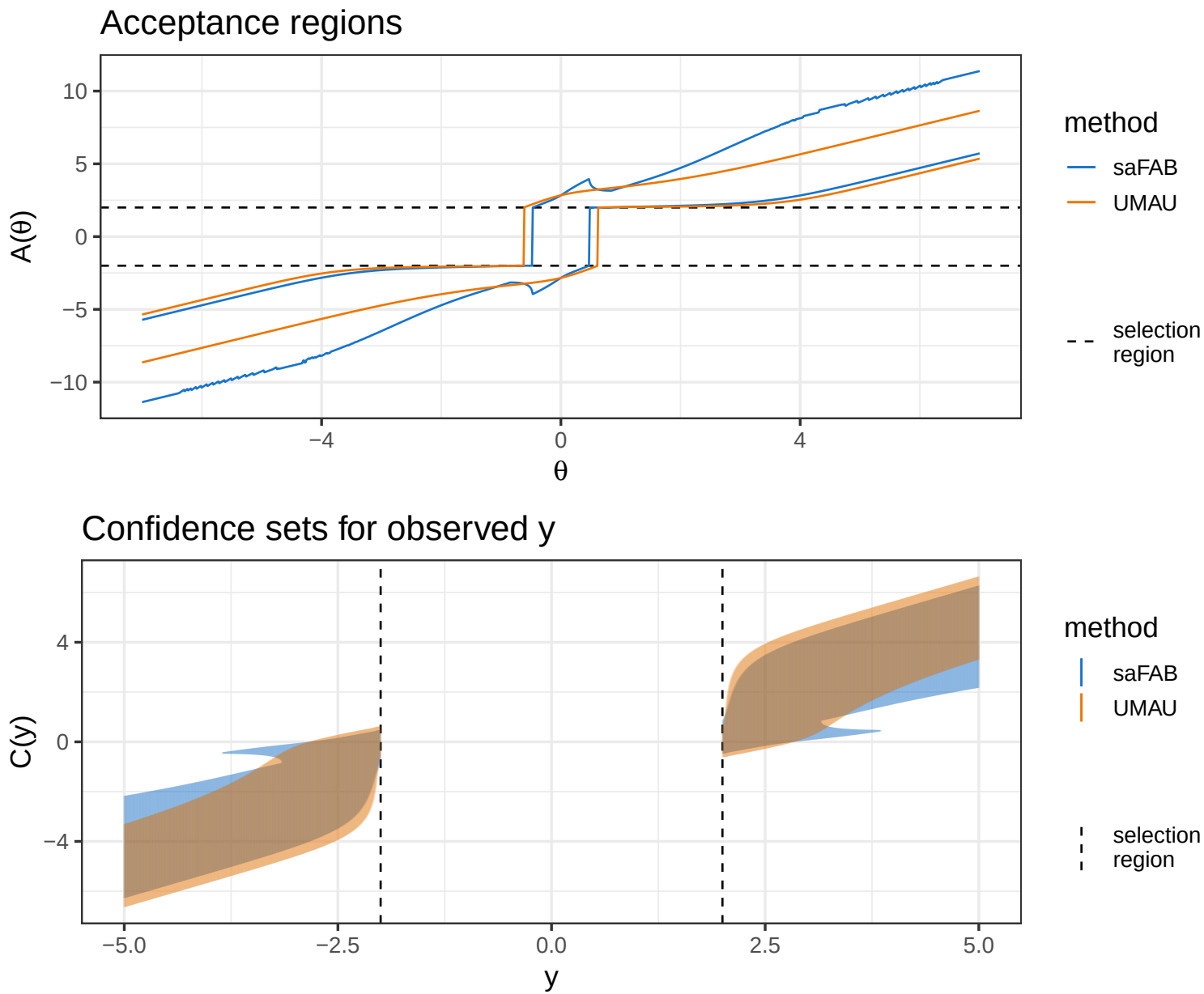
- Generate pairs of (θ_i, y_i)
- Construct 90% selection-adjusted confidence sets / credible intervals for $\{\theta_i : |y_i| > 2\}$

A toy example





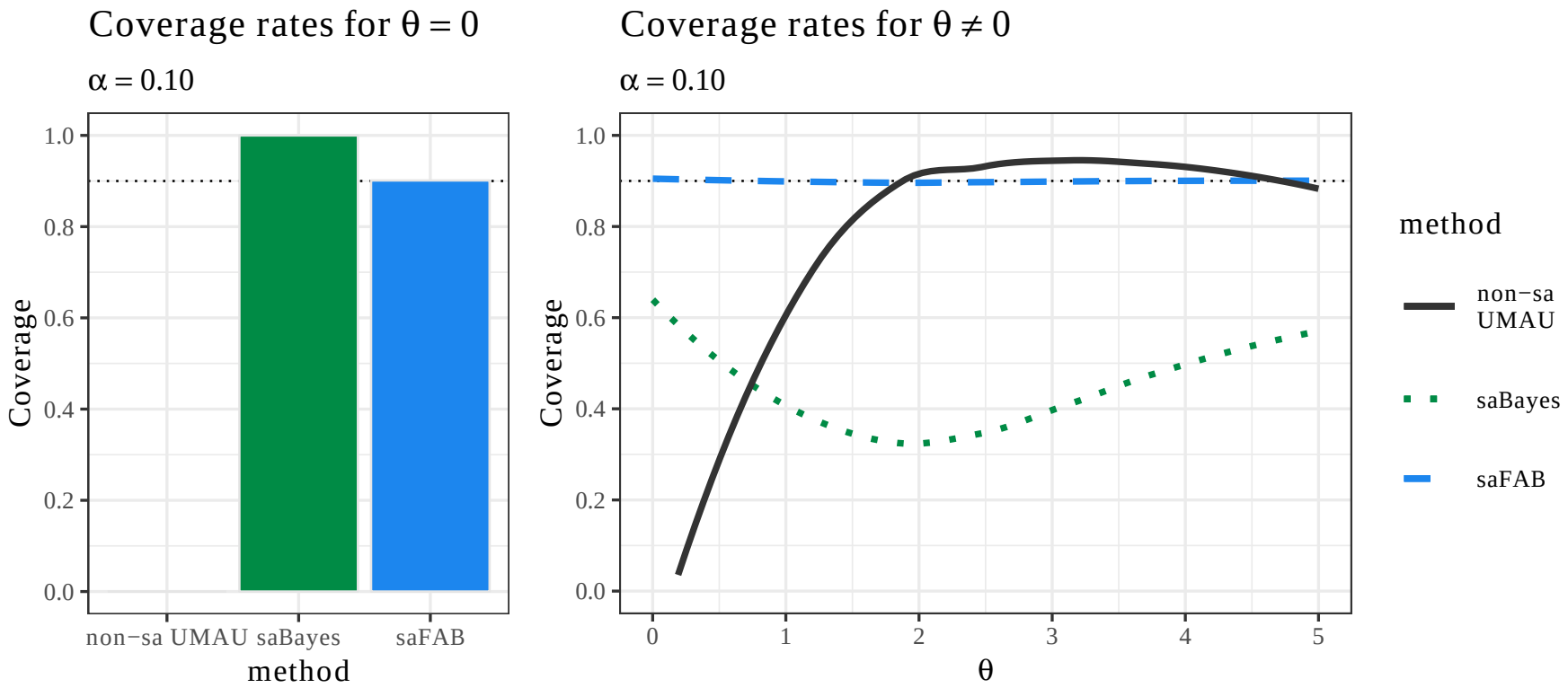
A toy example



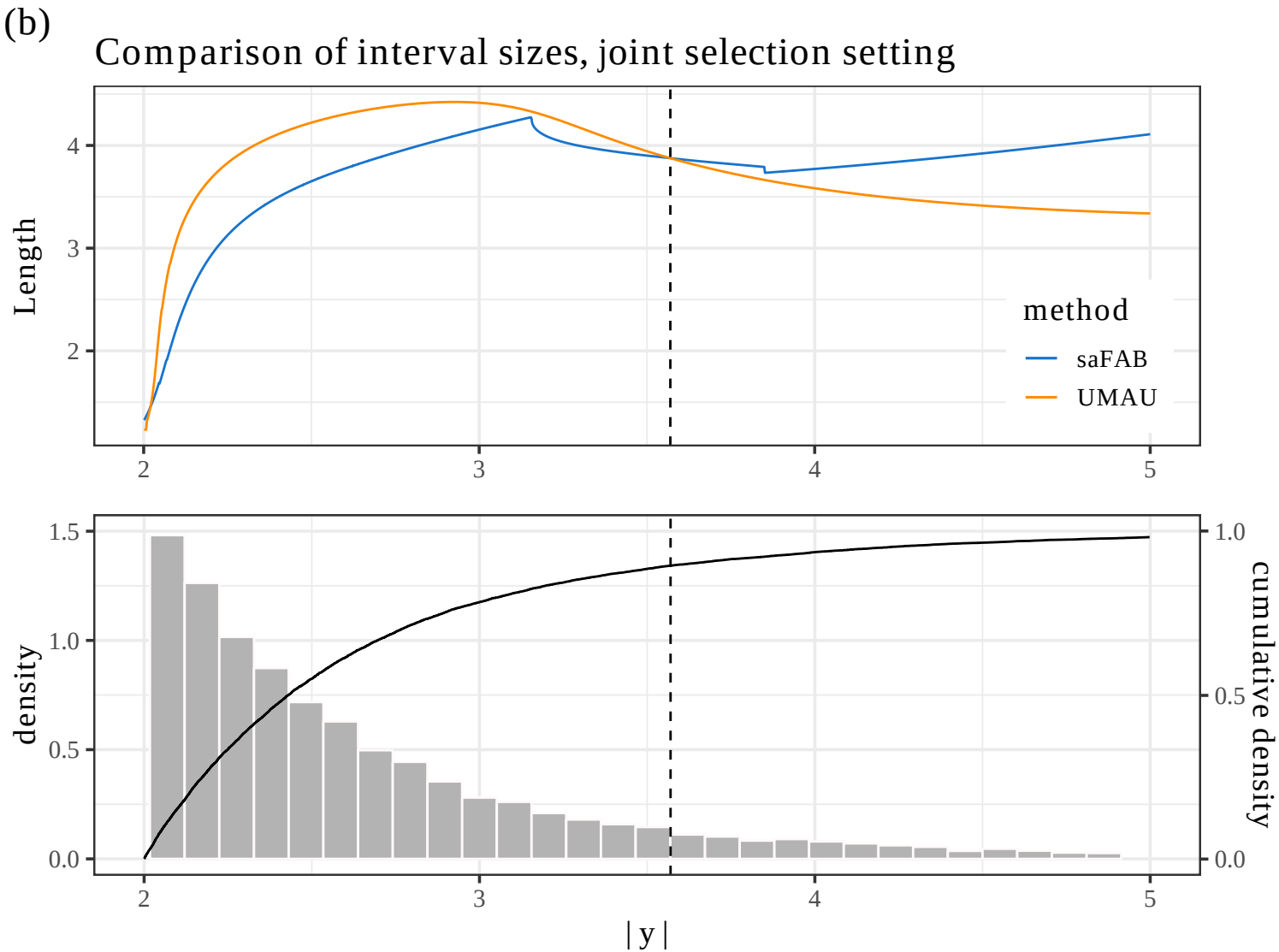
A toy example

Compare frequentist properties of three approaches:

(i) Unadjusted UMAU confidence sets ("non-sa UMAU")
(ii) Selection-adjusted Bayesian credible intervals ("saBayes")
(iii) saFAB



A toy example



Nonparametric empirical-Bayes saFAB

What if we don't know the prior?

Now, assume a generalized form of the two groups model:

$$\pi(\theta) = p \cdot \pi_1(\theta) + (1 - p) \cdot \delta_0(\theta)$$

Nonparametric empirical-Bayes saFAB

What if we don't know the prior?

Now, assume a generalized form of the two groups model:

$$\pi(\theta) = p \cdot \pi_1(\theta) + (1 - p) \cdot \delta_0(\theta)$$

- Using all data $y_1, \ldots, y_n$, use method of predictive recursion (Newton, 2002) to give estimates $\hat{p}$ and $\hat{\pi}_1(\theta)$
- Using the estimated $\hat{\pi}(\theta)$ as the "true" prior, construct $w(\theta)$ and proceed as before

Key fact: fidelity of capturing the "true" prior $\pi(\theta)$ affects efficiency of procedure, *but not coverage*.

Simulated examples

Compare performance of selective confidence sets

- Three variants of saFAB:
  - "Oracle": true prior is known
  - "Parametric empirical Bayes" (PEB): assume a parametric form of the prior, tune parameters with maximum marginal likelihood
  - "Nonparametric empirical Bayes" (NPEB): only assume the generalized two-groups prior, estimate p and non-null density $\pi_1(\theta)$
- sa UMAU confidence sets (Reid et al., 2014)

Performance metrics:

- Coverage rate
- Average size of confidence sets

Case 1: well-specified prior

$$(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2)$$
$$\theta_i \sim p \cdot \mathcal{N}(0, \tau^2) + (1-p) \cdot \delta_0$$

Same parameters as before ($\sigma^2 = 1, p = 0.2, \tau^2$). 100 batches of data, each with $n = 1000$ samples constructed as follows:

- Joint selection: generate n pairs of (θ_i, y_i)
- Conditional selection: Generate n draws of θ_i from the prior *once*. For each batch, generate $(y_i \mid \theta_i) \sim \mathcal{N}(\theta_i, \sigma^2)$

For both, same selection set $S = \{y : |y| > 2\}$

Case 1: well-specified prior

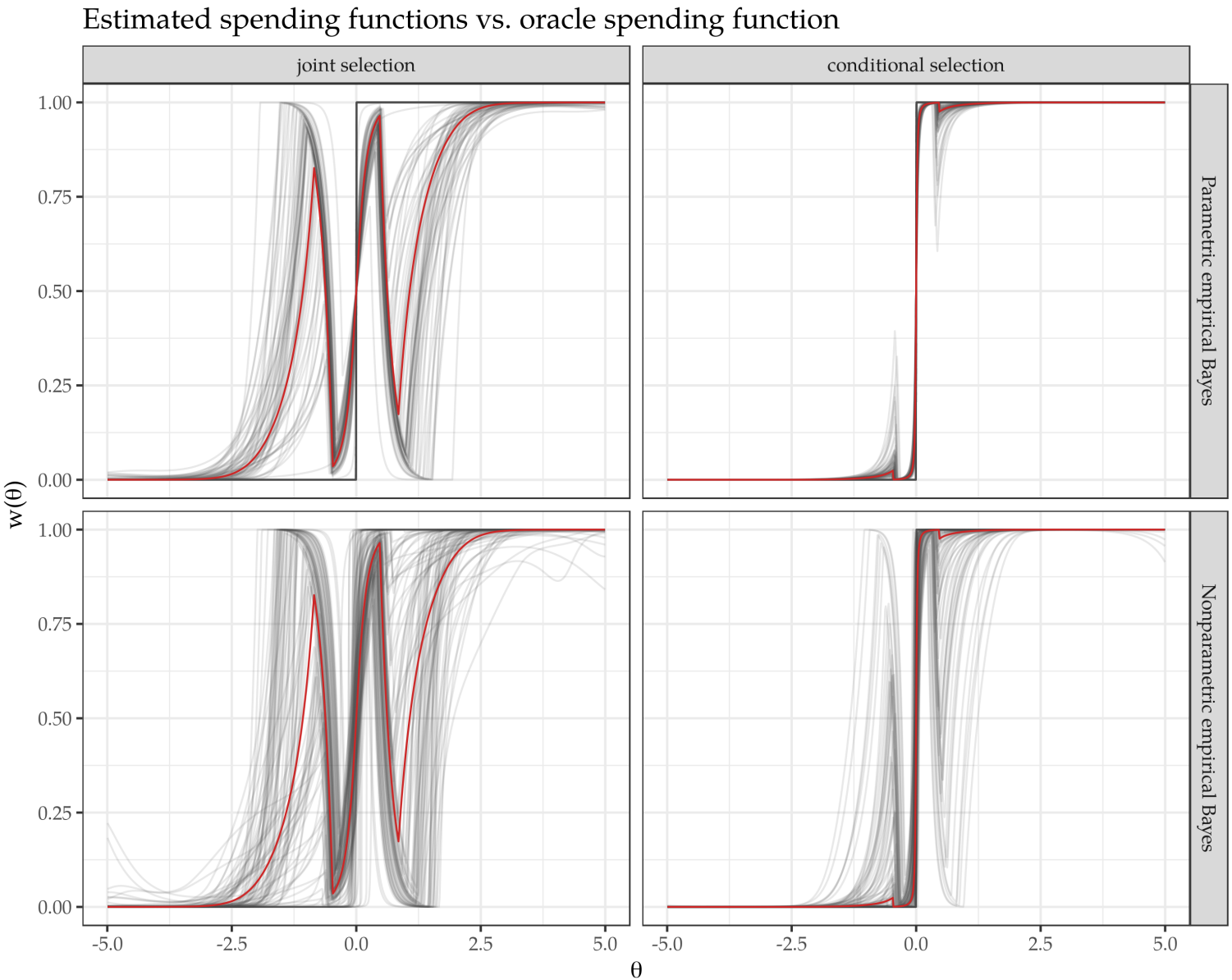
	Coverage	Average size	Rel. average size
Oracle	0.9022 (0.0297)	3.2801 (0.0099)	0.8755 (0.0027)
UMAU	0.9039 (0.0295)	3.7464 (0.0088)	1.0000 (0.0024)
PEB	0.9015 (0.0298)	3.2829 (0.0201)	0.8763 (0.0054)
NPEB	0.8891 (0.0314)	3.2892 (0.0285)	0.8780 (0.0076)

Table: Joint selection

	Coverage	Average size	Rel. average size
Oracle	0.8998 (0.0300)	3.3542 (0.0106)	0.9028 (0.0028)
UMAU	0.9056 (0.0292)	3.7153 (0.0079)	1.0000 (0.0021)
PEB	0.9153 (0.0278)	3.3948 (0.0122)	0.9137 (0.0033)
NPEB	0.8953 (0.0306)	3.3729 (0.0292)	0.9078 (0.0079)

Table: Conditional selection

Case 1: well-specified prior



Case 2: misspecified prior

For the "parametric empirical Bayes" saFAB procedure, still assume the point mass / Gaussian mixture.

However, consider two forms of prior misspecification

- **Bimodal non-null distribution:**

$$\theta_i \sim \begin{cases} \mathcal{N}(-\mu, \tau^2) & \text{w.p. } p/2 \\ \mathcal{N}(\mu, \tau^2) & \text{w.p. } p/2 \\ \delta_0 & \text{w.p. } 1-p \end{cases}$$

  with $p = 0.1, \mu = 4, \tau^2 = 1/4$

- **Skewed, non-null distribution:**

$$\theta_i \sim \begin{cases} \mu + \text{Exponential}(\lambda) & \text{w.p. } p \\ \delta_0 & \text{w.p. } 1-p \end{cases}$$

  with $p = 0.1, \mu = 1, \lambda = 1$

Case 2: misspecified prior

	Coverage	Average size	Rel. average size
Oracle	0.9041 (0.0294)	3.3012 (0.0013)	0.9030 (0.0004)
UMAU	0.8991 (0.0301)	3.6557 (0.0018)	1.0000 (0.0005)
PEB	0.9033 (0.0295)	3.3518 (0.0012)	0.9169 (0.0003)
NPEB	0.9038 (0.0295)	3.3059 (0.0015)	0.9043 (0.0004)

Table: Bimodal prior

	Coverage	Average size	Rel. average size
Oracle	0.8948 (0.0307)	3.2504 (0.0024)	0.8738 (0.0006)
UMAU	0.9003 (0.0300)	3.7199 (0.0023)	1.0000 (0.0006)
PEB	0.8984 (0.0302)	3.3496 (0.0064)	0.9004 (0.0017)
NPEB	0.8944 (0.0307)	3.2532 (0.0025)	0.8745 (0.0007)

Table: Skewed prior

Real data example: neural synchrony

- Data from Rob Kass & co. (CMU) on firing rates between pairs of neurons. (Kelly et al., 2010; Smith and Kohn, 2008)
- Have a test statistic z_i for each neuron pair i
- Previously analyzed using two-groups model by Scott et al. (2013)

Use nonparametric empirical Bayes, $S = \{z_i : |z_i| > 2\}$

Real data example: neural synchrony

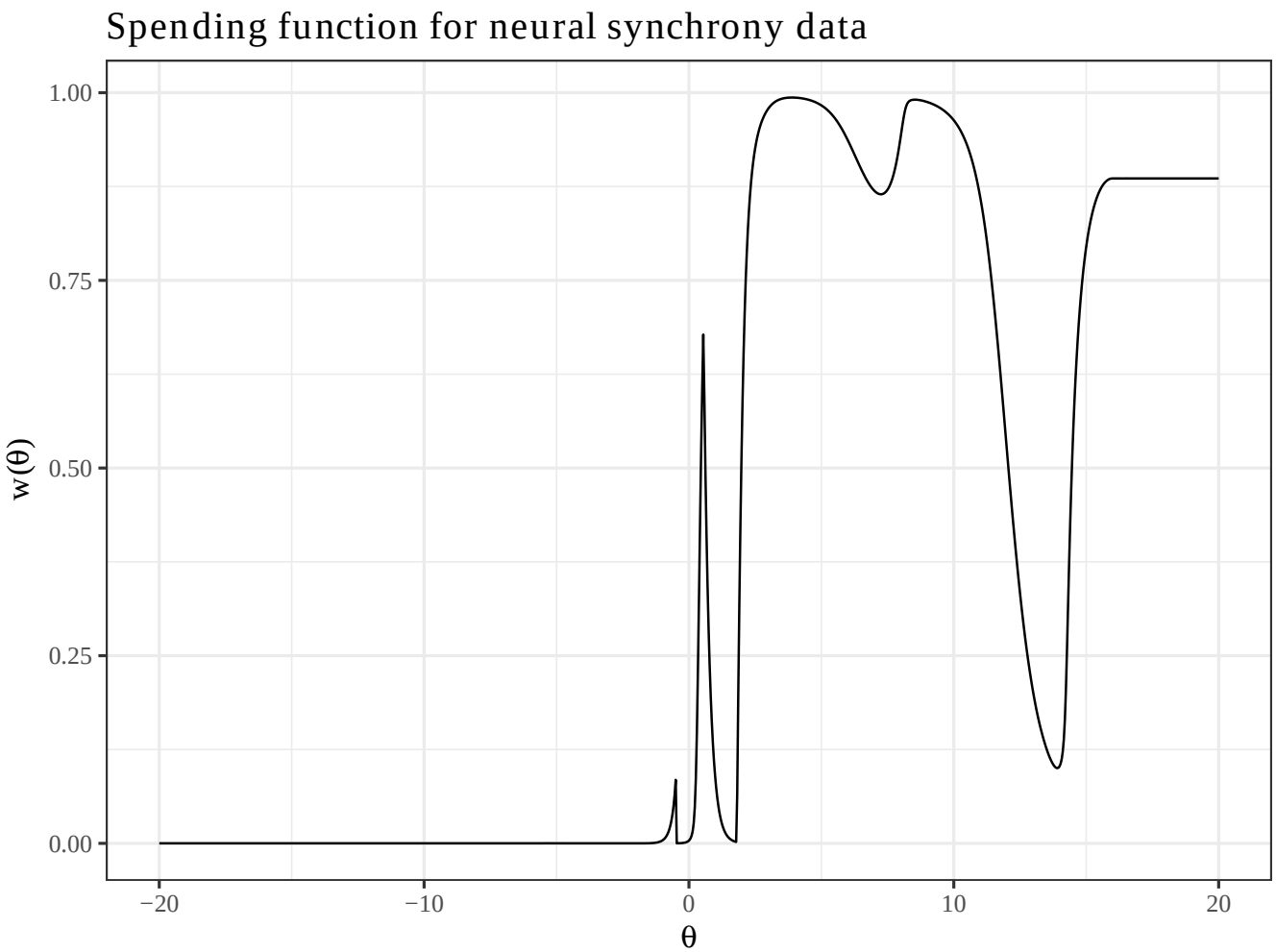
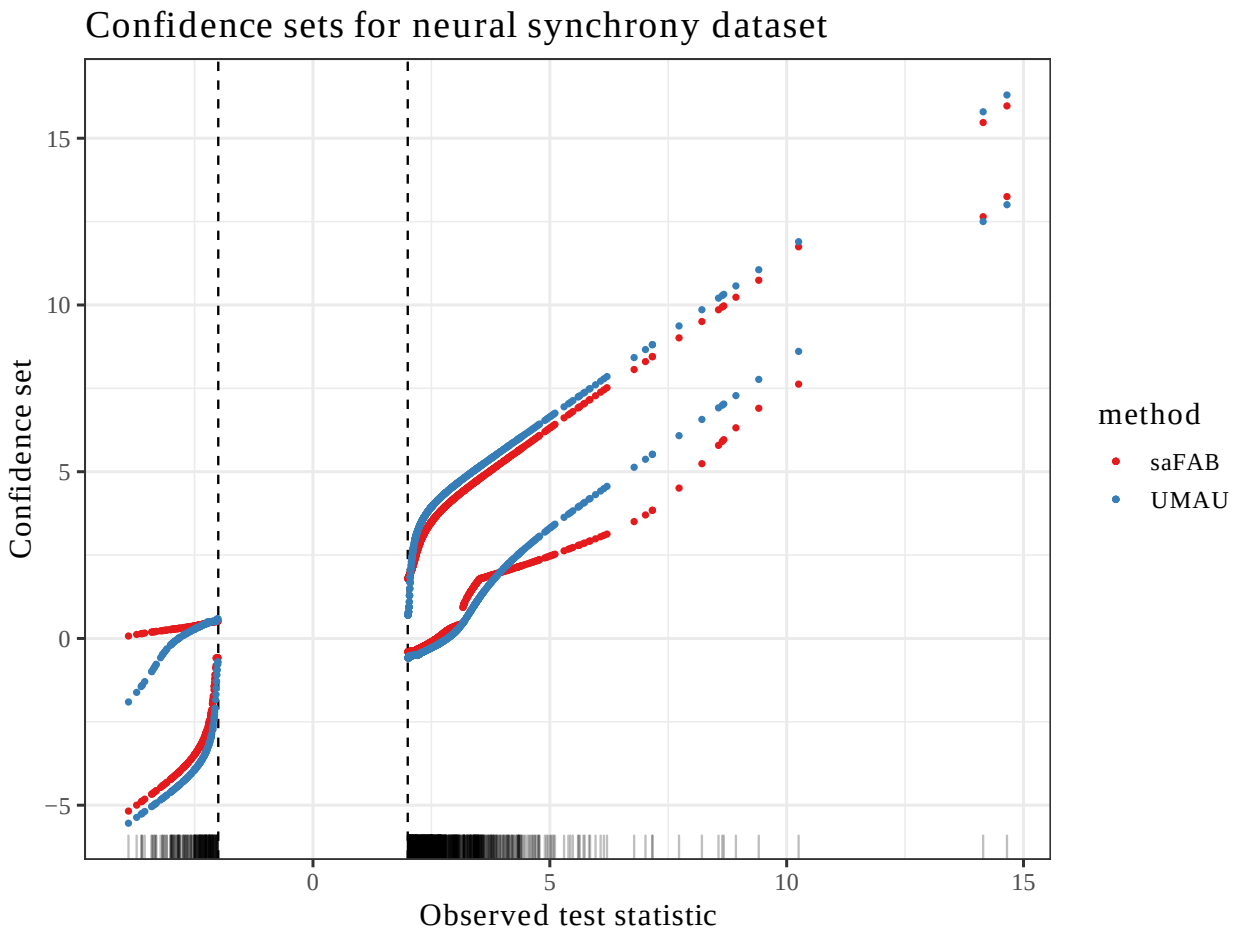


Figure: Estimated spending function for neural synchrony data from .

Real data example: neural synchrony



	Average size	Rel. average size
UMAU	3.8102	1.0000
saFAB	3.3671	0.8836

Recent work

Sparse means

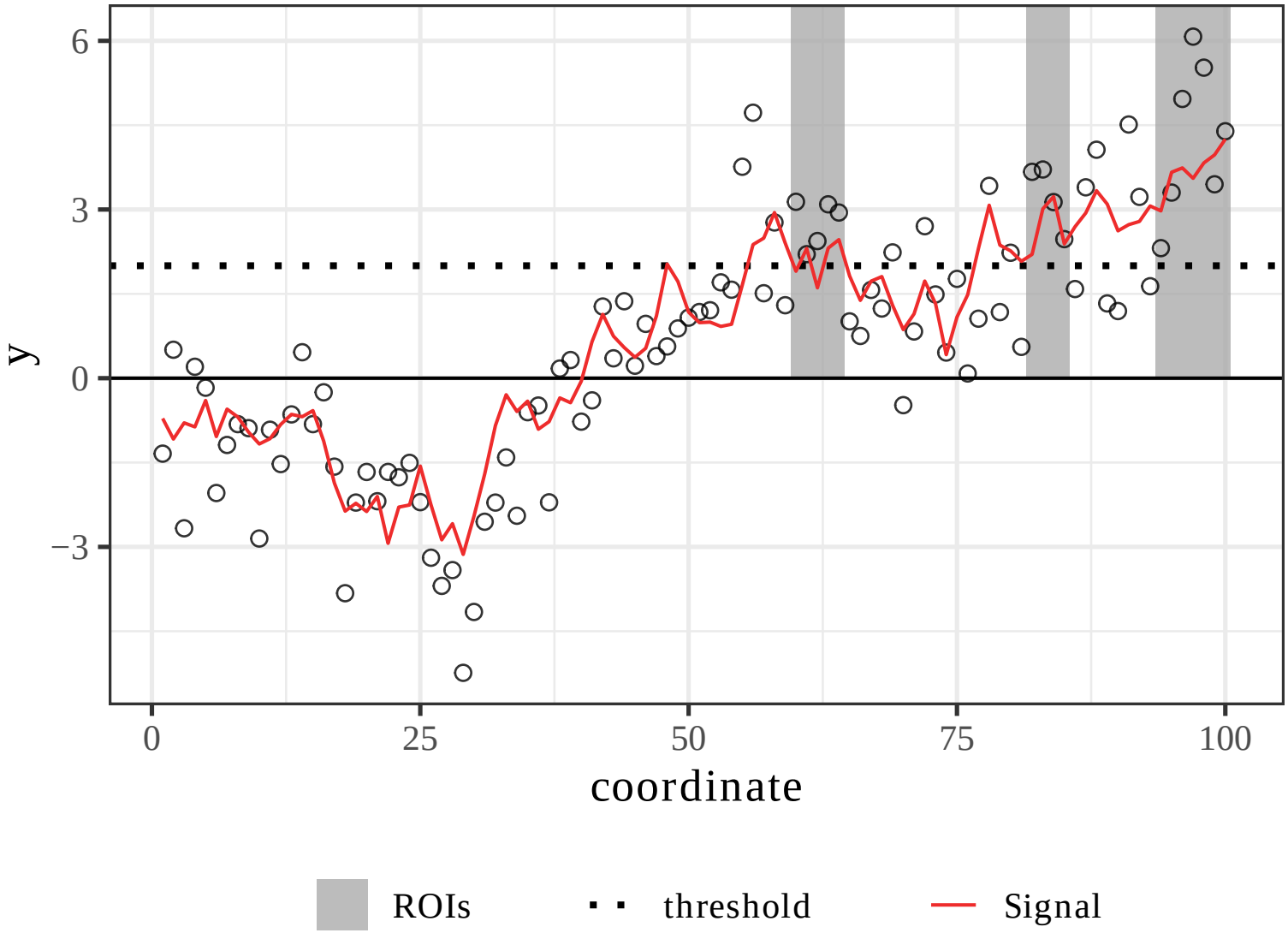
S. Woody and J.G. Scott. “Optimal post-selection inference for sparse signals: a nonparametric empirical-Bayes approach.” arXiv:1810.11042. 2018.

Submitted to *JASA Theory and Methods*.

Spatial hotspot detection

Ongoing

Detecting regions of interest



Example: Differentially methylated regions

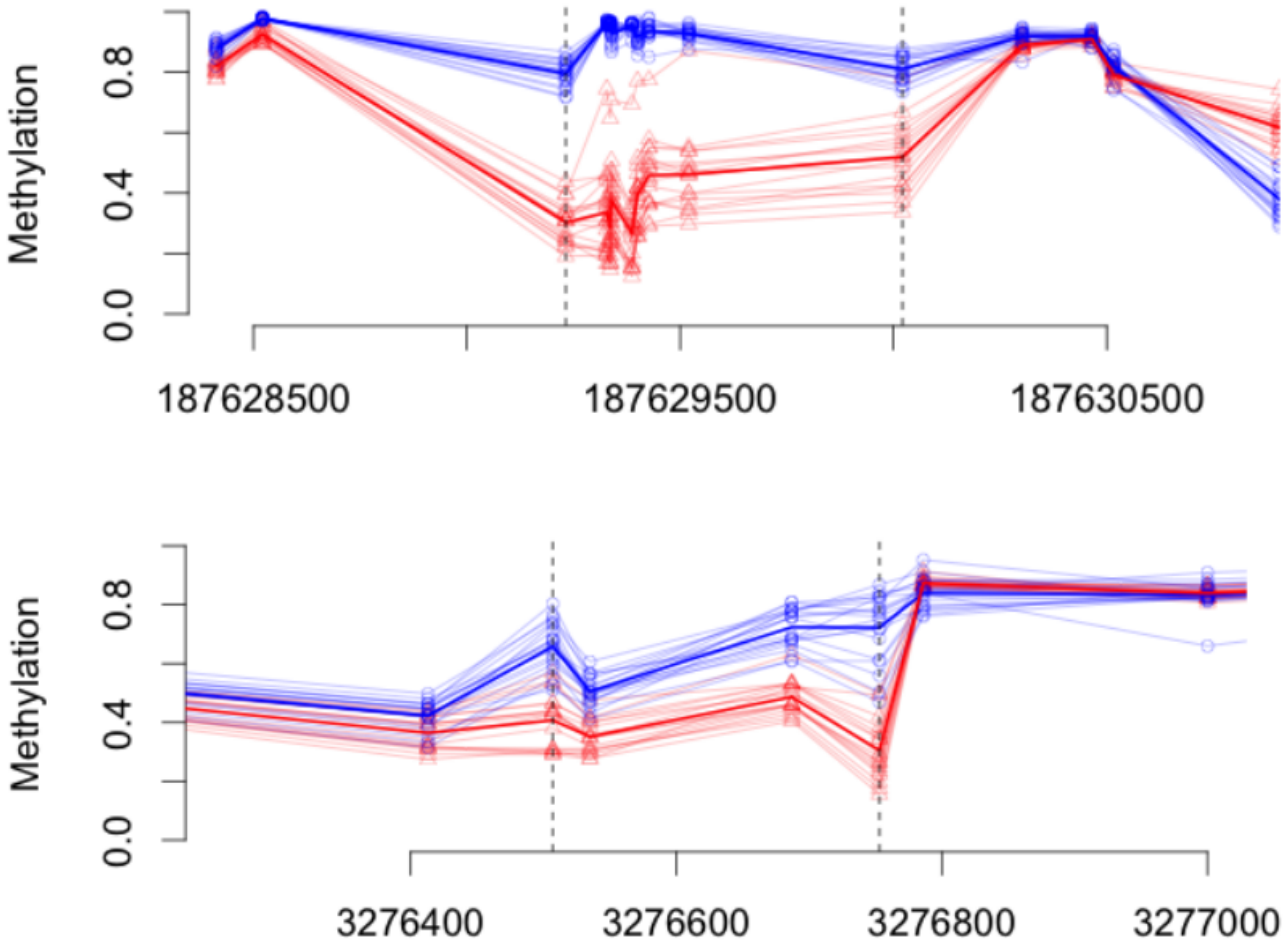


Figure: From Benjamini et al. (2016)

Set up

Observe a vector y associated with a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a latent spatial signal θ,

$$(y_v|\theta_v) \sim \mathcal{N}(\theta_v, \sigma^2),\ v \in \mathcal{V}$$

Detecting regions of interest (ROIs)

Denote an ROI as R, found following a three-step process:

(i) *Smooth* the noisy observations (optional), e.g. with a linear smoother,

$$\tilde{y} := Hy.$$

(ii) *Threshold* the smoothed observations at some value t.

(iii) *Merge* together contiguous regions where smoothed observations fall above the threshold. With a chain graph,

$$R = (a, a+1, \ldots, b-1, b) \text{ s.t. } \tilde{y}_i > t \; \forall \; i \in R$$

Key fact: Restrict inference to R *conditioned* on

$$\tilde{y}_R > t \Leftrightarrow H_R y > t$$

Detecting regions of interest

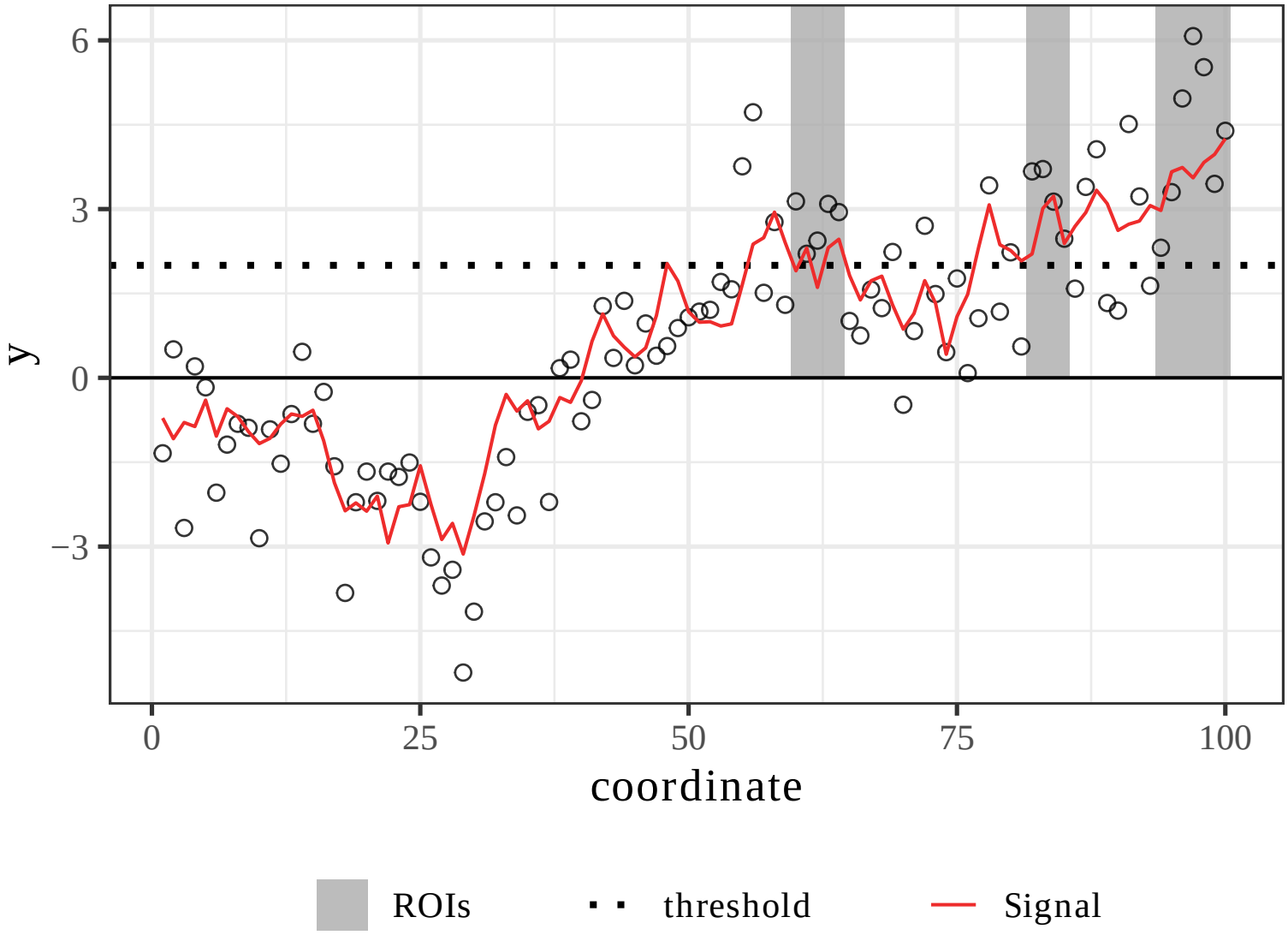


Figure: Threshold & merge

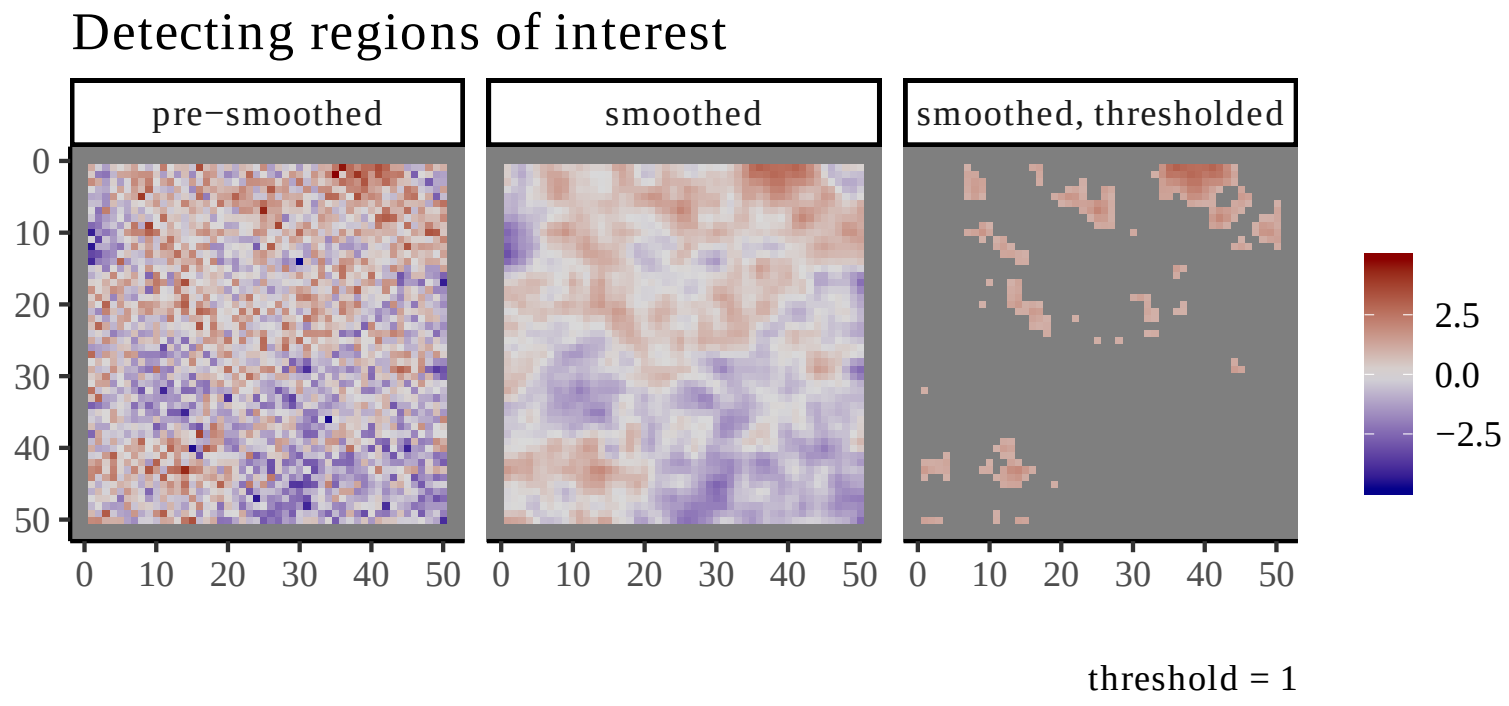


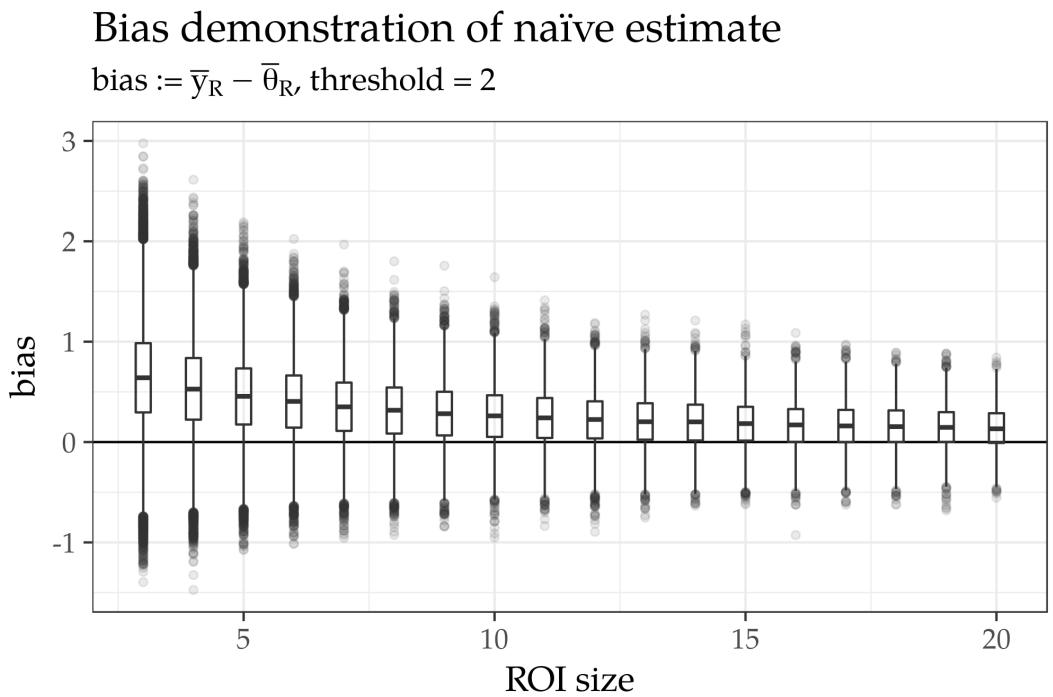
Figure: Smooth, threshold & merge

Target of inference

After detecting a region R, the goal is to provide inference for

$$\bar{\theta}_R := \frac{1}{|R|} \sum_{i \in R} \theta_i,$$

i.e. the *mean signal for the ROI*. The naïve estimate $\bar{y}_R$ will be *biased upwards*.



Target of inference

$$\bar{\theta}_R := \frac{1}{|R|} \sum_{i \in R} \theta_i,$$

Advantages:

- Autocorrelation implies that signals are approximately locally constant
- Turn a *spatial POSI problem* into a *scalar POSI problem*, i.e. framing it as a problem we've already solved

Post-selection inference with spatial ROIs

The selection-adjusted likelihood is

$$f_S(y \mid \theta) = \frac{\mathcal{N}(y \mid \theta, \sigma^2 \mathcal{I}) \cdot \mathbf{1}(H_R y > t)}{\int_{H_R y > t} \mathcal{N}(y \mid \theta, \sigma^2 \mathcal{I}) dy}.$$

May construct selective confidence sets as previously described

Bayesian inference

We use the centered ICAR prior for θ,

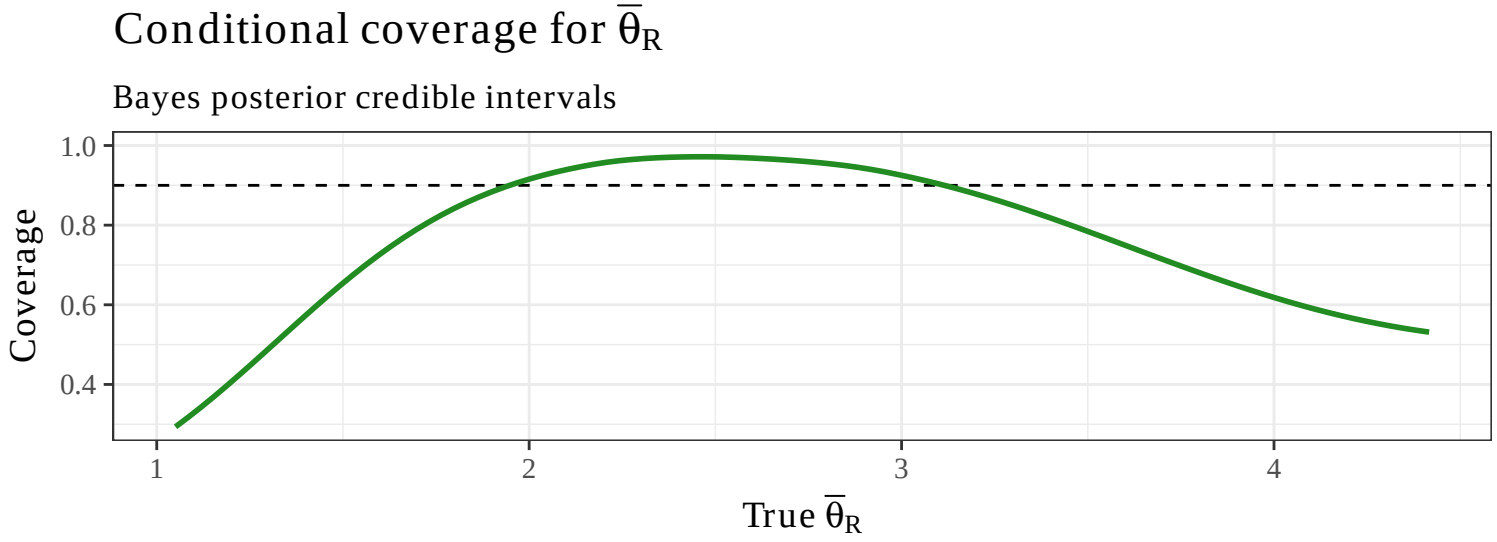
$$\pi(\theta) \propto \exp\left[-\frac{1}{2\tau^2}\sum_{(v,w)\in\mathcal{E}}(\theta_v-\theta_w)^2\right]\cdot\exp\left[-\frac{1}{2\lambda^2}\bar{\theta}^2\right],$$

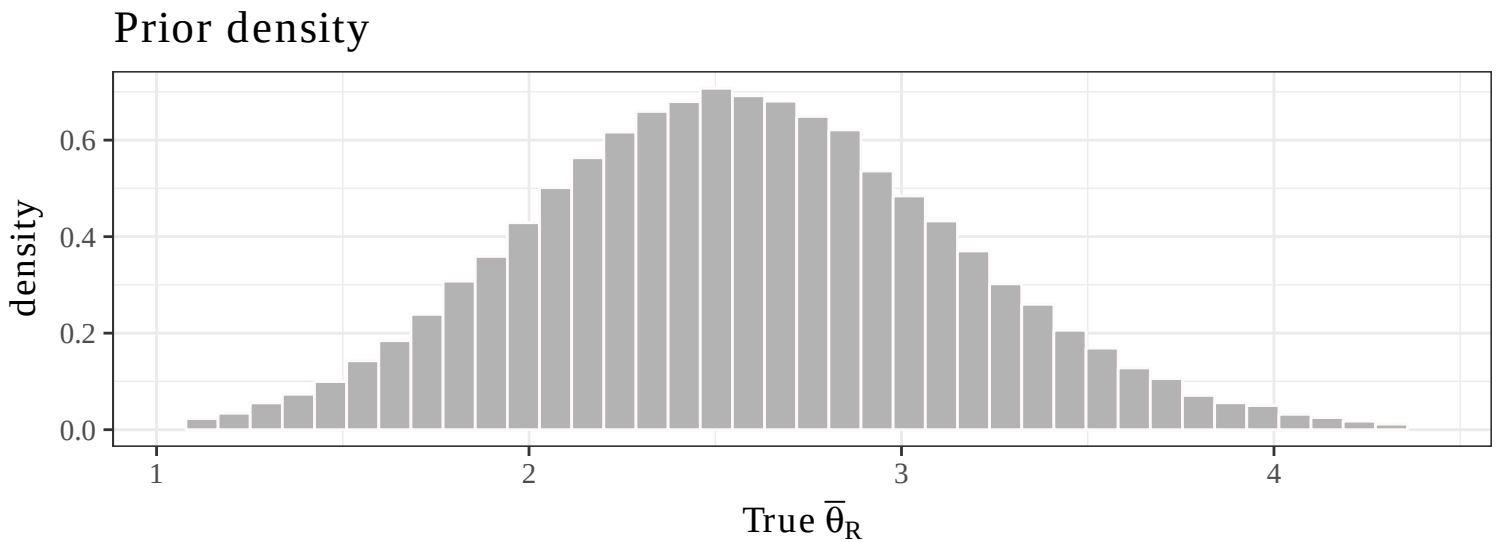
where $\bar{\theta}$ is the mean of the components of θ.

The sampling model is

$$(y_v \mid \theta_v) \sim \mathcal{N}(\theta_v, \sigma^2), \quad v \in \mathcal{V}$$

for the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.


Conditional coverage for $\overline{\theta}_R$
Bayes posterior credible intervals
Coverage
1.0
0.8
0.6
0.4
1
2
3
4
True $\overline{\theta}_R$


Prior density
density
0.6
0.4
0.2
0.0
1
2
3
4
True $\overline{\theta}_R$

Selection-adjusted confidence interval for ROI

Construct hypothesis tests for $\bar{\theta}_R$ around the the sampling distribution for the statistic $f_S(\bar{y}_R|\bar{\theta}_R)$ (see Benjamini et al., 2016).

Spatial selection-adjusted FAB procedure

Using $f_S(\bar{y}_R \mid \bar{\theta}_R)$ and marginal $m_S(\bar{y}_R)$, can construct spending function $w(\bar{\theta}_R)$ and family of biased hypothesis tests as before.

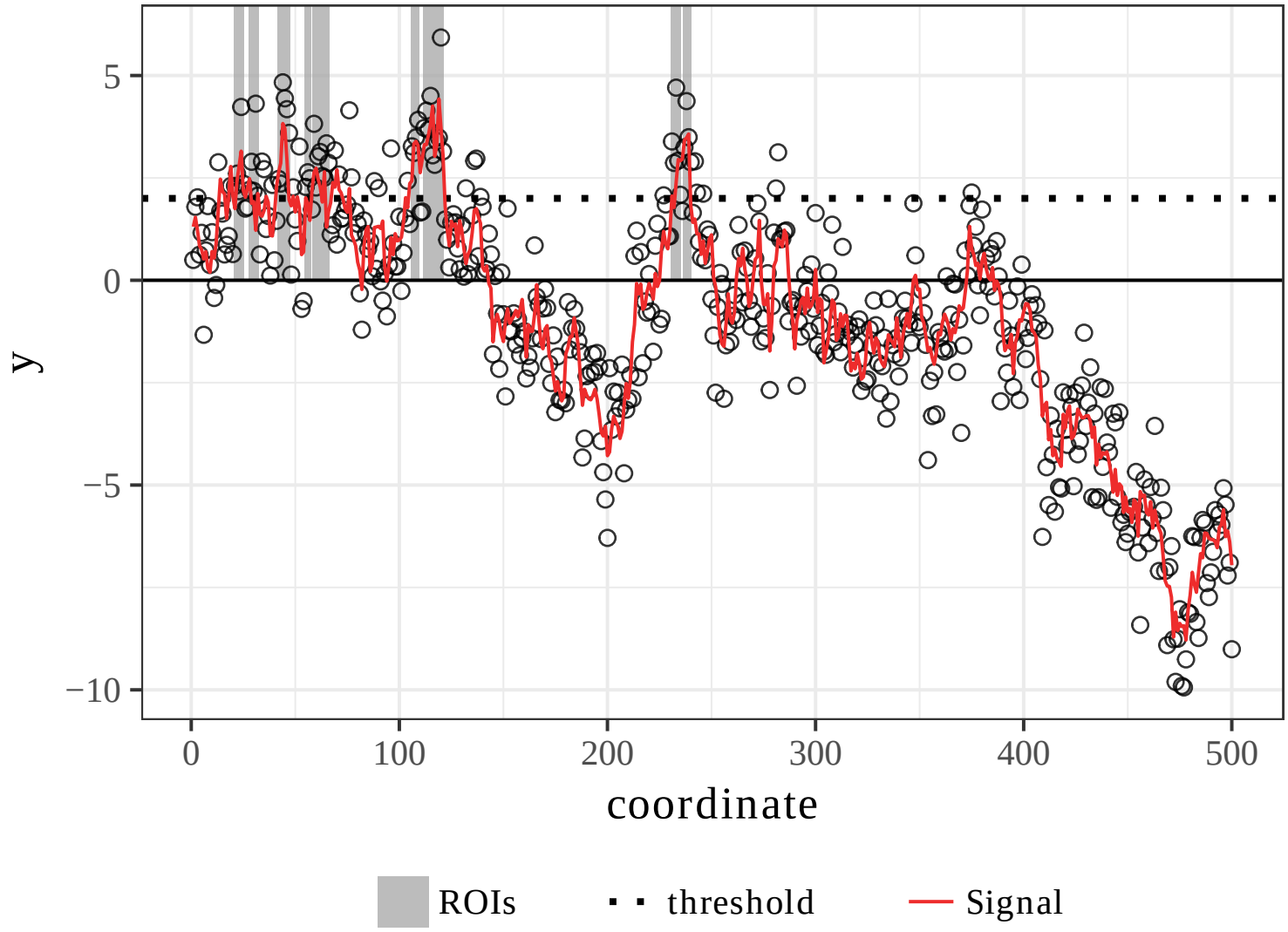
Simulation study

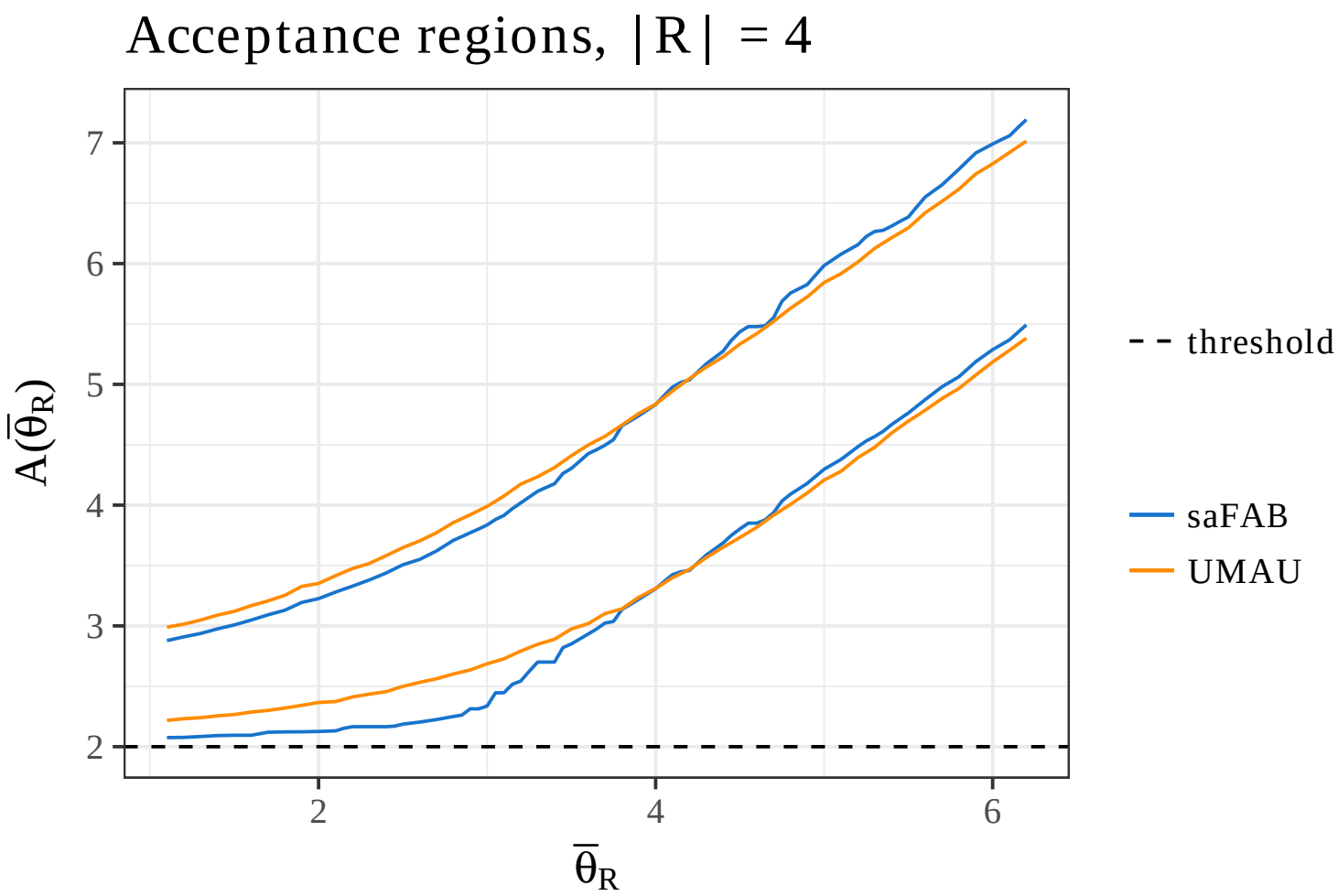
50,000 simulations performed as follows:

- Chain graph of length 500
- θ generated from ICAR prior with $\tau^2 = 0.25$ and $\lambda^2 = 1$
- $(y|\theta) \sim \mathcal{N}(\theta, \mathcal{I})$
- Threshold for detecting ROIs set to $t = 2$
- No smoothing step involved ($H = \mathcal{I}$)

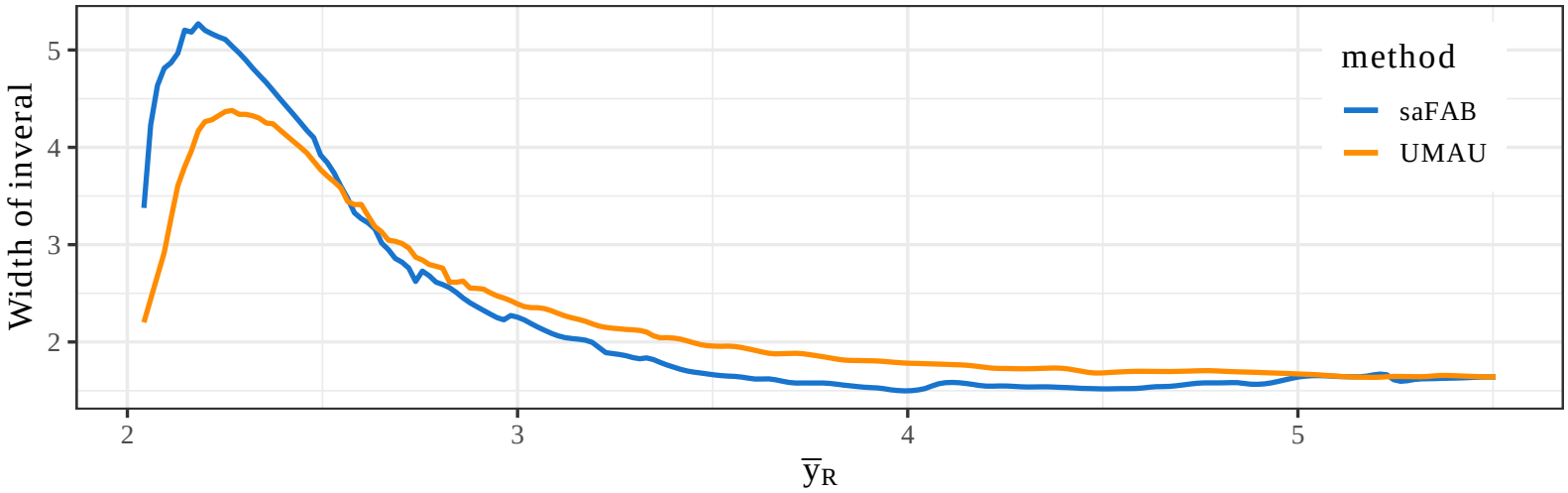
$F_S(\bar{y}_R \mid \bar{\theta}_R)$ and $M_S(\bar{y}_R)$ are approximated via Monte Carlo.

Detecting regions of interest

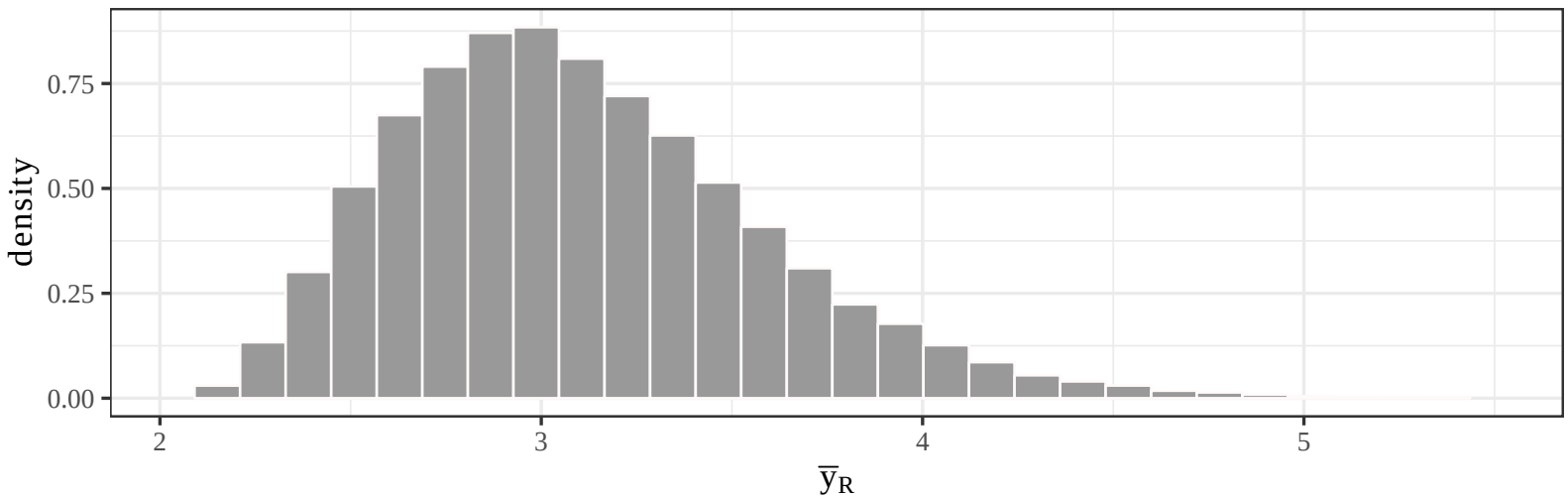


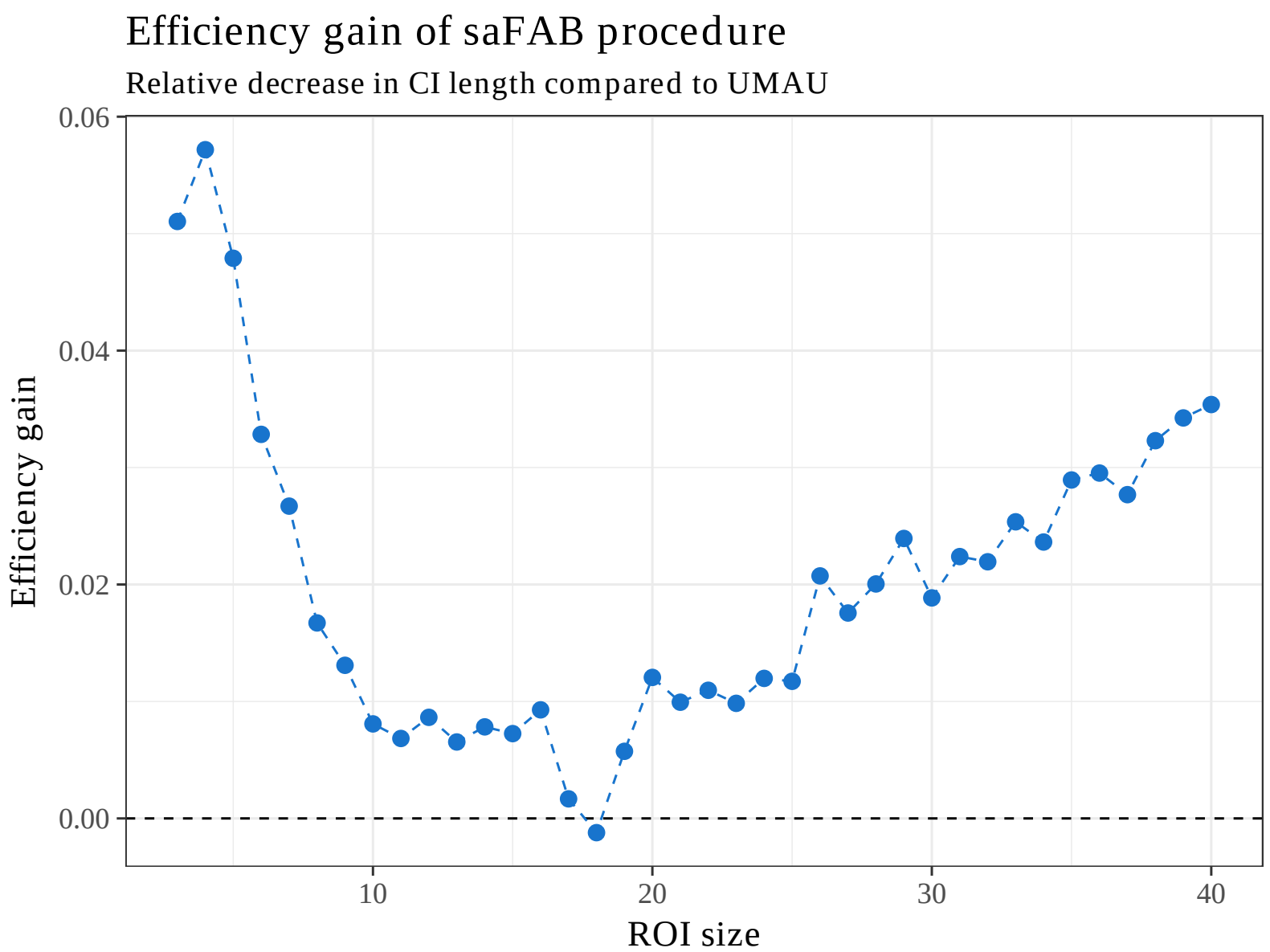

Acceptance regions, |R| = 4
7
6
5
4
3
2
A(θ̄R)
2
4
6
θ̄R
threshold
saFAB
UMAU

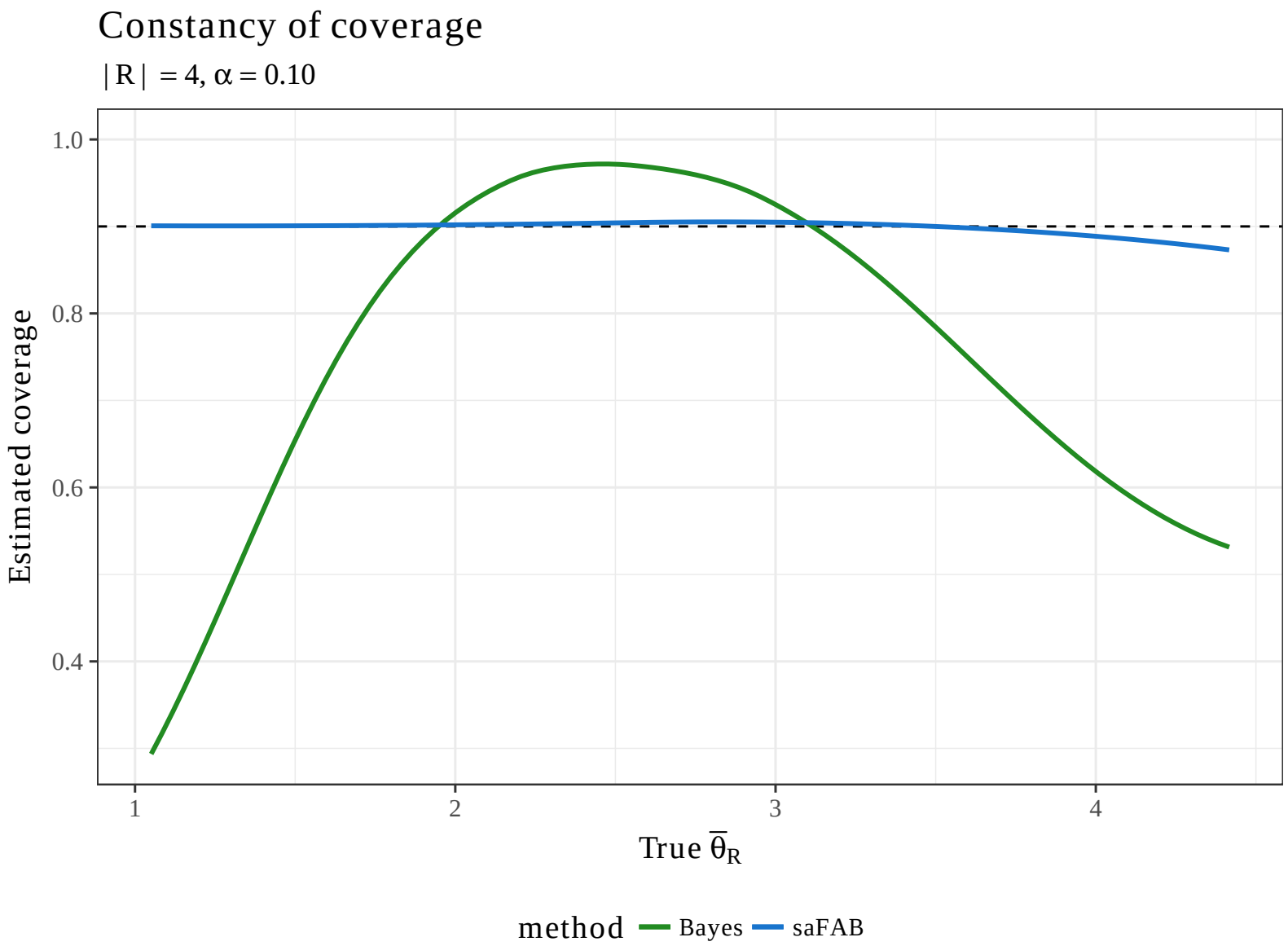
Comparison of interval widths for estimating $\overline{\theta}_R$, $|R|=4$



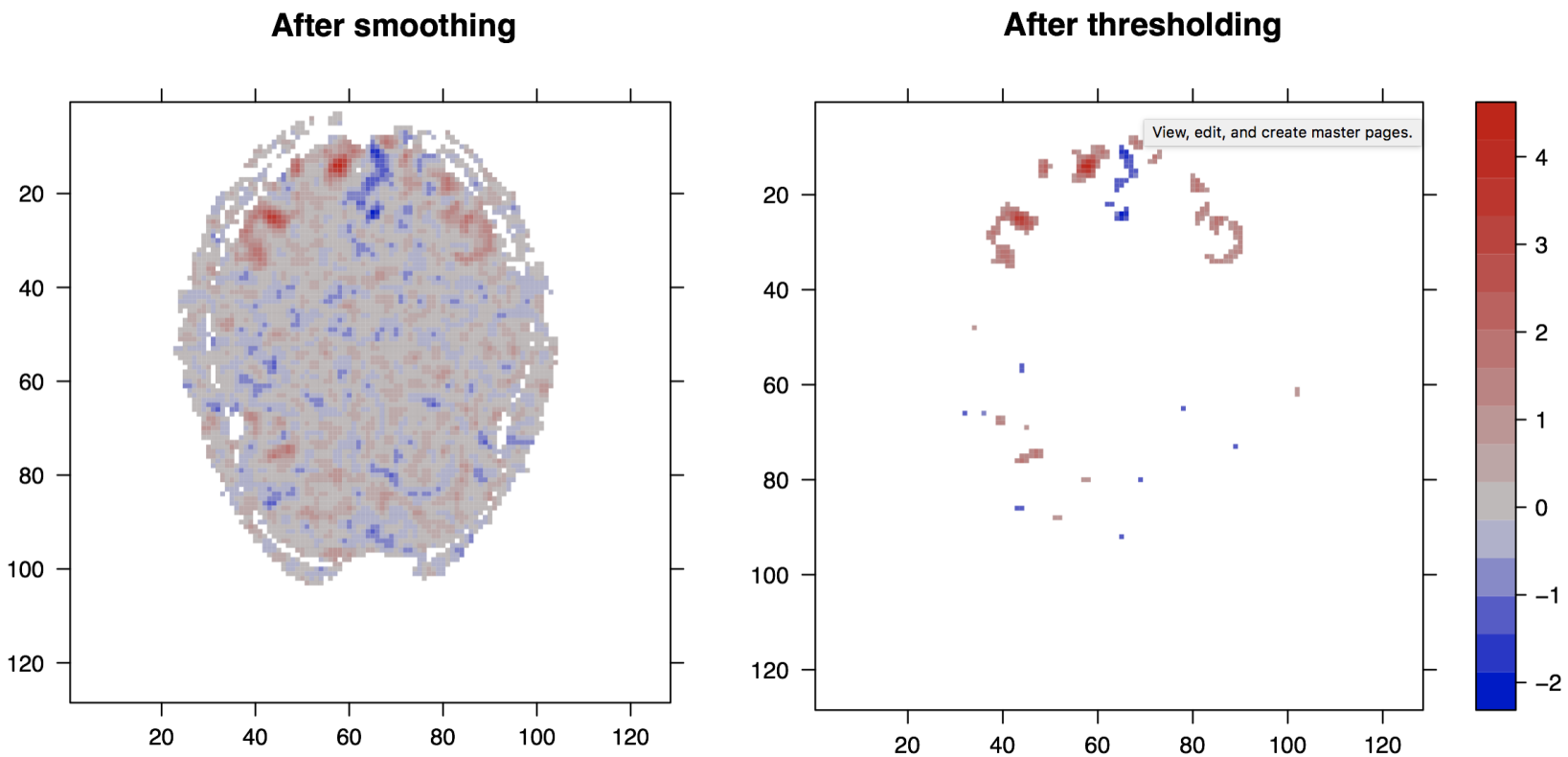
Marginal distribution of $\overline{y}_R$




Efficiency gain of saFAB procedure
Relative decrease in CI length compared to UMAU
Efficiency gain
0.06
0.04
0.02
0.00
10
20
30
40
ROI size


Constancy of coverage
|R| = 4, α = 0.10
Estimated coverage
1.0
0.8
0.6
0.4
1
2
3
4
True θ̄R
method
Bayes
saFAB

Future application: fMRI



Future directions

Interpreting complex models via posterior summarization; joint work with Profs. Carlos Carvalho and Jared Murray

Linear case

Consider linear model for large p,

$$y_i = X\beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Recent work has focused on variable selection by following a two-stage approach:

(1) Fit a (possibly sparse) Bayesian model on all coefficients
(2) Find the minimal subset of features that satisfactorily characterize predictions

Linear case

Hahn and Carvalho (2015): $\bar{\beta}$ is the posterior mean for β, solve the decision problem

$$\beta_\lambda := \arg\min_\gamma \|X\bar{\beta} - X\gamma\|_2^2 + \lambda\|\gamma\|_0$$

(related work by Goutis and Robert, 1998; Dupuis and Robert, 2003; Piironen et al., 2018):

What inferential statements can we make concerning the "selected" model, the nonzero components of β_λ?

Nonlinear case

Consider a general nonparametric regression,

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

A popular approach is BART (Chipman et al., 2010) a sum of trees model,

$$f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j),$$

allowing for interactions, and nonlinear, discontinuous effects.

Nonlinear case

There is a tradeoff between simple and interpretable, yet misspecified models, and more realistic and flexible, yet "black box" models

For instance, an easily interpretable model is the additive model

$$y_i = \gamma_0 + \sum_{j=1}^{p} \gamma_j(x_{ij}) + \epsilon_i$$

Nonlinear case

Key idea: estimate the best simple approximation $\gamma \in \Gamma$ to the "true" function $f \in \mathcal{F}$, solving the decision problem

$$\gamma = \arg\min_{\tilde{\gamma} \in \Gamma} d(f, \tilde{\gamma}) + p(\tilde{\gamma}), \tag{1}$$

for some distance $d(\cdot, \cdot)$ and penalty function $p(\cdot)$.

Nonlinear case

Key idea: estimate the best simple approximation $\gamma \in \Gamma$ to the "true" function $f \in \mathcal{F}$, solving the decision problem

$$\gamma = \arg\min_{\tilde{\gamma} \in \Gamma} d(f, \tilde{\gamma}) + p(\tilde{\gamma}), \tag{1}$$

for some distance $d(\cdot, \cdot)$ and penalty function $p(\cdot)$.

- Use the data once to approximate f using a Bayesian model
- For each posterior draw, solve (1) to get a posterior on this summarization
- This gives an interpretable *summary* to nonparametric fitted function f
- When is this summary adequate? When do we need to account for, e.g., interactions?

Conclusion

Email: spencer.woody@utexas.edu

Appendix

(additional slides)

Constructing the spending function

For a given prior $\pi(\theta)$ and selection-adjusted marginal $m_S(y)$, choose $w(\theta)$ to minimize the Bayes loss,

$$
\begin{aligned}
L(\pi, w(\theta)) &= \int R(\theta; w(\theta))\, \pi(\theta) \mathrm{d}\theta \\
&= \int \Pr(Y \in A(\tilde{\theta})) \mathrm{d}\tilde{\theta}. \\
\Rightarrow w(\theta) &= \arg\min_{w \in [0,1]} H_\theta(w), \\
H_\theta(w) &:= \Pr(y \in A(\theta)) \\
&= M_S\left[F_S^{-1}(\alpha w + 1 - \alpha \mid \theta)\right] - M_S\left[F_S^{-1}(\alpha w \mid \theta)\right].
\end{aligned}
$$

This is a variational optimization problem that can be solved:

- analytically in simple settings.
- by brute force (i.e. pointwise in θ) everywhere else.

Predictive recursion

$$
\begin{aligned}
(y_i \mid \theta_i) &\sim \mathcal{N}(\theta_i, \sigma^2) \\
\theta_i &\sim \Psi, \quad \Psi = \tilde{\pi}_1(\theta) + \pi_0 \delta_0,
\end{aligned}
$$

where $\tilde{\pi}_1(\theta) = p \cdot \pi_1(\theta)$ is a sub-density for signals, and $\pi_0 = 1 - p$ is the mass at zero for nulls. By marginalizing out θ we may reformulate the model as

$$
\begin{aligned}
y_i &\sim p \cdot f_1(y_i) + (1 - p) \cdot f_0(y_i) \\
f_0(z) &\sim \mathcal{N}(y; 0, \sigma^2) \\
f_1(z) &\sim \int_{\mathbb{R}} \mathcal{N}(y_i; \theta, \sigma^2) \pi(\theta) \mathrm{d}\theta.
\end{aligned}
$$

We use predictive recursion (Newton, 2002) to estimate the mixing density Ψ from the observations $y_1, \ldots, y_n$.

Begin with an intial guess $\Psi^{[0]}$ a sequence of weights $\gamma^{[i]} \in (0,1)$. For $i = 1, \ldots, n$, recursively compute the update

$$m^{[i-1]}(y_i) = \int_{\mathbb{R}} \mathcal{N}(y_i; u, \sigma^2) \Psi^{[i-1]}(\mathrm{d}u)$$

$$\Psi^{[i]}(\mathrm{d}u) = (1 - \gamma^{[i]}) \Psi^{[i-1]}(\mathrm{d}u) + \gamma^{[i]} \cdot \left\{ \frac{\mathcal{N}(y_i; u, \sigma^2) \Psi^{[i-1]}(\mathrm{d}u)}{m^{[i-1]}(y_i)} \right\}$$

Predictive recursion

Input : Data $y_1, \ldots, y_n$; null model $\mathcal{N}(0, \sigma^2)$; intial guess $\Psi^{[0]} = \tilde{\pi}_1^{[0]}(\theta) + \pi_0^{[0]}\delta_0$ with continuous subdensity $\tilde{\pi}_1^{[0]}(\theta)$ and a Dirac measure at zero of mass $\pi_0^{[0]}$

for $i = 1, \ldots, n$ **do**

$$m_0^{[i]} = \pi_0^{[i-1]} \cdot \mathcal{N}(y_i; 0, \sigma^2)$$

$$f_1^{[i]}(\theta) = \mathcal{N}(y_i; \theta, \sigma^2)\tilde{\pi}_1^{[i-1]}(\theta) \quad \text{(discrete grid)}$$

$$m_1^{[i]} = \int_{\mathbb{R}} f_1^{[i]}(\theta)\mathrm{d}\theta \quad \text{(trapezoid rule)}$$

$$\pi_0^{[i]} = (1 - \gamma^{[i]}) \cdot \pi_0^{[i-1]} + \gamma^{[i]} \cdot \left(\frac{m_0^{[i]}}{m_0^{[i]} + m_1^{[i]}} \right)$$

$$\pi_1^{[i]}(\theta) = (1 - \gamma^{[i]}) \cdot \pi_1^{[i-1]} + \gamma^{[i]} \cdot \left(\frac{f_1^{[i]}(\theta)}{m_0^{[i]} + m_1^{[i]}} \right)$$

end

Output: Estimates $p = 1 - \pi_0^{[n]}$ and $\pi_1(\theta) = \tilde{\pi}_1^{[n]}(\theta)/p$

References I

Yuval Benjamini, Jonathan Taylor, and Rafael A Irizarry. Selection corrected statistical inference for region detection with high-throughput assays. *bioRxiv*, 2016. doi: 10.1101/082321. URL https://www.biorxiv.org/content/early/2016/10/23/082321.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298, 03 2010. doi: 10.1214/09-AOAS285. URL https://doi.org/10.1214/09-AOAS285.

Jrome A. Dupuis and Christian P. Robert. Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111(1):77 – 94, 2003. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(02)00286-0. URL http://www.sciencedirect.com/science/article/pii/S0378375802002860. Special issue I: Model Selection, Model Diagnostics, Empirical Ba yes and Hierarchical Bayes.

W. Fithian, D. Sun, and J. Taylor. Optimal Inference After Model Selection. *ArXiv e-prints*, October 2014.

Constantinos Goutis and Chistian P. Robert. Model choice in generalised linear models: A bayesian approach via kullback-leibler projections. *Biometrika*, 85(1):29–37, 1998. doi: 10.1093/biomet/85.1.29. URL http://dx.doi.org/10.1093/biomet/85.1.29.

References II

P. Richard Hahn and Carlos M. Carvalho. Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015. doi: 10.1080/01621459.2014.993077. URL https://doi.org/10.1080/01621459.2014.993077.

Ryan C. Kelly, Matthew A. Smith, Robert E. Kass, and Tai Sing Lee. Local field potentials indicate network state and account for neuronal response variability. *Journal of Computational Neuroscience*, 29(3):567–579, Dec 2010. ISSN 1573-6873. doi: 10.1007/s10827-009-0208-9. URL https://doi.org/10.1007/s10827-009-0208-9.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 06 2016. doi: 10.1214/15-AOS1371. URL https://doi.org/10.1214/15-AOS1371.

A. Newton. On a nonparametric recursive estimator of the mixing distribution. *Sankhyā Ser. A*, pages 306–322, 2002.

J. Piironen, M. Paasiniemi, and A. Vehtari. Projective Inference in High-dimensional Problems: Prediction and Feature Selection. *ArXiv e-prints*, October 2018.

John W. Pratt. Shorter confidence intervals for the mean of a normal distribution with known variance. *Ann. Math. Statist.*, 34(2):574–586, 06 1963. doi: 10.1214/aoms/1177704170. URL https://doi.org/10.1214/aoms/1177704170.

References III

Stephen Reid, Jonathan Taylor, and Robert Tibshirani. Post-selection point and interval estimation of signal sizes in gaussian samples. *Canadian Journal of Statistics*, 45, 04 2014.

J. G. Scott, R. C. Kelly, M. A. Smith, P. Zhou, and R. E. Kass. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *ArXiv e-prints*, July 2013.

Matthew A. Smith and Adam Kohn. Spatial and temporal scales of neuronal correlation in primary visual cortex. *Journal of Neuroscience*, 28(48):12591–12603, 2008. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2929-08.2008. URL http://www.jneurosci.org/content/28/48/12591.

S. Woody and J. G. Scott. Optimal post-selection inference for sparse signals: a nonparametric empirical-Bayes approach. *ArXiv e-prints*, October 2018.

Daniel Yekutieli. Adjusted Bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):515–541, 2012. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2011.01016.x. URL http://dx.doi.org/10.1111/j.1467-9868.2011.01016.x.

C Yu and P D Hoff. Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, 105(2):319–335, 2018. doi: 10.1093/biomet/asy009. URL http://dx.doi.org/10.1093/biomet/asy009.