Feature Selection using e-values

SUBHABRATA ‘SUBHO’ MAJUMDAR (SPLUNK)

SNIGDHANSU ‘ANSU’ CHATTERJEE (UNIVERSITY OF MINNESOTA)

ICML International Conference On Machine Learning

Motivation

Feature selection is a fundamental problem in statistics and machine learning. There are two ways to do this: sparse penalized regression and best subset selection.

Motivation

Feature selection is a fundamental problem in statistics and machine learning. There are two ways to do this: sparse penalized regression and best subset selection.

Sparse methods have inferential and algorithmic issues. For example, Lasso estimates are biased and affected by feature correlations.

Motivation

Feature selection is a fundamental problem in statistics and machine learning. There are two ways to do this: sparse penalized regression and best subset selection.

Sparse methods have inferential and algorithmic issues. For example, Lasso estimates are biased and affected by feature correlations.

Best subset selection requires sifting through a model space that exponentially increases in size with model parameters. They are poorly explored for dependent or structured data models, such as mixed effect models.

The e-values method

We do best subset selection by fitting just the full model---the model with all p input features---and computing the model selection criterion at $p+1$ models.

The e-values method

We do best subset selection by fitting just the full model---the model with all p input features---and computing the model selection criterion at p+1 models.

Steps

1. Fit the full model and compute its e-value.

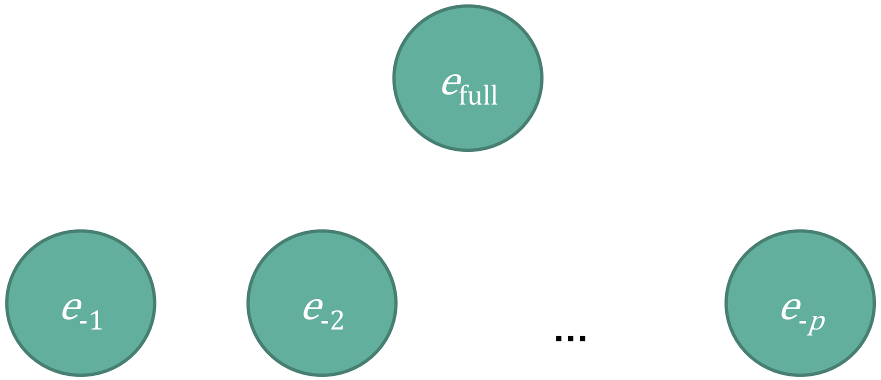
The e-values method

We do best subset selection by fitting just the full model---the model with all p input features---and computing the model selection criterion at $p+1$ models.

Steps

1. Fit the full model and compute its e-value.

2. Drop an input feature, compute the e-value of that dropped-feature model.



The e-values method

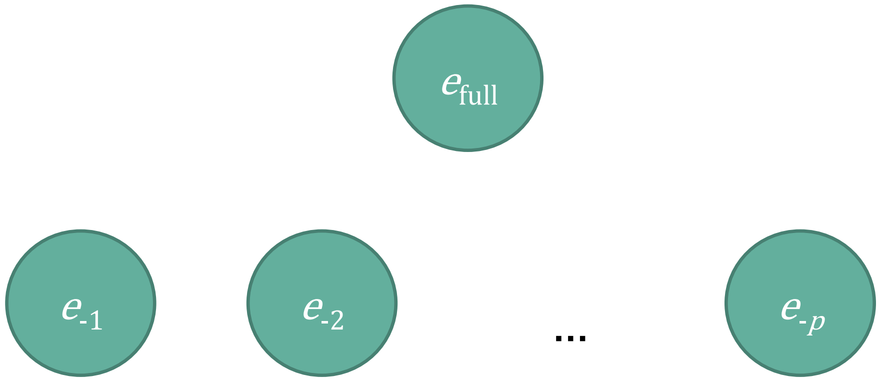
We do best subset selection by fitting just the full model---the model with all p input features---and computing the model selection criterion at p+1 models.

Steps

1. Fit the full model and compute its e-value.

2. Drop an input feature, compute the e-value of that dropped-feature model.

3. Collect input features dropping which causes the e-value to go down.



$$S = \{j: e_{-j} < e_{\text{full}}; 1 \leq j \leq p\}$$

Definition

Data depth functions $D(x, \mathbb{F})$ quantify the inlyingness of a point x in multivariate space with respect to a probability distribution $\mathbb{F}$.

Sampling distribution $(\mathbb{F}_{\mathcal{M}})$ of model $\mathcal{M}$ is the distribution of the model parameter estimate $\hat{\theta}_{\mathcal{M}}$, based on the random data samples the estimate is calculated from.

The **e-value** of model $\mathcal{M}$ is the mean data depth of its sampling distribution with respect to its full model sampling distribution:

$$e(M) = \mathbb{E}_{\hat{\theta}_{\mathcal{M}} \sim \mathbb{F}_{\mathcal{M}}} D\left(\hat{\theta}_{\mathcal{M}}, \mathbb{F}_{\text{full}}\right).$$

Only need to compute $\hat{\theta}_{\text{full}}$. For the j^{th} dropped-feature model, just make $\hat{\theta}_{\text{full},j} = 0$.

Generalized Bootstrap

Sampling distributions are approximated using **Generalized Bootstrap (GBS)**.

Parameter estimate	Expression
Original	$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \psi_i(\theta, Z_i)$
GBS version	$\hat{\theta}_B \approx \hat{\theta} - \frac{\tau_n}{\sqrt{n}} \left[\sum_{i=1}^{n} \psi_i''(\hat{\theta}, Z_i) \right]^{-1} \sum_{i=1}^{n} W_i \psi_i'(\hat{\theta}, Z_i)$

[1]S. Chatterjee and A. Bose, *The Annals of Statistics*, 33(1): 414-436, 2005

Generalized Bootstrap

Sampling distributions are approximated using **Generalized Bootstrap (GBS)**.

Parameter estimate	Expression
Original	$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \psi_i(\theta, Z_i)$
GBS version	$\hat{\theta}_B \approx \hat{\theta} - \frac{\tau_n}{\sqrt{n}} \left[\sum_{i=1}^{n} \psi_i''(\hat{\theta}, Z_i) \right]^{-1} \sum_{i=1}^{n} W_i \psi_i'(\hat{\theta}, Z_i)$

- $\psi_i(\theta, Z_i)$: Energy functions
- $i=1$: Samples may not be independent
- τ_n: Tuning parameter, optimized using a BIC-like criterion
- $\sum_{i=1}^{n} \psi_i''(\hat{\theta}, Z_i)$: Hessian
- W_i: i.i.d. weights with mean 0, variance 1
- $\psi_i'(\hat{\theta}, Z_i)$: Gradient

[1] S. Chatterjee and A. Bose, *The Annals of Statistics*, 33(1): 414-436, 2005

Generalized Bootstrap

Sampling distributions are approximated using **Generalized Bootstrap (GBS)**.

Parameter estimate	Expression
Original	$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \psi_i(\theta, Z_i)$
GBS version	$\hat{\theta}_B \approx \hat{\theta} - \frac{\tau_n}{\sqrt{n}} \left[\sum_{i=1}^{n} \psi_i''(\hat{\theta}, Z_i)\right]^{-1} \sum_{i=1}^{n} W_i \psi_i'(\hat{\theta}, Z_i)$

Energy functions: $\psi_i(\theta, Z_i)$

Samples may not be independent: $i=1$

Tuning parameter, optimized using a BIC-like criterion: τ_n

Hessian: $\sum_{i=1}^{n} \psi_i''(\hat{\theta}, Z_i)$

i.i.d. weights with mean 0, variance 1: W_i

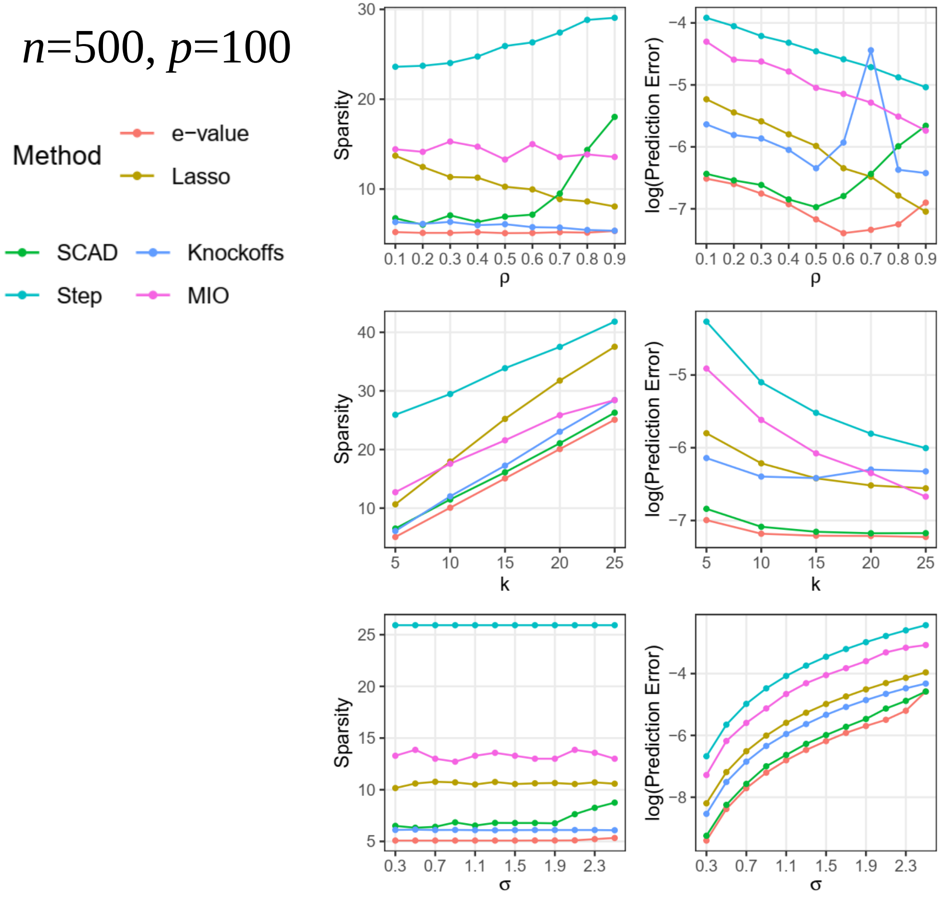
Gradient: $\psi_i'(\hat{\theta}, Z_i)$

FAST!
Only requires Monte-Carlo sampling of weights

[1] S. Chatterjee and A. Bose, *The Annals of Statistics*, 33(1): 414-436, 2005

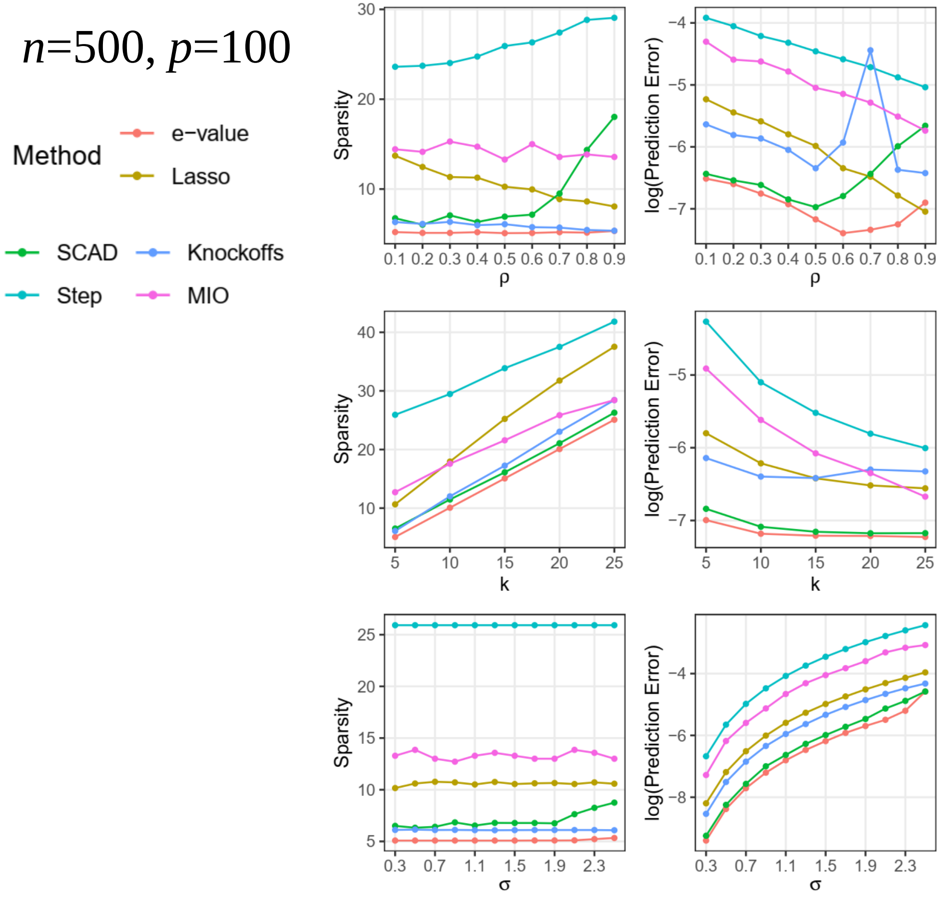
Experiments: linear model

n=500, p=100

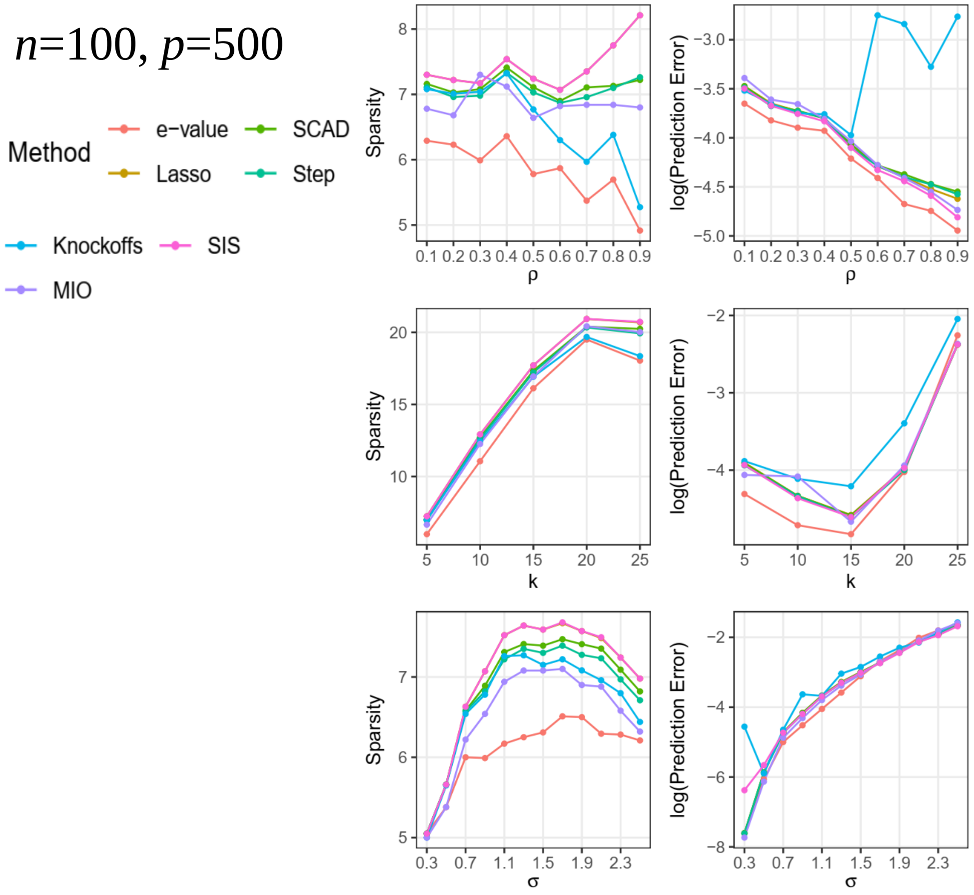


Experiments: linear model

n=500, p=100



n=100, p=500



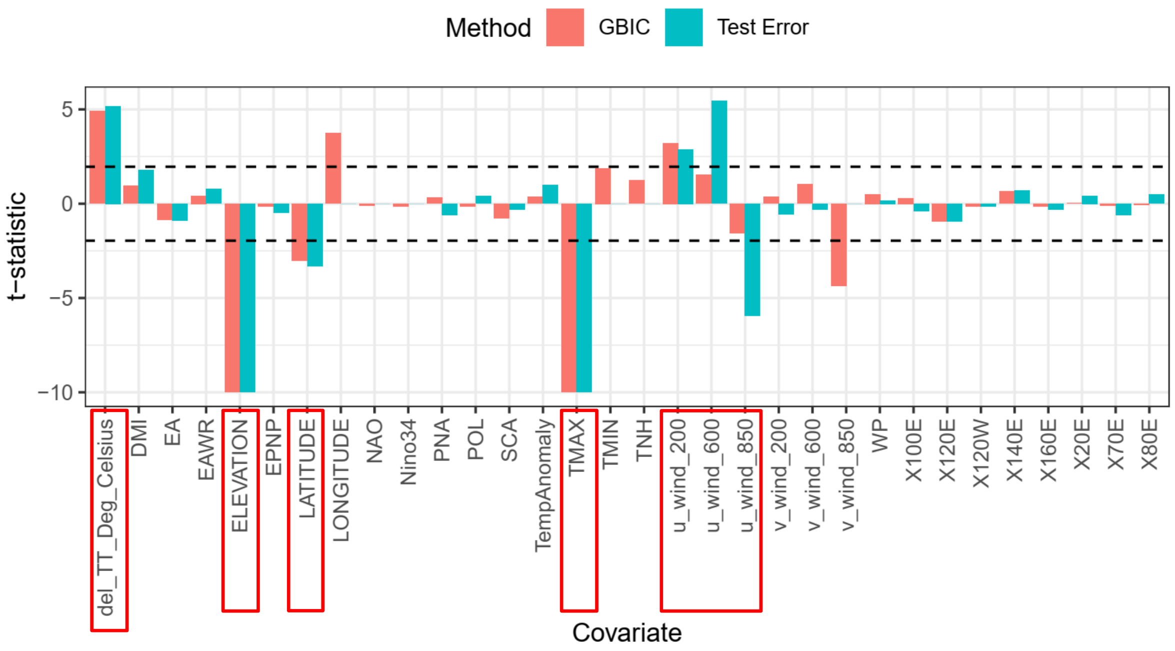
Experiments: linear mixed model

Method		Setting 1: $n_i = 5, m = 30$				Setting 2: $n_i = 10, m = 60$			
		FPR	FNR	Acc	MS	FPR	FNR	Acc	MS
e-value	$\delta = 0$	9.4	0.0	76	2.33	0.0	0.0	100	2.00
	$\delta = 0.01$	6.7	0.0	82	2.22	0.0	0.0	100	2.00
	$\delta = 0.05$	1.0	0.0	97	2.03	0.0	0.0	100	2.00
	$\delta = 0.1$	0.3	0.0	99	2.01	0.0	0.0	100	2.00
	$\delta = 0.15$	0.0	0.0	100	2.00	0.0	0.0	100	2.00
SCAD (Peng & Lu, 2012)	BIC	21.5	9.9	49	2.26	1.5	1.9	86	2.10
	AIC	17	11.0	46	2.43	1.5	3.3	77	2.20
	GCV	20.5	6	49	2.30	1.5	3	79	2.18
	$\sqrt{\log n / n}$	21	15.6	33	2.67	1.5	4.1	72	2.26
M-ALASSO (Bondell et al., 2010)		-	-	73	-	-	-	83	-
SCAD-P (Fan & Li, 2012)		-	-	90	-	-	-	100	-
rPQL (Hui et al., 2017)		-	-	98	-	-	-	99	-

Table 6.2. Performance comparison for mixed effect models. We compare e-values with a number of sparse penalized methods: (a) Peng & Lu (2012) that uses SCAD penalty and different methods of selecting regularization tuning parameters, (b) The adaptive lasso-based method of Bondell et al. (2010), (c) The SCAD-P method Fan & Li (2012), and (d) regularized Penalized Quasi-Likelihood Hui et al. (2017, rPQL). For comparison with Peng & Lu (2012), we present mean false positive (FPR) and false negative (FNR) rates, Accuracy (Acc), and Model Size (MS), i.e. the number of non-zero fixed effects estimated. To compare with other methods we only use Acc, since they did not report the rest of the metrics.

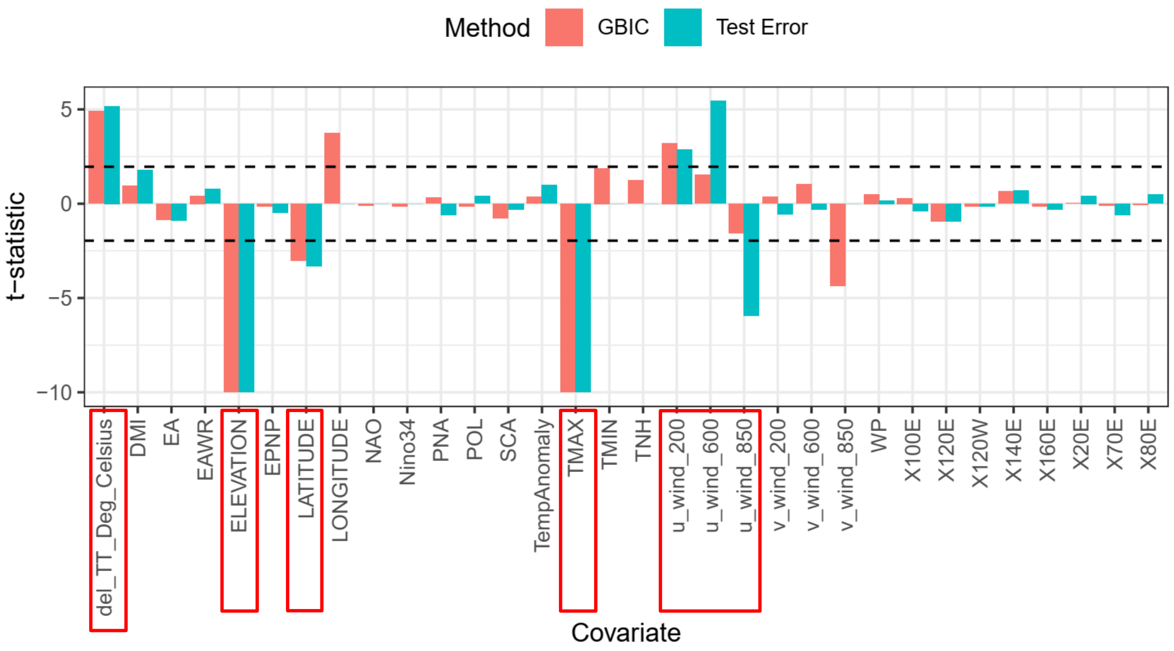
Real data experiments

Indian monsoon: our method isolates known factors instrumental behind amount of rainfall.

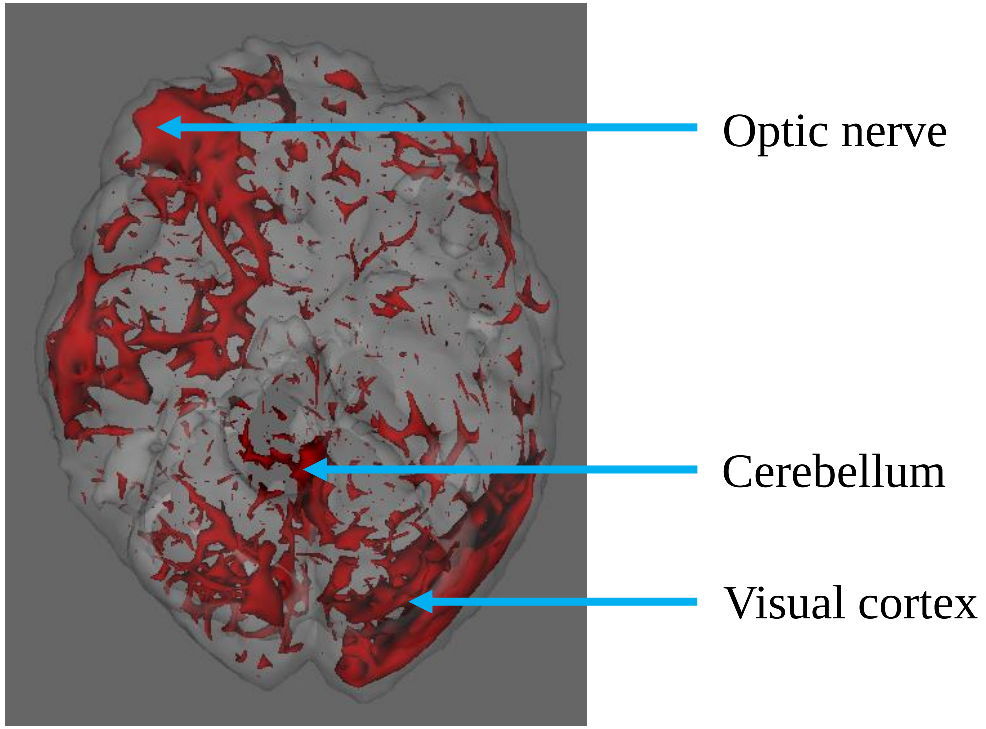


Real data experiments

Indian monsoon: our method isolates known factors instrumental behind amount of rainfall.



fMRI: our method detects activity in regions of brain responsible for visual perception.



Thank you!