From Probabilistic Models to Decision Theory and Back Again

Sanmi Koyejo

Stanford & University of Illinois at Urbana Champaign

Background

A Primer on (Bayes) Decision Theory

Cox and Hinkley (1979); Berger (1985)

Key Components:

- **Probabilistic Model:** $p(\boldsymbol{\beta})$ for model parameter $\boldsymbol{\beta}$
- **Decision Variable:** θ
- **Utility:** $\mathcal{U}(\theta, p)$ is score for selecting θ wrt. p

Step 1: Model Construction

e.g. Bayes posterior

$$p(\boldsymbol{\beta}|\mathcal{D}_n) \propto p(\mathcal{D}_n|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

Step 2: Decision Making

$$\theta^* = \arg\max_{\theta \in \Theta} \mathcal{U}(\theta, p)$$

A Primer on (Bayes) Decision Theory

Cox and Hinkley (1979); Berger (1985)

Key Components:

- **Probabilistic Model:** $p(\boldsymbol{\beta})$ for model parameter $\boldsymbol{\beta}$
- **Decision Variable:** θ
- **Utility:** $\mathcal{U}(\theta, p)$ is score for selecting θ wrt. p

Step 1: Model Construction

e.g. Bayes posterior

$$p(\boldsymbol{\beta}|\mathcal{D}_n) \propto p(\mathcal{D}_n|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

Step 2: Decision Making

$$\theta^* = \underset{\theta \in \Theta}{\arg\max}\, \mathcal{U}(\theta, p)$$

A Primer on (Bayes) Decision Theory

Cox and Hinkley (1979); Berger (1985)

Key Components:

- **Probabilistic Model:** $p(\boldsymbol{\beta})$ for model parameter $\boldsymbol{\beta}$
- **Decision Variable:** θ
- **Utility:** $\mathcal{U}(\theta, p)$ is score for selecting θ wrt. p

Step 1: Model Construction

e.g. Bayes posterior

$$p(\boldsymbol{\beta}|\mathcal{D}_n) \propto p(\mathcal{D}_n|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

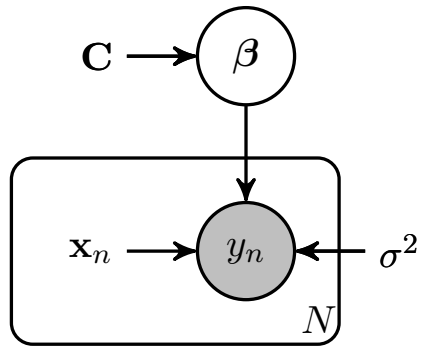
Step 2: Decision Making

$$\theta^* = \underset{\theta \in \Theta}{\arg\max}\, \mathcal{U}(\theta, p)$$

Example: Selecting a Point Estimate

- **Model:**

$$y_i|_{\boldsymbol{\beta},\mathbf{x}_i,\epsilon} = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon$$
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right), \ \boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{C}\right)$$



- **Utility:**

$$\mathcal{U}(\mathbf{w}, p) = -\mathrm{E}_{\boldsymbol{\beta}\sim p}\left[\, \|\mathbf{w} - \boldsymbol{\beta}\|_2^2 \,\right]$$

Step 1: Posterior Estimation

$$p(\boldsymbol{\beta}|\mathcal{D}_n) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

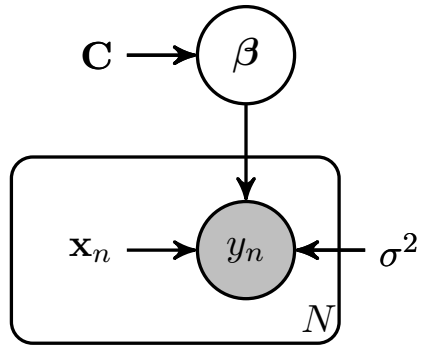
Step 2: Decision Making

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathcal{U}(\mathbf{w}, p(\boldsymbol{\beta}|\mathcal{D}_n))$$
$$= \mathrm{E}_{p(\boldsymbol{\beta}|\mathcal{D}_n)}\left[\, \boldsymbol{\beta} \,\right] = \boldsymbol{\mu}$$

Example: Selecting a Point Estimate

- **Model:**

$$y_i|_{\boldsymbol{\beta},\mathbf{x}_i,\epsilon} = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon$$
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right), \ \boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{C}\right)$$



- **Utility:**

$$\mathcal{U}(\mathbf{w}, p) = -\mathrm{E}_{\boldsymbol{\beta}\sim p}\left[\, \|\mathbf{w} - \boldsymbol{\beta}\|_2^2 \,\right]$$

Step 1: Posterior Estimation

$$p(\boldsymbol{\beta}|\mathcal{D}_n) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

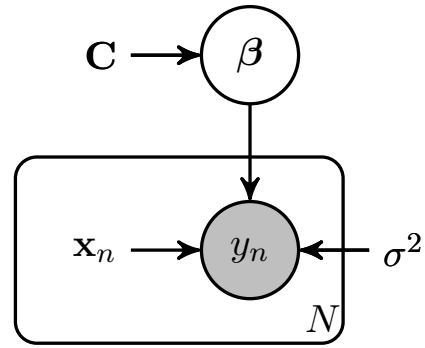
Step 2: Decision Making

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathcal{U}(\mathbf{w}, p(\boldsymbol{\beta}|\mathcal{D}_n))$$
$$= \mathrm{E}_{p(\boldsymbol{\beta}|\mathcal{D}_n)}\left[\, \boldsymbol{\beta} \,\right] = \boldsymbol{\mu}$$

Example: Selecting a Point Estimate

- **Model:**

$$y_i|_{\boldsymbol{\beta},\mathbf{x}_i,\epsilon} = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon$$
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right), \ \boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{C}\right)$$



- **Utility:**

$$\mathcal{U}(\mathbf{w}, p) = -\mathrm{E}_{\boldsymbol{\beta}\sim p}\left[\, \|\mathbf{w} - \boldsymbol{\beta}\|_2^2 \,\right]$$

Step 1: Posterior Estimation

$$p(\boldsymbol{\beta}|\mathcal{D}_n) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

Step 2: Decision Making

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathcal{U}(\mathbf{w}, p(\boldsymbol{\beta}|\mathcal{D}_n))$$
$$= \mathrm{E}_{p(\boldsymbol{\beta}|\mathcal{D}_n)}\left[\, \boldsymbol{\beta} \,\right] = \boldsymbol{\mu}$$

Outline

Part 1:

$$\theta^* = \arg\max_{\theta \in \Theta} \mathcal{U}(\theta, p)$$

- multilabel prediction with non-decomposable metrics
- application to decoding cognitive processes

Part 2:

$$\text{“}p^* = \arg\max_{p \in \mathcal{P}} \mathcal{U}(\tilde{\theta}, p)\text{”}$$

- incorporating user-defined utility as prior information
- application to structured probabilistic models

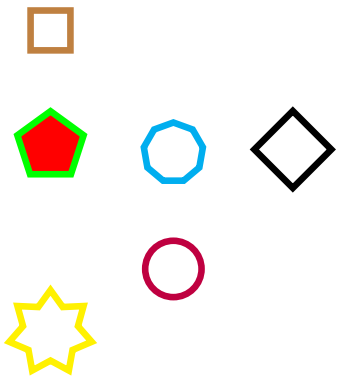
Part I:

Bayes Optimal Multilabel Classification

Joint work with:

Nagarajan Natarajan, Pradeep Ravikumar, Inderjit Dhillon, Russell A. Poldrack

Multilabel Classification

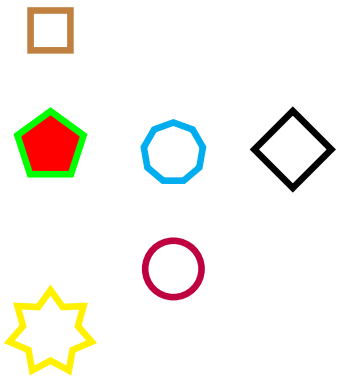


- **Multiclass**: each example has one class label
- **Multilabel**: each example has multiple class labels

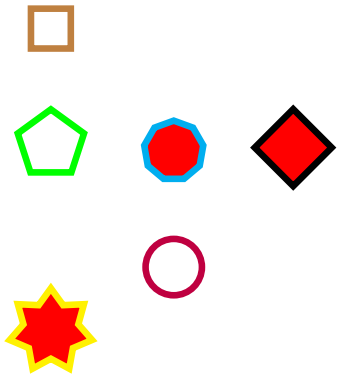
Problem Description

- Inputs: $X \in \mathcal{X}$, Labels: $Y \in [0,1]^M$ (with M labels)
- Output: classifier $\boldsymbol{\theta} : \mathcal{X} \mapsto [0,1]^M$ that maximizes utility $\mathcal{U}(\boldsymbol{\theta}, p)$

Multilabel Classification



- **Multiclass**: each example has one class label

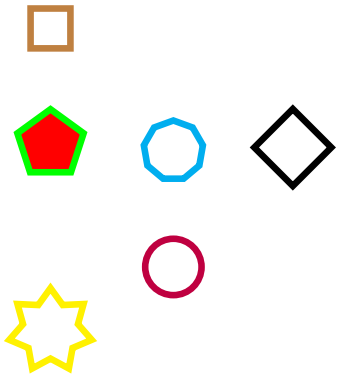


- **Multilabel**: each example has multiple class labels

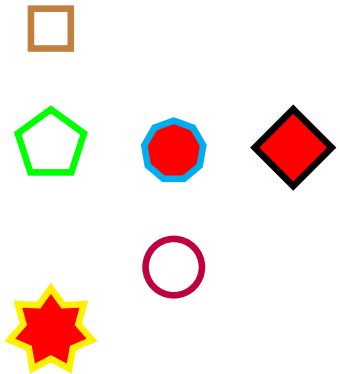
Problem Description

- Inputs: $X \in \mathcal{X}$, Labels: $Y \in [0,1]^M$ (with M labels)
- Output: classifier $\boldsymbol{\theta} : \mathcal{X} \mapsto [0,1]^M$ that maximizes utility $\mathcal{U}(\boldsymbol{\theta}, p)$

Multilabel Classification



- **Multiclass**: each example has one class label



- **Multilabel**: each example has multiple class labels

Problem Description

- Inputs: $X \in \mathcal{X}$, Labels: $Y \in [0,1]^M$ (with M labels)
- Output: classifier $\boldsymbol{\theta} : \mathcal{X} \mapsto [0,1]^M$ that maximizes utility $\mathcal{U}(\boldsymbol{\theta}, p)$

Example: Hamming Loss / Label Accuracy

- Measures misclassification error for each label (separately)

Utility Function

$$\mathcal{U}(\boldsymbol{\theta}) = \mathrm{E}_{X,Y\sim\mathbb{P}}\left[\sum_{m=1}^{M} 1_{[Y_m=\theta_m(X)]}\right] = \sum_{m=1}^{M} \mathbb{P}(Y_m = \theta_m(X))$$

Optimal Prediction

$$\theta_m^*(x) = \text{sign}\left(\mathbb{P}(Y_m = 1|x) - \frac{1}{2}\right) \;\forall\, x \in \mathcal{X}$$

Well known convex surrogates e.g. logistic loss, hinge loss (Bartlett et al., 2006)

Example: Hamming Loss / Label Accuracy

- Measures misclassification error for each label (separately)

Utility Function

$$\mathcal{U}(\boldsymbol{\theta}) = \mathrm{E}_{X,Y\sim\mathbb{P}}\left[\sum_{m=1}^{M} 1_{[Y_m=\theta_m(X)]}\right] = \sum_{m=1}^{M} \mathbb{P}(Y_m = \theta_m(X))$$

Optimal Prediction

$$\theta_m^*(x) = \mathsf{sign}\left(\mathbb{P}(Y_m = 1|x) - \frac{1}{2}\right) \;\forall\, x \in \mathcal{X}$$

Well known convex surrogates e.g. logistic loss, hinge loss (Bartlett et al., 2006)

Why does this matter?

$$\text{“}\theta^*_m(x) = \text{sign}\left(\mathbb{P}(Y_m = 1|x) - \frac{1}{2}\right)\text{”}$$

- defines the ideal point estimate (i.e. model summary) for a probabilistic model
- specifies the *target* for estimation algorithms i.e. what quantities must be estimated correctly?

When is θ^* difficult to estimate?

- optimal can be quite complicated function of p when the utility function is non-decomposable (cf. (Narasimhan et al., 2014)):

$$\mathcal{U}(\{Y_i, X_i\}) \neq \sum_i \mathcal{U}(Y_i, X_i)$$

This talk:

show that the optimal solution is indeed *simple* for a large family of multilabel metrics

When is θ^* difficult to estimate?

- optimal can be quite complicated function of p when the utility function is non-decomposable (cf. (Narasimhan et al., 2014)):

$$\mathcal{U}(\{Y_i, X_i\}) \neq \sum_i \mathcal{U}(Y_i, X_i)$$

This talk:

show that the optimal solution is indeed *simple* for a large family of multilabel metrics

Recall the Confusion Matrix for Binary Classification

	$Y=1$	$Y=0$
$\theta=1$	TP P($Y=1, \theta=1$)	FP P($Y=0, \theta=1$)
$\theta=0$	FN P($Y=1, \theta=0$)	TN P($Y=0, \theta=0$)

Binary classification metrics defined by confusion matrix

$$\mathcal{U}(\boldsymbol{\theta}) = \Psi\left(\begin{bmatrix} \mathsf{TP} & \mathsf{FP} \\ \mathsf{FN} & \mathsf{TN} \end{bmatrix}\right)$$

e.g. $\mathsf{ACC} = \mathsf{TP} + \mathsf{TN}, \quad \mathsf{WA} = w_1\mathsf{TP} + w_2\mathsf{TN}$

$$F_\beta = \frac{(1+\beta^2)\mathsf{TP}}{(1+\beta^2)\mathsf{TP} + \beta^2\mathsf{FN} + \mathsf{FP}}, \quad \mathsf{JAC} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP}}$$

Multilabel Confusion

Similar idea for multilabel classification, now across both labels m and examples n.

$$\widehat{\mathsf{TP}}(\boldsymbol{\theta})_{m,n} = 1_{\left[\theta_m(x^{(n)})=1, y_m^{(n)}=1\right]} \qquad \widehat{\mathsf{TN}}(\boldsymbol{\theta})_{m,n} = 1_{\left[\theta_m(x^{(n)})=0, y_m^{(n)}=0\right]}$$

$$\widehat{\mathsf{FP}}(\boldsymbol{\theta})_{m,n} = 1_{\left[\theta_m(x^{(n)})=1, y_m^{(n)}=0\right]} \qquad \widehat{\mathsf{FN}}(\boldsymbol{\theta})_{m,n} = 1_{\left[\theta_m(x^{(n)})=0, y_m^{(n)}=1\right]}$$

We focus on linear-fractional metrics e.g. Accuracy, F_β, Precision, Recall, Jaccard

Instance-averaged Utilities

Constructing multilabel metrics by averaging binary metrics

Average over labels for each example

$$\widehat{\mathsf{TP}}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathsf{TP}}(\boldsymbol{\theta})_{m,n},\ \widehat{\mathsf{FP}}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathsf{FP}}(\boldsymbol{\theta})_{m,n},$$

$$\Psi_{\mathsf{instance}} := \frac{1}{N}\sum_{n=1}^{N} \Psi(\widehat{\mathsf{TP}}_n, \widehat{\mathsf{FP}}_n, \widehat{\mathsf{TN}}_n, \widehat{\mathsf{FN}}_n).$$

Theorem (Koyejo et al., 2015)

$$\boldsymbol{\theta}^*_m(x) = \mathsf{sign}(\mathbb{P}(Y_m = 1|x) - \delta^*)\ \forall\, m \in [M].$$

- Only require $\mathbb{P}(Y_m = 1|x)$ i.e. label correlations do not affect optimal classification
- Shared threshold across labels

Instance-averaged Utilities

Constructing multilabel metrics by averaging binary metrics

Average over labels for each example

$$\widehat{\mathsf{TP}}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathsf{TP}}(\boldsymbol{\theta})_{m,n},\ \widehat{\mathsf{FP}}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathsf{FP}}(\boldsymbol{\theta})_{m,n},$$

$$\Psi_{\mathsf{instance}} := \frac{1}{N}\sum_{n=1}^{N} \Psi(\widehat{\mathsf{TP}}_n, \widehat{\mathsf{FP}}_n, \widehat{\mathsf{TN}}_n, \widehat{\mathsf{FN}}_n).$$

Theorem (Koyejo et al., 2015)

$$\boldsymbol{\theta}^*_m(x) = \mathsf{sign}(\mathbb{P}(Y_m = 1|x) - \delta^*)\ \forall\ m \in [M].$$

- Only require $\mathbb{P}(Y_m = 1|x)$ i.e. label correlations do not affect optimal classification
- Shared threshold across labels

Instance-averaged Utilities

Constructing multilabel metrics by averaging binary metrics

Average over labels for each example

$$\widehat{\mathsf{TP}}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathsf{TP}}(\boldsymbol{\theta})_{m,n},\ \widehat{\mathsf{FP}}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathsf{FP}}(\boldsymbol{\theta})_{m,n},$$

$$\Psi_{\mathsf{instance}} := \frac{1}{N}\sum_{n=1}^{N} \Psi(\widehat{\mathsf{TP}}_n, \widehat{\mathsf{FP}}_n, \widehat{\mathsf{TN}}_n, \widehat{\mathsf{FN}}_n).$$

Theorem (Koyejo et al., 2015)

$$\boldsymbol{\theta}^*_m(x) = \mathsf{sign}(\mathbb{P}(Y_m = 1|x) - \delta^*)\ \forall\ m \in [M].$$

- Only require $\mathbb{P}(Y_m = 1|x)$ i.e. label correlations do not affect optimal classification
- Shared threshold across labels

Instance-averaged Utilities

Constructing multilabel metrics by averaging binary metrics

Average over labels for each example

$$\widehat{\mathsf{TP}}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathsf{TP}}(\boldsymbol{\theta})_{m,n},\ \widehat{\mathsf{FP}}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathsf{FP}}(\boldsymbol{\theta})_{m,n},$$

$$\Psi_{\mathsf{instance}} := \frac{1}{N}\sum_{n=1}^{N} \Psi(\widehat{\mathsf{TP}}_n, \widehat{\mathsf{FP}}_n, \widehat{\mathsf{TN}}_n, \widehat{\mathsf{FN}}_n).$$

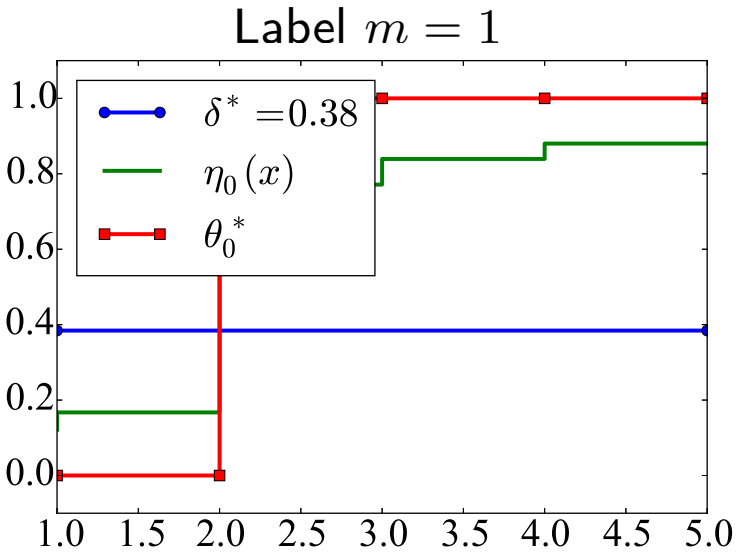
Theorem (Koyejo et al., 2015)

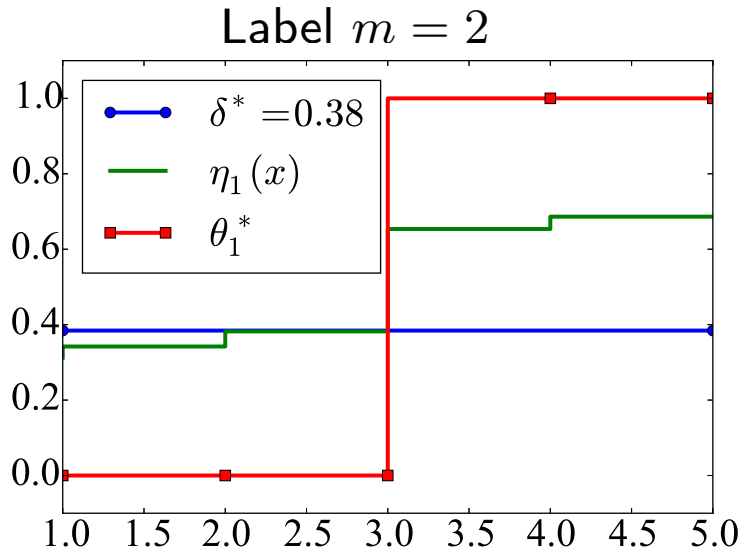
$$\boldsymbol{\theta}^*_m(x) = \mathsf{sign}(\mathbb{P}(Y_m = 1|x) - \delta^*)\ \forall\ m \in [M].$$

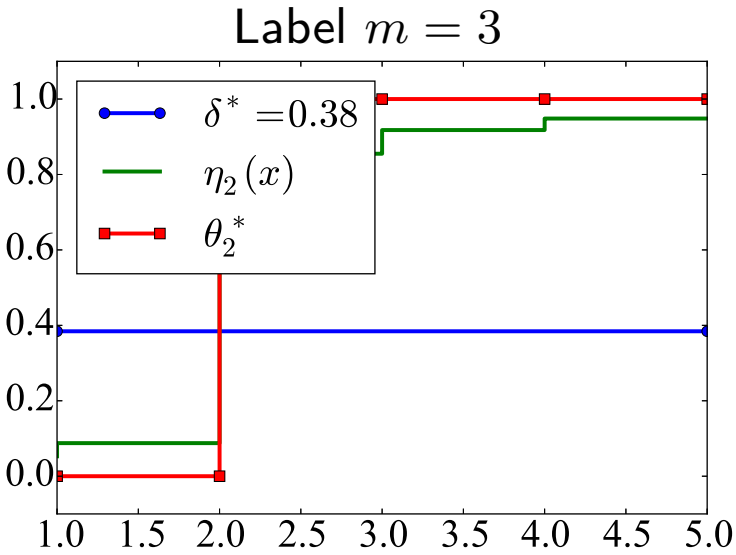
- Only require $\mathbb{P}(Y_m = 1|x)$ i.e. label correlations do not affect optimal classification
- Shared threshold across labels

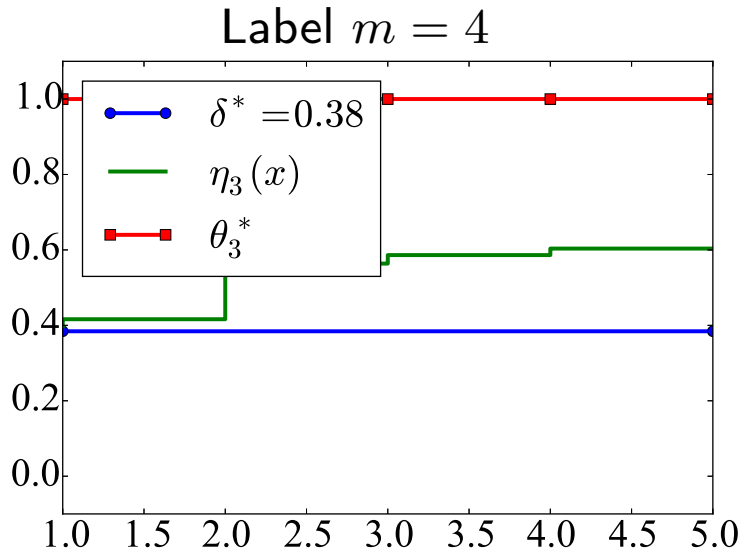
Simulated Data; Instance-averaged F1

Bayes Classifier Computed using Brute Force Search





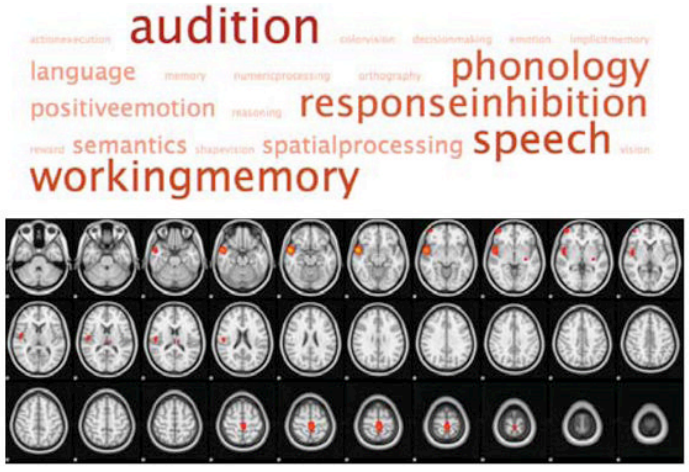


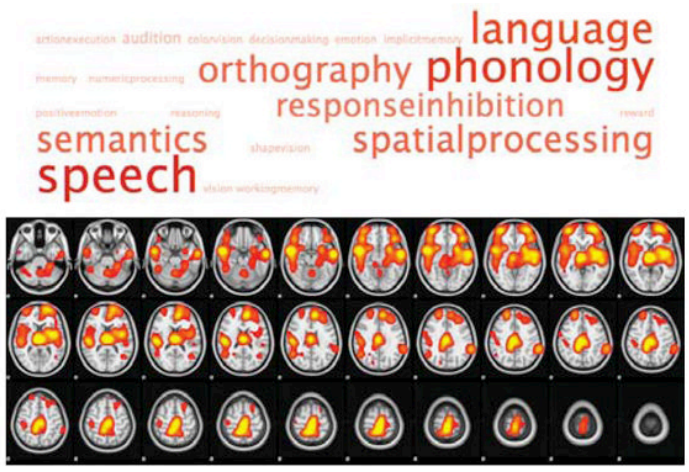


Application to Cognitive Neuroscience

Decoding Cognitive Processes

Cognitive Neuroscience Hypothesis: Brain function is decomposable into elementary cognitive processes. (Posner et al., 1988)

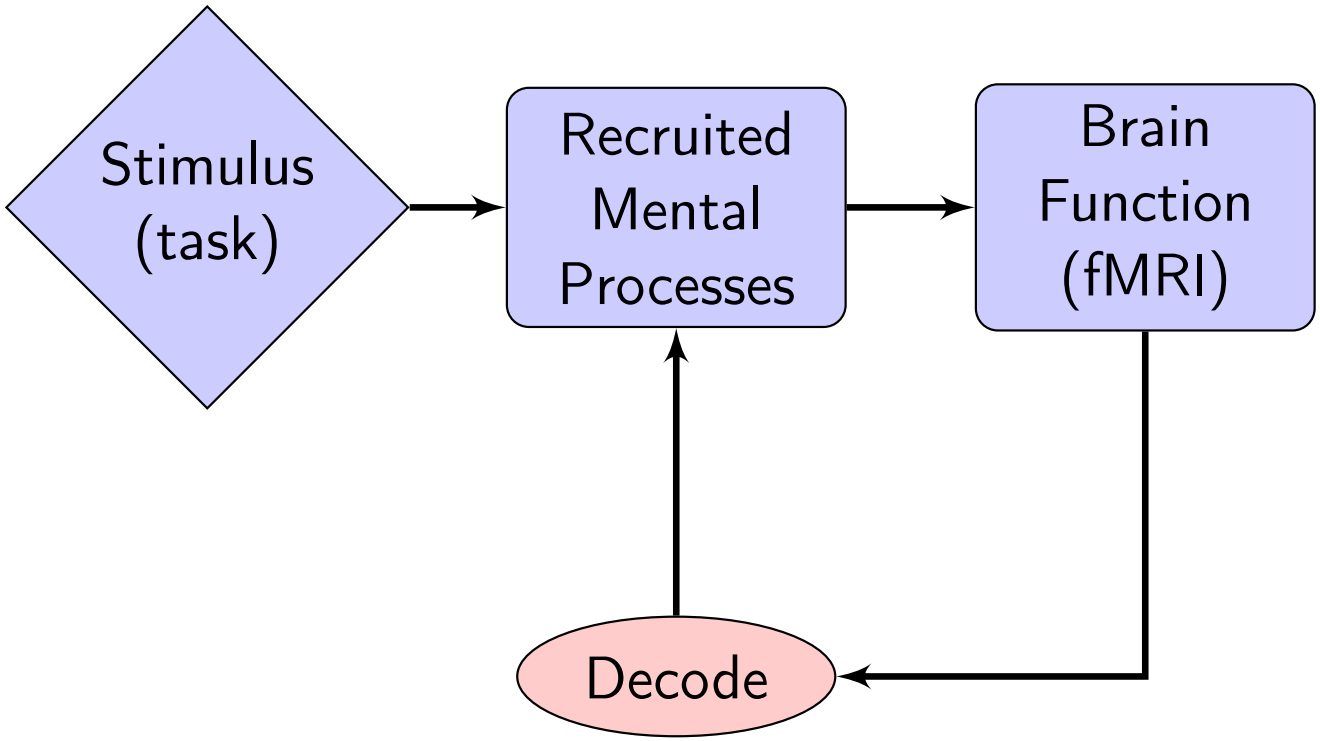




Poldrack et al. (2009)

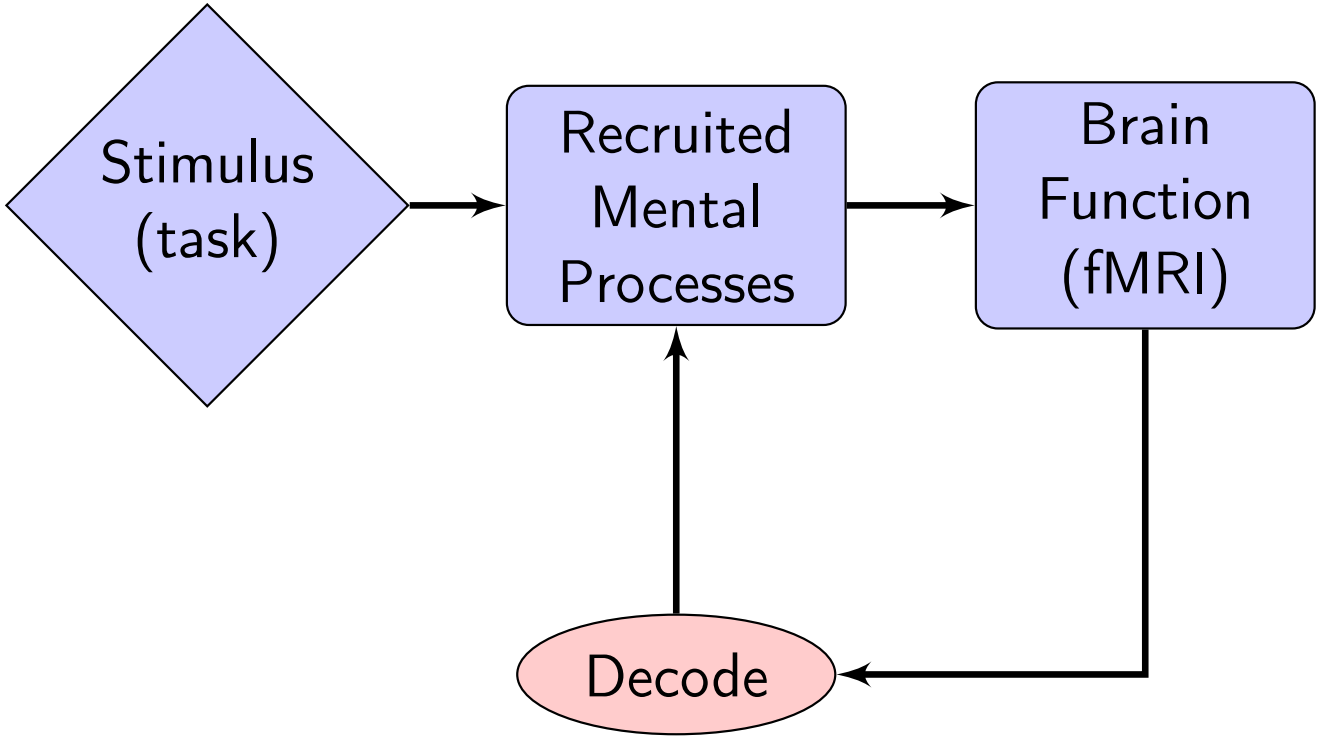
- Cognitive processes represent basis set of brain functions recruited for cognitive tasks (Poldrack, 2010).

Conceptual Block Diagram



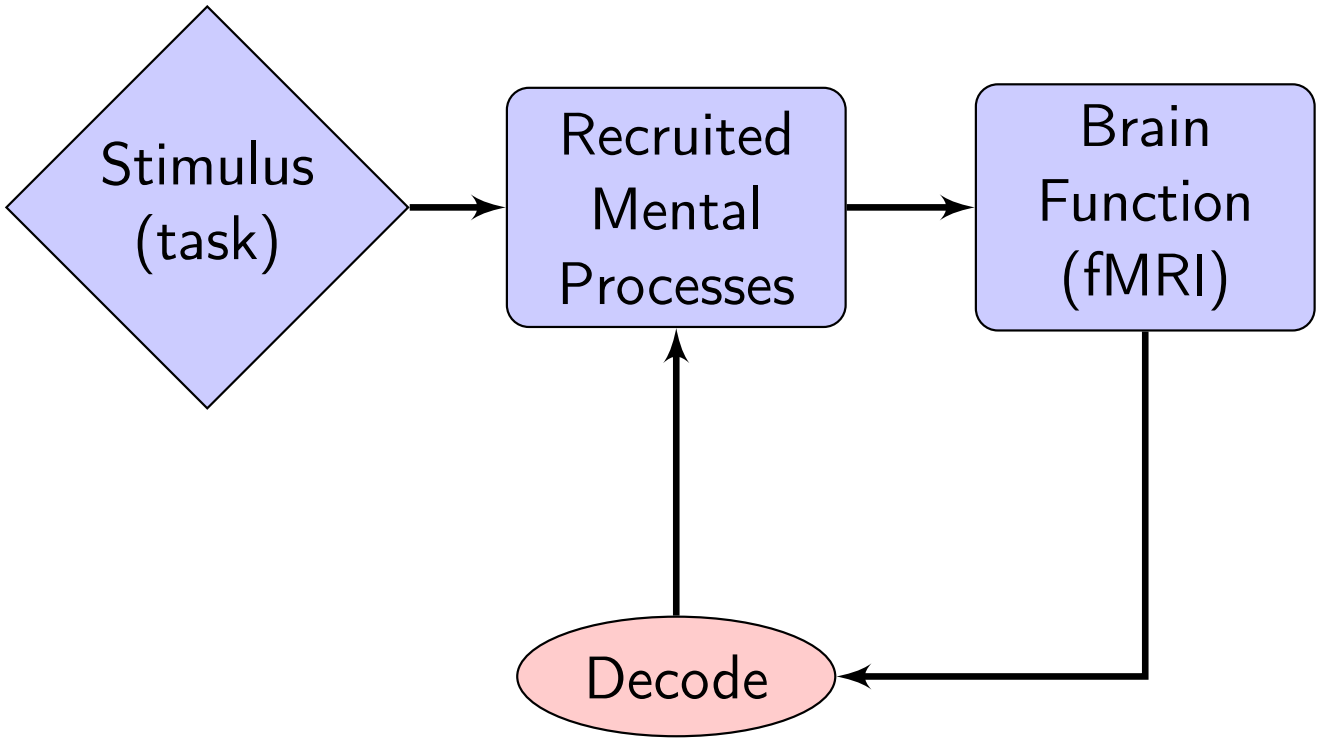
- *Motivating Question:* What are the set of cognitive processes associated with observed brain function?
- This is a multilabel classification problem

Conceptual Block Diagram



- *Motivating Question:* What are the set of cognitive processes associated with observed brain function?
- This is a multilabel classification problem

Conceptual Block Diagram



- *Motivating Question:* What are the set of cognitive processes associated with observed brain function?
- This is a multilabel classification problem

Datasets

- Open fMRI database
- Compiled 479 full brain z-statistic contrast images

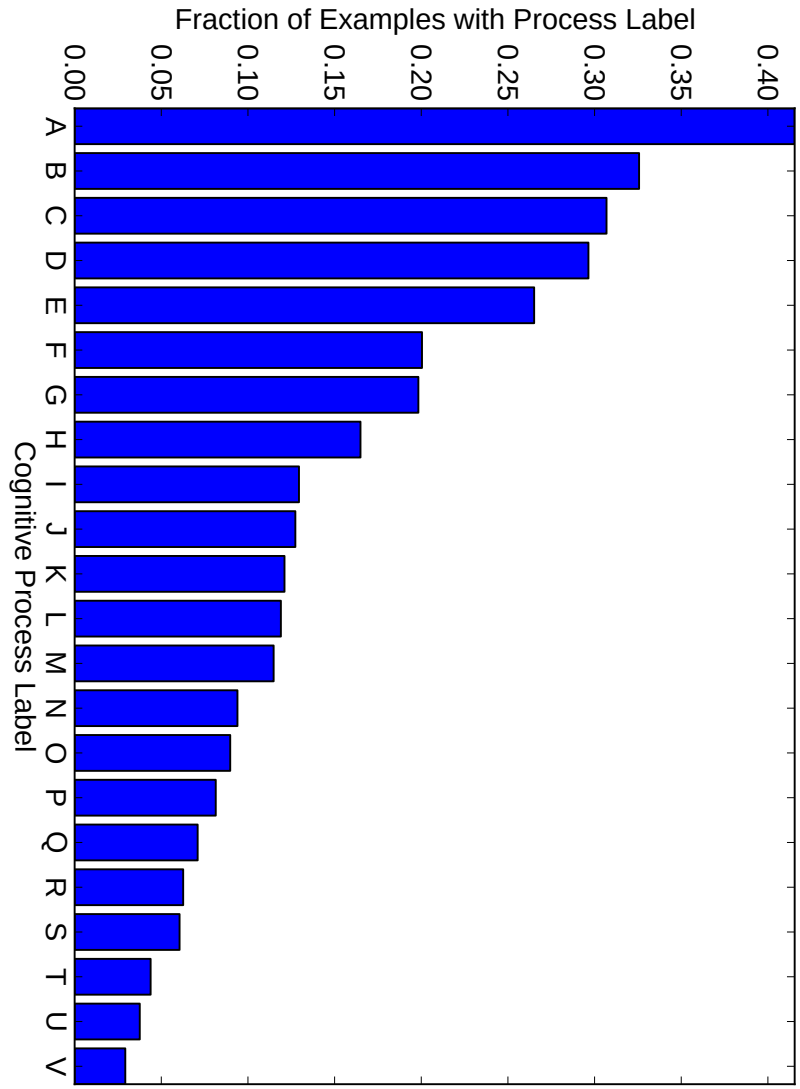
`openfmri.org`

- Open neuroimaging ontology
- Curated 26 cognitive process labels

`www.cognitiveatlas.org`

Label Proportions

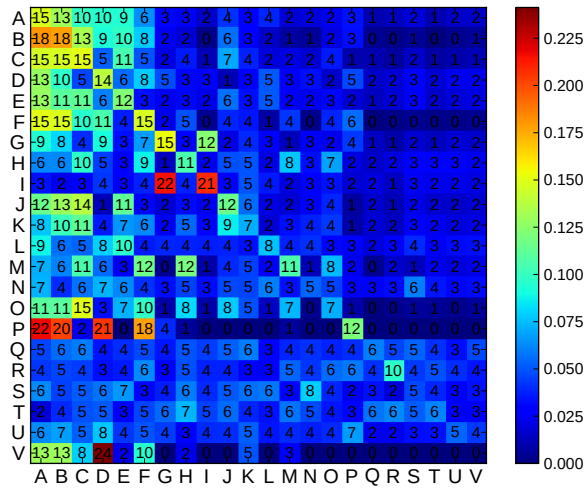
Code	Process Label
A	Vision
B	Action Execution
C	Decision Making
D	Orthography
E	Shape Vision
F	Audition
G	Phonology
H	Conflict
I	Semantics
J	Reinforcement Learning
K	Working Memory
L	Feedback
M	Response Inhibition
N	Reward
O	Stimulus-driven Attention
P	Speech
Q	Emotion Regulation
R	Mentalizing
S	Punishment
T	Error Processing
U	Memory Encoding
V	Spatial Attention



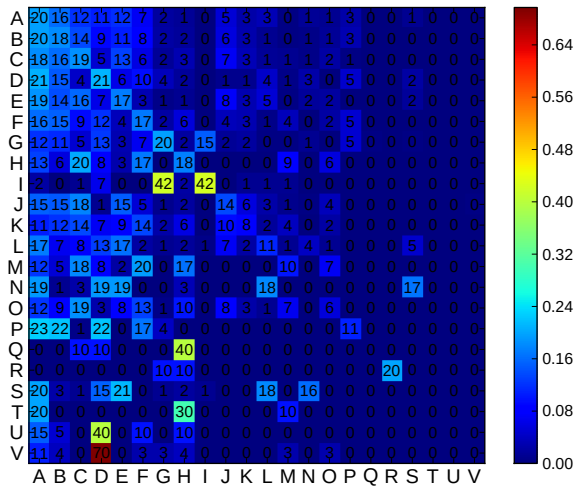
Micro-averaged Decoding of Cognitive Processes

	Accuracy	Precision	F_1	1 - Hamming Loss
SVM	0.43 (0.03)	**0.53 (0.03)**	0.51 (0.03)	0.79 (0.01)
Logistic	**0.44 (0.03)**	**0.53 (0.02)**	**0.52 (0.03)**	0.79 (0.01)
Ridge	0.34 (0.02)	0.47 (0.02)	0.39 (0.02)	**0.91 (0.00)**
Popularity	0.12 (0.01)	0.21 (0.02)	0.18 (0.02)	0.76 (0.01)

* statistically significant via permutation test, $p < 10^{-3}$



Logistic regression classifier, Hamming loss



Ridge regression classifier, Hamming loss

Conclusion: Part I

Optimal multilabel classification:

- Optimal classifiers for a large family of multilabel utility metrics have a simple threshold form: $\text{sign}(\mathbb{P}(Y_m = 1|x) - \delta)$
- Same results apply to other binary-list decision problems e.g. probabilistic clustering

Open questions:

- **Optimal classifiers for other metrics:** analysis does not cover some common metrics e.g. hierarchical losses
- **Label correlations:** exploring new classifiers that can take advantage of label correlations

Optimal multilabel classification:

- Optimal classifiers for a large family of multilabel utility metrics have a simple threshold form: $\text{sign}(\mathbb{P}(Y_m = 1|x) - \delta)$
- Same results apply to other binary-list decision problems e.g. probabilistic clustering

Open questions:

- **Optimal classifiers for other metrics:** analysis does not cover some common metrics e.g. hierarchical losses
- **Label correlations:** exploring new classifiers that can take advantage of label correlations

Part II:

Incorporating Utility into Probabilistic Models

$$\text{“}p^* = \arg\max_{q \in \mathcal{P}} \mathcal{U}(\tilde{\theta}, q)\text{“}$$

Joint work with Rajiv Khanna, Joydeep Ghosh and Russell A. Poldrack

- **Step 1:** start with user-defined $\mathcal{U}(\tilde{\theta}, \cdot)$ which summarizes "expert knowledge"
- **Step 2:** find a distribution $p_* \in \arg\max \mathcal{U}(\tilde{\theta}, \cdot)$

Example:

Mean structure (Jaynes, 1957; Jaakkola et al., 1999; Zhu et al., 2009; Ganchev et al., 2010):

$$\mathcal{U}(\tilde{\mathbf{m}}, q) = -\mathrm{E}_{\mathbf{x}\sim q}\left[\,\|\tilde{\mathbf{m}} - g(\mathbf{x})\|_2\,\right]$$

Applications:

- constructing new default priors that incorporate user information
- directly constraining the posterior (equiv. implicit prior) based on additional knowledge.

- **Step 1:** start with user-defined $\mathcal{U}(\tilde{\theta}, \cdot)$ which summarizes "expert knowledge"
- **Step 2:** find a distribution $p_* \in \arg\max \mathcal{U}(\tilde{\theta}, \cdot)$

Example:

Mean structure (Jaynes, 1957; Jaakkola et al., 1999; Zhu et al., 2009; Ganchev et al., 2010):

$$\mathcal{U}(\tilde{\mathbf{m}}, q) = -\mathrm{E}_{\mathbf{x}\sim q}\left[\,\|\tilde{\mathbf{m}} - g(\mathbf{x})\|_2\,\right]$$

Applications:

- constructing new default priors that incorporate user information
- directly constraining the posterior (equiv. implicit prior) based on additional knowledge.

- **Step 1:** start with user-defined $\mathcal{U}(\tilde{\theta}, \cdot)$ which summarizes "expert knowledge"
- **Step 2:** find a distribution $p_* \in \arg\max \mathcal{U}(\tilde{\theta}, \cdot)$

Example:

Mean structure (Jaynes, 1957; Jaakkola et al., 1999; Zhu et al., 2009; Ganchev et al., 2010):

$$\mathcal{U}(\tilde{\mathbf{m}}, q) = -\mathrm{E}_{\mathbf{x}\sim q}\left[\, \|\tilde{\mathbf{m}} - g(\mathbf{x})\|_2 \,\right]$$

Applications:

- constructing new default priors that incorporate user information
- directly constraining the posterior (equiv. implicit prior) based on additional knowledge.

- **Step 1:** start with user-defined $\mathcal{U}(\tilde{\theta}, \cdot)$ which summarizes "expert knowledge"
- **Step 2:** find a distribution $p_* \in \arg\max \mathcal{U}(\tilde{\theta}, \cdot)$

Example:

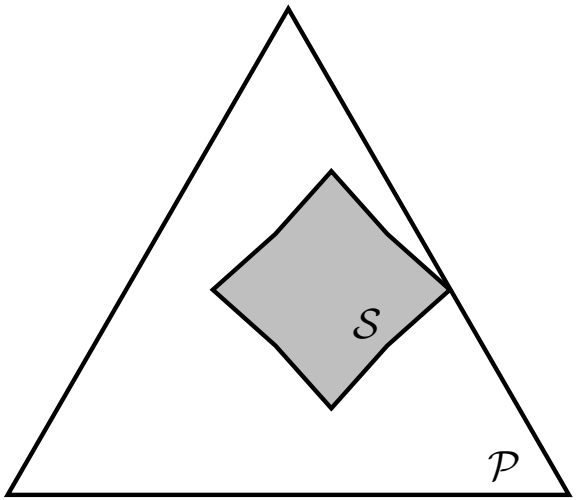
Mean structure (Jaynes, 1957; Jaakkola et al., 1999; Zhu et al., 2009; Ganchev et al., 2010):

$$\mathcal{U}(\tilde{\mathbf{m}}, q) = -\mathrm{E}_{\mathbf{x}\sim q}\left[\, \|\tilde{\mathbf{m}} - g(\mathbf{x})\|_2 \,\right]$$

Applications:

- constructing new default priors that incorporate user information
- directly constraining the posterior (equiv. implicit prior) based on additional knowledge.

Well-Posedness



Problem

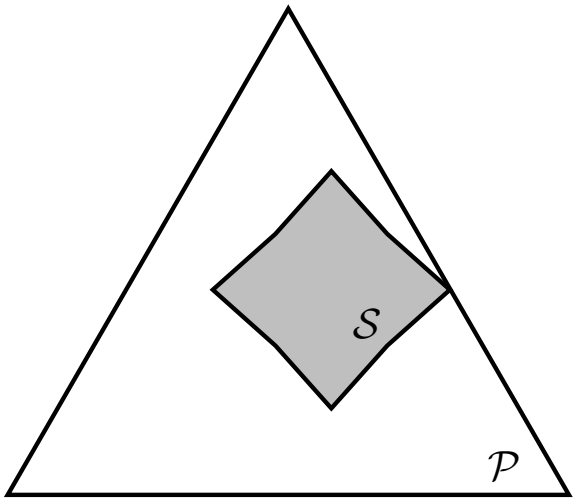
$\mathcal{S} = \{p \in \arg\min\ \mathcal{U}(\tilde{\theta}, \cdot)\}$ is too large!

Proposed Solution

Regularize distribution estimation using relative entropy:

$$q_* = \arg\min_{q \in \mathcal{P}}\ -\mathcal{U}(\tilde{\theta}, q) + \lambda \mathrm{KL}(q \| p)$$

Well-Posedness



Problem

$\mathcal{S} = \left\{p \in \arg\min\ \mathcal{U}(\tilde{\theta}, \cdot)\right\}$ is too large!

Proposed Solution

Regularize distribution estimation using relative entropy:

$$q_* = \arg\min_{q \in \mathcal{P}}\ -\mathcal{U}(\tilde{\theta}, q) + \lambda \mathrm{KL}(q \| p)$$

Background: Relative Entropy

KL Divergence (Kullback, 1959)

$$\mathrm{KL}(p\|q) = \mathrm{E}_p\left[\log \frac{p}{q}\right]$$

- change in *information* between p and q (Jaynes, 1957)
- "natural" distance on probability manifold (Csiszar, 1975)
- the basis for many approximate posterior estimation methods
- $\mathrm{KL}(p\|q) \geq 0$, $\mathrm{KL}(p\|q) = 0$ iff. $p = q$.

From Regularization to Information Projection

$$q_* = \arg\min_{q \in \mathcal{P}} \; -\mathcal{U}(\tilde{\theta}, q) + \lambda \mathrm{KL}(q \| p)$$

$$\equiv q_* = \arg\min_{q \in \mathcal{P}} \mathrm{KL}(q \| p) \text{ s.t. } \mathcal{U}(\tilde{\theta}, q) \geq \epsilon$$

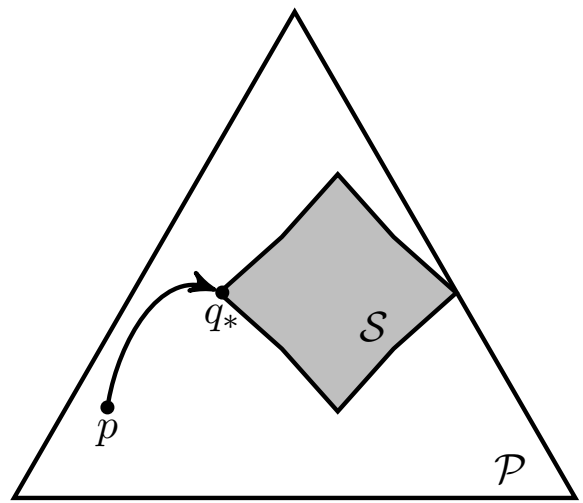
q_* is the information projection of p to $\mathcal{S} = \{q \,|\, \mathcal{U}(\tilde{\theta}, q) \geq \epsilon\}$

From Regularization to Information Projection

$$q_* = \arg\min_{q \in \mathcal{P}} \; -\mathcal{U}(\tilde{\theta}, q) + \lambda \mathrm{KL}(q \| p)$$

$$\equiv q_* = \arg\min_{q \in \mathcal{P}} \mathrm{KL}(q \| p) \text{ s.t. } \mathcal{U}(\tilde{\theta}, q) \geq \epsilon$$

q_* is the information projection of p to $\mathcal{S} = \{q \,|\, \mathcal{U}(\tilde{\theta}, q) \geq \epsilon\}$



Probability Restriction for Sparse Variables

Sparsity-Encouraging Utility

Consider the utility which encourages sparsity level $\tilde{k}$ for p:

$$\mathcal{U}(\tilde{k}, p) \propto |\mathsf{supp}(p)|\, 1_{\left[|\mathsf{supp}(p)| \leq \tilde{k}\right]}$$

The resulting information projection equivalent to:

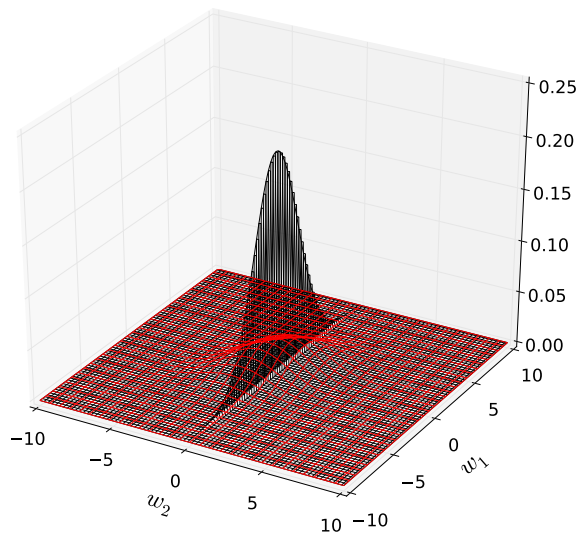
$$\max_{k<\tilde{k}} \left\{ \min_{q \in \mathcal{P}} \mathrm{KL}(q\|p) \text{ s.t. } |\mathsf{supp}(q)| = k \right\}$$

Theorem (Koyejo et al., 2014a)

Let $\mathcal{S} = \{q | \text{ supp}(q) = \text{A}\}$, the information projection of p to $\mathcal{S}$ is the restriction of p to the domain A.

$$p_{\text{A}} = \arg\min_{q \in \mathcal{S}} \text{KL}(q \| p) = \begin{cases} \frac{p(w)}{\int_{\text{A}} p(w) dw} & w \in \text{A}, \\ 0 & \text{otherwise.} \end{cases}$$

Example



Let $d = 2$, $\theta = \{1\}$

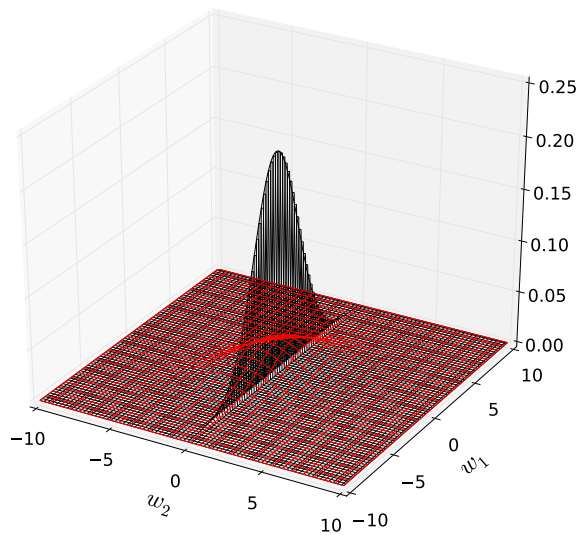
- $\text{A}_{\{1\}} = \{[w_1, w_2] \text{ s.t. } w_2 = 0\}$
- $\mathcal{S}_{\{1\}} = \{q | \text{ supp}(q) = \text{A}_{\{1\}}\}$
- $p_{\{1\}}(w) \propto p(w) \; \forall \; w \in \text{A}_{\{1\}}$

Theorem (Koyejo et al., 2014a)

Let $\mathcal{S} = \{q | \text{ supp}(q) = \mathsf{A}\}$, the information projection of p to $\mathcal{S}$ is the restriction of p to the domain A.

$$p_{\mathsf{A}} = \arg\min_{q \in \mathcal{S}} \mathrm{KL}(q \| p) = \begin{cases} \frac{p(w)}{\int_{\mathsf{A}} p(w) dw} & w \in \mathsf{A}, \\ 0 & \text{otherwise.} \end{cases}$$

Example



Let $d = 2$, $\theta = \{1\}$

- $\mathsf{A}_{\{1\}} = \{[w_1, w_2] \text{ s.t. } w_2 = 0\}$
- $\mathcal{S}_{\{1\}} = \{q | \text{ supp}(q) = \mathsf{A}_{\{1\}}\}$
- $p_{\{1\}}(w) \propto p(w) \; \forall \; w \in \mathsf{A}_{\{1\}}$

Scoring the Sparse Restriction

Restriction Score

Let $\theta \in [d]$ so $\mathcal{S}_\theta = \{q | \text{ supp}(q) = \mathsf{A}_\theta\}$

$$J(\theta) = -\min_{q \in \mathcal{S}_\theta} \mathrm{KL}(q \| p) = -\mathrm{KL}(p_\theta \| p)$$

- measures "information" retained after θ variables selected

Variable selection

Estimate the subset of k most *important* variables

$$\theta^* = \arg\max_{|\theta|=k} J(\theta)$$

- Expensive search e.g. $d = 10000, k = 100$ requires $\mathcal{O}(10^{241})$ evaluations ($10^{78} - 10^{82}$ est. atoms in the known universe)

Scoring the Sparse Restriction

Restriction Score

Let $\theta \in [d]$ so $\mathcal{S}_\theta = \{q | \text{ supp}(q) = \mathsf{A}_\theta\}$

$$J(\theta) = -\min_{q \in \mathcal{S}_\theta} \mathrm{KL}(q \| p) = -\mathrm{KL}(p_\theta \| p)$$

- measures "information" retained after θ variables selected

Variable selection

Estimate the subset of k most *important* variables

$$\theta^* = \arg\max_{|\theta|=k} J(\theta)$$

- Expensive search e.g. $d = 10000, k = 100$ requires $\mathcal{O}(10^{241})$ evaluations ($10^{78} - 10^{82}$ est. atoms in the known universe)

Efficient Variable Selection

Theorem (Koyejo et al., 2014a)

$J(\theta)$ is monotone submodular wrt θ for any bounded density p.

Greedy

Input: $k, \theta = \emptyset$
while $|\theta| < k$ **do**
  foreach $i \in [d] \backslash \theta$
  $f_i = J(\theta \cup i) - J(\theta)$
  $\theta = \theta \cup \{\arg\max f_i\}$
end while
Return: θ.

- *Greedy* is guaranteed to get $1 - \frac{1}{e}$ close to the *optimal* selection (Nemhauser et al., 1978)

Efficient Variable Selection

Theorem (Koyejo et al., 2014a)

$J(\theta)$ is monotone submodular wrt θ for any bounded density p.

Greedy

Input: $k, \theta = \emptyset$
while $|\theta| < k$ **do**
  foreach $i \in [d] \backslash \theta$
  $f_i = J(\theta \cup i) - J(\theta)$
  $\theta = \theta \cup \{\arg\max f_i\}$
end while
Return: θ.

- *Greedy* is guaranteed to get $1 - \frac{1}{e}$ close to the *optimal* selection (Nemhauser et al., 1978)

Efficient Variable Selection

Theorem (Koyejo et al., 2014a)

$J(\theta)$ is monotone submodular wrt θ for any bounded density p.

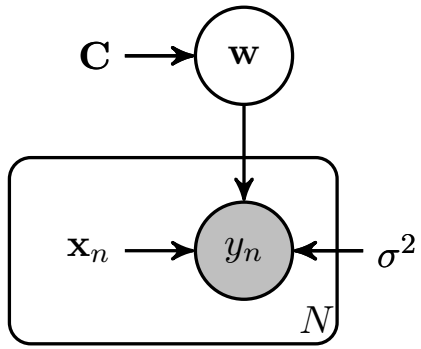
Greedy

Input: $k, \theta = \emptyset$
while $|\theta| < k$ **do**
  foreach $i \in [d] \backslash \theta$
  $f_i = J(\theta \cup i) - J(\theta)$
  $\theta = \theta \cup \{\arg\max f_i\}$
end while
Return: θ.

- *Greedy* is guaranteed to get $1 - \frac{1}{e}$ close to the *optimal* selection (Nemhauser et al., 1978)

Regression Model

$$y_i|_{\mathbf{w},\mathbf{x}_i,\epsilon} = \mathbf{w}^\top \mathbf{x}_i + \epsilon$$
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right), \ \mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{C}\right)$$



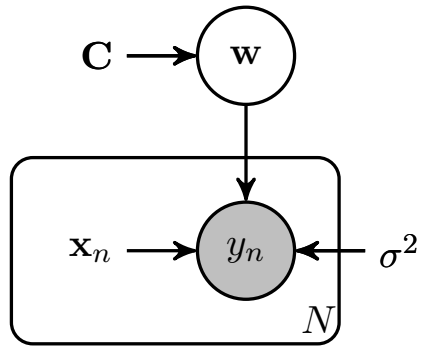
Variable Selection

- Let $P(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ and $P_\theta(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\mathbf{m}_\theta, \mathbf{S}_\theta\right)$

$$J(\theta) \propto \underbrace{(\boldsymbol{\mu} - \mathbf{m}_\theta)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_\theta)}_{\text{mean approx.}} - \underbrace{\log \frac{|(\mathbf{S}_\theta^{-1})|}{|\boldsymbol{\Sigma}^{-1}|}}_{\text{Cov. approx.}}$$

Regression Model

$$y_i|_{\mathbf{w},\mathbf{x}_i,\epsilon} = \mathbf{w}^\top \mathbf{x}_i + \epsilon$$
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right), \ \mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{C}\right)$$



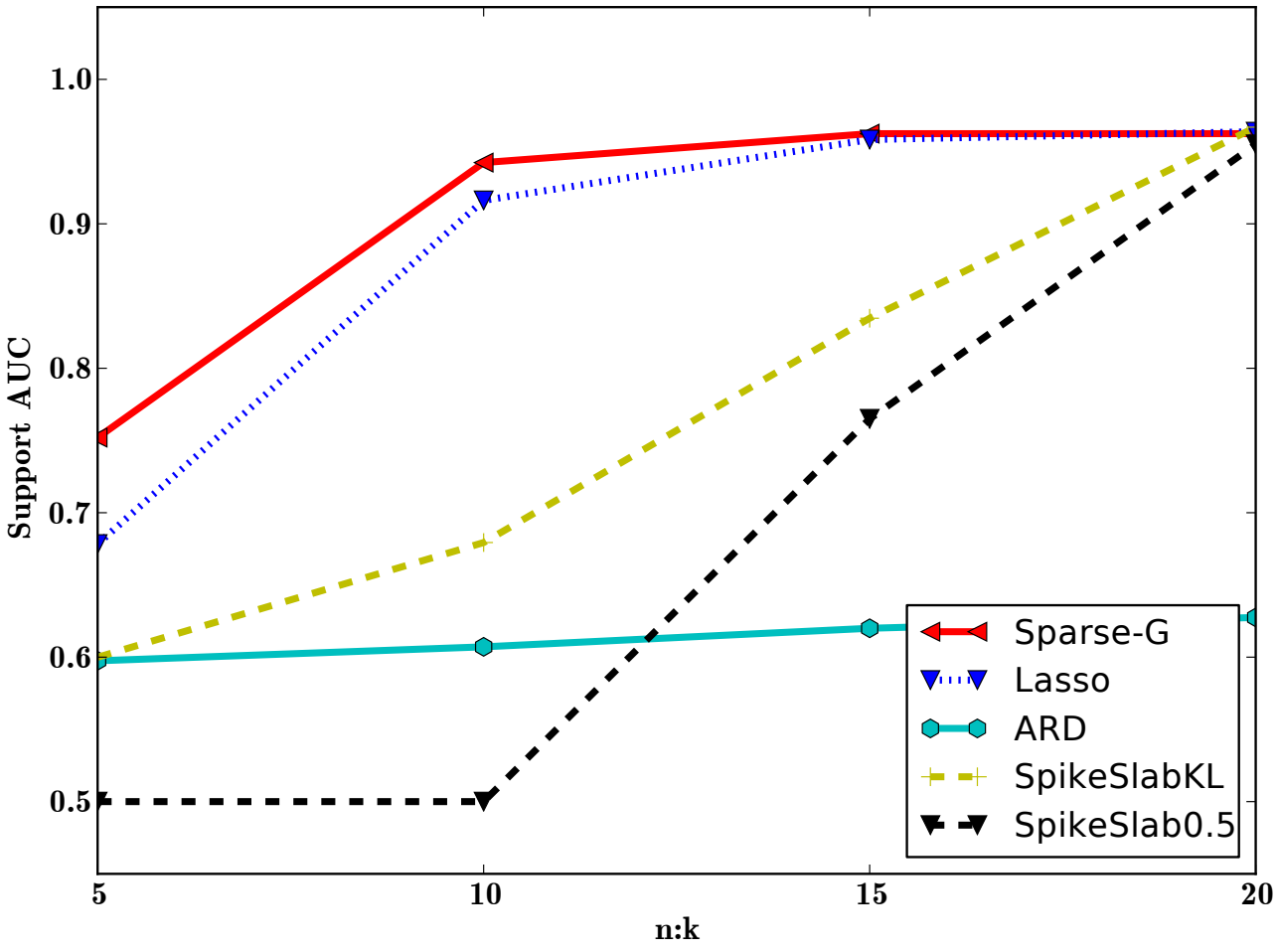
Variable Selection

- Let $P(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ and $P_\theta(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\mathbf{m}_\theta, \mathbf{S}_\theta\right)$

$$J(\theta) \propto \underbrace{(\boldsymbol{\mu} - \mathbf{m}_\theta)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_\theta)}_{\text{mean approx.}} - \underbrace{\log \frac{|(\mathbf{S}_\theta^{-1})|}{|\boldsymbol{\Sigma}^{-1}|}}_{\text{Cov. approx.}}$$

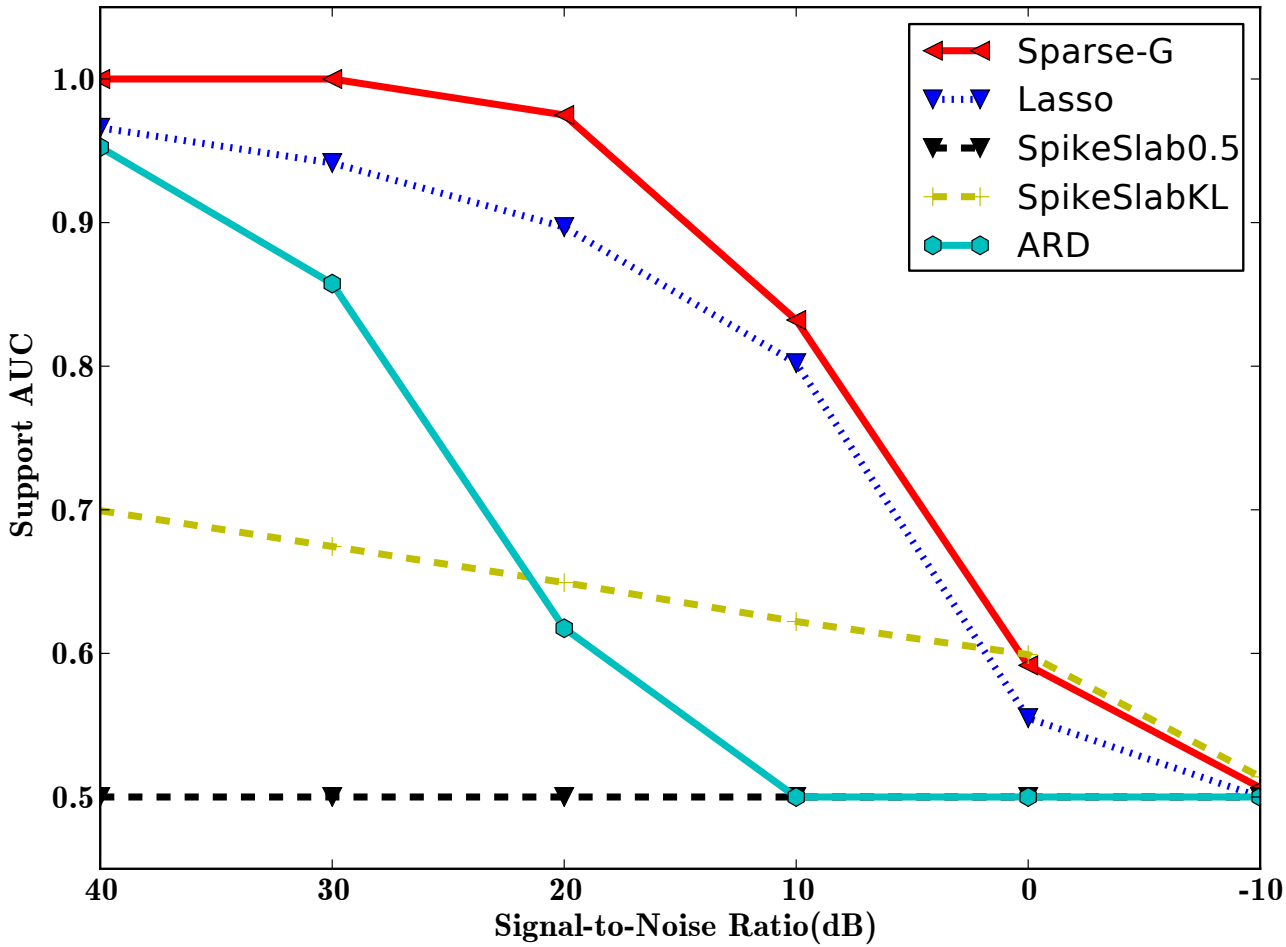
Simulated Data Results: Support Recovery

k=20, d=10,000, SNR $= 20$dB , $n = 100, \ldots, 400$



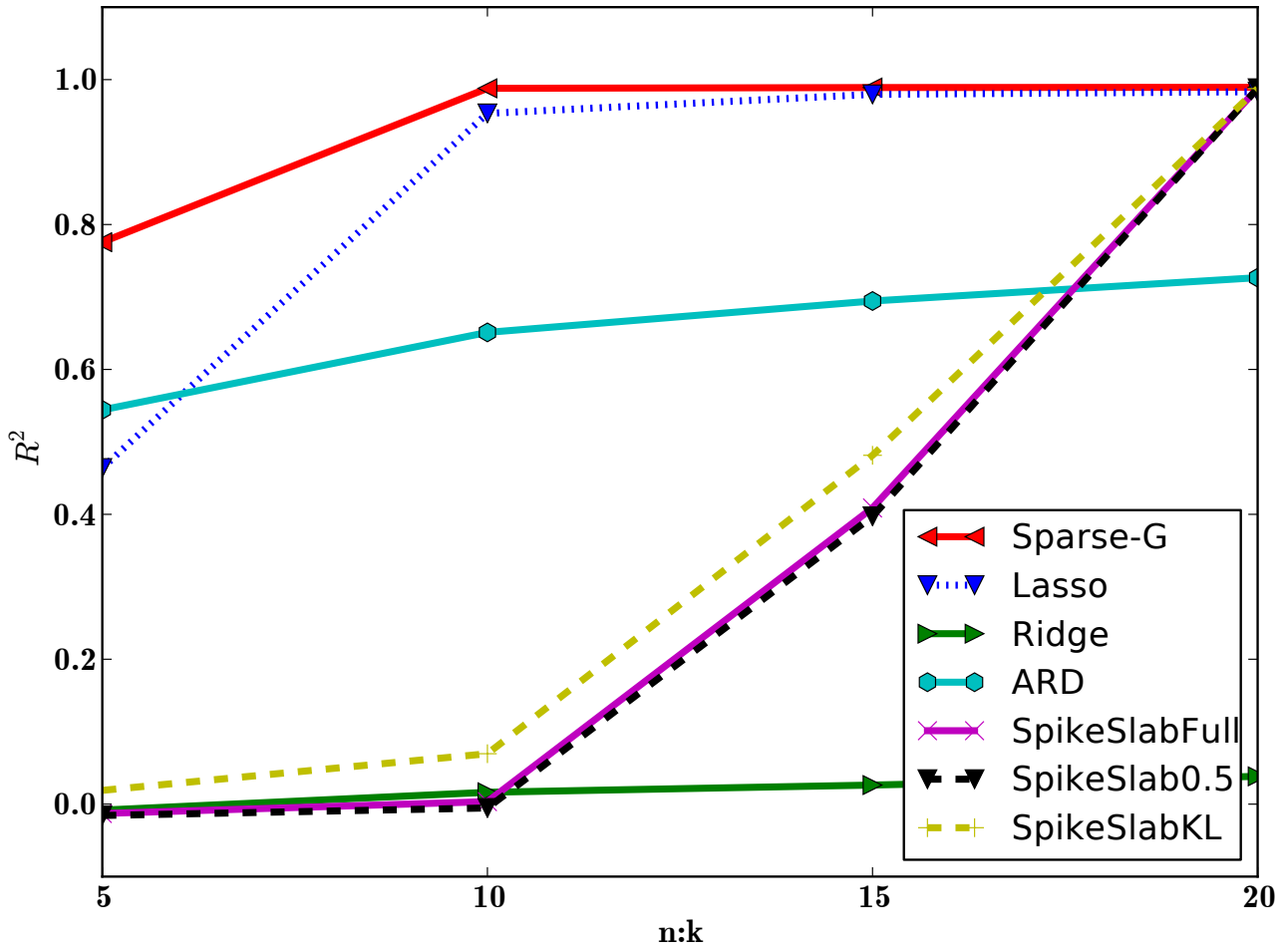
Simulated Data Results: Support Recovery

k=20, d=10,000, SNR $= 40\text{dB} \ldots -10\text{dB}$, $n = 200$



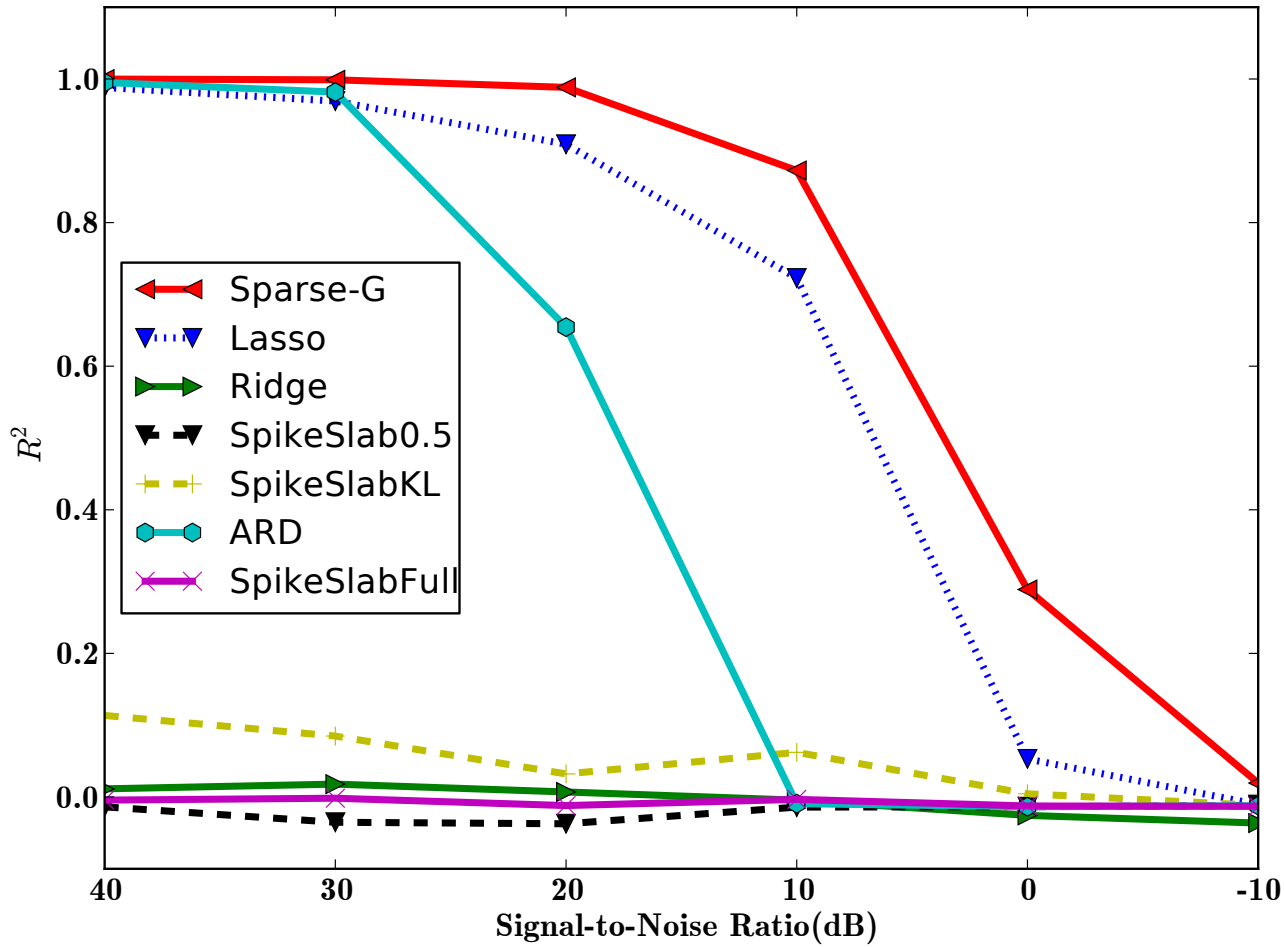
Simulated Data Results: Regression

k=20, d=10,000, SNR $= 20$dB , $n = 100, \ldots, 400$



Simulated Data Results: Regression

k=20, d=10,000, SNR $= 40\text{dB} \ldots -10\text{dB}$, $n = 200$



Probabilistic Sparse Canonical Correlation Analysis

Khanna et al. (2016)[Under Review]

Canonical Correlation Analysis

- Canonical correlation analysis is useful or joint analysis of multi-view datasets
- Let $\mathbf{X} \in \mathbb{R}^{n \times d_1}, \mathbf{Y} \in \mathbb{R}^{n \times d_2}$
- Canonical Correlation Analysis (CCA) seeks factors $u \in \mathbb{R}^{d_1}, v \in \mathbb{R}^{d_2}$ which "explain" the cross-correlation

$$\mathbf{u}_*, \mathbf{v}_* = \arg\max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^\top \mathbf{X}^\top \mathbf{Y} \mathbf{v}}{|\mathbf{u}^\top \mathbf{X} \mathbf{u}||\mathbf{v}^\top \mathbf{Y} \mathbf{v}|}$$

Probabilistic CCA

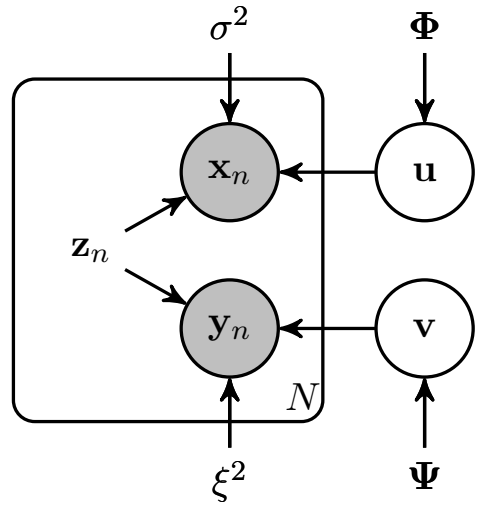
Bach and Jordan (2005)

$$\mathbf{x}_n|_{\mathbf{u},\mathbf{z}_n,\epsilon} = \mathbf{u}^\top \mathbf{z}_n + \epsilon$$

$$\mathbf{y}_n|_{\mathbf{v},\mathbf{z}_n,\epsilon} = \mathbf{v}^\top \mathbf{z}_n + \nu$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}), \ \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Phi})$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2), \ \nu \sim \mathcal{N}(0, \xi^2)$$



Sparse Probabilistic CCA

Model fit using Expectation Maximization (EM)

E-Step

- select elements separately from $\mathbf{u}, \mathbf{v}$ subject to a budget constraint.
- equivalent constraint set is a *Partition Matroid*.

Probabilistic CCA

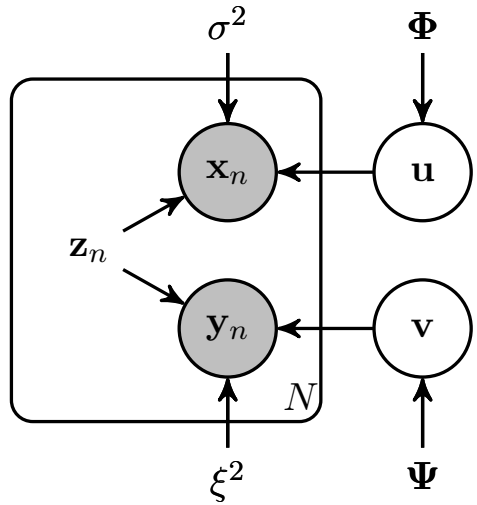
Bach and Jordan (2005)

$$\mathbf{x}_n|_{\mathbf{u},\mathbf{z}_n,\epsilon} = \mathbf{u}^\top \mathbf{z}_n + \epsilon$$

$$\mathbf{y}_n|_{\mathbf{v},\mathbf{z}_n,\epsilon} = \mathbf{v}^\top \mathbf{z}_n + \nu$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}), \ \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Phi})$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2), \ \nu \sim \mathcal{N}(0, \xi^2)$$



Sparse Probabilistic CCA

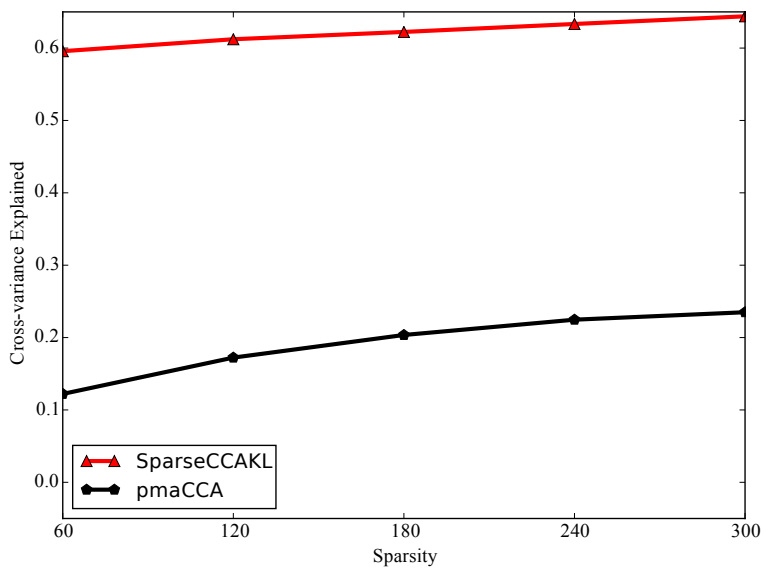
Model fit using Expectation Maximization (EM)

E-Step

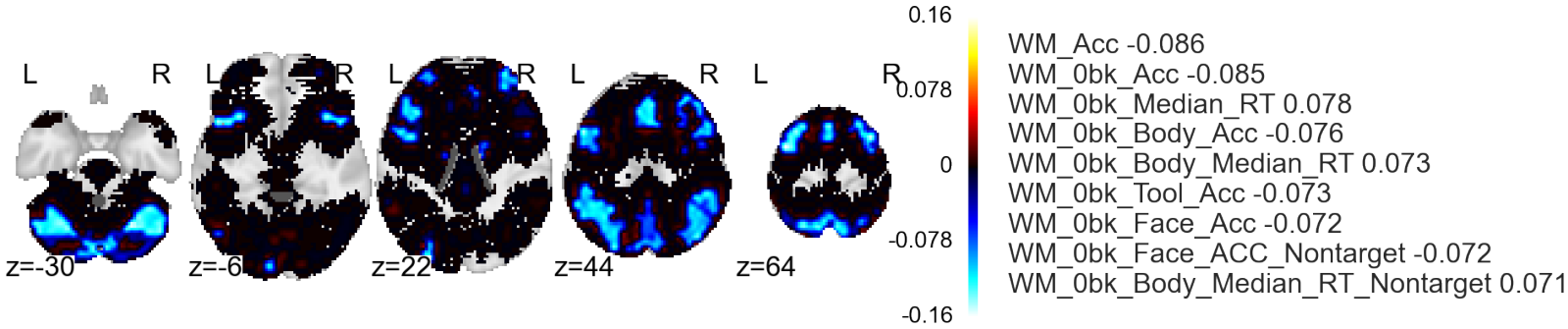
- select elements separately from $\mathbf{u}, \mathbf{v}$ subject to a budget constraint.
- equivalent constraint set is a *Partition Matroid*.

Human Connectome Project Data (Essen et al., 2013)

- Investigating association between human brain function and human behavior
- Joint decomposition of task brain images and behavioral variables
- $n = 497$ adult subjects. Each subject has $d_1 = 380$ behavioral variables, $d_2 = 27000$ voxels

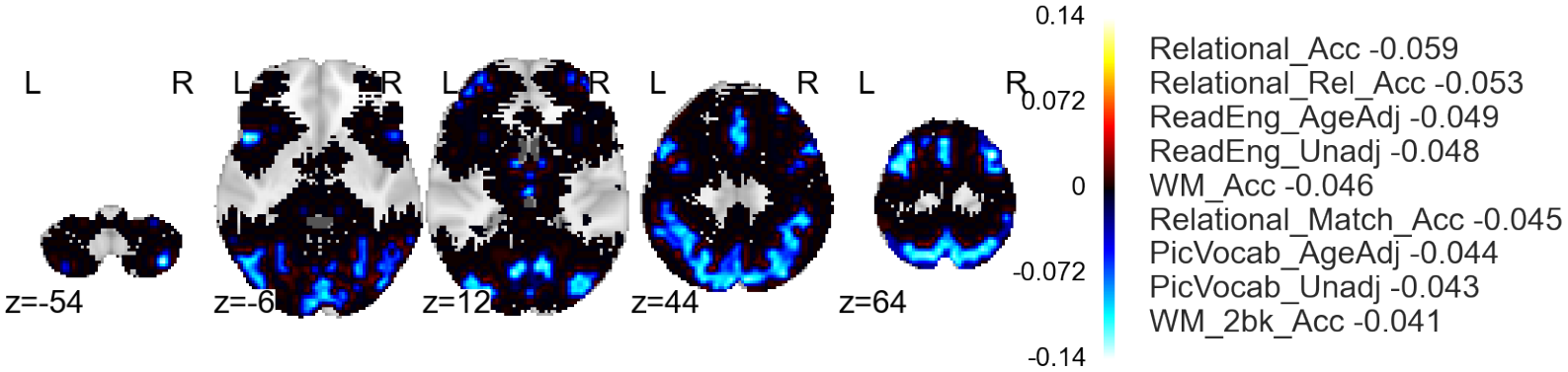


2 Back vs 0 Back contrast (measures working memory)



Neural support is seen in a number of frontal and parietal regions and cerebellum, consistent with cognitive control systems usually engaged by the task. Behavioral correlates including both reaction time and accuracy on the task, showing greater neural engagement associated with slower and less accurate performance.

REL vs MATCH contrast (measures relational processing)



Neural support is observed in frontal, parietal, and occipital cortex. Behavioral correlates captured both performance on this particular task, as well as independent measures related to higher cognitive functions including working memory capacity, vocabulary, and reading.

Conclusion: Part II

Probabilistic inference for structured variables:

- proposed regularized approach for incorporating user-constructed utility into probabilistic model
- discussed applications to constructing structured distributions via restriction
- efficient submodular optimization when applied to *sparsity*

Analysis of neuroimaging data:

- **key hypothesis:** *joint* spatial smoothness and sparsity
- performance meets or exceeds state of the art
- **open problems:** construct fMRI priors based on DWI

Probabilistic inference for structured variables:

- proposed regularized approach for incorporating user-constructed utility into probabilistic model
- discussed applications to constructing structured distributions via restriction
- efficient submodular optimization when applied to *sparsity*

Analysis of neuroimaging data:

- **key hypothesis:** *joint* spatial smoothness and sparsity
- performance meets or exceeds state of the art
- **open problems:** construct fMRI priors based on DWI

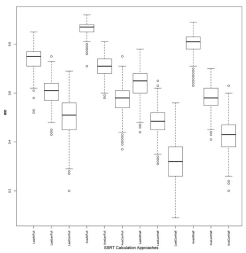
web: http://sanmik.github.io
email: sanmi@illinois.edu

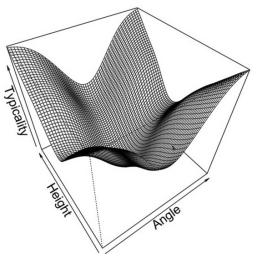
Other Research

Behavioral

Disease symptoms
(10s of scales)

Neurocognitive
measures
(100s of variables)





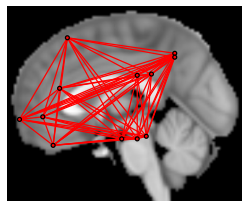
Imaging

Structural MRI
(200K vertices)

Functional MRI
(200K voxels X
150-1000 timepoints)

Diffusion MRI
(200K voxels X
50 diffusion directions)

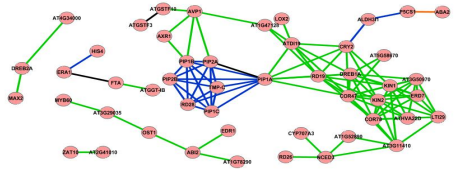
Dense connectome
(~2^{10} connections)

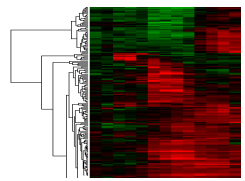


Transcriptome

Microarray/RNA-seq
(~21K genes)

Gene set
enrichment analysis
(100s of pathways)



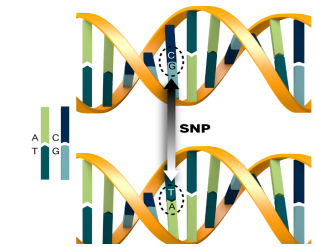


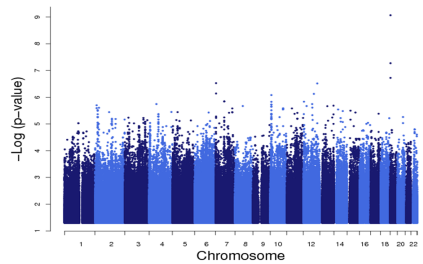
Genome

SNP (~1M variants)

WGS (2.2B bases)

mutational load
(~21K genes)





Development of Principled, Scalable, Effective Methods

Exploratory and Predictive Modeling:

- high dimensional data analysis (multiscale, multi-modal)
- structured variables with complicated interactions

Scalability:

- efficient algorithms, parallel processing

Decision theory:

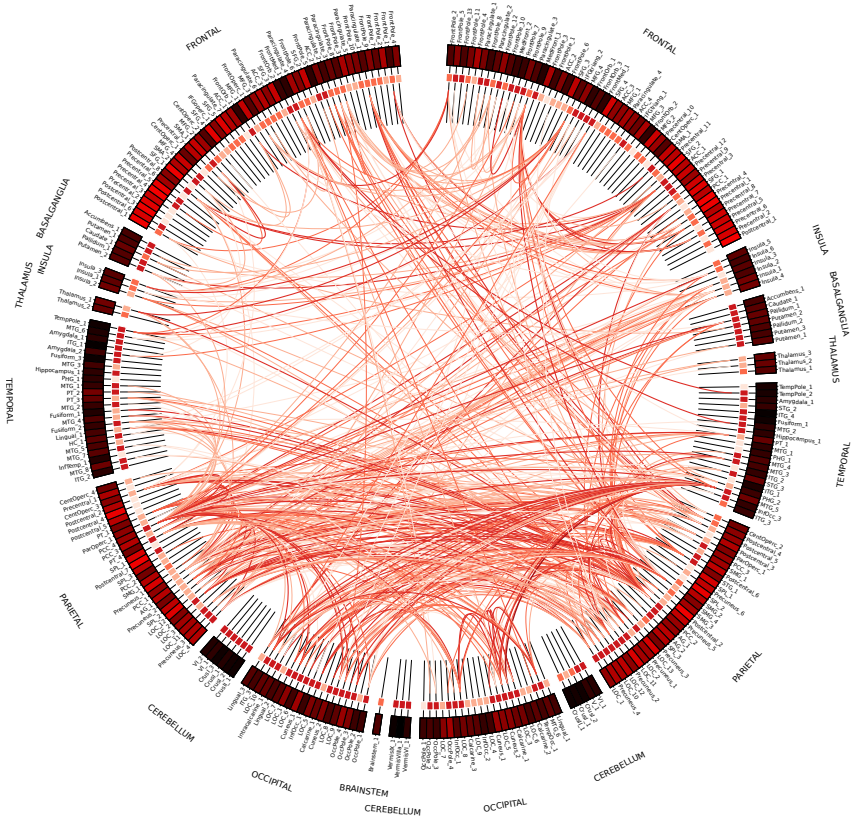
- optimal predictive decision making
- variable selection, large scale hypothesis testing

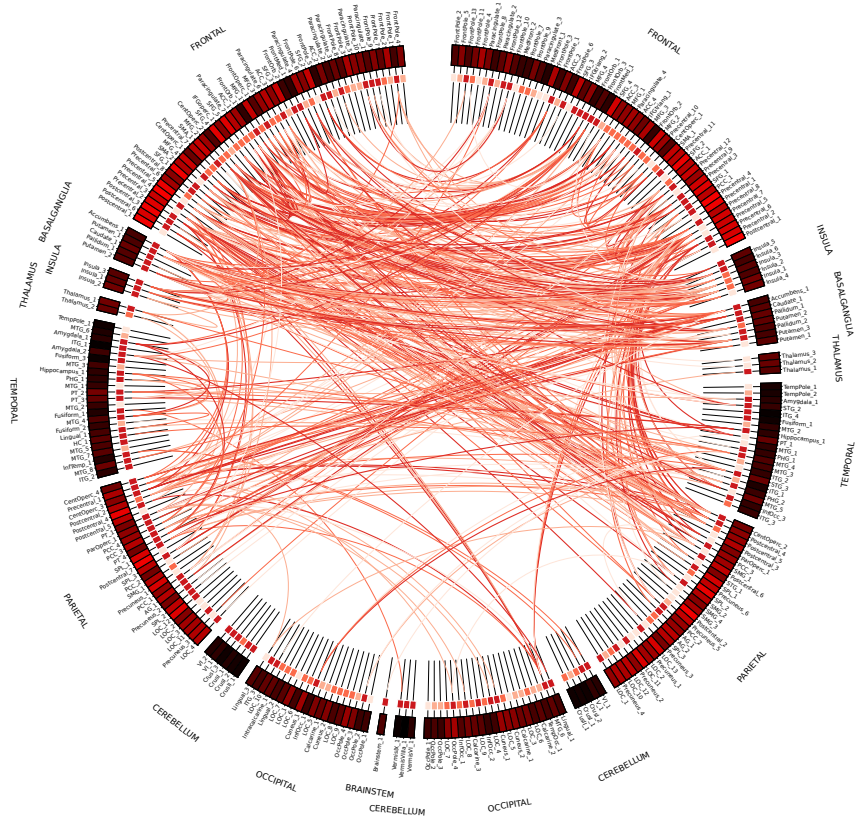
Theoretical Analysis:

- statistical guarantees e.g consistency, sample complexity

Exploratory Analysis of Imaging, Genetics and Behavior

Koyejo, McKay, Knowles, Blangero, Glahn, and Poldrack (2014b)

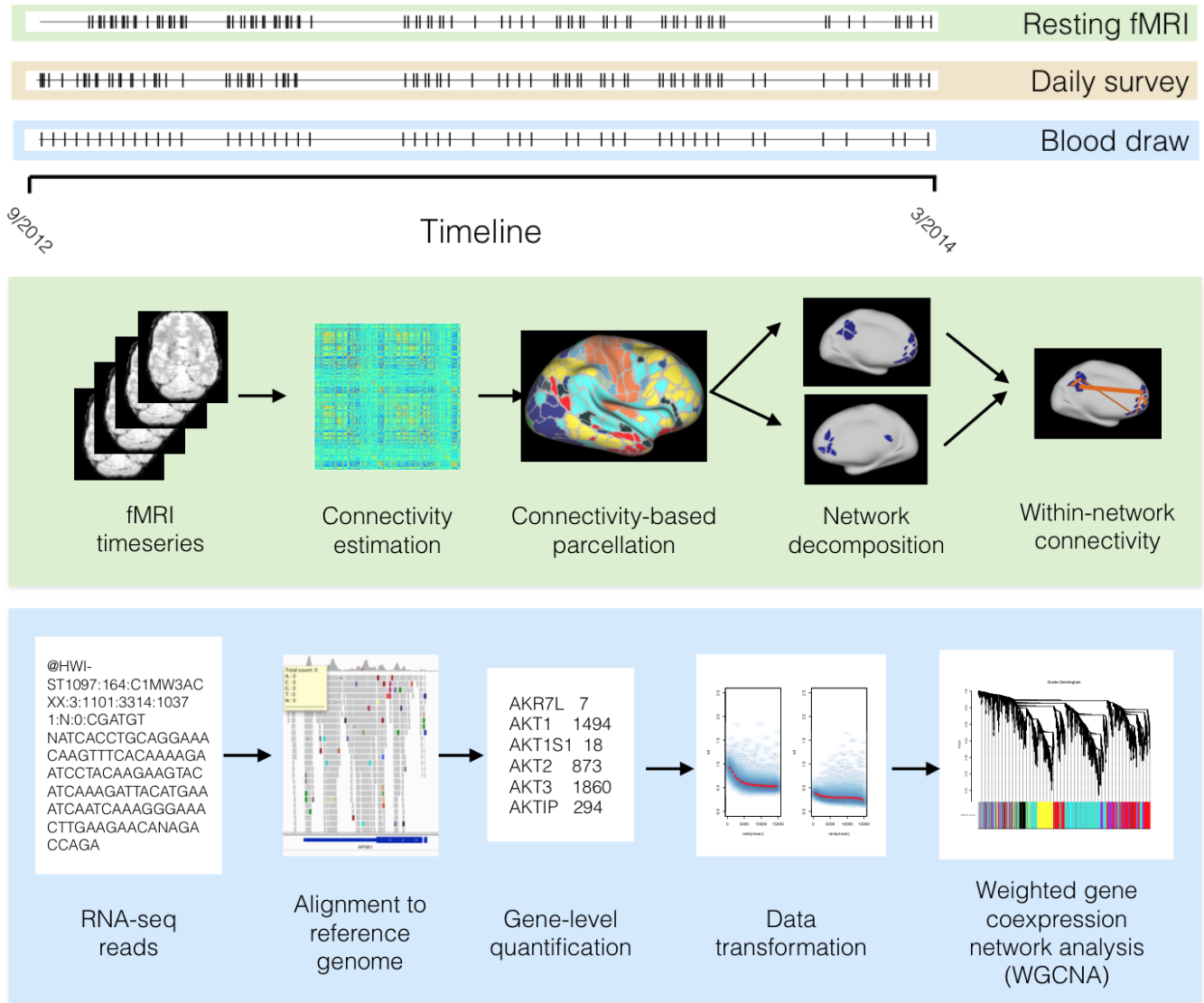




Behav. h^2 (p-value)	**DTI h^2 (p-value)**
0.175 (0.036)	0.229 (0.010)
0.224 (0.007)	0.598 (4.488×10^{-10})
0.223 (0.008)	0.823 (8.525×10^{-19})

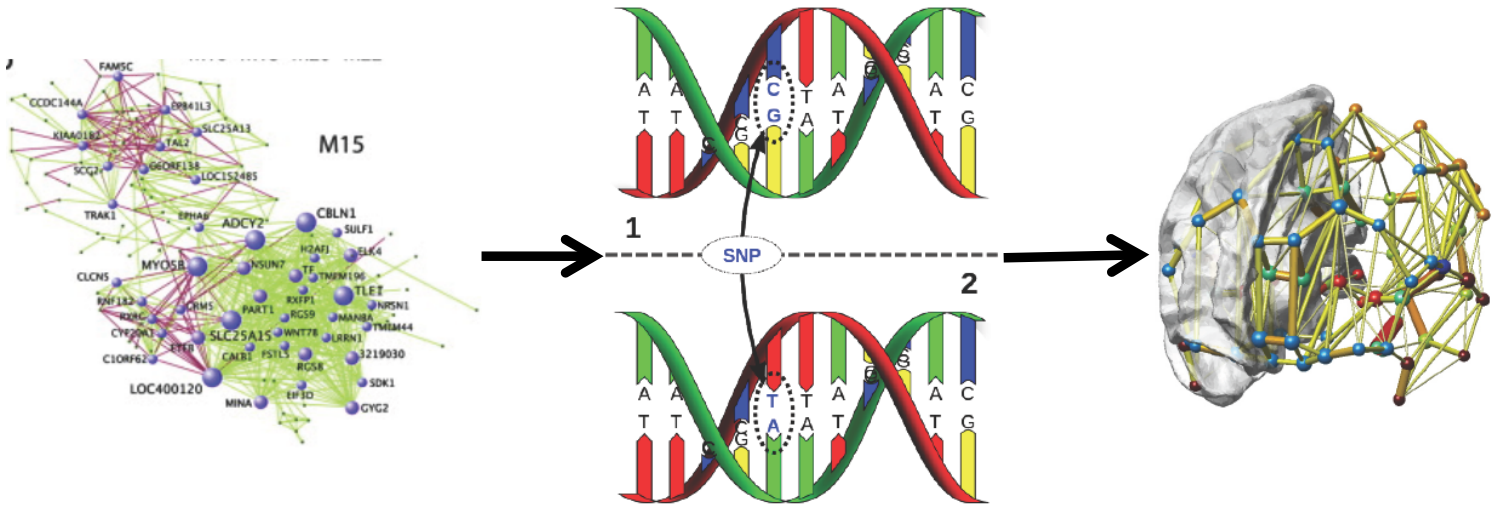
MyConnectome

Poldrack et al. (2015)

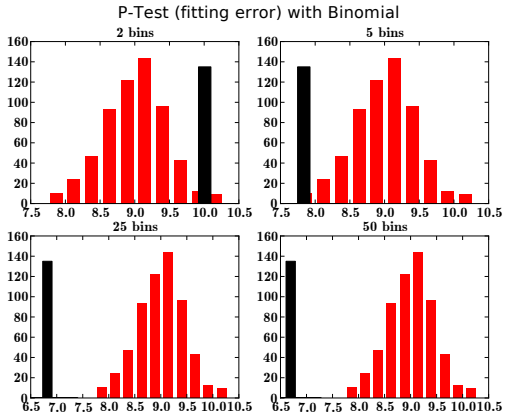


Optimal Predictive Decision Making

- Analysis of optimal predictors and consistent estimators for structured outputs e.g multilabel prediction, graph prediction

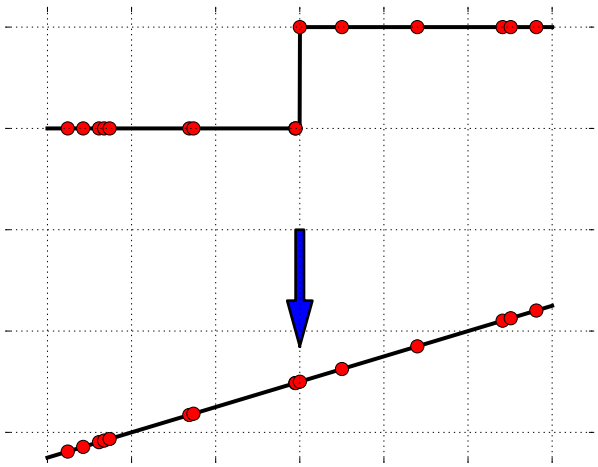


High Dimensional Estimation and Predictive Modeling with Aggregated Data



Bhowmik, Ghosh, and Koyejo (2015)
Bhowmik, Ghosh, and Koyejo (2016)

Ranking



Acharyya, Koyejo, and Ghosh (2012)
Koyejo, Acharyya, and Ghosh (2013)

References

References I

Sreangsu Acharyya, Oluwasanmi Koyejo, and Joydeep Ghosh. Learning to rank with Bregman divergences and monotone retargeting. In *Proceedings of the 28th conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

Francis Bach and Michael Jordan. A probabilistic interpretation of canonical correlation analysis. techreport 688, 2005. URL `http://www.di.ens.fr/~{}fbach/probacca.pdf`.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

Avradeep Bhowmik, Joydeep Ghosh, and Oluwasanmi Koyejo. Generalized linear models for aggregated data. In *Proceedings of the 18th International conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Avradeep Bhowmik, Joydeep Ghosh, and Oluwasanmi Koyejo. Sparse parameter recovery from aggregated data. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1090–1099, 2016.

Mark S Cohen and Susan Y Bookheimer. Localization of brain function using magnetic resonance imaging. *Trends in neurosciences*, 17(7):268–277, 1994.

David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. CRC Press, 1979.

I. Csiszar. i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):pp. 146–158, 1975.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The wu-minn human connectome project: An overview. *NeuroImage*, 80:62 – 79, 2013. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2013.05.041. URL `http://www.sciencedirect.com/science/article/pii/S1053811913005351`. Mapping the Connectome.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, 11:2001–2049, 2010.

Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In *NIPS*, 1999.

E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review Online Archive (Prola)*, 106(4):620–630, 1957.

References II

Rajiv Khanna, Joydeep Ghosh, Russell A. Poldrack, and Oluwasanmi Koyejo. Information projection and approximate inference for structured sparse variables. 2016. Submitted.

Oluwasanmi Koyejo, Sreangsu Acharyya, and Joydeep Ghosh. Retargeted matrix factorization for collaborative filtering. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 49–56, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2409-0. doi: 10.1145/2507157.2507185.

Oluwasanmi Koyejo, Rajiv Khanna, Joydeep Ghosh, and Russell Poldrack. On prior distributions and approximate inference for structured variables. In *Advances in Neural Information Processing Systems*, pages 676–684, 2014a.

Oluwasanmi Koyejo, David Reese McKay, Emma E.M. Knowles, John Blangero, David Glahn, and Russell A. Poldrack. Exploratory analysis of imaging and behavioral phenotypes with sparse CCA. In *Organization for Human Brain Mapping (Abstract)*, 2014b.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3303–3311. Curran Associates, Inc., 2015. URL `http://papers.nips.cc/paper/5883-consistent-multilabel-classification.pdf`.

Solomon Kullback. *Information Theory and Statistics*. Dover, 1959.

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Advances in Neural Information Processing Systems*, pages 1493–1501, 2014.

G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

R A Poldrack. Subtraction and beyond: The logic of experimental designs for neuroimaging. In S. J. Hanson and M. Bunzl, editors, *Foundational Issues in Human Brain Mapping*, pages 147–160. MIT Press, Cambridge, MA, 2010.

Russell A Poldrack, Yaroslav O Halchenko, and Stephen José Hanson. Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, 20: 1364–1372, 2009.

References III

Russell A. Poldrack, Timothy Laumann, Oluwasanmi Koyejo, Ashleigh Hover Brenda Gregory, Mei-Yen Chen, Jeffrey Luci, Sung Jun Joo, Scott Hunicke-Smith Ryan Boyd, Zack Booth Simpson, Thomas Caven, James M. Shine Vanessa Sochat, Evan Gordon, Abraham Snyder, Babatunde Adeyemo, Steven E. Petersen, David Glahn, D. Reese Mckay, Joanne E. Curran, Harald H. H. Goring, Melanie A. Carless, John Blangero, Laurie Frick, Edward M. Marcotte, and Jeanette A. Mumford. Phenome-wide dynamics of mind, brain, and body: The myconnectome project. 2015. Submitted.

M I Posner, S E Petersen, P T Fox, and M E Raichle. Localization of cognitive operations in the human brain. *Science*, 240(4859):1627–31, Jun 1988.

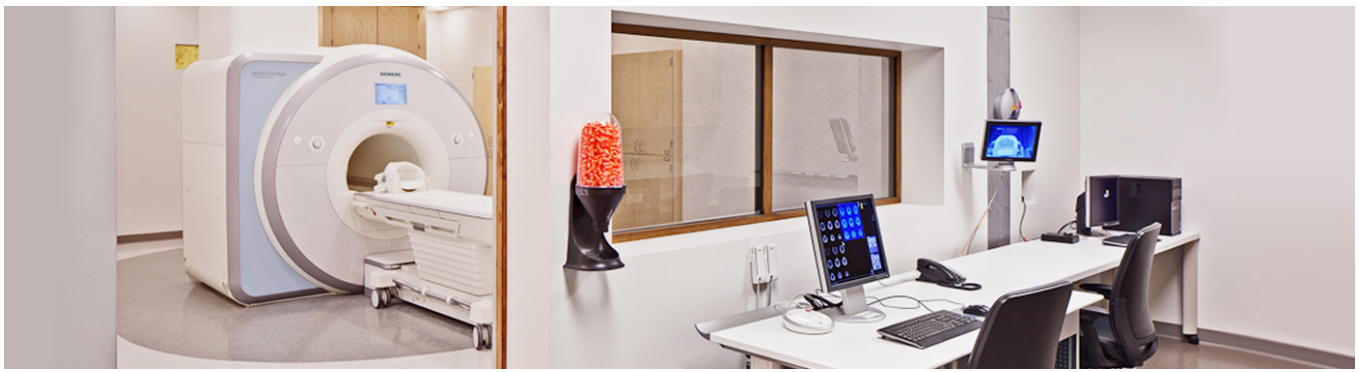
Corey N White, Eliza Congdon, Jeanette A Mumford, Katherine H Karlsgodt, Fred W Sabb, Nelson B Freimer, Edythe D London, Tyrone D Cannon, Robert M Bilder, and Russell A Poldrack. Decomposing decision components in the stop-signal task: A model-based approach to individual differences in inhibitory control. *Journal of Cognitive Neuroscience*, 2014.

Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *ICML*, 2009.

Backup Slides

Functional Magnetic Resonance Imaging (fMRI)

- measures blood flow correlated with neuronal activity

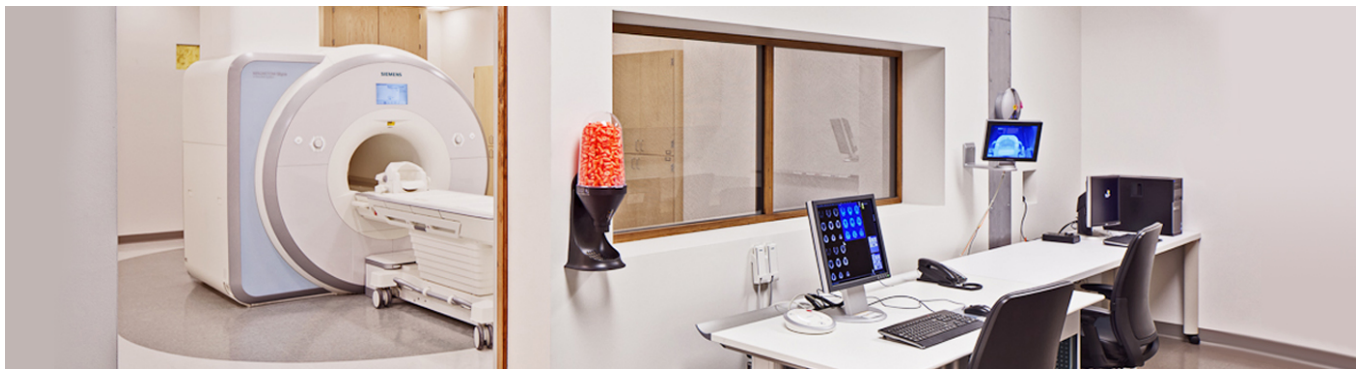


Common paradigms:

- **task fMRI:** measurements of participant during experiment
- **resting state fMRI:** measurement while participant is at "rest"

Functional Magnetic Resonance Imaging (fMRI)

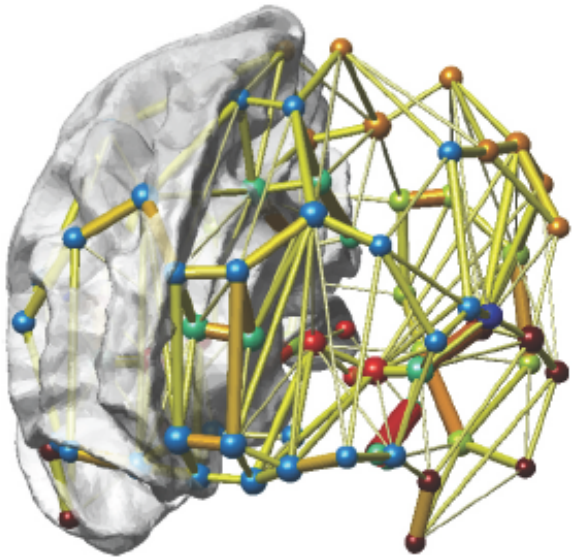
- measures blood flow correlated with neuronal activity



Common paradigms:

- **task fMRI:** measurements of participant during experiment
- **resting state fMRI:** measurement while participant is at "rest"

Spatial Smoothness

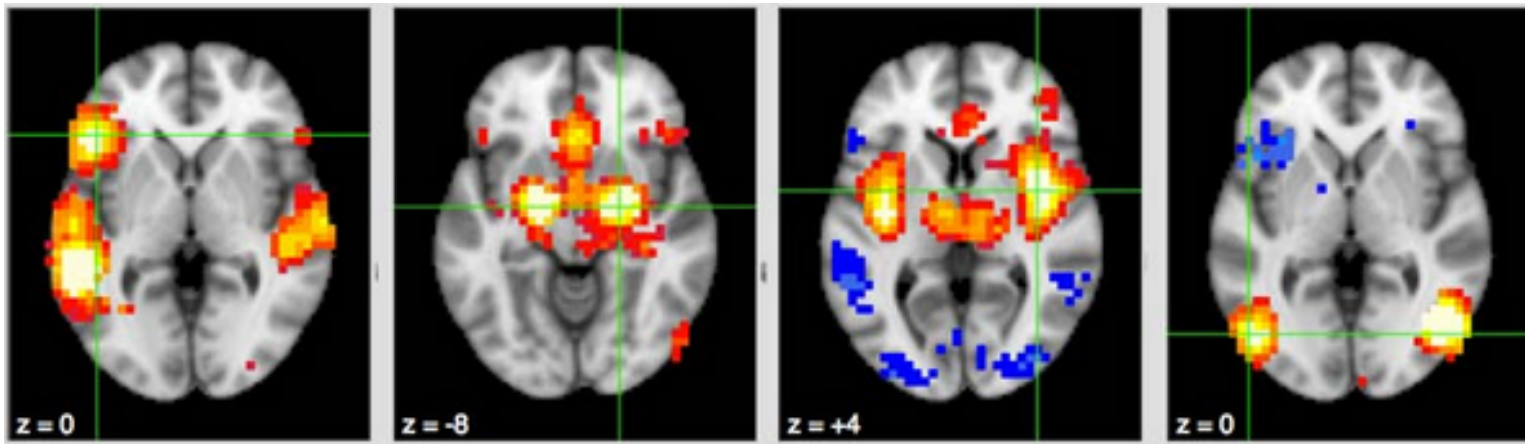


- **hypothesis:** fMRI signal is spatially smooth, due to anatomy, prepossessing . . .

- spatial smoothness is captured by spatially correlated priors e.g. pair-wise Markov random field (MRF)

Sparsity

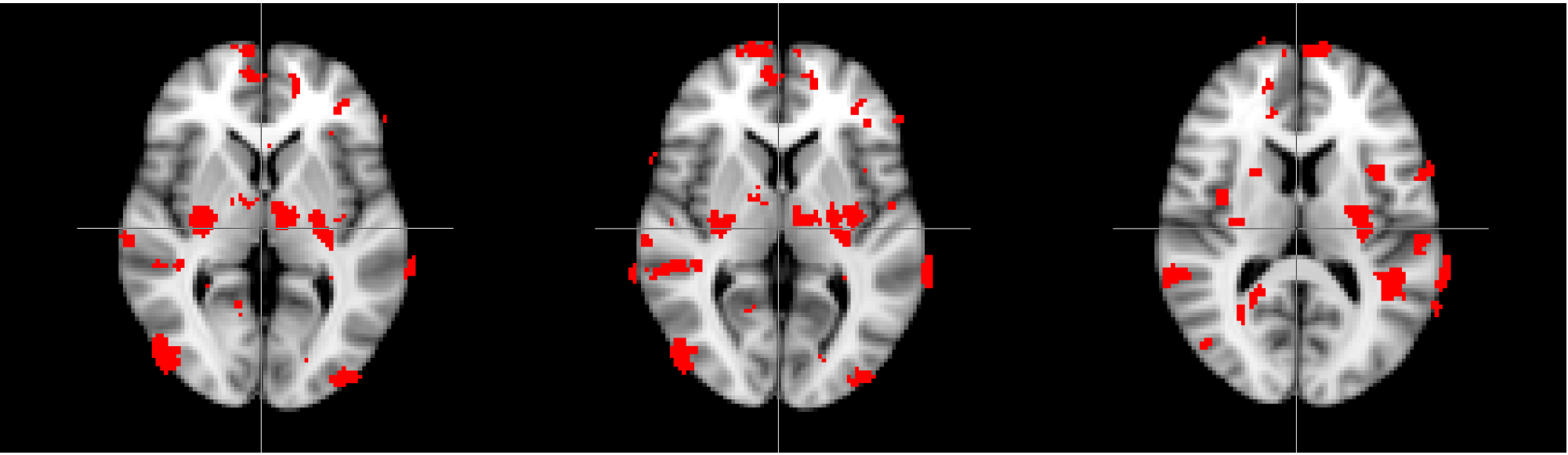
- **hypothesis:** mental processes are spatially localized (Cohen and Bookheimer, 1994)



- **motivation:** jointly *sparse* and *spatially smooth* model i.e. clustered sparsity

FMRI stop signal task (White et al., 2014)

- **stop signal task:** designed to measure impulse control



$z = 2$ $z = 5$ $z = 12$

- $n = 126, d = 10,000, k^* = 300$
- recovered regions correlated with stop signal reaction time e.g. include orbitofrontal cortex, dorsolateral prefrontal cortex, putamen, anterior cingulate, parietal cortex (Koyejo et al., 2014a)