Experimental Design on a Budget for Sparse Linear Models and Applications

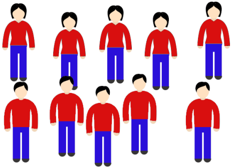
Sathya N. Ravi[⋆] Vamsi K. Ithapu[⋆] Sterling C. Johnson [‡] Vikas Singh[†,⋆]

[⋆] Computer Sciences [†] Biostatistics and Medical Informatics
[‡] William S Middleton Memorial VA Hospital

University of Wisconsin Madison

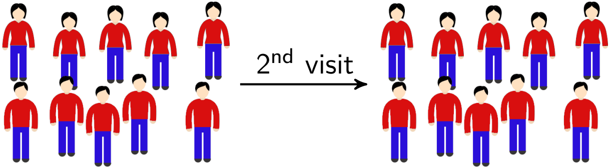
June 20, 2016

Longitudinal Study



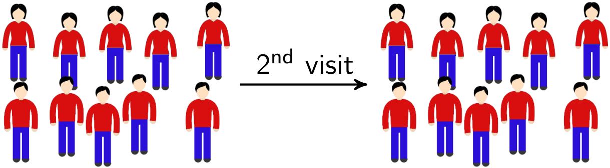
Observe: (x, s_t)

Longitudinal Study



Observe: (x, s_t) (x, s_{t+1})

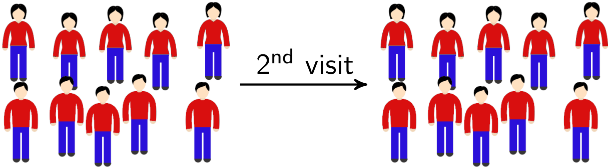
Longitudinal Study



Observe: (x, s_t) (x, s_{t+1})

- $y = s_{t+1} - s_t$

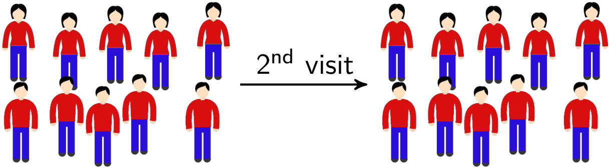
Longitudinal Study



Observe: (x, s_t) (x, s_{t+1})

- $y = s_{t+1} - s_t$
- Assume $f : x \to y$.

Longitudinal Study



Observe: (x, s_t) (x, s_{t+1})

- $y = s_{t+1} - s_t$
- Assume $f : x \to y$. Specific examples of x, y?

Covariates/Response in Alzheimer's Disease Studies

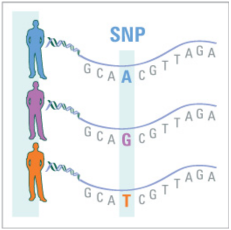
Examples of Covariates $\mathbf{x}$:

Imaging

Covariates/Response in Alzheimer's Disease Studies

Examples of Covariates $\mathbf{x}$:

Imaging

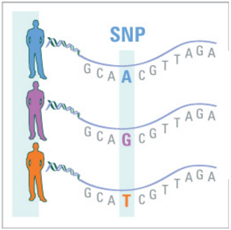


Genetics

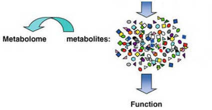
Covariates/Response in Alzheimer's Disease Studies

Examples of Covariates $\mathbf{x}$:

Imaging



Genetics

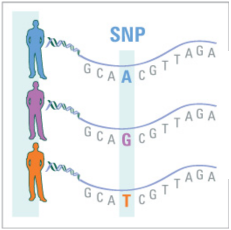


Metabolomics

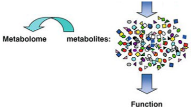
Covariates/Response in Alzheimer's Disease Studies

Examples of Covariates $\mathbf{x}$:

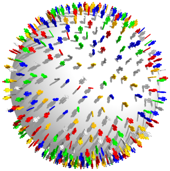
Imaging



Genetics



Metabolomics

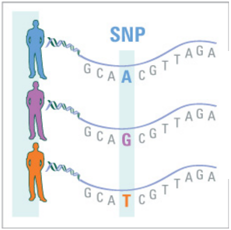


Demographics

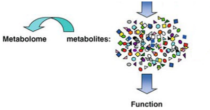
Covariates/Response in Alzheimer's Disease Studies

Examples of Covariates **x** :

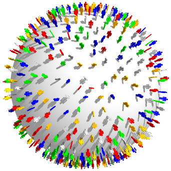
Imaging



Genetics



Metabolomics



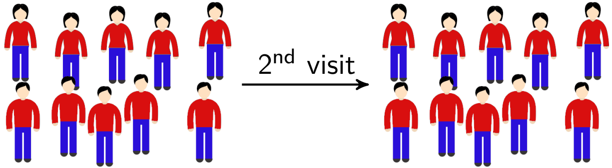
Demographics

Examples of Response **y** : Clinical Dementia Rating (CDR), Alzheimer's Disease Assessment Score (ADAS) etc..

Financial implications

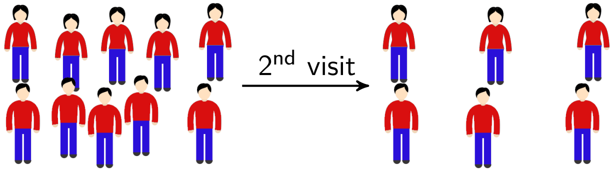
Financial implications

500 **patients with cost** $\$150$**/visit** $= \$150 \times 500 = \$75000$.



Financial implications

300 **patients with cost** $\$150$/**visit** = $\$150 \times 300 = \$45000$.



Savings=$\$30000$.

How to choose x_i's in a meaningful way?

How to choose x_i's in a meaningful way?

- Active Learning approaches

How to choose x_i's in a meaningful way?

NOT allowed to get y_i's on a cost to know basis.

- Active Learning approaches

How to choose x_i's in a meaningful way?

- Active Learning approaches
- Variational approaches

How to choose x_i's in a meaningful way?

- Active Learning approaches
- Variational approaches

Compute integrals, not scalable to high dimensions/large data setting.

How to choose x_i's in a meaningful way?

- Active Learning approaches
- Variational approaches
- Sensing Matrix

How to choose x_i's in a meaningful way?

- Active Learning approaches
- Variational approaches
- Sensing Matrix

Covariate data agnostic, not applicable here.

How to choose x_i's in a meaningful way?

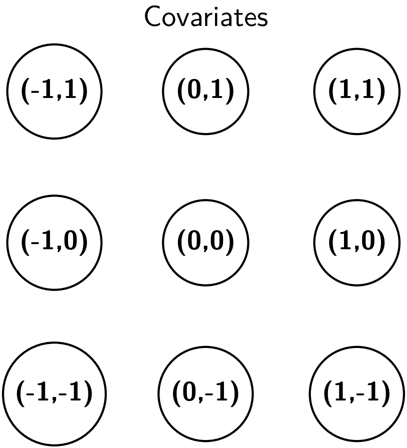
- Active Learning approaches
- Variational approaches
- Sensing Matrix
- Experimental Design

How to choose x_i's in a meaningful way?

- Active Learning approaches
- Variational approaches
- Sensing Matrix
- Experimental Design ✓

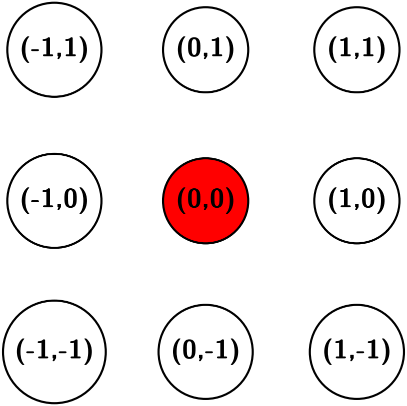
Well studied when f is linear. Often solved using Convex Optimization.

Example in 2−D: Choose 4 points



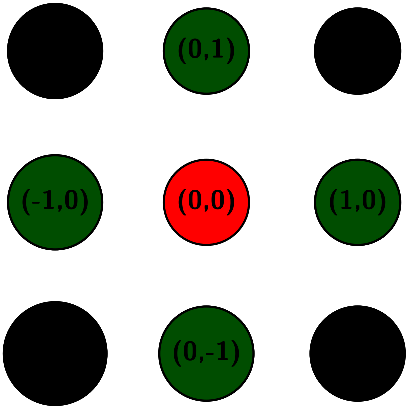
Example in 2−D: Choose 4 points

Origin is not useful.



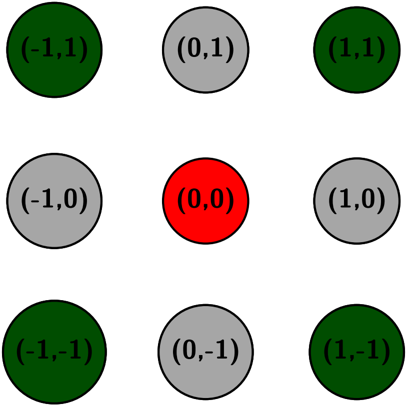
Example in 2−D: Choose 4 points

Not a great Choice!



Example in 2−D: Choose 4 points

Ah! Much better.



$D-$Optimal Design

$$\min_{\mathcal{S} \subseteq N} \log \det \left(\sum_{i \in \mathcal{S}} x_i {x_i}^T \right)^{-1} \quad \text{s.t.} \quad |\mathcal{S}| \leq B$$

$D-$Optimal Design

Information Matrix

$$\min_{\mathcal{S} \subseteq N} \log \det \left(\sum_{i \in \mathcal{S}} x_i {x_i}^T \right)^{-1} \quad \text{s.t.} \quad |\mathcal{S}| \leq B$$

$D-$Optimal Design

$$\min_{\mathcal{S}\subseteq N} \log\det\left(\sum_{i\in\mathcal{S}} x_i {x_i}^T\right)^{-1} \quad \text{s.t.} \quad |\mathcal{S}| \leq B$$

Information Matrix

−Shannon Differential Information Content

$D-$Optimal Design

Information Matrix

$$\min_{\mathcal{S}\subseteq N} \log\det\left(\sum_{i\in\mathcal{S}} x_i {x_i}^T\right)^{-1} \quad \text{s.t.} \quad |\mathcal{S}| \leq B$$

Full set of subjects

–Shannon Differential Information Content

$D-$Optimal Design

$$\min_{\mathcal{S}\subseteq N} \log\,\det \left(\sum_{i\in\mathcal{S}} x_i {x_i}^T\right)^{-1} \quad \text{s.t.} \quad |\mathcal{S}| \leq B$$

Information Matrix

Full set of subjects

–Shannon Differential Information Content

- $A-$Optimal: Replace $\log\det$ by trace and so on. $\log\det$ is just a placeholder.

MINLO Formulation

$$\min_{\mu_i \in \{0,1\}} \log\det\left(\sum_{i \in \mathcal{S}} \mu_i x_i {x_i}^T\right)^{-1} \quad \text{s.t.} \quad 1^T \mu \leq B$$

Convex Formulation

$$\min_{\mu_i \in [0,1]} \log\det\left(\sum_{i\in\mathcal{S}} \mu_i x_i x_i^T\right)^{-1} \quad \text{s.t.} \quad 1^T\mu \le B$$

Relax

Convex Formulation

$$\min_{\mu_i \in [0,1]} \log\det\left(\sum_{i\in\mathcal{S}} \mu_i x_i x_i^T\right)^{-1} \quad \text{s.t.} \quad {\color{blue}c^T}\mu \leq B$$

Costs

Relax

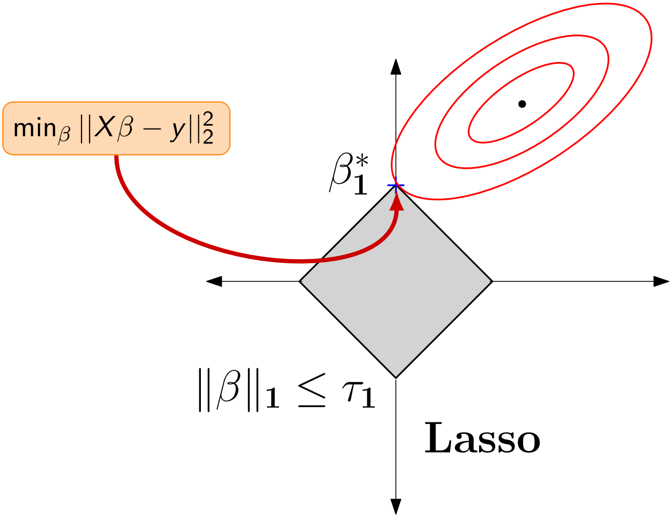
Convex Formulation

$$\min_{\mu_i \in [0,1]} \log \det \left(\sum_{i \in \mathcal{S}} \mu_i x_i x_i^T \right)^{-1} \quad \text{s.t.} \quad c^T \mu \leq B$$

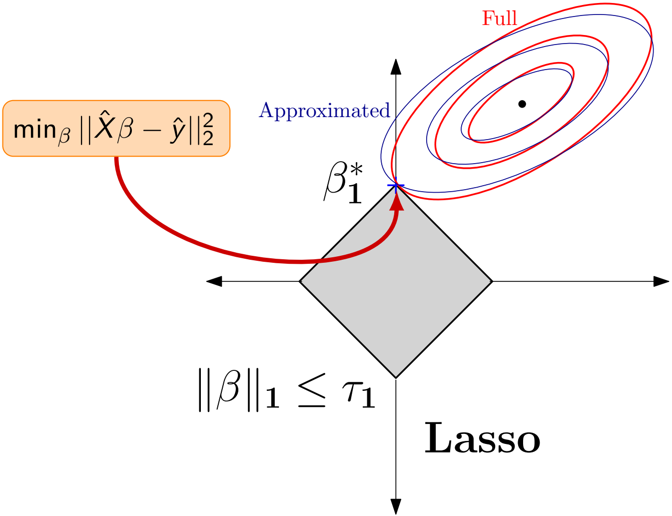
Goal: Extend to f sparse linear

Sparse f: Feature Selection

Look at constrained formulations



Look at constrained formulations



ED–S

$$\min_{\mu \in [0,1]^n} \log\det \left(\sum_{i=1}^{n} \mu_i x_i x_i^T \right)^{-1} + \lambda R_\Gamma(\Gamma_\mu) \quad \text{s.t.} \quad c^T \mu \leq B$$

ED–S

Linearity

$$\min_{\mu \in [0,1]^n} \log\det \left(\sum_{i=1}^{n} \mu_i x_i x_i^T \right)^{-1} + \lambda R_\Gamma(\Gamma_\mu) \quad \text{s.t.} \quad c^T \mu \leq B$$

ED–S

Linearity

$$\min_{\mu\in[0,1]^n} \log\det\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} + \lambda R_\Gamma(\Gamma_\mu) \quad \text{s.t.} \quad c^T\mu \le B$$

Regularization Parameter

Curvature

ED–S

$$\min_{\mu \in [0,1]^n} \log\det \left(\sum_{i=1}^{n} \mu_i x_i x_i^T \right)^{-1} + \lambda R_\Gamma(\Gamma_\mu) \quad \text{s.t.} \quad c^T \mu \leq B$$

- Γ denotes the set of eigenvectors of the full set.
- R measures the distance between the eigenvectors of the full and the selected subset.
- Explicitly...

ED–S

$$\min_{\mu \in [0,1]^n} \log\det \left(\sum_{i=1}^{n} \mu_i x_i {x_i}^T \right)^{-1} + \lambda R_\Gamma(\Gamma_\mu) \quad \text{s.t.} \quad c^T \mu \leq B$$

- Γ denotes the set of eigenvectors of the full set.
- R measures the distance between the eigenvectors of the full and the selected subset.
- Explicitly...

ED–S

$$\min_{\mu\in[0,1]^n} \log\det\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1} + \lambda R_\Gamma(\Gamma_\mu) \quad \text{s.t.} \quad c^T\mu \le B$$

- Γ denotes the set of eigenvectors of the full set.
- R measures the distance between the eigenvectors of the full and the selected subset.
- Explicitly...

ED–Spectral (ED–S) with the top eigenvector γ

$$\min_{\mu,u} \log\det\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1} + \lambda||\gamma - u||_2^2 + u^T\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1} u$$

$$\text{s.t. } 0 \le \mu \le 1, \quad c^T\mu \le B, \quad ||u||_2^2 = 1$$

ED–Spectral (ED–S) with the top eigenvector γ

Quadratic over linear

$$\min_{\mu,u} \log\det\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} + \lambda||\gamma - u||_2^2 + u^T\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} u$$

$$\text{s.t. } 0 \le \mu \le 1, \quad c^T\mu \le B, \quad ||u||_2^2 = 1$$

ED–Spectral (ED–S) with the top eigenvector γ

Quadratic over linear

$$\min_{\mu, u} \log\det\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} + \lambda||\gamma - u||_2^2 + u^T\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} u$$

$$\text{s.t. } 0 \le \mu \le 1, \quad c^T\mu \le B, \quad ||u||_2^2 = 1$$

Nonconvex

ED–S More generally

$$\min_{\mu,U} \log\det\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1} + \sum_{j=1}^{k}\lambda_j||\gamma_j - u_j||_2^2 + \mathrm{tr}\left(U^T\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1}U\right)$$

$$\text{s.t. } 0 \le \mu \le 1, \quad c^T\mu \le B, \quad U^TU = I$$

ED–S More generally

Convex Optimization with Orthogonality Constraints

$$\min_{\mu,U} \log\det\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1} + \sum_{j=1}^{k}\lambda_j||\gamma_j - u_j||_2^2 + \mathrm{tr}\left(U^T\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1}U\right)$$

$$\text{s.t. } 0 \le \mu \le 1, \quad c^T\mu \le B, \quad U^TU = I$$

ED–S Algorithm

Alternating Minimization or Batch Coordinate Descent Algorithm

$$\min_{\mu,U} \log\det\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1} + \sum_{j=1}^{k}\lambda_j||\gamma_j - u_j||_2^2 + \mathrm{tr}\left(U^T\left(\sum_{i=1}^{n}\mu_i x_i x_i^T\right)^{-1}U\right)$$

$$\text{s.t. } 0 \le \mu \le 1, \quad c^T\mu \le B, \quad U^TU = I$$

(i) Fix U, update μ

(ii) Fix μ, update U

Practical implications

- Formulated geometric intuition into a well studied optimization problem.

Synopsis

Practical implications

- Formulated geometric intuition into a well studied optimization problem.
- # Decision variables increased to $pk + n$

Practical implications

- Formulated geometric intuition into a well studied optimization problem.
- # Decision variables increased to $pk + n$
- Still nonconvex! Can we do better?

Restricted Isometry Property (RIP)

Definition

Let $X \in \mathbb{R}^{n \times p}$. For $s \geq 0$, the s-restricted isometry constant δ_s of X is the smallest nonnegative number δ such that $(1-\delta)||\beta||_2 \leq ||X\beta||_2 \leq (1+\delta)||\beta||_2$ for all $s-$sparse β, i.e., $||\beta||_0 \leq s$. If $\delta_s < 1$, then X is a $s-$restricted isometry ($s-$RIP).

Restricted Isometry Property (RIP)

Definition

Let $X \in \mathbb{R}^{n \times p}$. For $s \geq 0$, the s-restricted isometry constant δ_s of X is the smallest nonnegative number δ such that $(1-\delta)||\beta||_2 \leq ||X\beta||_2 \leq (1+\delta)||\beta||_2$ for all $s-$sparse β, i.e., $||\beta||_0 \leq s$. If $\delta_s < 1$, then X is a $s-$restricted isometry ($s-$RIP).

Theorem (Candes et al.)

Suppose X has $4s-$RIP constant $\delta_{4s} \leq \frac{1}{4}$. Let $\beta_0 \in \arg\min_\beta\{||\beta||_0 : X\beta = y\}$ and $\beta_1 \in \arg\min_\beta\{||\beta||_1 : X\beta = y\}$. If $||\beta_0||_0 \leq s$, then $\beta_1 = \beta_0$.

Restricted Isometry Property (RIP)

Definition

Let $X \in \mathbb{R}^{n \times p}$. For $s \geq 0$, the s-restricted isometry constant δ_s of X is the smallest nonnegative number δ such that $(1-\delta)||\beta||_2 \leq ||X\beta||_2 \leq (1+\delta)||\beta||_2$ for all $s-$sparse β, i.e., $||\beta||_0 \leq s$. If $\delta_s < 1$, then X is a $s-$restricted isometry ($s-$RIP).

Checking RIP is NP Hard! (Bandeira et al.)

Theorem (Candes et al.)

Suppose X has $4s-$RIP constant $\delta_{4s} \leq \frac{1}{4}$. Let $\beta_0 \in \arg\min_\beta\{||\beta||_0 : X\beta = y\}$ and $\beta_1 \in \arg\min_\beta\{||\beta||_1 : X\beta = y\}$. If $||\beta_0||_0 \leq s$, then $\beta_1 = \beta_0$.

Relax, RIP is only sufficient

"Certainly, the RIP is not the only property of a design (sensing) matrix X which allows to obtain good error bounds for ℓ_1-recovery of sparse signals"–Nemirovski

Relax, RIP is only sufficient

"Certainly, the RIP is not the only property of a design (sensing) matrix X which allows to obtain good error bounds for ℓ_1-recovery of sparse signals"–Nemirovski

Coherence:

$$\max_{i=1,\ldots,n} \operatorname{diag}\left(X\left(X^TX\right)^{-1}X^T\right)$$

ED–Incoherent (ED–I)

$$\min_{\mu} \log\det \left(\sum_{i=1}^{n} \mu_i x_i x_i^T \right)^{-1} + \lambda \max_{i} \left(\mu_i e_i^T X \left(\sum_{i=1}^{n} \mu_i x_i x_i^T \right)^{-1} X^T e_i \right)$$

$$\text{s.t. } \mu \in \{0,1\}^n, \quad c^T \mu \leq B$$

ED–I

$$\min_{\mu} \log\det\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} + \lambda \max_{i}\left(\mu_i^2 e_i^T X \left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} X^T e_i\right)$$

$$\text{s.t. } \mu \in \{0,1\}^n, \quad c^T\mu \le B$$

ED–I

Set $\mu_i e_i^T X = z_i$

$$\min_{\mu} \log\det\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} + \lambda \max_{i}\left(z_i^T\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} z_i\right)$$

$$\text{s.t. } \mu \in \{0,1\}^n, \quad c^T\mu \le B$$

ED–I

Quadratic over linear

$$\min_{\mu} \log\det\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} + \lambda \max_i \left(z_i^T \left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} z_i \right)$$

$$\text{s.t. } \mu \in [0,1]^n, \quad c^T\mu \leq B, \quad \mu_i e_i^T X = z_i \ \forall\ i = 1, \ldots, n$$

Relax

ED–I

$$\min_{\mu} \log\det\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} + \lambda \max_{i}\left(\mu_i^2 e_i^T X \left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} X^T e_i\right)$$

$$\text{s.t.} \;\; \mu \in [0,1]^n, \quad c^T \mu \leq B$$

Convex Optimization Problem with no extra variables!

ED–I Coordinate Descent Algorithm

$$\min_{\mu} \log\det\left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} + \lambda \max_i \left(\mu_i^2 e_i^T X \left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} X^T e_i\right)$$

$$\text{s.t. } \mu \in [0,1]^n, \quad c^T \mu \le B$$

for $t = 1, 2, ..., T$ do
    for $k = 1, 2, ..., n$ do
        Pick $i \in \{0, ..., n\}$, $\mu_i \leftarrow \mu_i - \eta \nabla_{\mu_i} f$
    end for
    Project μ
end for
Output: μ^*

How do we round (the fractional) μ^* with guarantees?

How do we round (the fractional) μ^* with guarantees?

Is Pipage Rounding scheme applicable?

- For any feasible μ, need a vector ν, $\delta, \tau > 0$ s.t. $\mu + \delta\nu$ or $\mu - \tau\nu$ is more integral.
- f must be convex in the direction ν.
- Need μ s.t. $f(\mu) \leq \kappa \cdot \text{opt}$

How do we round (the fractional) μ^* with guarantees?

Is Pipage Rounding scheme applicable?

- For any feasible μ, need a vector ν, $\delta, \tau > 0$ s.t. $\mu + \delta\nu$ or $\mu - \tau\nu$ is more integral.
- f must be convex in the direction ν. ✓
- Need μ s.t. $f(\mu) \leq \kappa \cdot \text{opt}$ ✓

Convexity

How do we round (the fractional) μ^* with guarantees?

Is Pipage Rounding scheme applicable?

- For any feasible μ, need a vector ν, $\delta, \tau > 0$ s.t. $\mu + \delta\nu$ or $\mu - \tau\nu$ is more integral. ✓
- f must be convex in the direction ν. ✓
- Need μ s.t. $f(\mu) \leq \kappa \cdot \text{opt}$ ✓

Choose $e_{i_f} - e_{j_f}$

How do we round (the fractional) μ^* with guarantees?

Is Pipage Rounding scheme applicable?

- For any feasible μ, need a vector ν, $\delta, \tau > 0$ s.t. $\mu + \delta\nu$ or $\mu - \tau\nu$ is more integral. ✓
- f must be convex in the direction ν. ✓
- Need μ s.t. $f(\mu) \leq \kappa \cdot \text{opt}$ ✓

Efficiently implementable/Derive approximation ratios. ✓

Experiments

Evaluations setup

- Evaluate the consistency of both the models: Do ED-I and ED-S agree?
- Sensitivity analysis with respect to the budget.
- Error in feature selection with respect to the full dataset.

Experiments

Evaluations setup

- Evaluate the consistency of both the models: Do ED-I and ED-S agree?
- Sensitivity analysis with respect to the budget.
- Error in feature selection with respect to the full dataset.

Experiments

Evaluations setup

- Evaluate the consistency of both the models: Do ED-I and ED-S agree?
- Sensitivity analysis with respect to the budget.
- Error in feature selection with respect to the full dataset.

Experiments

Evaluations setup

- Evaluate the consistency of both the models: Do ED-I and ED-S agree?
- Sensitivity analysis with respect to the budget.
- Error in feature selection with respect to the full dataset.

Need f to be sparse linear

Experiments

Evaluations setup

- Evaluate the consistency of both the models: Do ED-I and ED-S agree?
- Sensitivity analysis with respect to the budget.
- Error in feature selection with respect to the full dataset.

Need f to be sparse linear – use standard datasets.

Standard Dataset

prostate

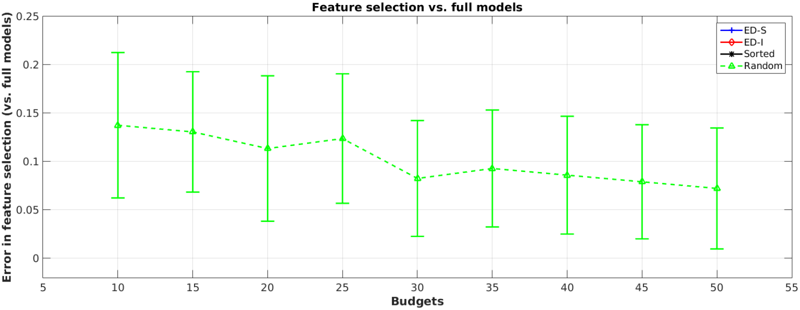
x : cancer volume, prostate weight, age, Gleason score etc.

y : prostate specific antigen
n : 97

Standard Dataset

prostate

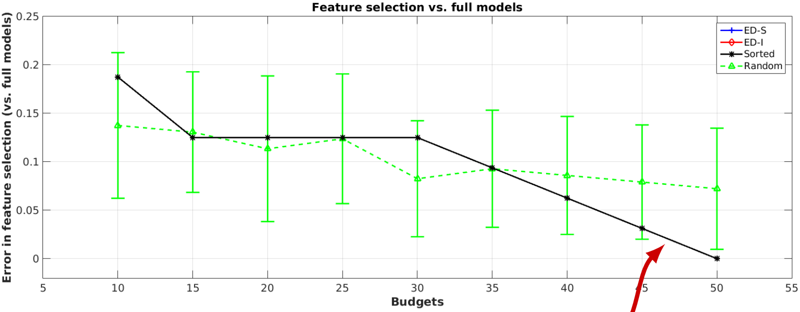
Sensitivity with Budgets



Standard Dataset

prostate

Sensitivity with Budgets

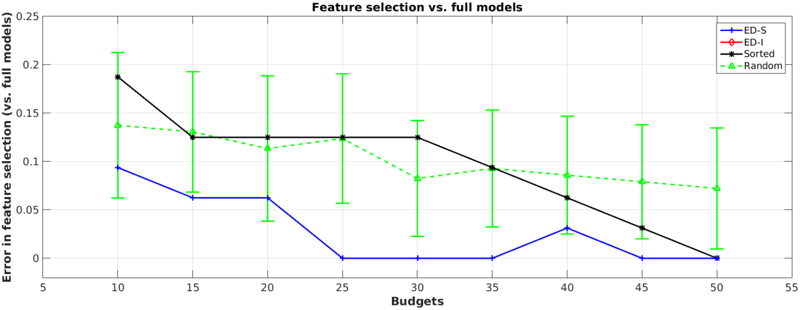


sorted $\approx$ $k-$means

Standard Dataset

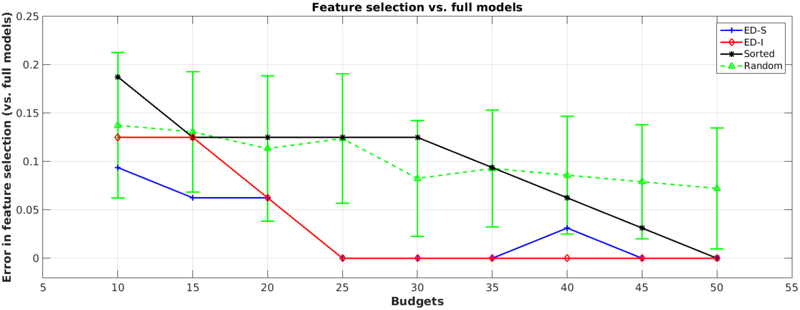
prostate

Sensitivity with Budgets



prostate

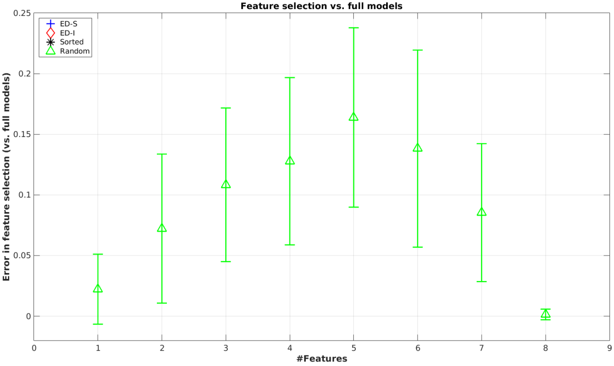
Sensitivity with Budgets



Standard Dataset

prostate

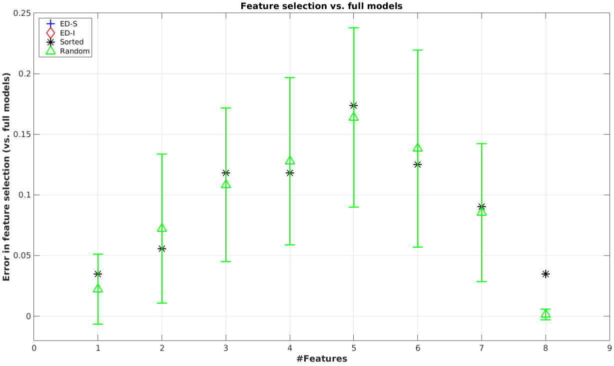
Feature selection errors



Standard Dataset

prostate

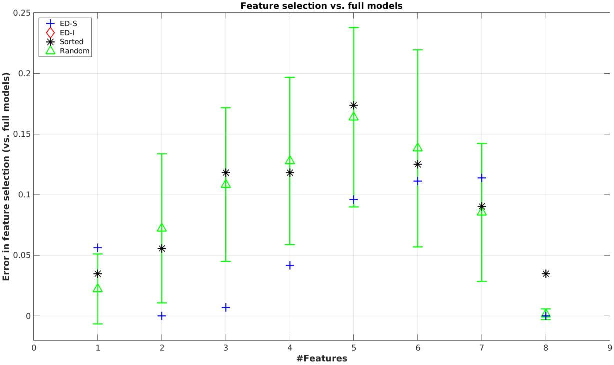
Feature selection errors



Standard Dataset

prostate

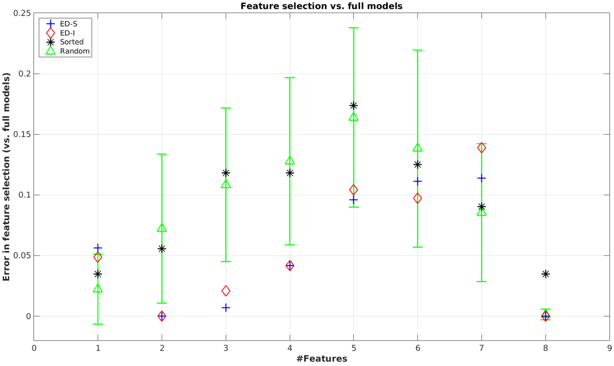
Feature selection errors



Standard Dataset

prostate

Feature selection errors

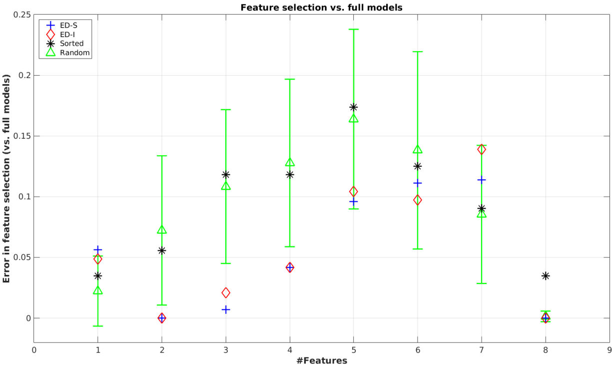


Standard Dataset

Similar results on lars dataset

prostate

Feature selection errors



Neuroimaging Dataset

ADNI: $x_i \in \mathbb{R}^{118}$, $n = 985$

x : Image derived ROI summaries from PET

y : 1. ADAS, 2. CDR

Neuroimaging Dataset

ADNI: $x_i \in \mathbb{R}^{118}$, $n = 985$

x : Image derived ROI summaries from PET

y : 1. ADAS, 2. CDR

Offers significant cost savings.

Neuroimaging Dataset

ADNI: $x_i \in \mathbb{R}^{118}$, $n = 985$

x : Image derived ROI summaries from PET

y : 1. ADAS, 2. CDR

Offers significant cost savings.

But first...

- Classical non-nested hypothesis testing (like F, χ^2) are not applicable. Have to use goodness of fit measures.
- Full dataset might not be the *ground truth*. Have to use change in goodness of fit measures.

Neuroimaging Dataset

ADNI: $x_i \in \mathbb{R}^{118}, n = 985$

x : Image derived ROI summaries from PET

y : 1. ADAS, 2. CDR

Offers significant cost savings.

But first...

- Classical non-nested hypothesis testing (like F, χ^2) are not applicable. Have to use goodness of fit measures.
- Full dataset might not be the *ground truth*. Have to use change in goodness of fit measures.

Neuroimaging Dataset

ADNI: $x_i \in \mathbb{R}^{118}$, $n = 985$

x : Image derived ROI summaries from PET

y : 1. ADAS, 2. CDR

Offers significant cost savings.

But first...

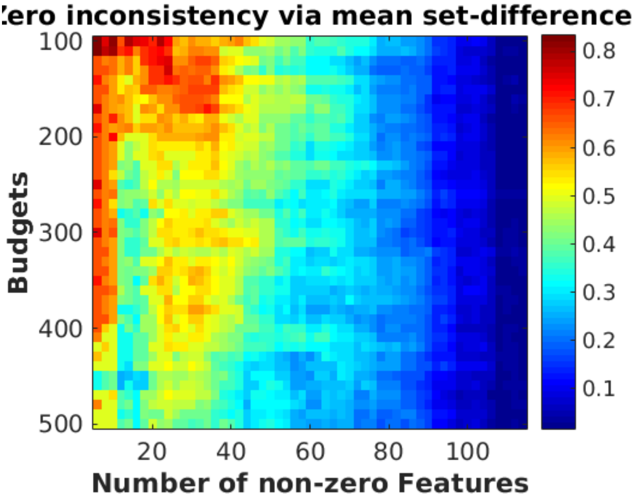
- Classical non-nested hypothesis testing (like F, χ^2) are not applicable. Have to use goodness of fit measures.
- Full dataset might not be the *ground truth*. Have to use change in goodness of fit measures.

Neuroimaging Dataset

ADNI

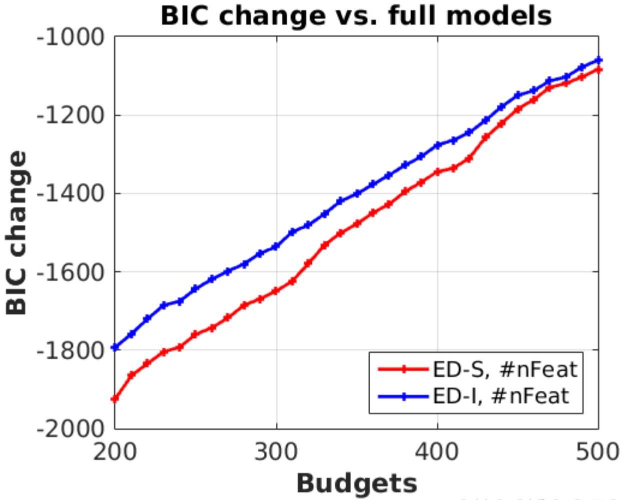
Neuroimaging Dataset

ADNI



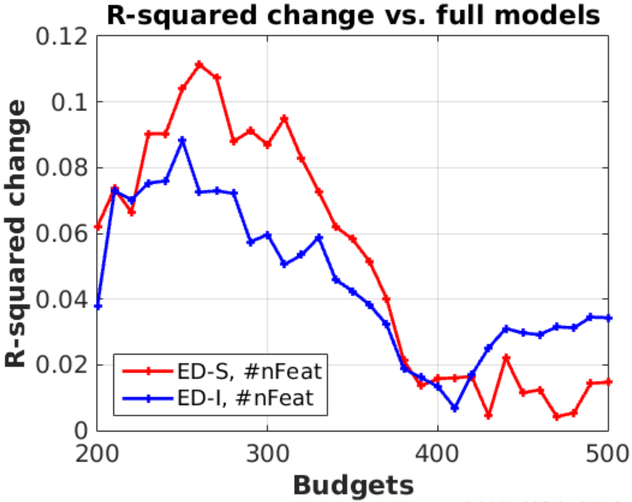
Neuroimaging Dataset

ADNI

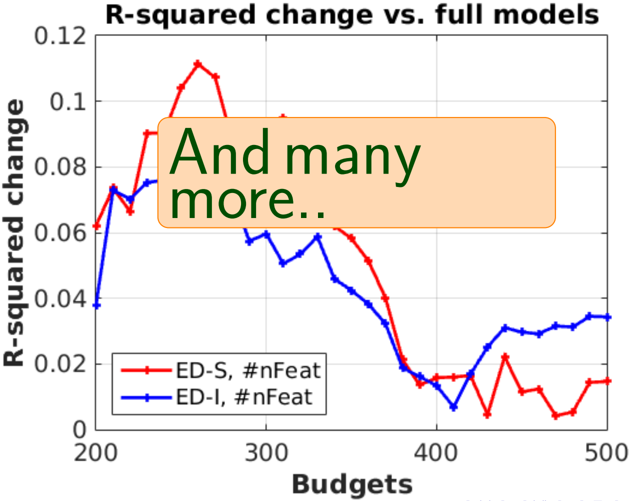


Neuroimaging Dataset

ADNI



ADNI



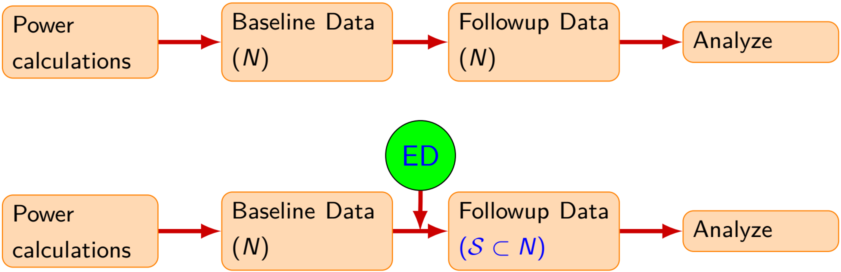
And many more..

Conclusions

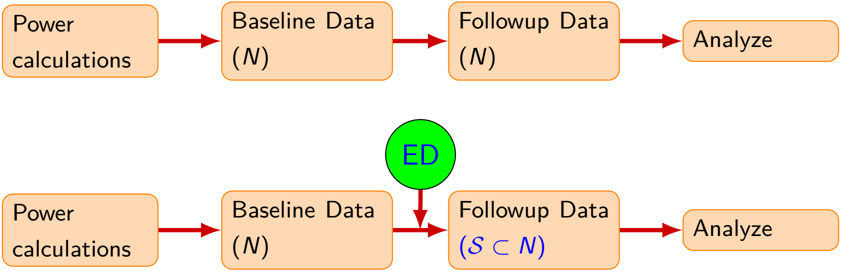
Conclusions

Power calculations → Baseline Data (N) → Followup Data (N) → Analyze

Conclusions



Conclusions



- Proposed algorithmically efficient formulations for Experimental Design for Sparse Linear Model which are effective practically as well.
- Future direction: Theory for adaptive statistical power calculations using this framework.

The end...

Thank you!
Questions?

Code is publicly available at https://github.com/sravi-uwmadison/Exp_design_sparse.
See you at the poster session (today 3 pm - 7 pm)!

We are supported by NIH R01 AG040396 (VS), NSF CAREER RI 1252725, NIH R01 AG027161 (SCJ), NSF CCF 1320755, UW ADRC (P50 AG033514) and UW CPCP (U54 AI117924).