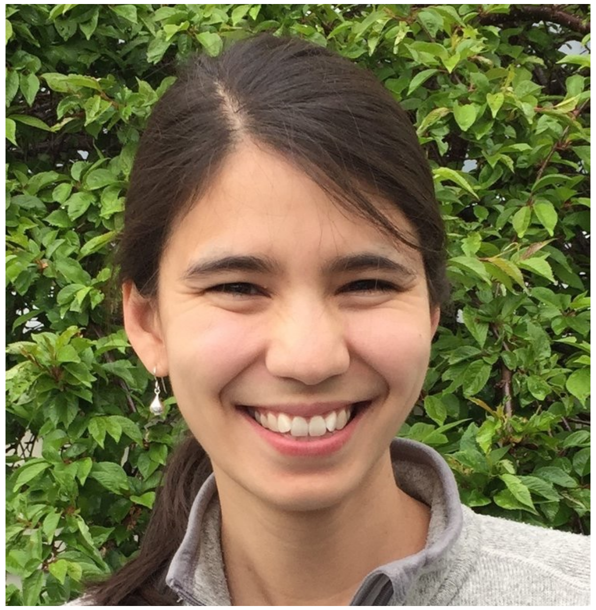
High-dimensional hierarchical modeling with exchangeability of effects across covariates

Brian L. Trippe
Postdoctoral Fellow
Columbia University

Tamara Broderick

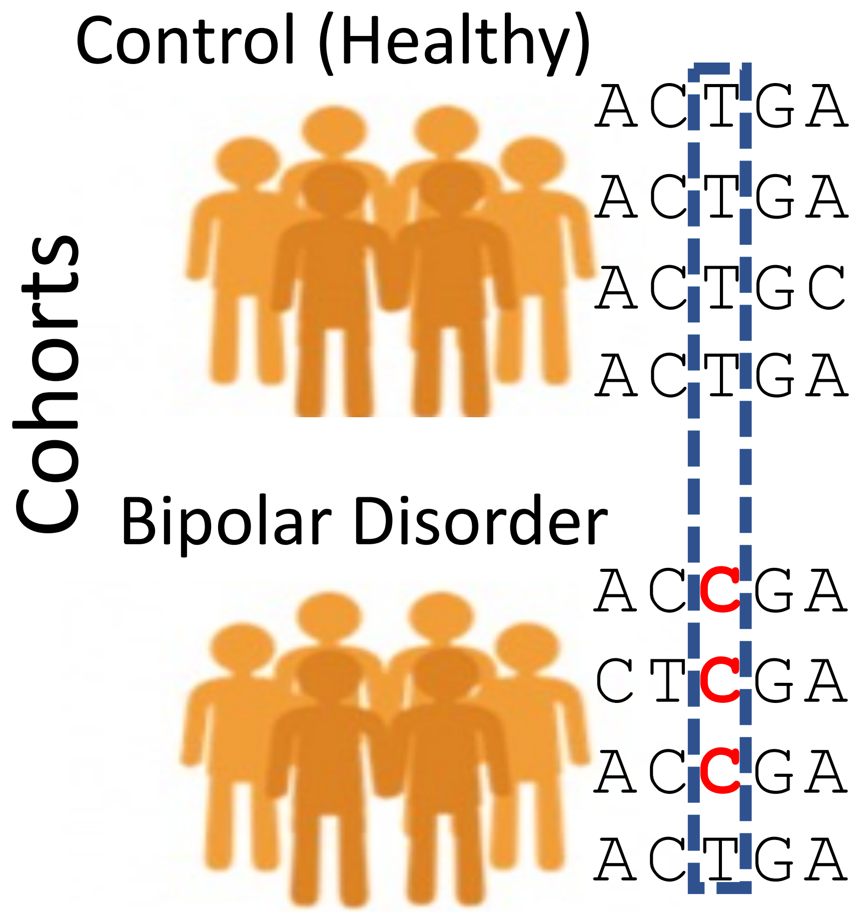


Hilary Finucane

Hierarchical Linear Modeling in High Dimensions

Example: How do differences in genetics impact Bipolar disorder?

Goal: Understand the many contributing factors → linear models

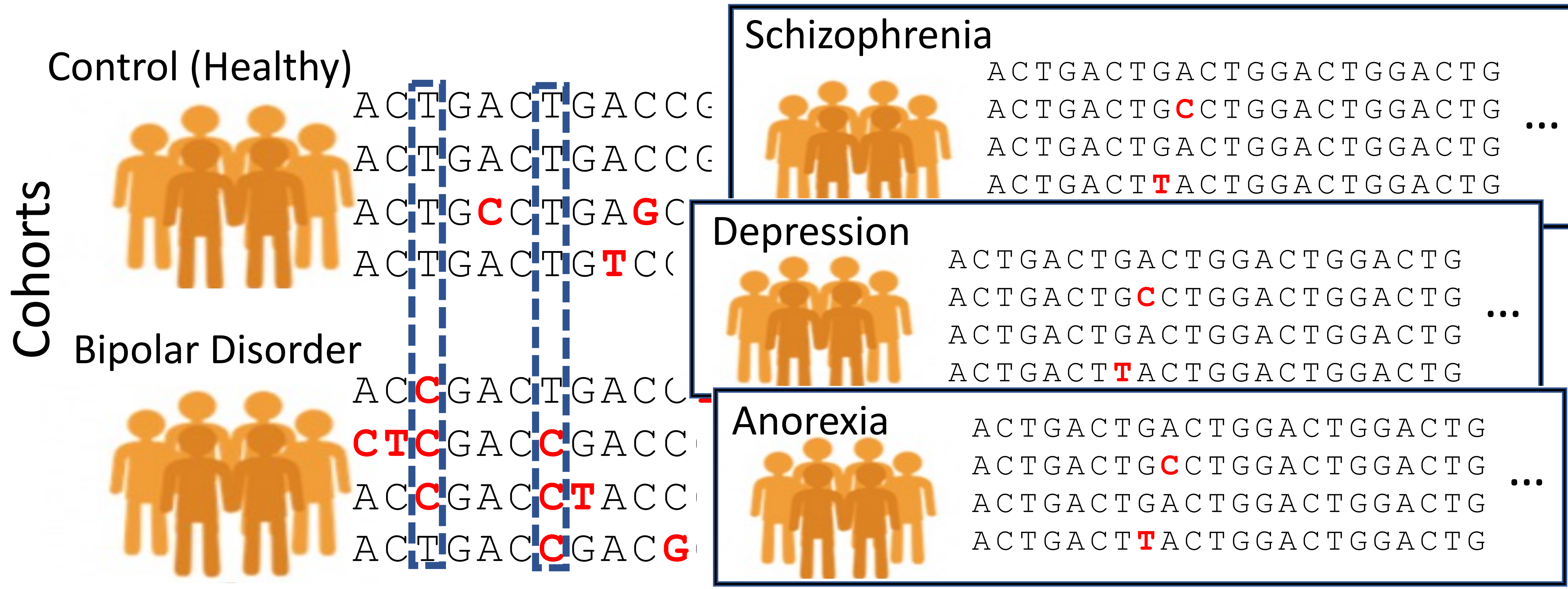


Challenges:

Hierarchical Linear Modeling in High Dimensions

Example: How do differences in genetics impact Bipolar disorder?

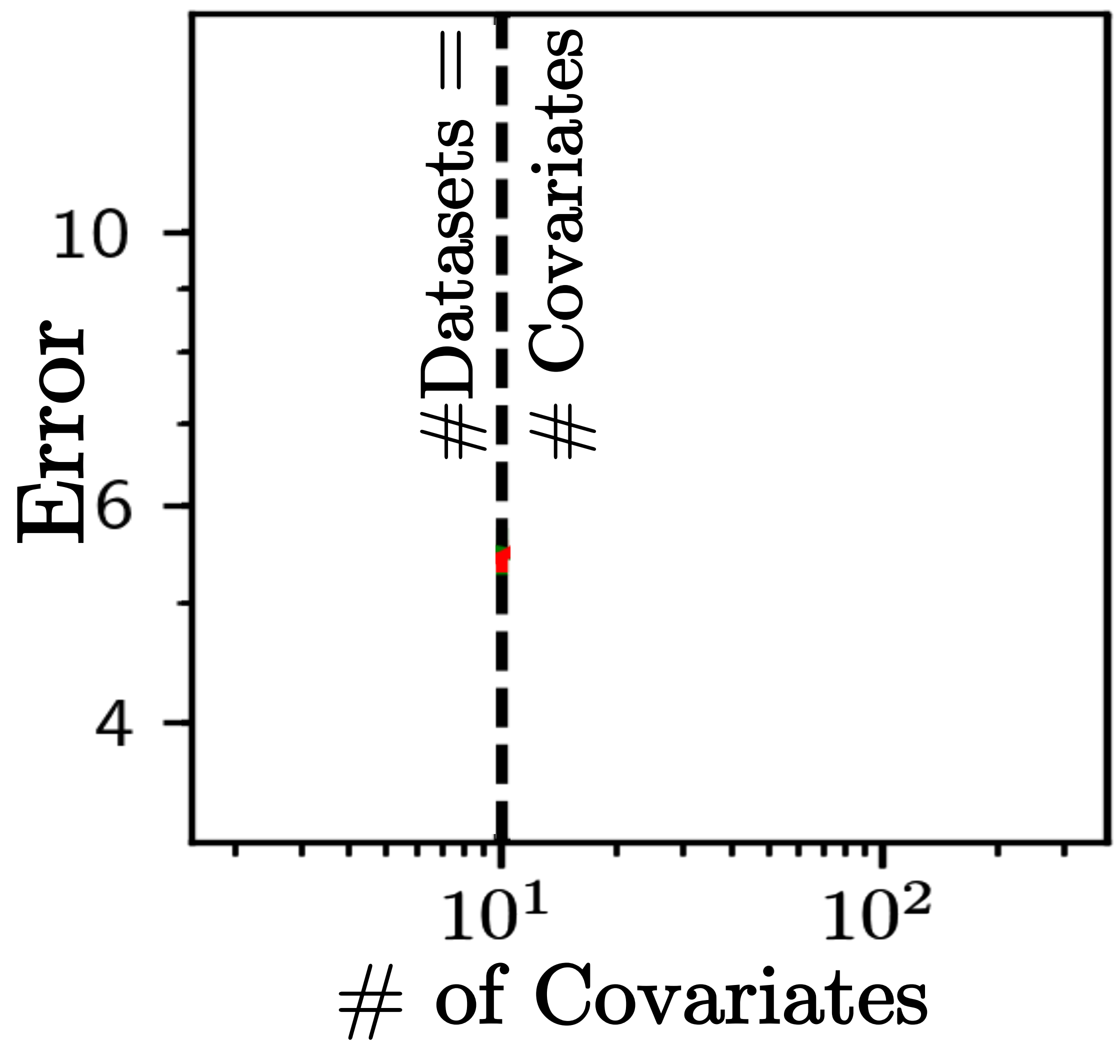
Goal: Understand the many contributing factors → linear models



Challenges: Uncertainty, multiple groups of data → hierarchical Bayes

This Talk: In high-dimensions (#Covariates > # Datasets)

Hierarchical Linear Modeling in High Dimensions



This Talk: In high-dimensions (#Covariates > # Datasets)

1. Standard approach (e.g. `lme4`) fails (worse than non-hierarchical!)
2. Unconventional use of exchangeability is more intuitive & accurate

Roadmap

- Background & Notation
  - Linear models
  - Bayesian inference
  - Modeling in high dimensions
- Our method: exchangeability of effects across covariates (rather than within datasets)
- Fast algorithms for inference in the new model
- Benefits of our method in high dimensions (theory and empirics)

Background and Notation: Linear Modeling

Example in education: Relate student participation in free lunch program to academic performance.

For each student $n = 1,2,\ldots,N$

"Effect"

Change in Performance ("Response") → $$Y_n = X_n \beta + \epsilon_n$$ ← Other Factors ("Residual")

Participation ("Covariate") $= \begin{cases} 1 & \text{if in program} \\ 0 & \text{otherwise} \end{cases}$

What if we have data from multiple schools?
(e.g. in Cambridge, Boston and Dallas)

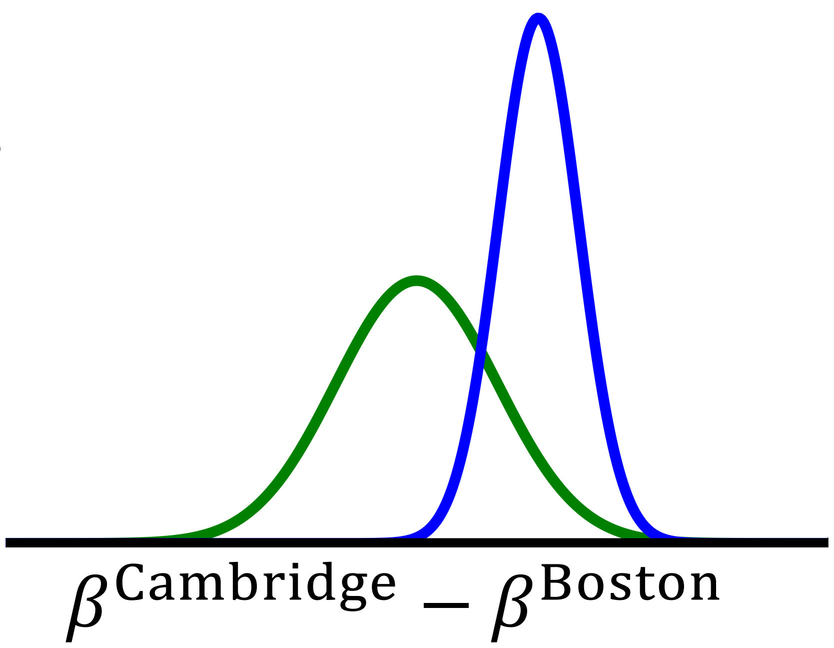
Analysis Options:

1. Combine all data together -- ignores differences
2. Analyze independently -- worse performance if data limited
3. Partial pooling via hierarchical Bayesian modeling

Background and Notation: Bayesian Inference

Prior $p(\boldsymbol{\beta} \mid \boldsymbol{Y}) \propto p(\boldsymbol{\beta})\, p(\boldsymbol{Y} \mid \boldsymbol{\beta})$ **Likelihood**

- Subjective beliefs before seeing data → probabilities
- Codify assumptions about dataset similarity



Posterior

- Bayes Rule: update beliefs after seeing data
- Computational step (requires algorithms)

Empirical Bayes

- Use data to automate choice of prior
- "Learn" extent of partial pooling, less subjective

Background and Notation: Multiple Covariates

What if we have multiple covariates for each student?

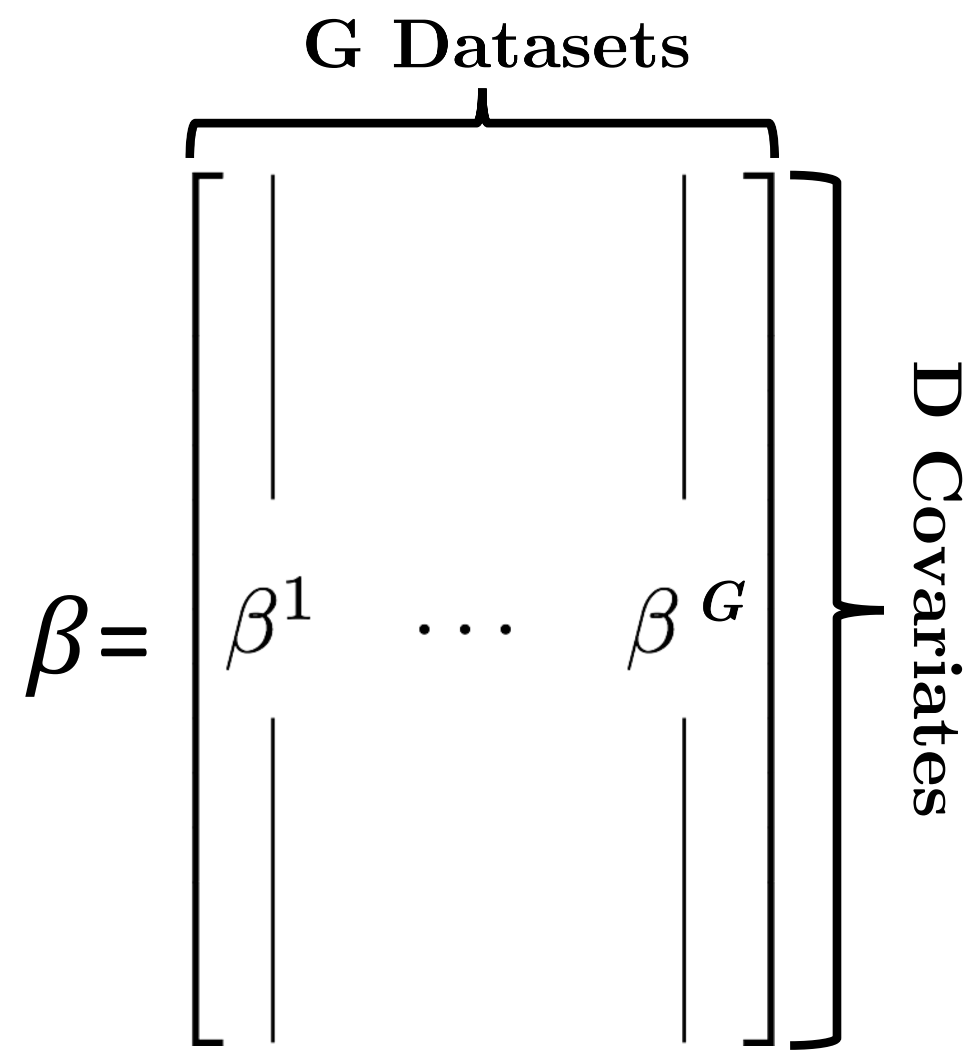
- E.g. playing a sport, past performance, demographics

- For each school $g = 1,2,\ldots,G$ and each student $n = 1,2,\ldots,N^g$

$$Y_n^g = \sum_{d=1}^{D} X_{n,d}^g \beta_d^g + \epsilon_n^g$$

Response: Y_n^g; Covariates: $X_{n,d}^g$; Effects: β_d^g; Residual: ϵ_n^g

D = # Covariates (student attributes)
G = # Datasets (schools)
N^g = # Samples in dataset g (students)



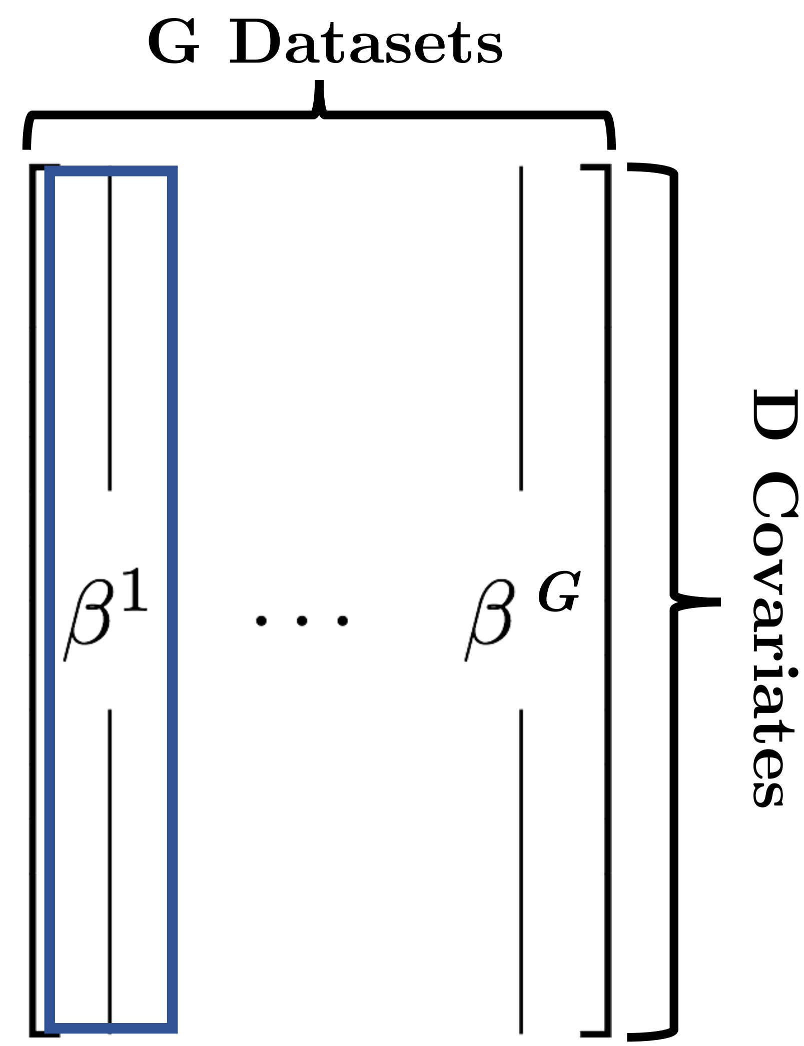
Question: What prior do we put on this matrix?

Choosing $p(\beta)$: Exchangeability Across Datasets vs. Covariates

Standard approach (Lindley and Smith, 1972)

- Assume exchangeability across datasets
- Model correlations in β across covariates

  ← Γ (D×D matrix)



<u>More formally:</u> Assume "exchangeability"

β is *a priori* exchangeable across datasets if for every G-permutation σ,

$$p(\beta^1, \beta^2, \dots, \beta^G) = p\left(\beta^{\sigma(1)}, \beta^{\sigma(2)}, \dots, \beta^{\sigma(G)}\right).$$

- **De Finetti**: model β^g's as *conditionally i.i.d.*
- Convenient choice: $\beta^g \sim N(\xi, \Gamma)$

  (via empirical Bayes)
- Ubiquitous in software (`lme4`) and pedagogy

  [Bates et al., 2015] **[Gelman, et al., 2013]**

Limitations when D>>G

- Less intuitive (Cambridge, Boston & Dallas are not equally similar)
- O(D^2) parameters [statistical & computational]
- Poor estimation accuracy

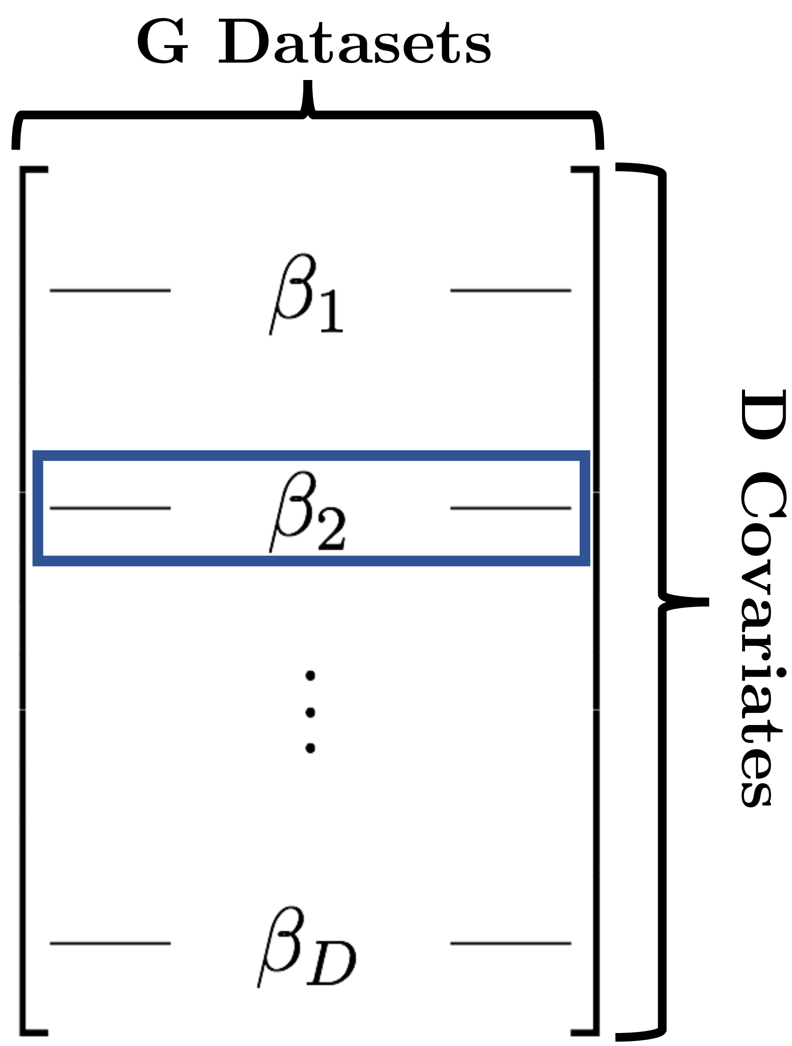
Choosing $p(\beta)$: Exchangeability Across Datasets vs. Covariates

Standard approach (Lindley and Smith, 1972)

- Assume exchangeability across datasets
- Model correlations in β across covariates ← Γ (D×D matrix)
- Specific choice: $\beta^g \sim N(\xi, \Gamma)$

Our approach [TFB2021]

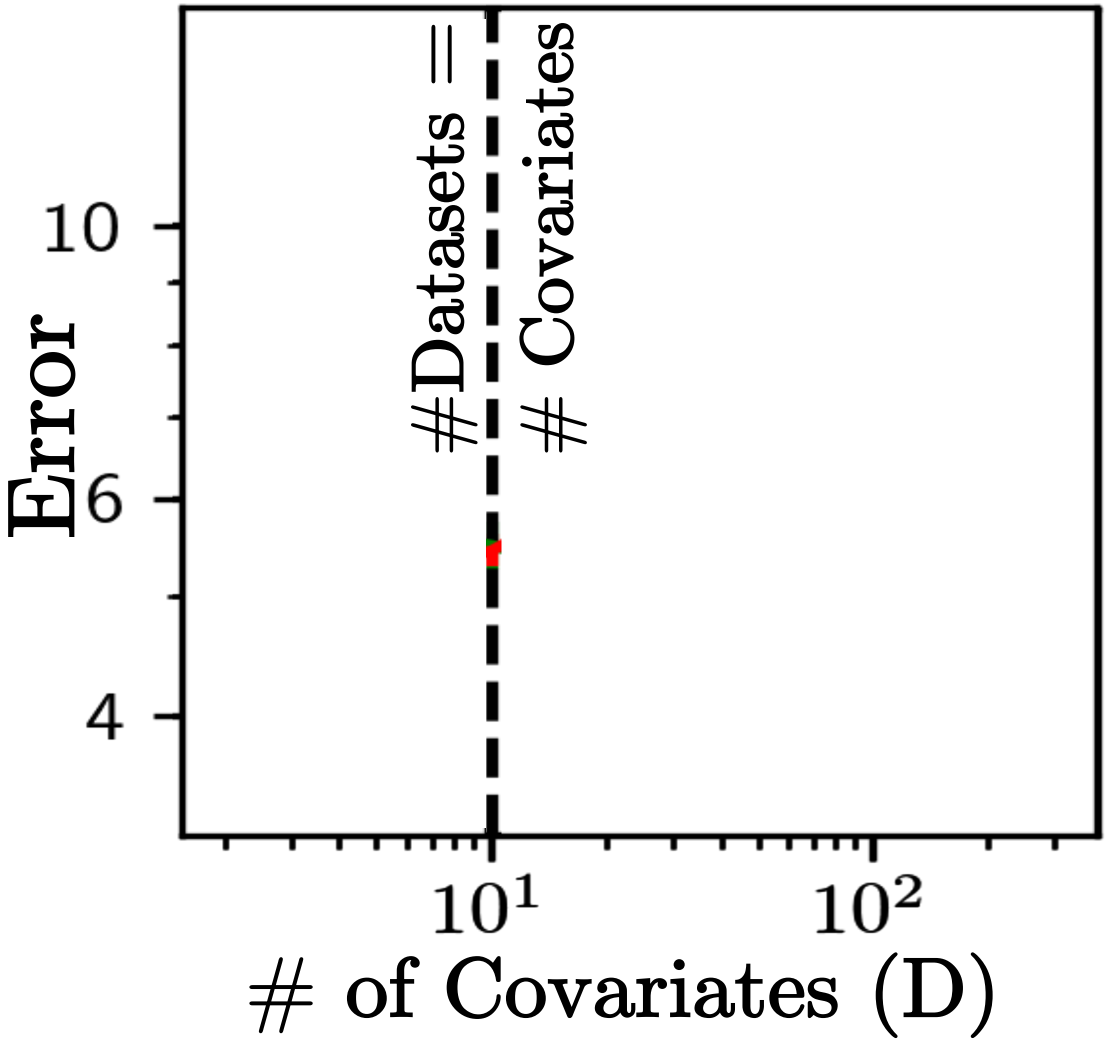
- Assume exchangeability across *covariates*
- Model correlations in β across *datasets* ← Σ (G×G matrix)
- Specific choice: $\beta_d \sim N(\mu, \Sigma)$



Details to fill in to use the new model:

- Need practical algorithms: posterior inference, empirical Bayes
- Need theory & experiments: justify whether this is effective

Choosing $p(\beta)$: Correlations Across Datasets vs. Covariates



In high dimensions (D>G)

- **Standard approach** does worse than independent analyses.
  - `lme4` does not run when D>G
- **Exchangeable across covariates** effectively shares information.

- **Though conceptually similar, different dependence on dimension**

Roadmap

- Background & Notation
  - Linear models
  - Bayesian inference
  - Modeling in high dimensions
- Our method: models correlations across datasets (rather than within datasets)
- Fast algorithms for inference in the new model
- Benefits of our method in high dimensions (theory and empirics)

Inference under Exchangeability Across Covariates

Prior:

For each covariate d:

$$\beta_d \sim N(0, \Sigma)$$

Likelihood:

For dataset g and datapoint n:

$$Y_n^g \mid \beta^g \sim N(X_n^g \beta^g, \sigma^2)$$

D = # Covariates
G = # Datasets
N^g = # Samples in dataset g

Posterior:

Estimate β as posterior mean:

$$\hat{\beta}_{\text{ECov}} = \int \beta \; p(\beta \mid Y)\, d\beta$$

G Datasets

$$\begin{bmatrix} \text{—} & \beta_1 & \text{—} \\ \text{—} & \beta_2 & \text{—} \\ & \vdots & \\ \text{—} & \beta_D & \text{—} \end{bmatrix}$$

D Covariates

Gaussian conjugacy → analytic form for $\hat{\beta}_{\text{ECov}}$

G·D Effects

$$\begin{bmatrix} \hat{\beta}^1_{\text{ECov.}} \\ \vdots \\ \hat{\beta}^G_{\text{ECov.}} \end{bmatrix} = \left(\sigma^2 \begin{bmatrix} \Sigma_{1,1} I_D & \cdots & \Sigma_{1,G} I_D \\ \vdots & \ddots & \vdots \\ \Sigma_{G,1} I_D & \cdots & \Sigma_{G,G} I_D \end{bmatrix}^{-1} + \begin{bmatrix} X^{1\top} X^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X^{G\top} X^G \end{bmatrix} \right)^{-1} \begin{bmatrix} X^{1\top} Y^1 \\ \vdots \\ X^{G\top} Y^G \end{bmatrix}$$

[Bishop, 2006 – Chapter 3.3]

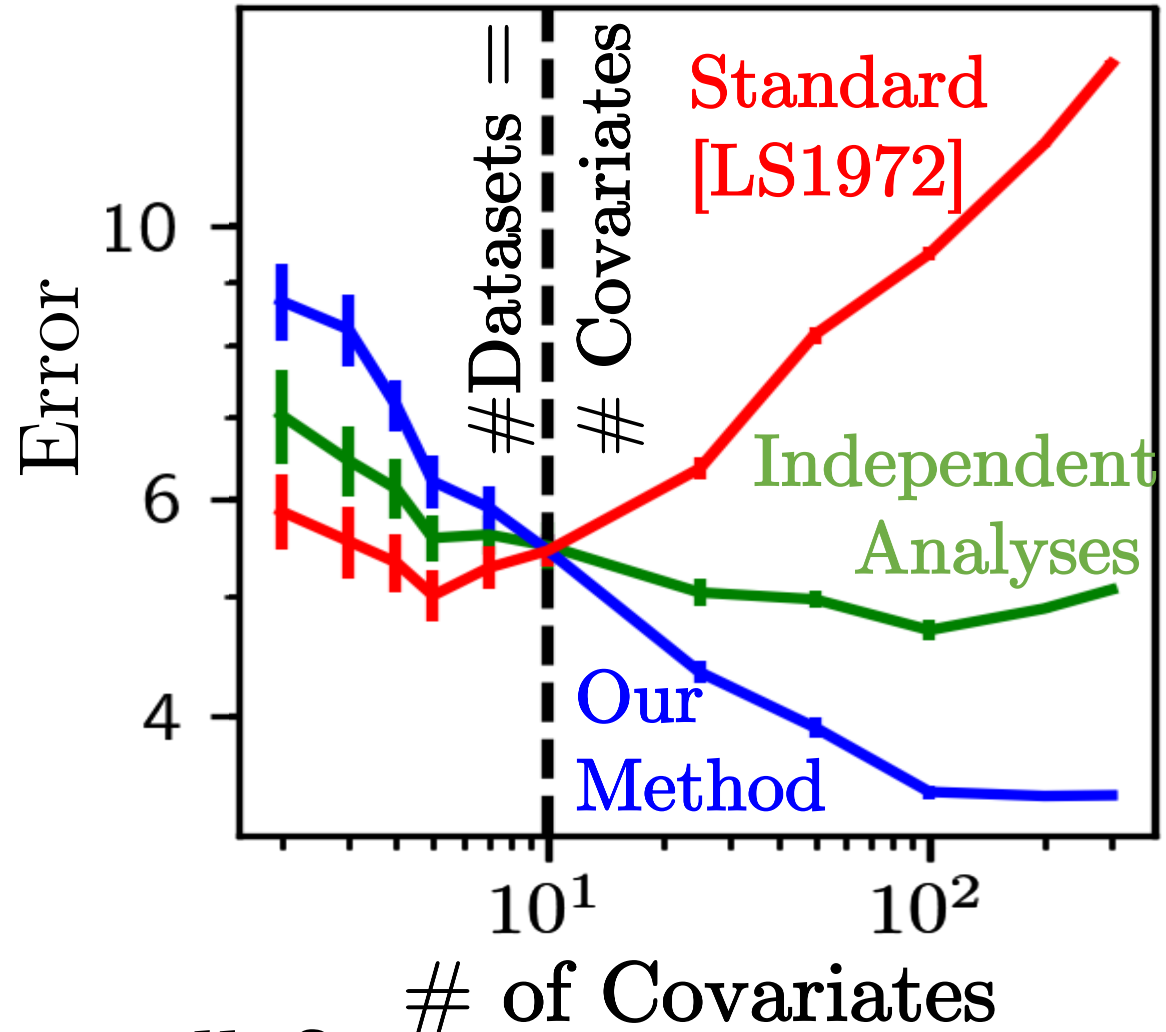
- Catch: High-dimensional linear system with G·D parameters
- We show: tractable via the conjugate gradient algorithm

Use empirical Bayes to estimate dataset relatedness

$$\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \boldsymbol{p}(\boldsymbol{Y}^1, \boldsymbol{Y}^2, \ldots, \boldsymbol{Y}^G \mid \boldsymbol{\Sigma})$$

- We develop an expectation maximization algorithm

Is this new method better in high dimensions?



Estimation Procedure

1. Empirical Bayes (EM to choose Σ).
2. Posterior Inference: $\hat{\beta}_{\text{ECov}} = \mathbb{E}[\beta | Y; \Sigma]$

Simulation Set-Up

1. Draw effects from prior: $\beta_d \sim N(0, \Sigma)$
2. Sample $Y \sim p(Y|\beta)$, many times
3. Estimate $\underbrace{\mathbb{E}_{Y|\beta}\left[\sum_g \sum_d (\hat{\beta}_d^g - \beta_d^g)^2\right]}_{\text{"Risk": } R(\hat{\beta}, \beta)}$

How do we assess if this works more generally?

~~**Idea 1:** Simulate for various β.~~ Infinitely many β – can't try them all!

Idea 2: Use real data. We'll get there, but same problem.

Idea 3: Use theory! Under what conditions on β can we *prove* $R(\hat{\beta}_{\text{ECov}}, \beta)$ is small?

Challenge: $R(\hat{\beta}_{\text{ECov}}, \beta)$ is the integral of non-differentiable function of a matrix.

Is this new method better in high dimensions?

Theorem (Domination over Least Squares) [TFB2021]:
If $D > 2G + 2$, and each X^g is well-conditioned, then for any β

$$R(\hat{\beta}_{\text{ECov}}, \beta) < R(\hat{\beta}_{\text{LeastSquares}}, \beta) < R(\hat{\beta}_{\text{EData}}, \beta).$$

- In high dimensions, $\hat{\beta}_{\text{ECov}}$ does well
  - Better to capture correlations across datasets
- Our approach reduces risk regardless of β!

Still unresolved: Risk improvement size? Boost from combining groups?

- Consider $R(\hat{\beta}_{\text{ECov}}, \beta) - R(\hat{\beta}_{\text{ECovIndep}}, \beta)$

[$\hat{\beta}_{\text{ECov}}$ on each dataset separately]

But... $R(\hat{\beta}_{\text{ECov}}, \beta)$ depends on non-central Wishart eigenvalues – Intractable!

Make comparison tractable by reformulating problem:

- Consider asymptotics in # of covariates ($D \to \infty$).
- Bayesian analysis: $\beta_d \sim N(0, \Sigma^*)$, consider $R_{\Sigma^*}(\hat{\beta}) := \mathbb{E}[R(\hat{\beta}, \beta)]$

Theorem (Asymptotic Gain of Joint Modeling) [TFB2021]:

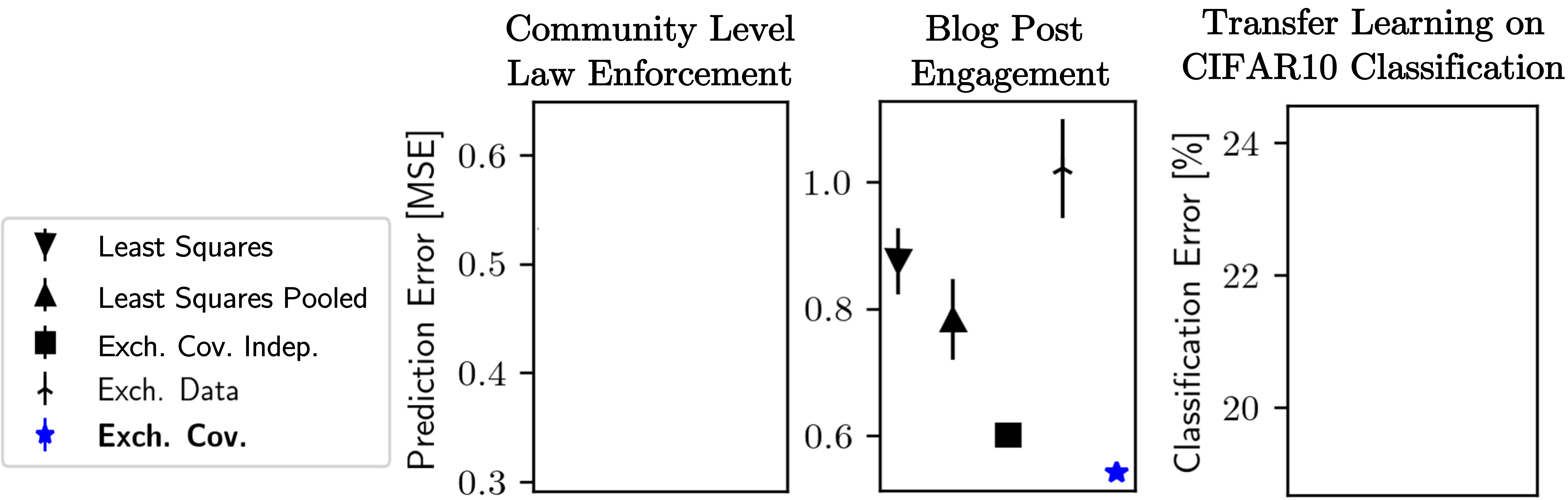
$$\lim_{D \to \infty} \frac{R_{\Sigma^*}(\hat{\beta}_{\text{ECovIndep}}) - R_{\Sigma^*}(\hat{\beta}_{\text{ECov}})}{D} \geq 0 \frac{||diag(\Sigma^*)^{\downarrow} - \lambda(\Sigma^*)^{\downarrow}||_2^2}{(1 + ||\Sigma^*||_2)^3} \geq 0$$

- Distance between the eigenvalues vs. diagonals of Σ^* determines sharing.

Exch. Cov. Performance in Diverse Applications

Challenge: in real data – can't check accuracy of effect estimation.

- We use prediction performance as a proxy for estimation
- We evaluate Mean Squared Error [MSE] with 5-fold cross validation



N ≈ 500 Communities
D = 117 Demographic Attributes
G = 4 Regions/Charges

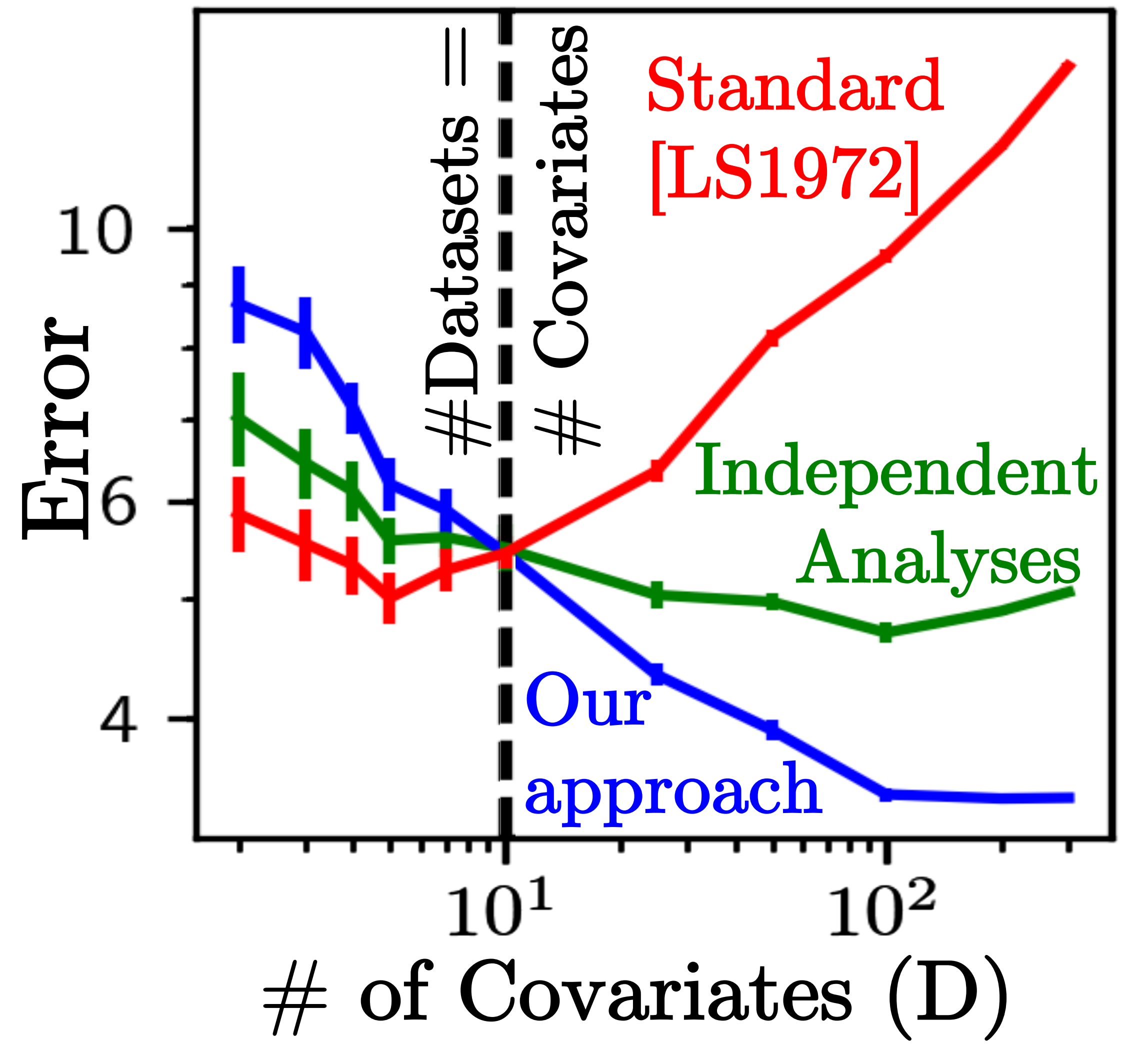
N ≈ 500 Blog posts
D = 279 Characteristics of posts
G = 3 Corpora

N ≈ 100 Images
D = 128 Dimensions of latent embedding
G = 8 Tasks

In diverse applications, exchangeability across covariates improves predictions.

Conclusions

Today: I showed modeling correlations across datasets performs better in high dimensions.



Primary Reference:

Trippe, Finucane, Broderick (2021) "For high-dimensional hierarchical models, consider exchangeability of effects across covariates instead of across datasets" In Neural Information Processing Systems

Contact me: blt2114@columbia.edu