NIPS 2010

A Dirty Model for Multitask Learning

Ali Jalali, Pradeep Ravikumar, Sujay Sanghavi and Chao Ruan

UT Austin

Motivation

- Modern Settings: High-Dimensional Problems
  - number of observations $n \ll$ number of variables p
  - Biology, Vision, Nanotechnology, Financial Analysis, ...

- Low-Dimensional Structure only hope for consistency?
  - Sparsity, Block Sparsity, Low-Rank, Graphical model Structure

This talk:

- Superposition of structures: also low-dimensional but can model surprisingly "dirty" data.

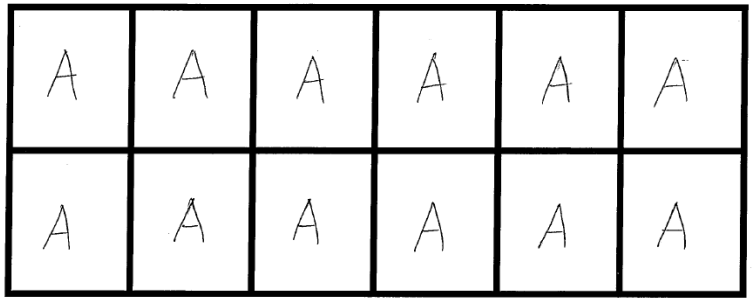
Multitask Learning

- Multiple tasks with some "shared" structure

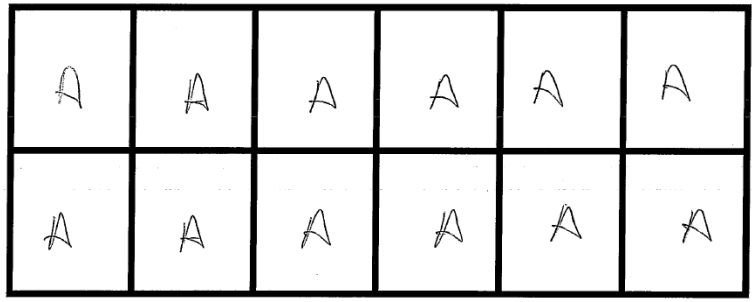
Problem:

- Learn tasks jointly (as opposed to separately)
  - e.g. Optical Character Recognition (OCR)

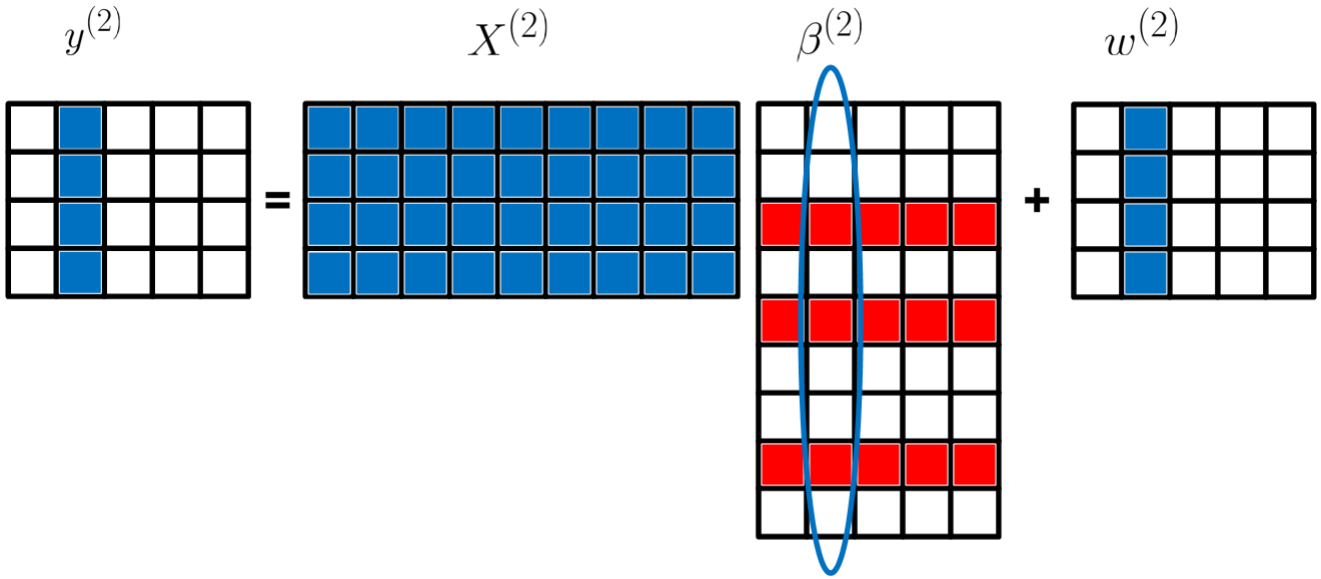
Writer 1



Writer 2

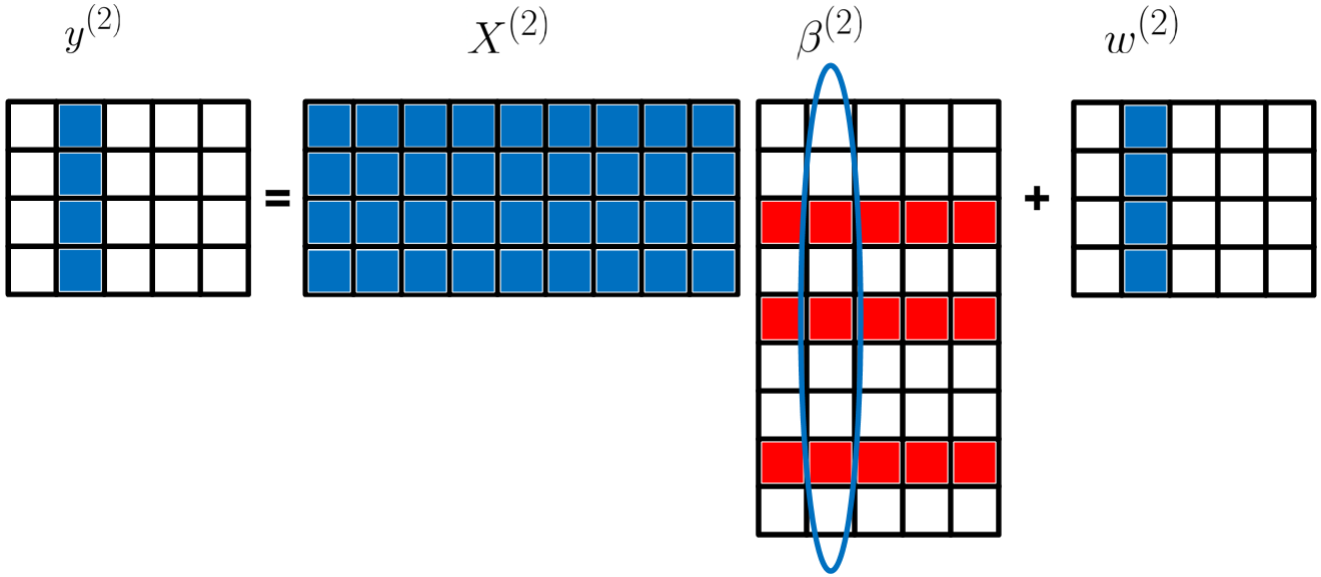


Multiple Linear Regressions



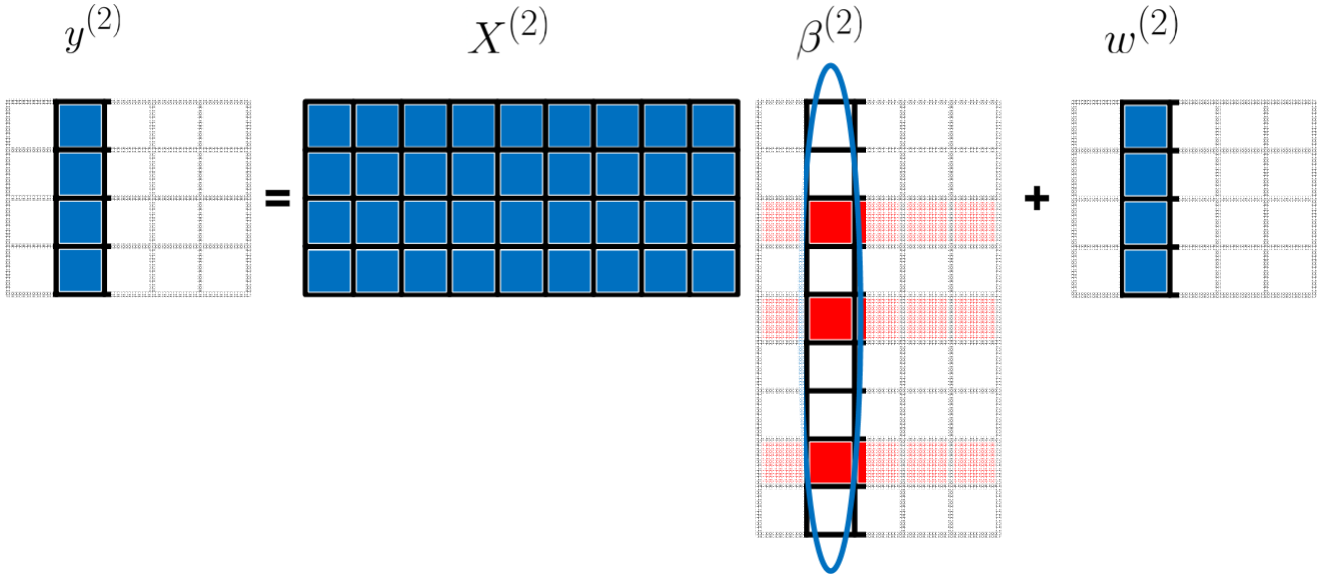
- Linear Model: $y^{(k)} = X^{(k)}\beta^{(k)} + w^{(k)}$ for all tasks $1 \leq k \leq r$
- **Problem:** Estimate β given n_k samples of $X_i^{(k)}$, $y_i^{(k)}$

Leveraging Sparsity



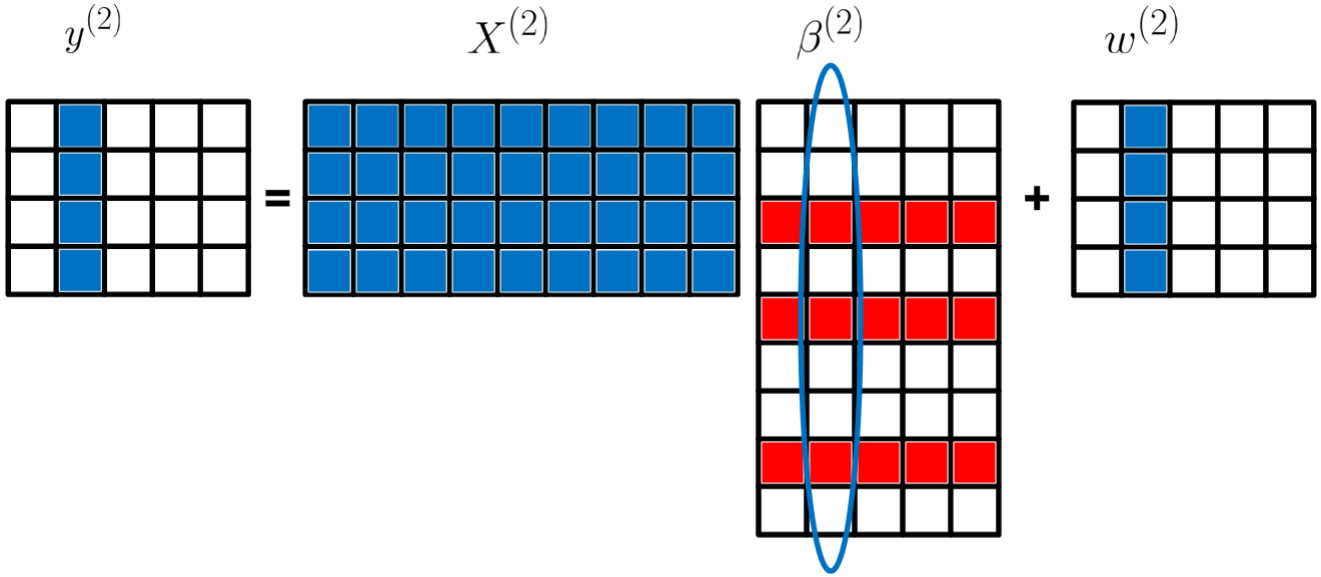
- Sparse structure: few features are relevant
- LASSO $$\min_{\beta^{(k)}} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| y_i^{(k)} - X_i^{(k)} \beta^{(k)} \right\|_2^2 + \lambda_k \left\| \beta^{(k)} \right\|_1$$

Leveraging Sparsity



- Sparse structure: few features are relevant
- LASSO $$\min_{\beta^{(k)}} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| y_i^{(k)} - X_i^{(k)} \beta^{(k)} \right\|_2^2 + \lambda_k \left\| \beta^{(k)} \right\|_1$$

Leveraging block-sparsity

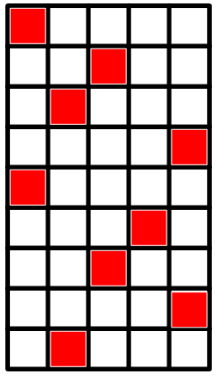


- Block-sparse structure: shared sparsity
- Group LASSO $$\min_{\beta} \sum_{k=1}^{r} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| y_i^{(k)} - X_i^{(k)} \beta^{(k)} \right\|_2^2 + \lambda \left\| \beta \right\|_{1,\infty}$$

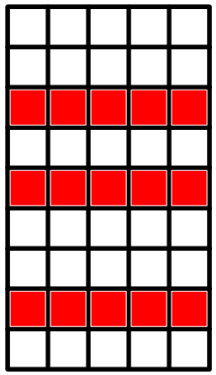
where $\|\beta\|_{1,\infty} = \sum_j \max_k \left| \beta_j^{(k)} \right|$ (sum of maximum of rows)

Existing Methods Performance

- LASSO
  - Does not use shared sparsity

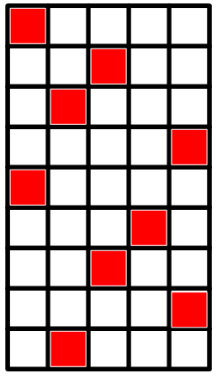


- Group LASSO
  - Does not model individual sparsity

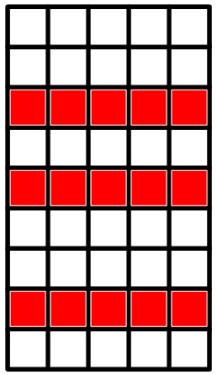


Existing Methods Performance

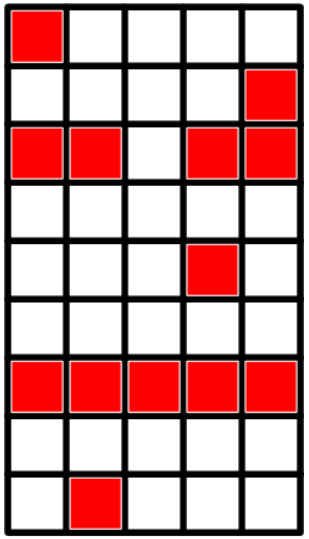
- LASSO
  - Does not use shared sparsity



- Group LASSO
  - Does not model individual sparsity



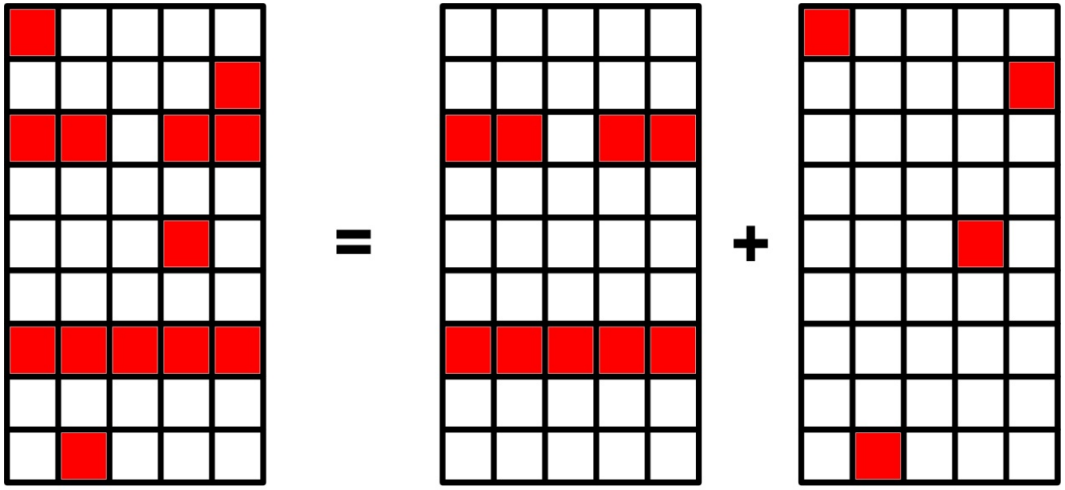
How about more realistic cases?



β

Dirty Statistical Model

- Superposition of parameters with diff. structures



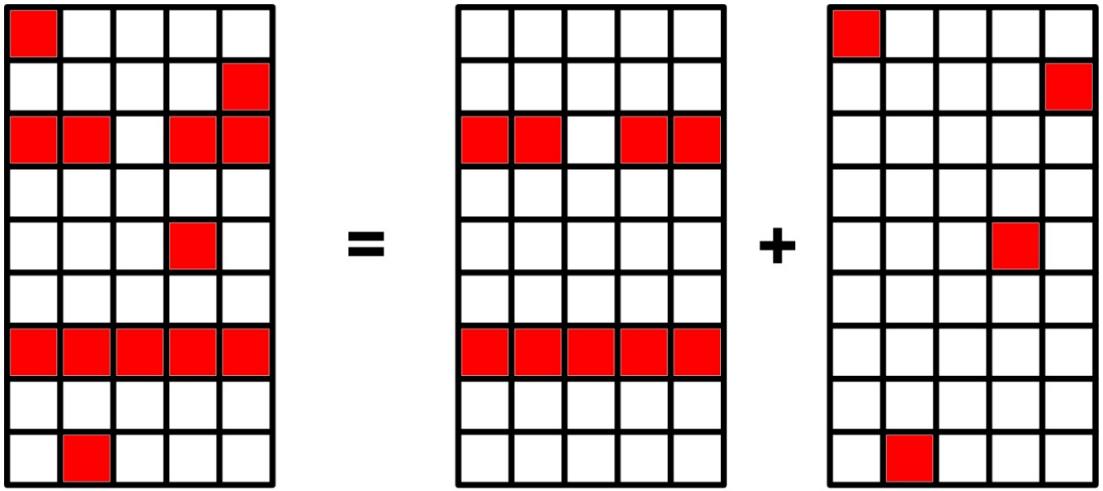
$$\beta = \mathbf{B} + \mathbf{S}$$

$$\|B\|_{1,\infty} \qquad \|S\|_{1,1}$$

shared features

non-shared features

Dirty Statistical Model



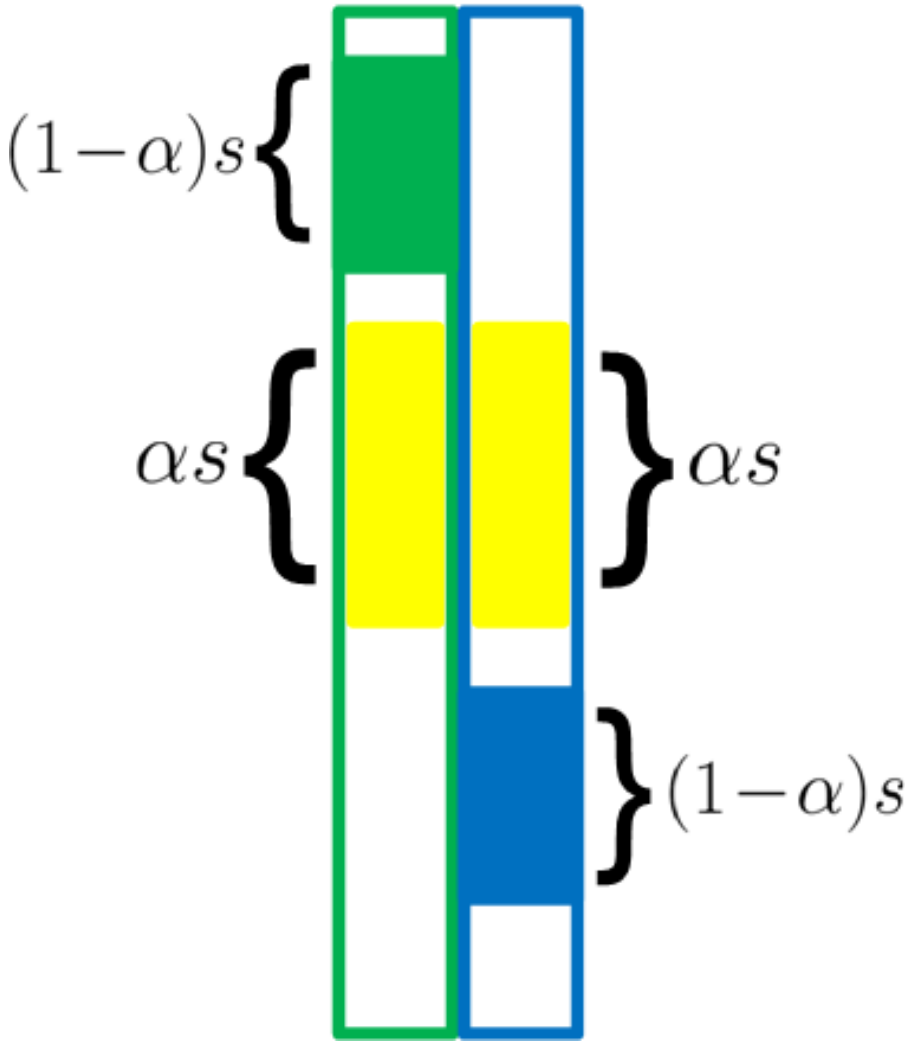
Algorithm:

$$\min_{B,S} \sum_{k=1}^{r} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| y_i^{(k)} - X_i^{(k)} \left(B^{(k)} + S^{(k)} \right) \right\|_2^2 + \lambda_b \|B\|_{1,\infty} + \lambda_s \|S\|_{1,1}$$

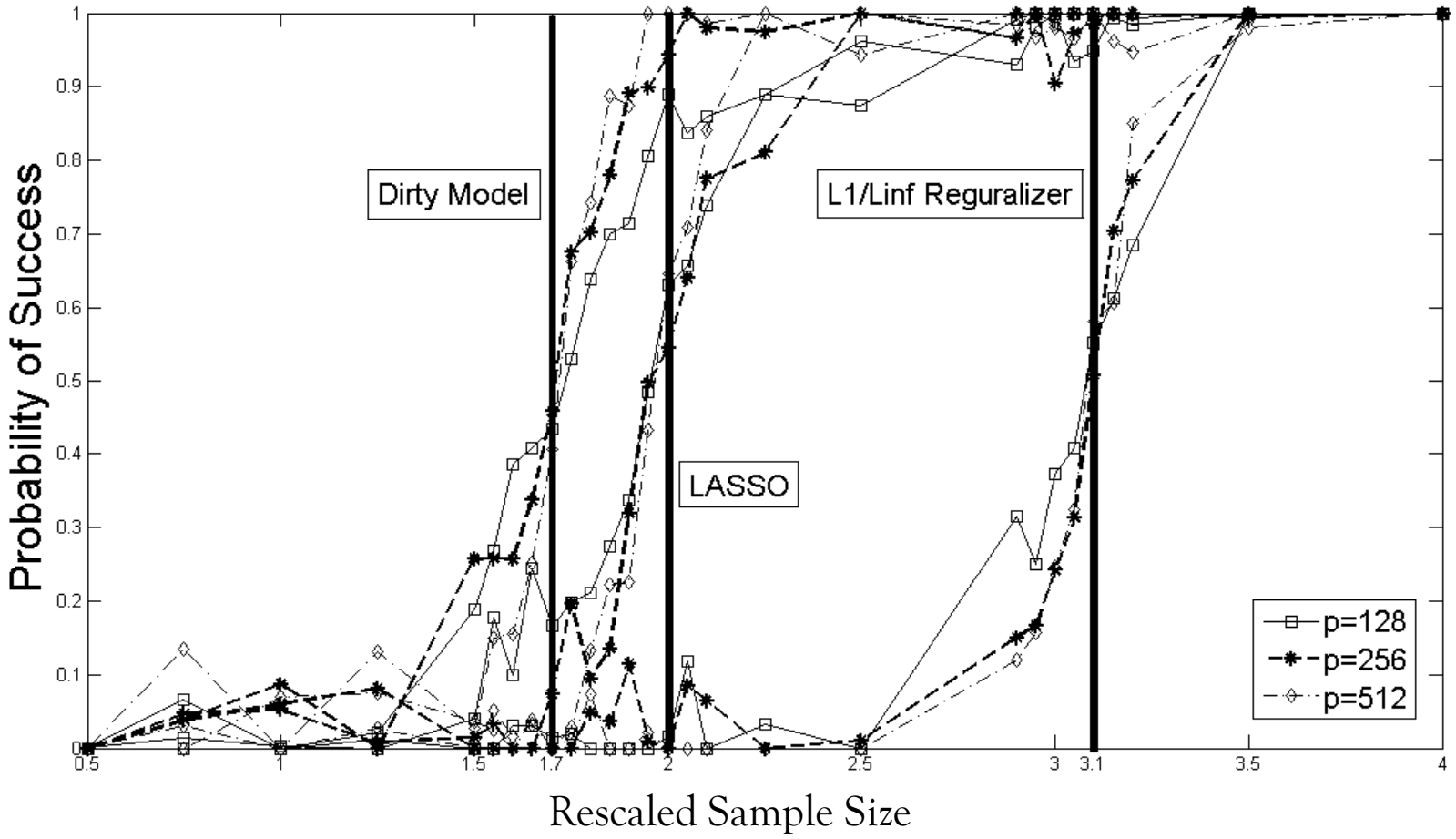
output $\hat{\beta} = \hat{B} + \hat{S}$

Two Tasks Case

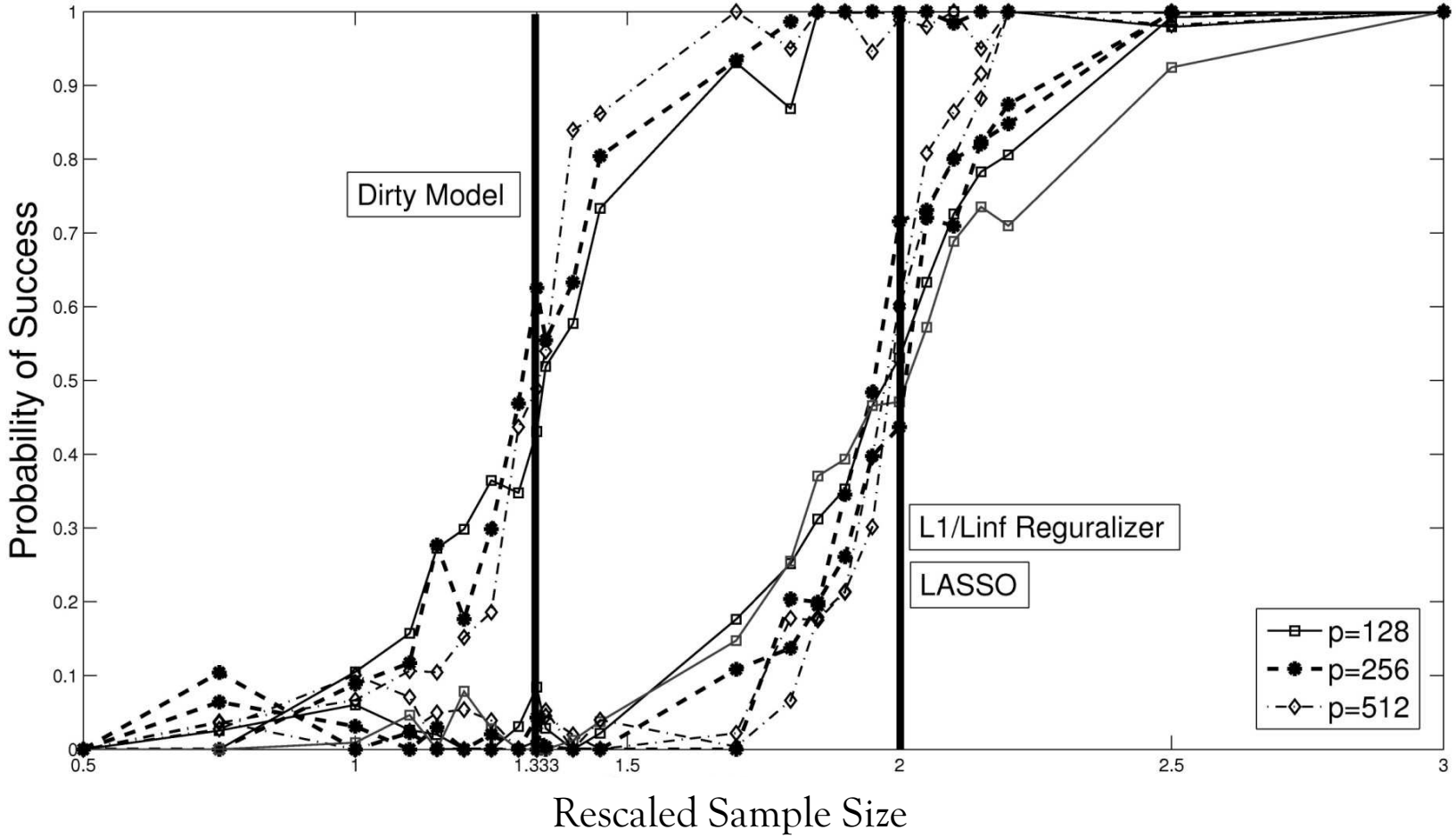
- Each task depends on "s" features
- α-portion of the features overlaps



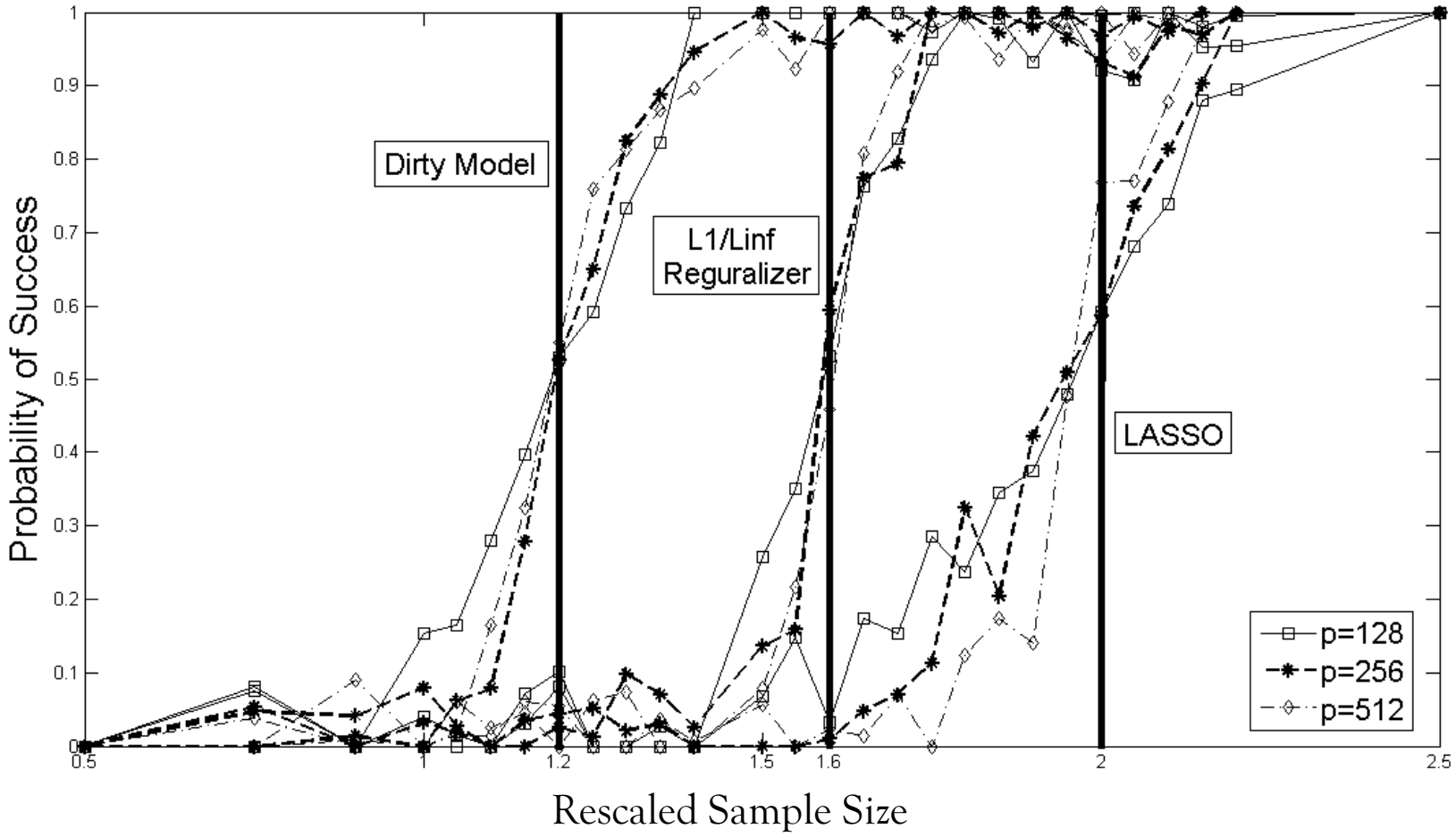
Little overlap: $\alpha = 0.3$



Critical overlap: $\alpha = 2/3$



High overlap: $\alpha = 0.8$



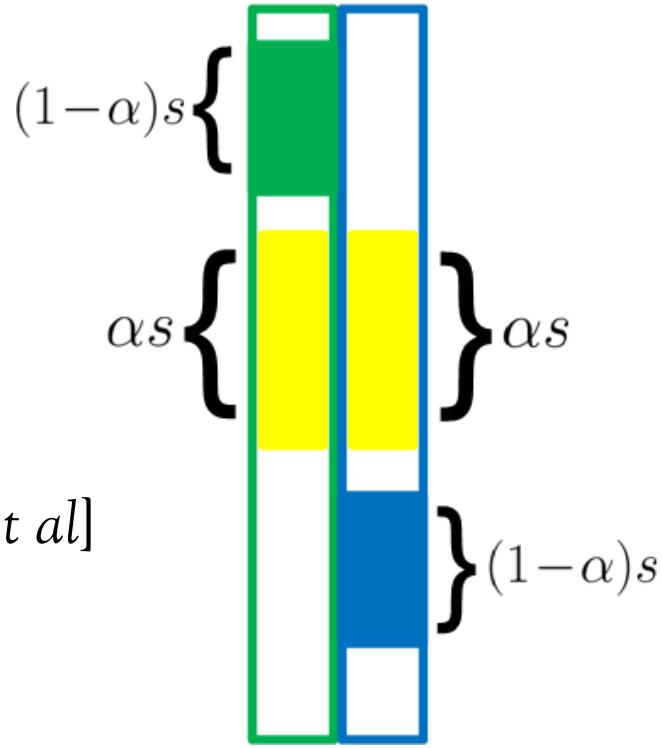
Phase Transition for Two Tasks

- LASSO:

$$\frac{n}{s\log(p)} \approx 2$$ [Wainwright]

- Group LASSO:

$$\frac{n}{s\log(p)} \approx 4-3\alpha$$ [Negahban *et al*]



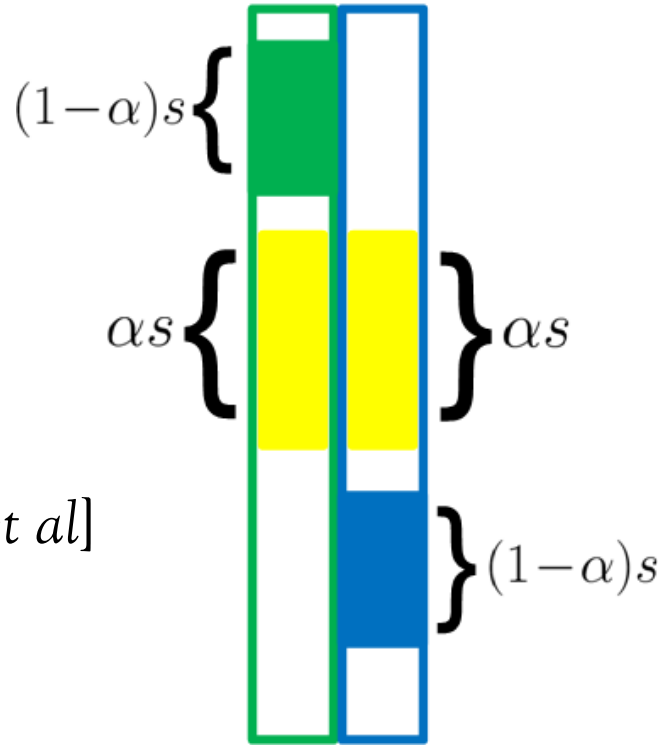
Phase Transition for Two Tasks

- LASSO:

$$\frac{n}{s\log(p)} \approx 2$$ [Wainwright]

- Group LASSO:

$$\frac{n}{s\log(p)} \approx 4-3\alpha$$ [Negahban *et al*]



Consequences:

$\alpha < 2/3$ :: Lasso is better
$\alpha > 2/3$ :: Group-Lasso is better

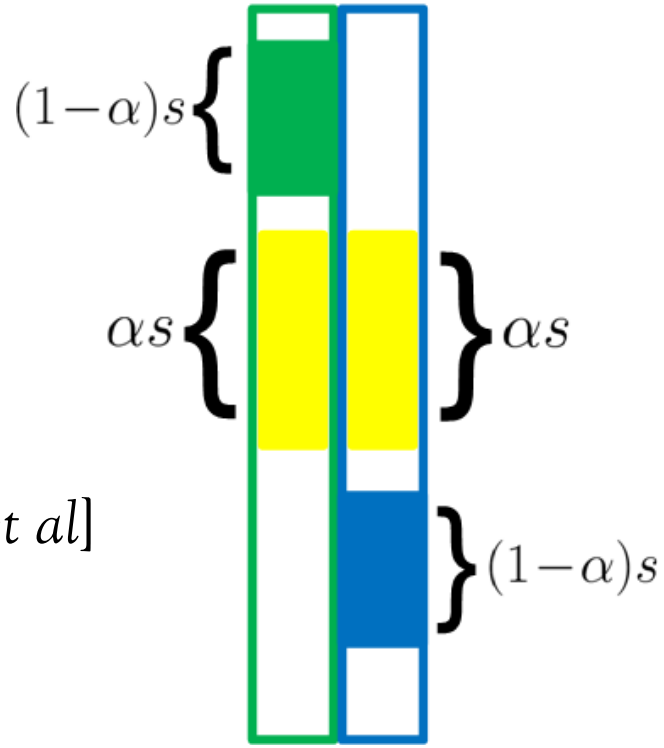
Phase Transition for Two Tasks

- LASSO:

$$\frac{n}{s\log(p)} \approx 2$$ [Wainwright]

- Group LASSO:

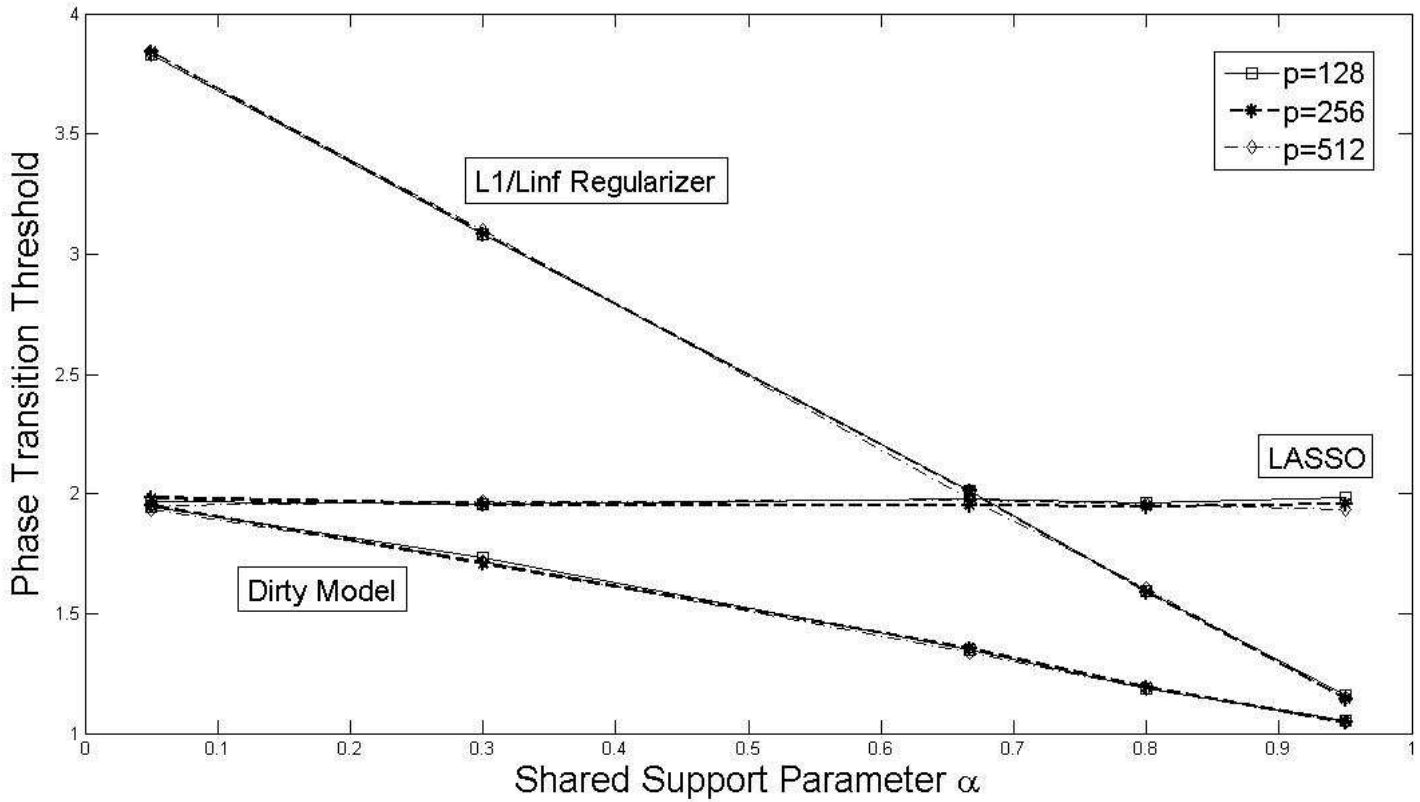
$$\frac{n}{s\log(p)} \approx 4-3\alpha$$ [Negahban *et al*]



- Dirty Model:

Theorem. $\frac{n}{s\log(p)} \approx 2-\alpha$

Comparison



Theorem. $\frac{n}{s\log(p)} \approx 2-\alpha$

General r-Task Case

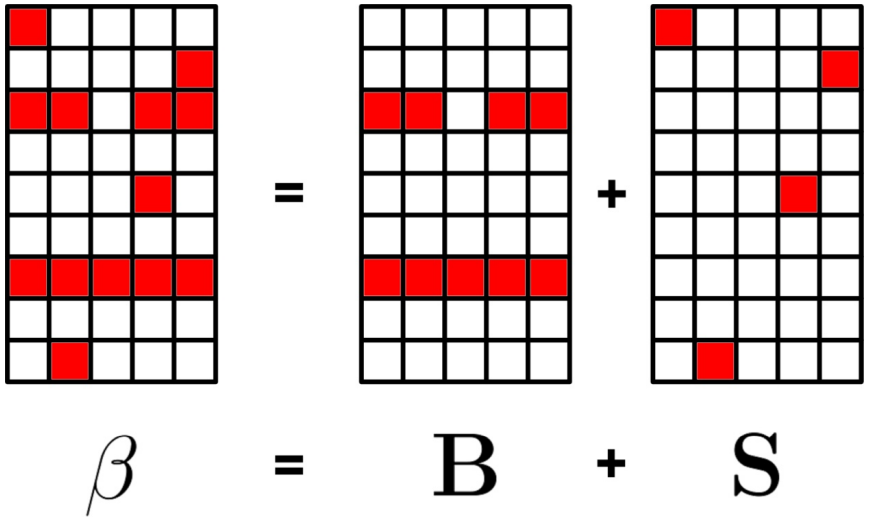
- Model: $y^{(k)} = X^{(k)}\beta^{(k)} + w^{(k)}$
- Algorithm: $\min_{B,S} \sum_{k=1}^{r} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| y_i^{(k)} - X_i^{(k)} \left(B^{(k)} + S^{(k)} \right) \right\|_2^2 + \lambda_b \|B\|_{1,\infty} + \lambda_s \|S\|_{1,1}$

Theorem (Gaussian Design): There exists a suitable choice of (λ_b, λ_s) for which our algorithm recovers the **exact sign support** of β with high probability provided that

$$n \geq K s \log(p)$$

where K only depends on $X^{(k)}$ and r.

General Dirty Models



- **Dirty Models**: "Superposition of Simple Structures"
  - Sparse + Low-Rank
    - Latent Variables, Graph Clustering, PCA with corruptions, etc
  - Block-Sparse + Low-Rank
    - Collaborative Filtering, PCA with Outliers, etc

Summary

- Multi-task learning is challenging when there is partial overlap across tasks
  - relevant structure is neither sparsity nor block-sparsity

- A superposition of simple structures, i.e., dirty model surprisingly useful for modeling such "dirty" structure.

- For the multi-task learning problem, dirty model outperforms solo-structured lasso and group-lasso.