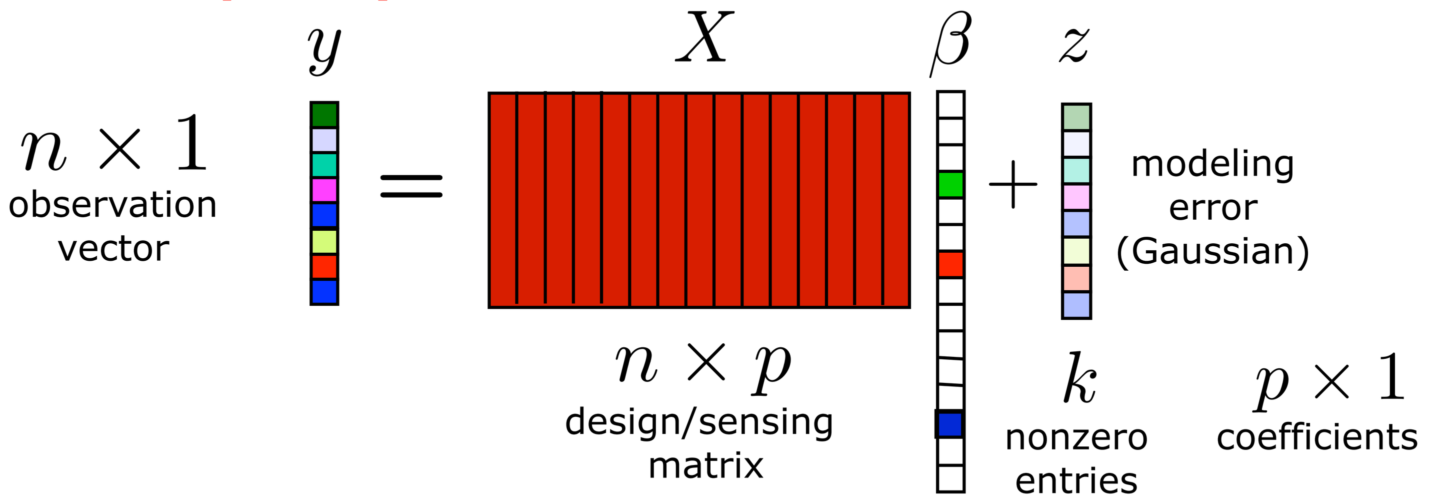
Average-Case Analysis of High-Dimensional Block-Sparse Recovery and Regression for Arbitrary Designs

Waheed U. Bajwa — Rutgers University
Marco F. Duarte — UMass Amherst
Robert Calderbank — Duke University

From Sparsity to Group Sparsity

Sparsity-Aware Linear Inference Problems



- ***Linear Regression***: Estimate $X\beta$ from $y = X\beta + \text{noise}$
  - Lasso: $\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2 + 2\lambda\sigma\|\beta\|_1$
- ***Sparse Recovery*** (Compressive Sensing): Recover β from $y = X\beta$
  - Basis Pursuit: $\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ subject to } y = X\beta$

Modeling Correlated Coefficients via Group Sparsity

Vector β partitioned into r groups of size m: $\beta = [\beta_1^T \ \beta_2^T \ \dots \ \beta_r^T]^T$
Matrix X partitioned into r submatrices: $X = [X_1 \ X_2 \ \dots \ X_r]$

- Group Lasso: $\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2 + 2\lambda\sigma\sqrt{m}\sum_{i=1}^{r}\|\beta_i\|_2$
- Block Basis Pursuit: $\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{r}\|\beta_i\|_2 \text{ subject to } y = X\beta$

Existing theoretical guarantees rely on ***subdictionaries*** (column submatrices of sensing/design matrix) being well conditioned and either:

- directly check conditioning with *combinatorial computation*, or
- indirectly check conditioning and provide *pessimistic bounds*.

It is possible to avoid these issues by switching to a ***statistical performance*** measurement setting:

- endowing β with a "***uniform***" distribution over all sparse vectors,
- considering the ***average-case subdictionary conditioning***.

Recent theoretical guarantees for problems with standard sparsity that are valid with high probability and require only simple matrix metrics

Random Group Subdictionaries

Average-Case Subdictionary Conditioning Metrics

- Intra-Block Coherence: $\mu_I := \max_{1 \le i \le r} \|X_i^T X_i - I_m\|_2$
- Inter-Block Coherence: $\mu_B := \max_{1 \le i,j \le r} \|X_i^T X_j\|_2$
- Spectral Norm: $\|X\|_2 = \max_{\alpha \in \mathbb{R}^p} \|X\alpha\|_2 / \|\alpha\|_2$

Block Incoherence Condition (BIC): $\mu_I \le c_1, \ \mu_B \le c_2 / \log p$

Blocks are each close to orthogonal and are sufficiently incoherent with one another

> ***Theorem (Average-Case Subdictionary Conditioning):*** Assume that X satisfies the BIC and S is a k-subset drawn from $\{1,\dots,r\}$ uniformly at random. Then, as long as $k \le c_0 r / (\|X\|_2^2 \log p)$ the singular values of the block subdictionary $X_S = [X_i : i \in S]$ satisfy $\sigma_i(X_S) \in [\sqrt{1/2}, \sqrt{3/2}]$, $i = 1,\dots,km$, with probability over the choice of set S of at least $1-2p^{-4\log 2}$.

- The size of the largest well-conditioned subdictionaries scales inversely with the sensing/design matrix spectral norm
- The coherence measures do not affect the size of the well-conditioned subdictionaries (other than through the BIC)

"Uniform" Distribution over Block Sparse Signals

- Block support of β distributed uniformly among k-subsets of $\{1,\dots,r\}$
- Entries of β have zero median (i.e., its entries have positive and negative signs with equal probabilities): $\mathbb{E}[\text{sign}(\beta)] = 0$
- Nonzero blocks of β have statistically independent "directions":

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{S}} \left(\overline{\text{sign}}(\beta_i) \in \mathcal{A}_i\right)\right) = \prod_{i \in \mathcal{S}} \mathbb{P}\left(\overline{\text{sign}}(\beta_i) \in \mathcal{A}_i\right)$$

where $\mathcal{A}_i \subset \mathbb{S}^{m-1}$, the m-dimensional sphere, and $\overline{\text{sign}}(\beta_i) = \frac{\beta_i}{\|\beta_i\|_2}$

Average-Case Performance Guarantees

Block-Sparse Recovery

Theorem: Assume that β is drawn "uniformly" over the set of k-block sparse signals and X satisfies the BIC. As long as $k \le c_0 r / (\|X\|_2^2 \log p)$, basis pursuit will return $\widehat{\beta} = \beta$ with probability of at least $1-4p^{-4\log 2}$.

Comparison to basis pursuit/standard sparsity: $k \le c_0 n / (\|X\|_2^2 \log p)$ if *coherence* (max. column inner product) $\mu(X) \le c / \log p$ [Tropp 2008]
Matched performance for $m = 1$

Block-Sparse Linear Regression

Theorem: Assume that β is drawn "uniformly" over the set of k-block sparse signals and X satisfies the BIC. As long as $k \le c_0 r / (\|X\|_2^2 \log p)$, the estimate returned by the group lasso with $\lambda = \sqrt{2 \log p}$ obeys $\|X\beta - X\widehat{\beta}\|_2^2 \le Cmk\sigma^2 \log p$ with probability of at least $1 - p^{-1}(2\pi \log p)^{-1/2} - 8p^{-4\log 2}$.

Comparison to standard lasso: same performance only if coefficients are *independent* (even within each group) [Candès and Plan 2009]
Without independence, we require $k \le c_0 r / (\|X\|_2^2 m \log p)$

Discussion

- As long as dictionary coherences are sufficiently small, $\|X\|_2$ is the only matrix metric affecting size of well-conditioned subdictionaries
- Tight frames provide smallest spectral norm $\|X\|_2^2 \approx p/n$; therefore, largest well-conditioned subdictionaries obey $k = \mathcal{O}(n / \log p)$
- Results translate to Multiple Measurement Vector (MMV) setting: Kronecker-structured matrices with translatable norm/coherence metrics [Candès and Plan 2009]

Full version of this paper: http://arxiv.org/pdf/1309.5310/

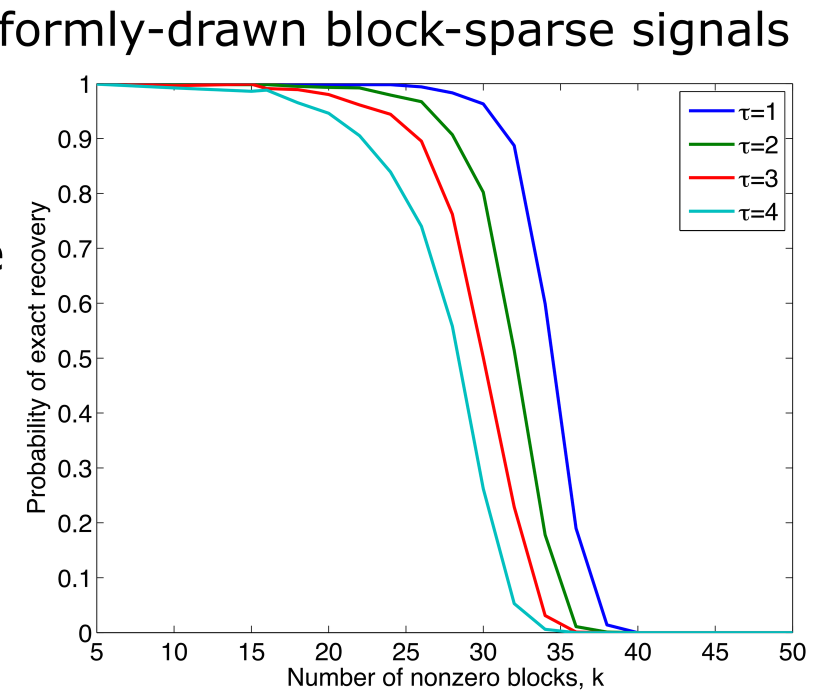
Numerical Results

Sparsity-Aware Linear Inference Simulations

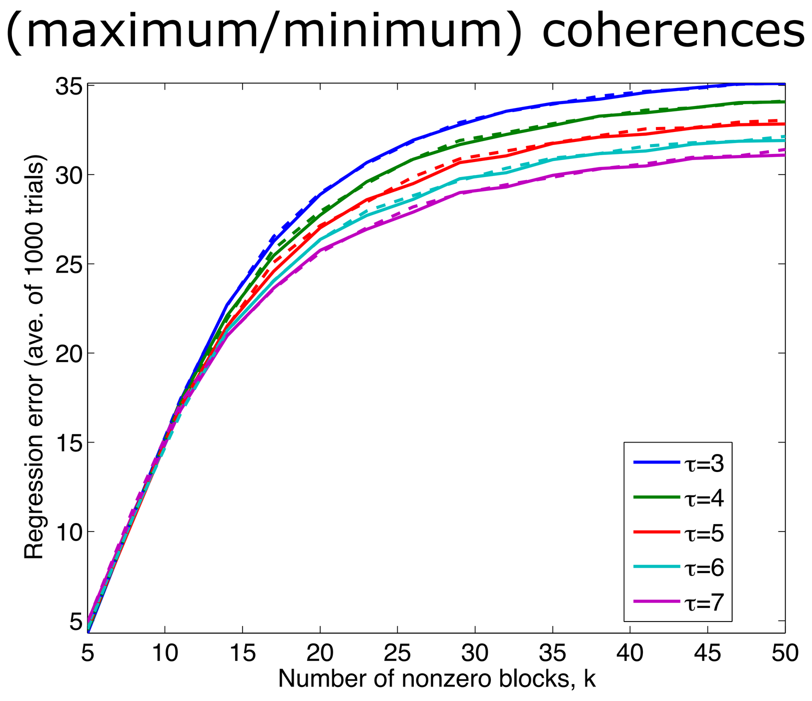
- Generated 2000 random matrices with normalized columns, $p = 5000$, $m = 10$, $r = 500$, $n = 858$ [Rao, Recht, Nowak 2012]
- Scaled matrix spectral norm (via SVD) with multiplier set $\tau \in \mathcal{T}$
- Resulting $2000|\mathcal{T}|$ design/sensing matrices had columns renormalized, block coherence metrics computed
- Average performance over 1000 uniformly-drawn block-sparse signals

Experiment 1: Measure sparse recovery performance for matrices selected to have matching coherence values among several spectral norm multipliers

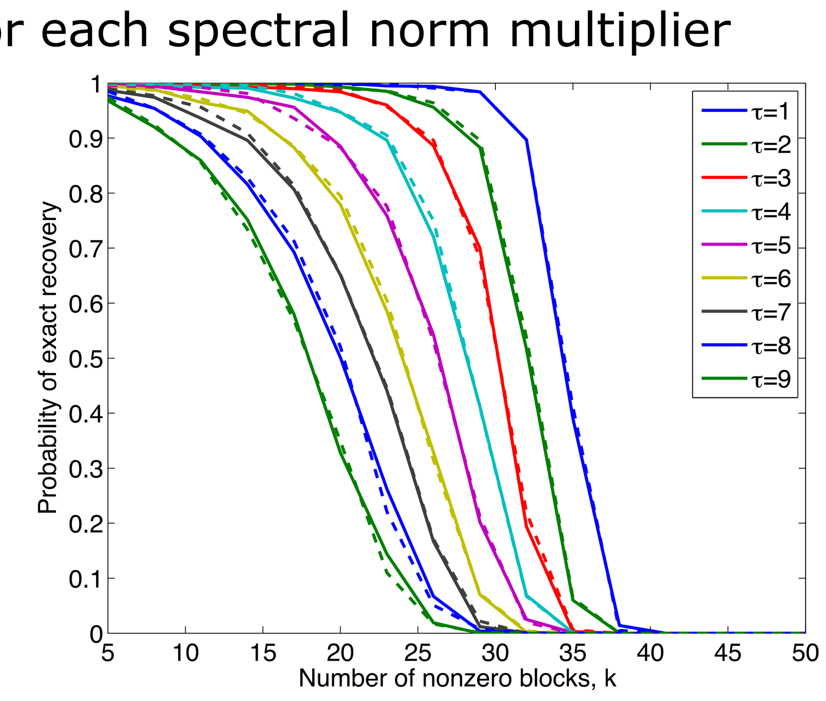
τ	1	2	3	4
$\|X_\tau\|_2$	3.3963	6.7503	10.0547	13.2034
$\mu(X_\tau)$	0.1992	0.2026	0.2000	0.2207
$\mu_B(X_\tau)$	0.2973	0.3431	0.5573	0.8490
$\mu_I(X_\tau)$	0.1992	0.2026	0.2177	0.3787



Experiment 2: Measure performance for matrices with extremal (maximum/minimum) coherences for each spectral norm multiplier



Sparse Regression



Sparse Recovery

Dashed: Lowest Coherence; Solid: Highest Coherence