Composite Loss Functions and Multivariate Regression; Sparse PCA

G. Obozinski, B. Taskar, and M. I. Jordan (2009). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, to appear.

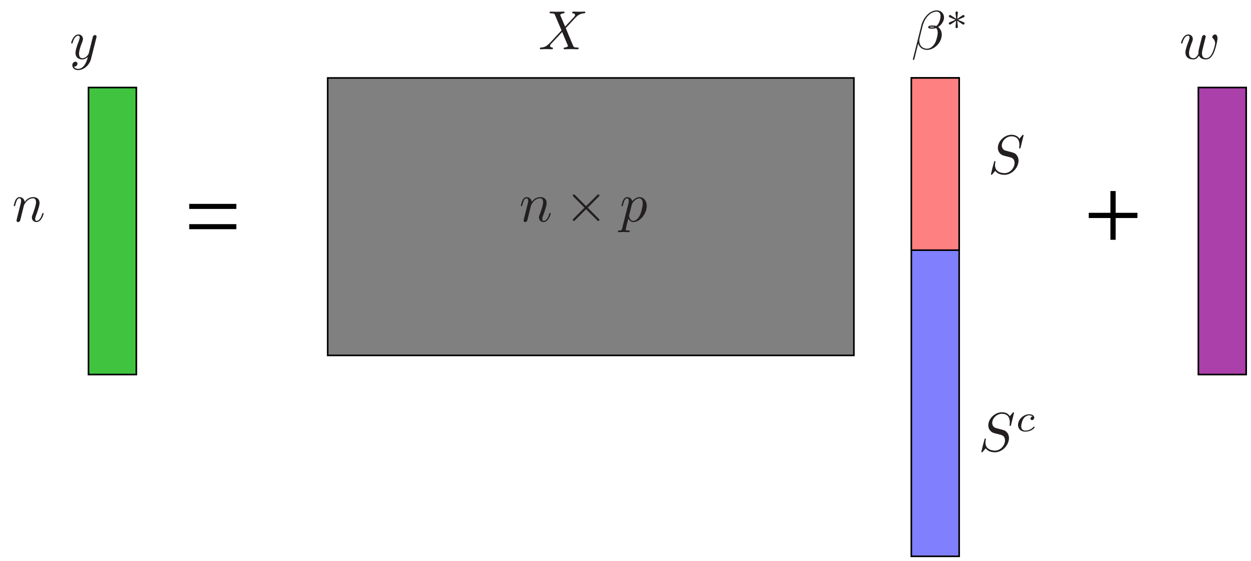
G. Obozinski, M. J. Wainwright, and M. I. Jordan (2009). Union support recovery in multivariate regression. *Annals of Statistics*, under review.

A. Amini and M. J. Wainwright (2009). High-dimensional analysis of semidefinite relaxations for sparse PCA. *Annals of Statistics*, to appear.

Introduction

- classical asymptotic theory of statistical inference:
  - number of observations $n \to +\infty$
  - model dimension p stays fixed
- not suitable for many modern applications:
  - $\{$ images, signals, systems, networks $\}$ frequently large ($p \approx 10^3 - 10^8$)...
  - interesting consequences: might have $p = \Theta(n)$ or even $p \gg n$
- curse of dimensionality: frequently impossible to obtain consistent procedures unless $p/n \to 0$
- can be saved by a lower *effective dimensionality*, due to some form of complexity constraint

Example: Sparse linear regression

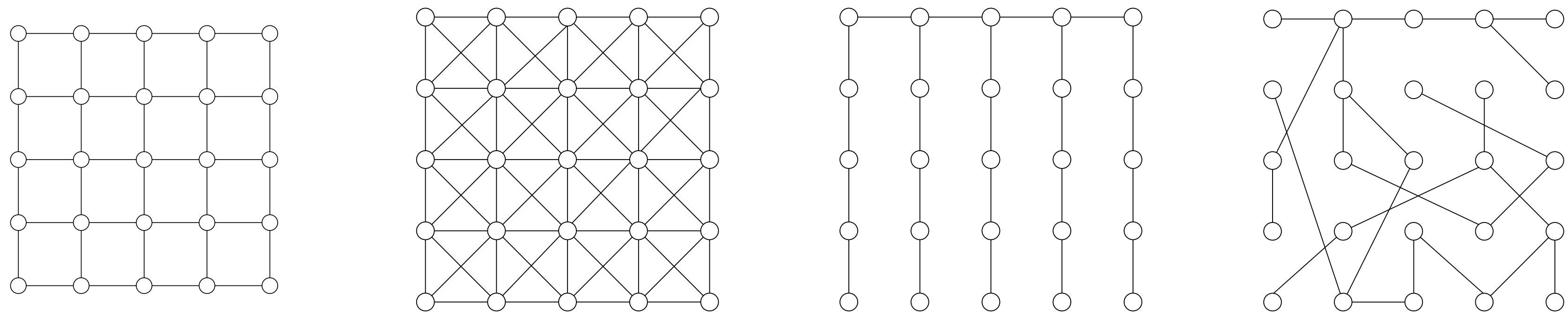


- vector $\beta^* \in \mathbb{R}^p$ with at most $k \ll p$ non-zero entries

- observation model: noisy linear observations $y = X\beta^* + w$
  $X \in \mathbb{R}^{n \times p}$: design matrix
  $w \in \mathbb{R}^{n \times 1}$: noise vector

- various applications (database sketching, imaging, genetic testing...)

Example: Graphical model selection



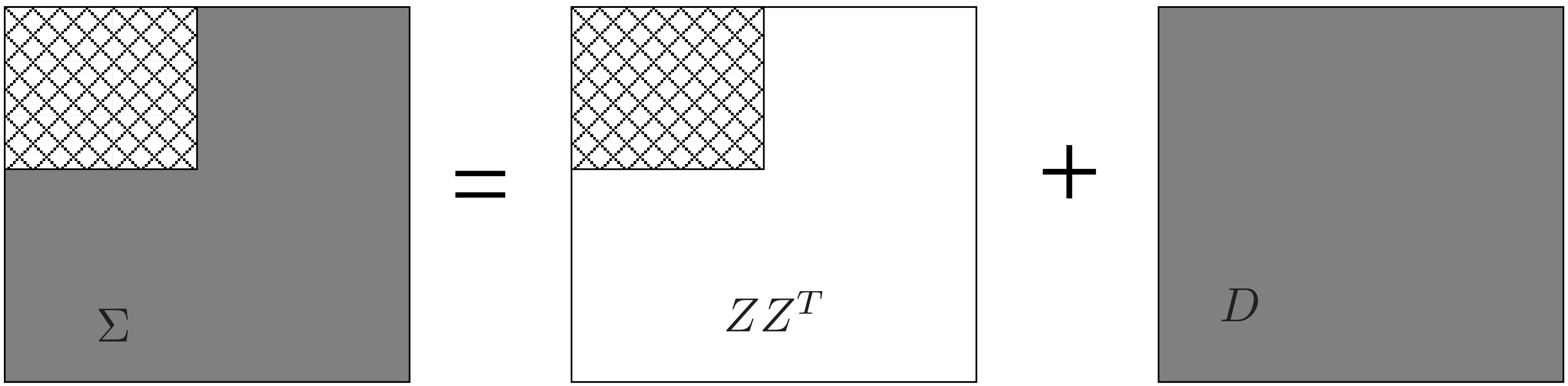
- consider m-dimensional random vector $Z = (Z_1, \ldots, Z_m)$:

$$\mathbb{P}(Z_1, \ldots, Z_m; \beta) \quad \propto \quad \exp \left\{ \sum_{(i,j) \in E} \beta_{ij} Z_i Z_j \right\}.$$

- given n independent and identically distributed (i.i.d.) samples of $\vec{Z}$, identify underlying graph $G = (V, E)$

- lower effective dimensionality: graphs with $k \ll p := \binom{m}{2}$ edges

Example: Sparse principal components analysis



Set-up: Covariance matrix $\Sigma = ZZ^T + D$, where leading eigenspace Z has sparse columns.

Goal: Produce an estimate $\widehat{Z}$ based on samples $X^{(i)}$ with covariance matrix Σ.

Some issues in high-dimensional inference

- Consider some fixed loss function, and a fixed level δ of error.

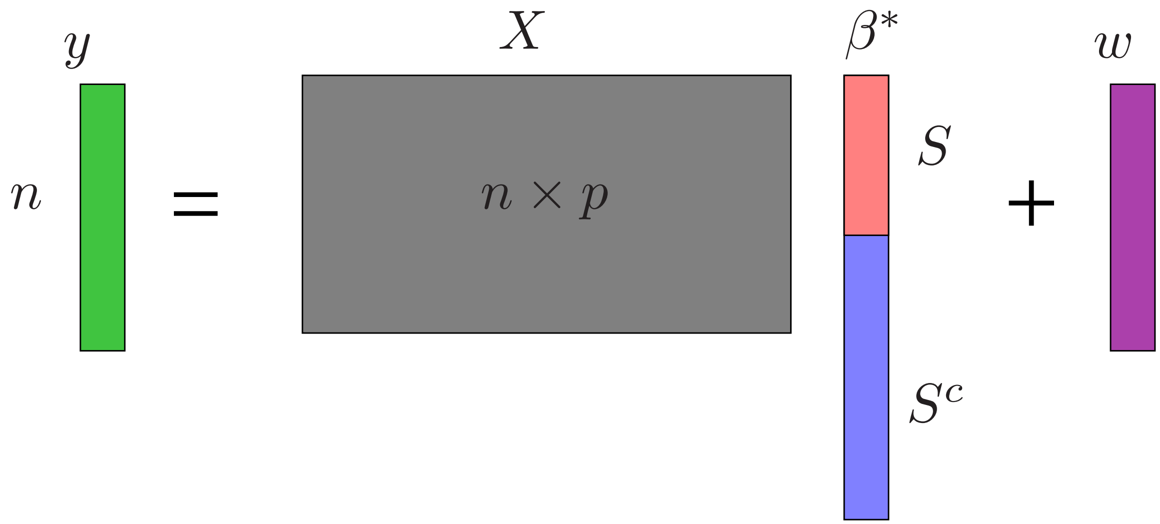
- Given particular (polynomial-time) algorithms
  - for what sample sizes n do they succeed/fail to achieve error δ?
  - when does more computation reduce minimum # samples needed?

Outline

1. Multivariate regression in high dimensions
   (a) Practical limitations: scaling laws for second-order cone programs
   (b) SOCP vs. Lasso: when does more computation reduce statistical error?

2. Sparse principal component analysis in high dimensions
   (a) Thresholding methods
   (b) Semidefinite programming

Optimization-based estimators in (sparse) regression



$$\text{Regularized QP:} \quad \widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \Big\{ \underbrace{\frac{1}{2n}\|y - X\beta\|_2^2}_{\text{Data term}} + \rho_n \underbrace{R(\beta)}_{\text{Regularizer}} \Big\}.$$

$R(\beta) = \|\beta\|_2$	Ridge regression	(Tik43, HoeKen70)
$R(\beta) = \|\beta\|_1$	convex ℓ_1-constrained QP	(CheDonSau96; Tibs96)
$R(\beta) = \|\beta\|_0$	Subset selection: combinatorial, NP-hard	(Nat95)
$R(\beta) = \|\beta\|_a$, $a \in (0,1)$	Non-convex ℓ_a regularization	

Different loss functions

Given an estimate $\widehat{\beta}$, how to assess its performance?

1. Predictive loss: compute expected error $\mathbb{E}[\|\widetilde{y} - X\widehat{\beta}\|_2^2]$
   - goal is to construct model with good predictive power
   - β^* itself of secondary interest (need not be uniquely determined)

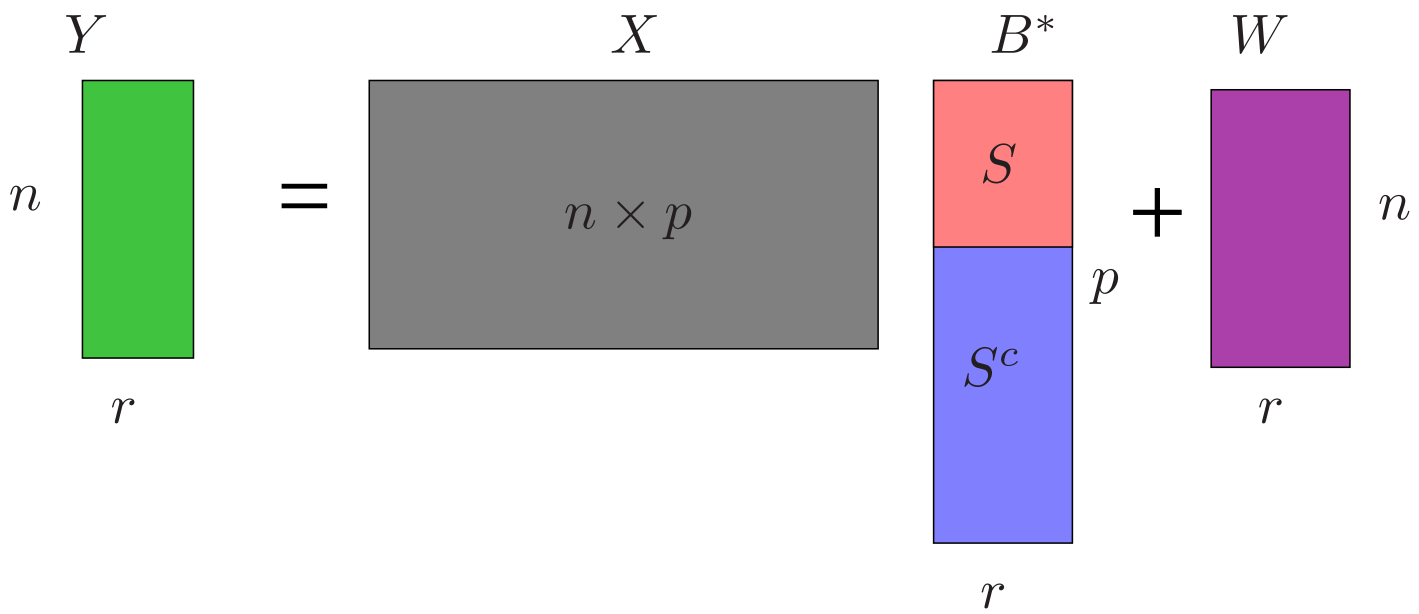
2. ℓ_2-loss $\mathbb{E}[\|\widehat{\beta} - \beta^*\|_2^2]$
   - appropriate when B^* is of primary interest (signal recovery, compressed sensing, denoising etc.)

3. Support recovery criterion: define estimated support

$$S(\widehat{\beta}) \quad = \quad \{i = 1, \ldots, p \mid \widehat{\beta}_i \neq 0\},$$

   and measure probability $\mathbb{P}[S(\widehat{\beta}) \neq S(\beta^*)]$.
   - useful for feature selection, dimensionality reduction, model selection
   - can be used as a pre-processing step for estimation in ℓ_2-norm

§1. Multivariate regression in high dimensions



- signal B^* is a $p \times r$ matrix: partitioned into **non-zero rows** S and **zero rows** S^c
- observe n noisy projections, defined via **design matrix** $X \in \mathbb{R}^{n \times p}$ and **noise matrix** $W \in \mathbb{R}^{n \times r}$
- matrix $\mathbf{Y} \in \mathbb{R}^{\mathbf{n} \times \mathbf{r}}$ **of observations**
- high-dimensional scaling: allow parameters $(n, p, r, |S|)$ to scale

Block regularization and second-order cone programs

(Obozinski, Taskar & Jordan, 2009)

- for fixed parameter $q \in [1, \infty]$, estimate B^* via:

$$\widehat{B} \in \arg\min_{B \in \mathbb{R}^{p \times r}} \Big\{ \underbrace{\frac{1}{2n} \|Y - XB\|_F^2}_{\text{Data term } \sum_{j=1}^{n} \sum_{\ell=1}^{r} \left[Y_{j\ell} - (XB)_{j\ell} \right]^2} + \rho_n \underbrace{|||B|||_{1,q}}_{\sum_{i=1}^{p} \|(B_{i1}, \ldots, B_{ir})\|_q} \Big\}.$$

- regularization constant $\rho_n > 0$ to be chosen by user
- different cases:
  - $q = 1$: elementwise ℓ_1 norm (constrained QP)
  - $q = 2$: second-order cone program (SOCP)
  - $q = \infty$: block ℓ_1/ℓ_∞ max-norm (constrained QP)

- in all cases, efficiently solvable (e.g., by interior point methods)
- generalization of the Lasso (Tibshirani, 1996; Chen et al., 1998),
- special case of the CAP family (Zhao, Rocha, & Yu, 2006); see also (Turlach et al., 2005; Yuan & Lin, 2006, Nardi & Rinaldo, 2008)

Two strategies

Goal: Model selection consistency: recover union of supports

$$S(B^*) \quad := \quad \{i \in \{1,2,\ldots,p\} \mid \|B^*_{i1},\ldots,B^*_{ir}\|_2 \neq 0\}.$$

Different methods:

- *Lasso-based recovery:*
  1. Solve a separate Lasso (ℓ_1-constrained QP) for each column $\ell = 1,\ldots,r$, yielding column vector $\widehat{\beta}_\ell \in \mathbb{R}^p$.
  2. Estimate row support $\widehat{S}_{\text{Lasso}} = \big\{i \in \{1,2,\ldots,p\} \mid \widehat{\beta}_{i\ell} \neq 0 \text{ for some } \ell\big\}$.
- *SOCP-based recovery:*
  1. Solve a single SOCP, obtaining matrix estimate $\widehat{B} \in p \times r$.
  2. Estimate support $\widehat{S}_{\text{SOCP}} = \big\{i \in \{1,\ldots,p\} \mid \|(\widehat{B}_{i1},\ldots,\widehat{B}_{ir}\|_2 \neq 0\big\}$.

Trade-offs:

- Lasso (QP) cheap to solve, but method ignores coupling among columns
- SOCP more expensive, but block-regularizer better tailored to matrix structure

Scaling law for high-dimensional SOCP recovery

(Obozinski, Wainwright & Jordan, 2009)

- SOCP method: $\widehat{B} \in \arg \min_{B \in \mathbb{R}^{p \times r}} \{ \frac{1}{2n} \|Y - XB\|_F^2 + \rho_n |||B|||_{1,2} \}$.
- Parameters: Problem dimension p; number of non-zero rows k
- Design matrix X: i.i.d. rows from sub-Gaussian distribution, with "suitable" covariance Σ

Theorem: If the *rescaled sample size*

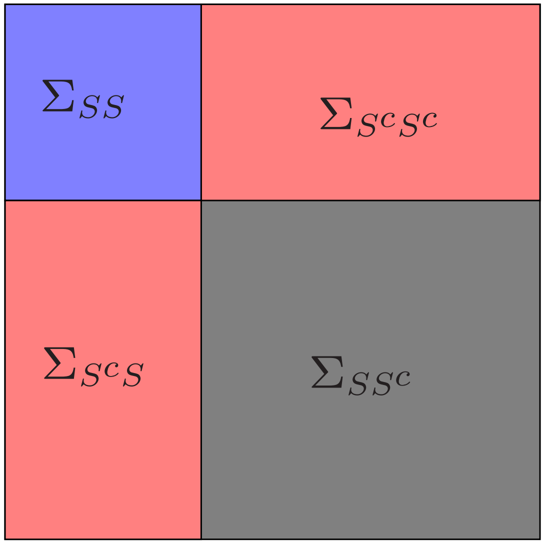
$$\theta_{\mathrm{SOCP}}(n, p, k, \; B^*) \; := \; \frac{n}{\Psi(B_S^*; \Sigma_{SS}) \;\; \log(p-k)}$$

is greater than a critical threshold $\theta_\ell(\Sigma; \sigma^2)$, then for suitable ρ_n we have with probability greater than $1 - 2\exp(c_2 \log k)$:

(a) the SOCP has a unique solution $\widehat{B}$ s.t. $\widehat{S}(\widehat{B}) \subseteq S(B^*)$, and

(b) It includes all rows i with $\|B_i^*\|_2 \geq c_3 \sqrt{\frac{\max\{k, \log(p-k)\}}{n}}$.

Assumptions on design covariance



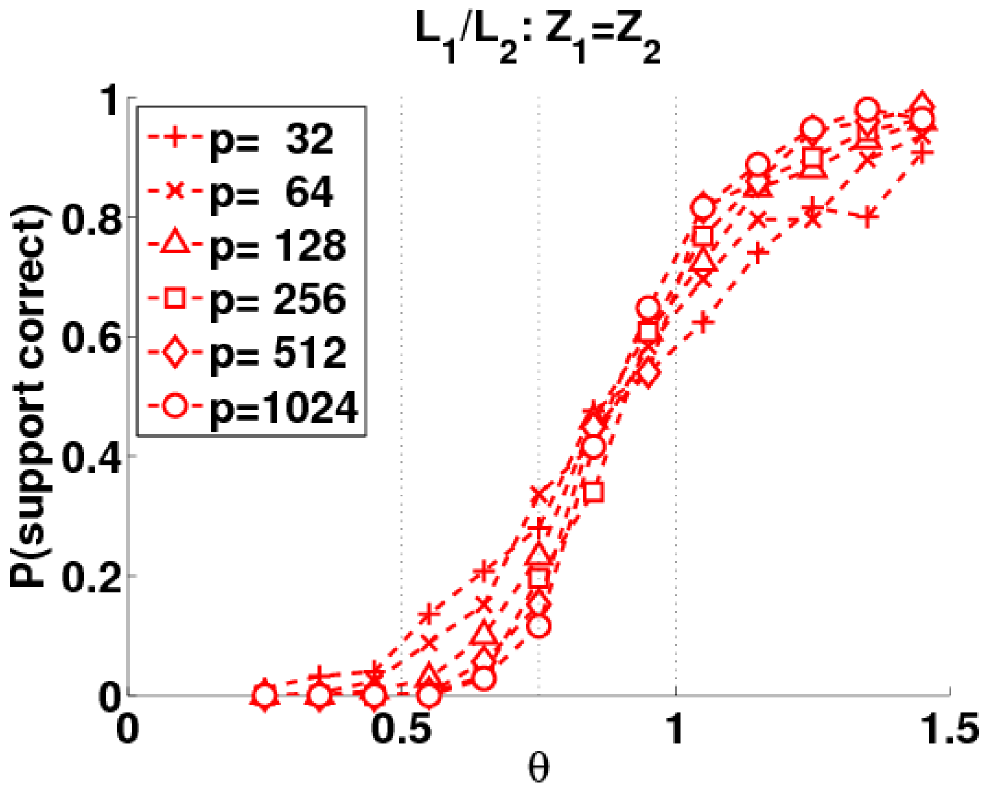
- support set $S = \{i \mid \beta_i^* \neq 0\}$
- complement $S^c := \{1, \ldots p\} \backslash S$.
- random design matrix $X \in \mathbb{R}^{n \times p}$
- rows drawn i.i.d., cov. Σ, sub-Gaussian

1. **Bounded eigenspectrum:** $\lambda(\Sigma_{SS}) \in [C_{min}, C_{max}]$.
2. **Mutual incoherence/irrepresentability:** There exists an $\nu \in (0, 1]$ such that

$$|||\Sigma_{S^cS}(\Sigma_{SS})^{-1}|||_{\infty,\infty} \leq \quad 1 - \nu.$$

Example: if $\Sigma_{SS} = I$, then require $\max_{j \in S^c} \sum_{i \in S} |\Sigma_{ji}| \leq 1 - \nu$.

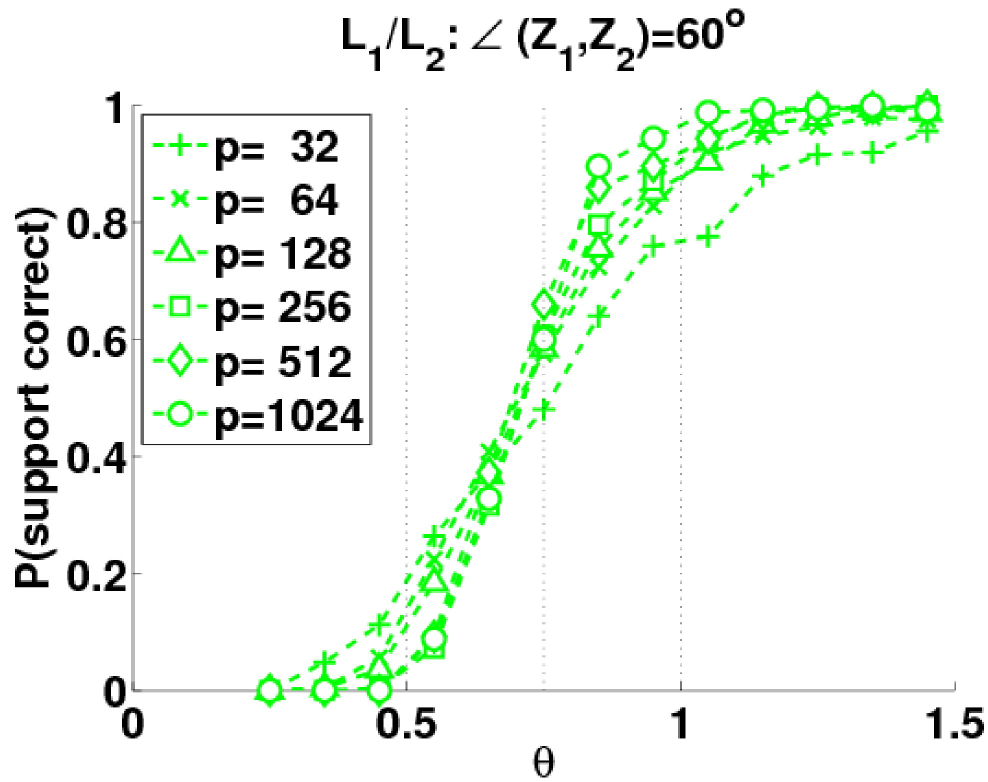
Order parameter captures threshold (Angle 0°)



Prob. success versus rescaled sample size

$$\theta_{\mathrm{SOCP}}(n, p, k, B^*) = \frac{n}{\Psi(B^*_S; \Sigma_{SS}) \log(p-k)}.$$

Order parameter captures threshold (Angle 60°)



Prob. success versus rescaled sample size

$$\theta_{\text{SOCP}}(n, p, k, B^*) = \frac{n}{\Psi(B^*_S; \Sigma_{SS}) \log(p-k)}.$$

Sparsity overlap function Ψ

- form gradient matrix $Z(B_S^*) := \nabla \|\!| B_S \|\!|_{1,2} \Big|_{B_S = B_S^*} \in \mathbb{R}^{k \times r}$
- equivalent to renormalizing B_S^* to have unit ℓ_2-norm rows
- form $r \times r$ Gram matrix:

$$G \quad = \quad Z^T (\Sigma_{SS})^{-1} Z$$

  with $G_{a,b} = \langle\!\langle Z_a,\ Z_b \rangle\!\rangle_{(\Sigma_{SS})^{-1}}$
- sparsity overlap function is max. eigenvalue of G:

$$\Psi(B_S^*; \Sigma_{SS}) \quad = \quad \|\!| G \|\!|_2.$$

- measures relative alignments of the renormalized columns of B^*
- **Special case:** Univariate regression ($r = 1$): $Z(\beta_S^*) = k$ for any vector β_S^*

Concrete examples ($k = 4, r = 2$)

Aligned columns

$$B_S^* = \begin{bmatrix} 2 & 2 \\ 10 & 10 \\ 1 & 1 \\ 7 & 7 \end{bmatrix} \qquad Z(B_S^*) = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

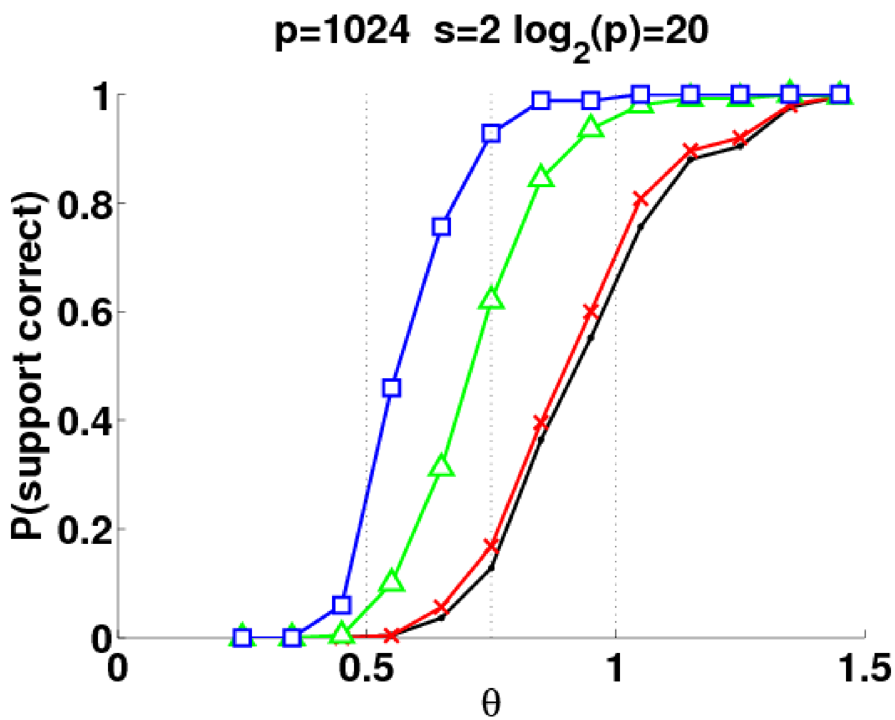
$$G = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \qquad \|G\|_2 = 4$$

Orthogonal columns

$$B_S^* = \begin{bmatrix} 2 & 2 \\ 10 & 10 \\ 1 & -1 \\ 7 & -7 \end{bmatrix} \qquad Z(B_S^*) = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$G = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \qquad \|G\|_2 = 2$$

Empirical illustration of sparsity-overlap Ψ



- Orthogonal regression: Columns $Z_1 \perp Z_2$
- Intermediate angle: Columns at 60°
- Aligned regression: Columns parallel
- Ordinary Lasso: solve problems separately.

SOCP versus ordinary QP

Corollary: If $\Sigma_{SS} = I_{k \times k}$, SOCP always dominates ordinary QP, with relative statistical efficiency:

$$1 \leq \underbrace{\frac{\max\limits_{\ell=1,\ldots r} k_\ell \log(p - k_\ell)}{\Psi(B_S^*; I) \log(p-k)}}_{\text{(QP sample size)/(SOCP sample size)}} \leq r$$

- increased statistical efficiency of SOCP: dependent on orthogonality properties of rescaled columns B_S^*
- up to a factor $1/r$ reduction in number of samples required
- most pessimistic case: no gain for disjoint supports, SOCP can be worse in some cases (if $\Sigma_{SS} \neq I$)

Proof sketch of sufficient conditions

Direct analysis :

Given n observations of $\beta^* \in \mathbb{R}^p$ with $|S(\beta^*)| = k$, oracle decoder performs following two steps:

1. For each subset S of size k, solve the quadratic program:

$$f(S) \quad = \quad \min_{\beta_S \in \mathbb{R}^k} \|Y - X_S \beta_S\|_2^2.$$

2. Output the subset $\widehat{S} = \arg \min_{|S|=k} f(S)$.

- by symmetry of ensemble, may assume that fixed subset S is chosen
- for sets U different from true set S, consider range of *non-overlaps* $t := |U \backslash S| \in \{1, \ldots, k\}$
- number of subsets with non-overlap t given by $N(t) = \binom{k}{t-k} \binom{p-k}{t}$

Error exponents for random projections

- union bound yields upper bound on error probability $\mathbb{P}[\text{error} \;\mid\; S \text{ true}]$:

$$\sum_{t=1}^{k} \binom{k}{k-t} \binom{p-k}{t} \,\mathbb{P}[\text{error on subset with non-overlap } t]$$

- orthogonal projection $\Pi_U^{\perp} := I_{n \times n} - X_U \left[X_U^T X_U\right]^{-1} X_U^T$
- optimal decoder chooses U incorrectly over S if and only if

$$\Delta(U) \quad = \quad \underbrace{\left\| \Pi_U^{\perp} \left(X_{S \backslash U} \beta^*_{S \backslash U} + W \right) \right\|^2}_{\text{effective noise in } U^{\perp}} - \underbrace{\left\| \Pi_S^{\perp} W \right\|^2}_{\text{effective noise in } S^{\perp}} \; < \; 0$$

- use large deviations to establish that

$$\mathbb{P}[\Delta(U) < 0] \quad \leq \quad \exp\left(-n \; F(\|\beta^*_{S \backslash U}\|^2; t) \right).$$

Proof sketch of necessary conditions

- Fano's inequality applied to a restricted ensemble, assuming *fixed choice* of β^*:

$$\beta_i^*[U] \quad = \quad \begin{cases} \beta_{min} & \text{if } i \in U \\ 0 & \text{otherwise.} \end{cases}$$

- by Fano's inequality, probability of success upper bounded as

$$1 - \mathbb{P}[\text{error}] \quad \leq \quad \frac{I(Y;\beta^*)}{\log(M-1)} - o(1),$$

where
  - $I(Y;\beta^*)$: mutual information between β^* and observation vector Y
  - $M = \binom{p}{k}$: number of competing models

- some work to establish the upper bound holds w.h.p. for X:

$$I(Y, \beta^* \mid X) \quad \leq \quad \frac{n}{2} \log \left[1 + (1 - \frac{k}{p}) k \beta_{min}^2 \right]$$

§2. High-dimensional analysis of sparse PCA

- principal components analysis (PCA): classical method for dimensionality reduction
- high-dimensional version: eigenvectors from sample covariance $\widehat{\Sigma}$ based on n samples in p dimensions
- in general, high-dimensional PCA inconsistent unless $p/n \to 0$ (Joh01, JohLu04)
- natural to investigate more structured ensembles for which consistency still possible even with $p/n \to +\infty$:
  - sparse eigenvector recovery (JolEtal03, JohLu04, ZouEtAl06)
  - sparse covariance matrices (LevBic06,ElKar07)

Spiked covariance ensembles

- sequences $\{\Sigma_p\}$ of spiked population covariance matrices:

$$\Sigma_p = \sum_{i=1}^{M} \alpha_i \beta_i \beta_i^T + \Gamma_p, \qquad \text{with leading eigenvectors } (\beta_i, i = 1, \ldots M).$$

- past work on identity spiked ensembles ($\Gamma_p = I_p$) (Joh01; JohLu04)
- different sparsity models:
  - hard sparsity model: β has exactly k non-zero coefficients
  - weak ℓ_q-sparsity: β belongs to the ℓ_q- "ball":

$$\mathbb{B}_q(R_q) \quad = \quad \{z \in \mathbb{R}^p \mid \sum_{i=1}^{p} |z_i|^q \leq R_q\}.$$

- given n i.i.d. samples $\{X_i\}_{i=1}^n$ with $\mathbb{E}[X_i] = 0$ and $\mathrm{cov}(X_i) = \Sigma_p$

SDP relaxation of sparse PCA

(D'Asprémont, El Ghaoui, Jordan & Lanckriet, 2006)

- Courant-Fischer variational principle for maximum eigenvalue/vector (PCA):

$$\lambda_{\max}(Q) \quad = \quad \max_{\|z\|_2=1} z^T Q z.$$

- equivalent/exact semidefinite program (SDP) of max. eigenvector:

$$\lambda_{\max}(Q) \quad = \quad \max_{Z \succeq 0, \text{trace}(Z)=1} \text{trace}\left(Z\, Q\right).$$

- *SDP relaxation* of sparse PCA:

$$\widehat{Z} \quad = \quad \arg \max_{Z \succeq 0, \text{trace}(Z)=1} \left\{ \text{trace}\left(Z\, Q\right) - \rho_n \Big(\sum_{i,j} |Z_{ij}| \Big) \right\},$$

with regularization parameter $\rho_n > 0$ chosen by user.

Rates in spectral norm

- given n samples from spiked identity model $\Sigma_p = \alpha z z^T + \sigma^2 I_p$
- eigenvector z in weak ℓ_q-ball $\mathbb{B}_q(R_q)$
- SDP relaxation: $\widehat{Z} \in \arg \min_{Z \succeq 0, \text{trace}(Z)=1} \big\{ - \text{trace}(Z \widehat{\Sigma}) + \rho_n \sum_{i,j} |Z_{ij}| \big\}$.

Theorem: (AmiWai08b) Suppose that we apply the SDP to the sample covariance $\widehat{\Sigma}$ with regularization parameter $\rho_n = f(\alpha, \sigma^2) \sqrt{\frac{\log p}{n}}$. Then with probability greater than $1 - c_1 \exp(-c_2 \log p) \rightarrow 0$, we have:

$$\|\widehat{Z} - z z^T\|_2 \leq C \; R_q \; \big(\frac{\log p}{n}\big)^{\frac{1}{2(1+q)}}.$$

Example (Hard sparsity): $q = 0$, and radius $R_q = k$ ($\#$ non-zeros)

$$\|\widehat{Z} - z z^T\|_2 \quad \leq \quad C \sqrt{\frac{k^2 \log p}{n}}.$$

Comparison to some known results

- Estimating sparse covariance matrices (BicLev07)
  - Thresholding estimator $T_{\lambda_n}(\widehat{\Sigma})$ achieves rate:

$$|||T_{\lambda_n}(\widehat{\Sigma}) - \Sigma|||_2 \quad \leq \quad C R_q \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}.$$

  - by matrix perturbation results, for "well-separated" eigenvalues, same rate applies to leading eigenvector
  - agrees with SDP result for $q = 0$, but slower rate for $q > 0$

- Minimax rates for $q \in (0, 2)$: (PauJoh08)
  - with $\mathrm{sign}\langle \widehat{z},\, z\rangle = 1$:

$$\min_{\widehat{z}} \max_{z \in \mathbb{B}_q(R_q)} \mathbb{E}[\|\widehat{z} - z\|_2^2] \quad \geq \quad C\, R_q \left(\frac{\log p}{n}\right)^{1-\frac{q}{2}}.$$

  - same rate as normal sequence model (DonJoh94)
  - SDP rate is slower, but approaches minimax rate as $q \to 0$

Model selection consistency for hard sparsity ($q = 0$)

Goal: Given spiked model with k-sparse eigenvector ($z_i = \pm\frac{1}{\sqrt{k}}$), recover support set $S(z) = \{i \in \{1, 2, \ldots, p\} \mid z_i \neq 0\}$ exactly.

Methods:

1. Diagonal thresholding: Complexity $\mathcal{O}(np + p \log p)$ (JohLu04)

   (a) Form sample covariance $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$.

   (b) Extract top k order statistics $\widehat{\Sigma}_{(11)}, \ldots, \widehat{\Sigma}_{(kk)}$, and estimate support $\widehat{S}(D)$ by rank indices.

2. SDP-based recovery: Complexity $\mathcal{O}(np + p^4 \log p)$ (AspLanGhaJor08)

   (a) Solve SDP with $\rho_n = \alpha/(2\sigma^2 k)$.

   (b) Given solution $\widehat{Z}$, estimate support

$$\widehat{S} \quad := \quad \big\{i \in \{1, \ldots, p\} \mid \widehat{Z}_{ij} \neq 0 \quad \text{for some } j\big\}.$$

Sharp threshold for diagonal thresholding

Model: $\Sigma_p = \alpha z z^T + \sigma^2 I_p$

Parameters:
- $p \equiv$ model dimension
- $k \equiv$ number of non-zeroes in spiked eigenvector

Proposition: (AmiWai08a) If $k = \mathcal{O}(p^{1-\delta})$ for any $\delta \in (0,1)$, diagonal thresholding for support recovery controlled by *rescaled sample size*

$$\theta_{\text{thr}}(n,p,k) \quad := \quad \frac{n}{k^2 \log(p-k)}.$$

I.e., there are constants $0 < \tau_\ell^*(\alpha, \sigma^2) \leq \tau_u^*(\alpha, \sigma^2) < \infty$ such that

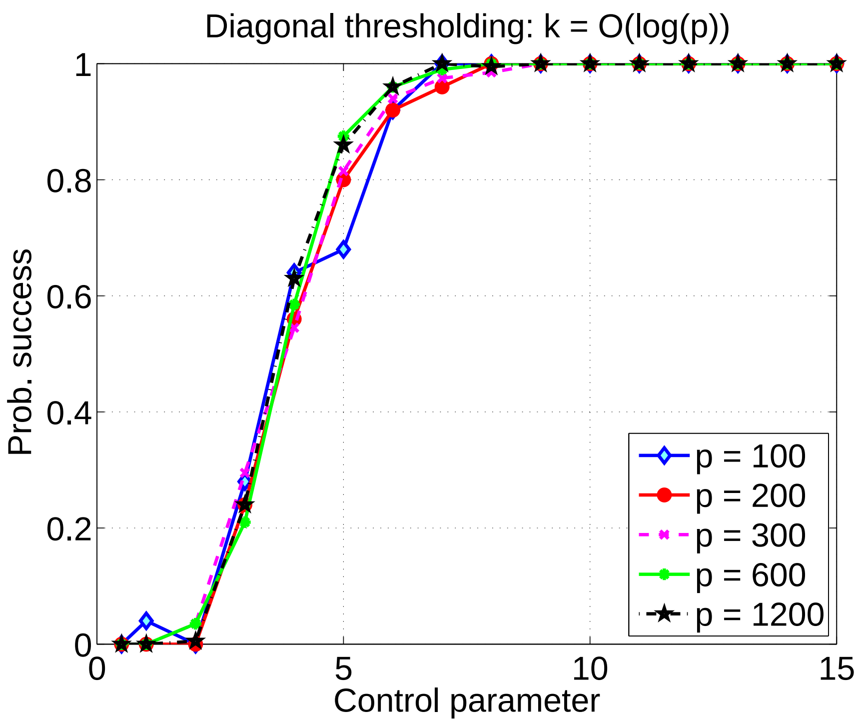
(a) **Success:** If $n > \tau_u^* k^2 \log(p-k)$, then

$$\mathbb{P}[\widehat{S}(D) = S(\beta)] \geq 1 - c_1 \exp\big(-c_2 k^2 \log(p-k)\big) \;\rightarrow\; 1.$$

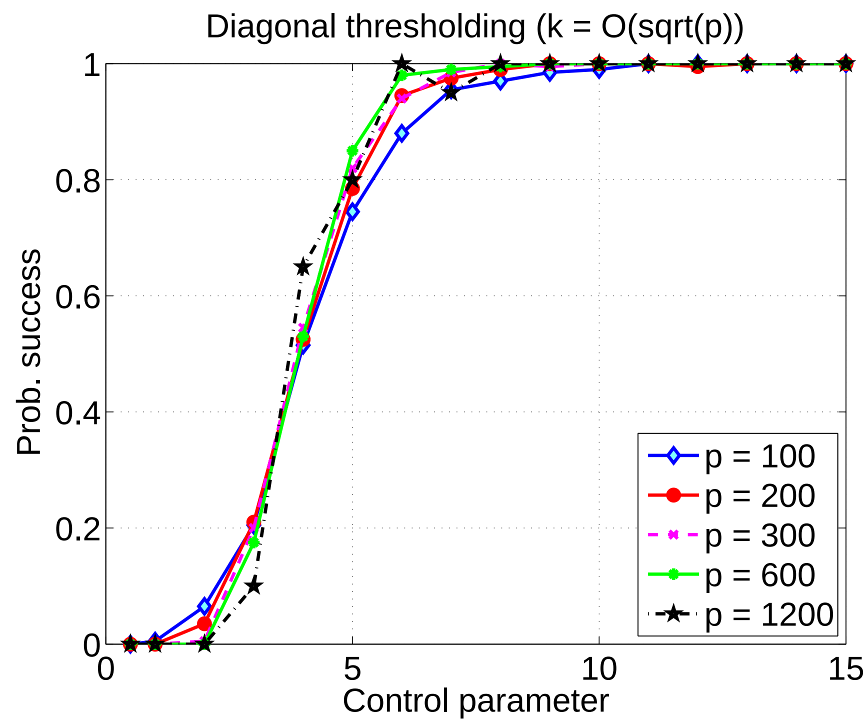
(b) **Failure:** If $n \leq \tau_\ell^* \, k^2 \log(p-k)$, then

$$\mathbb{P}[\widehat{S}(D) = S(\beta)] \leq c_1 \exp\left(-c_2(\log(p-k))\right) \;\rightarrow\; 0.$$

Performance of diagonal thresholding



(a) Log. sparsity



(b) Square-root sparsity

Probability of success $\mathbb{P}[S(D) = S(\beta^*)]$ versus rescaled sample size

$$\theta_{\text{thr}}(n, p, k) = \frac{n}{k^2 \log(p - k)}$$

Eigenvector support recovery via SDP relaxation

- spiked identity model $\Sigma_p = \alpha z z^T + \sigma^2 I_p$ with k-sparse eigenvector z
- SDP relaxation: $\widehat{Z} \in \arg \min_{Z \succeq 0, \text{trace}(Z)=1} \{ -\text{trace}(Z\widehat{\Sigma}) + \rho_n \sum_{i,j} |Z_{ij}| \}$.

Theorem: (AmiWai08a) Suppose that we solve the SDP with $\rho_n = \alpha/(2\sigma^2 k)$. Then there are constants θ_{wr} and θ_{crit} such that

(a) For sample sizes such that $\theta_{\text{thr}}(n, p, k) = \frac{n}{k^2 \log(p-k)} > \theta_{\text{wr}}$, the SDP has a rank one solution w.h.p., and

(b) For problem sequences such that $k = \mathcal{O}(\log p)$, and

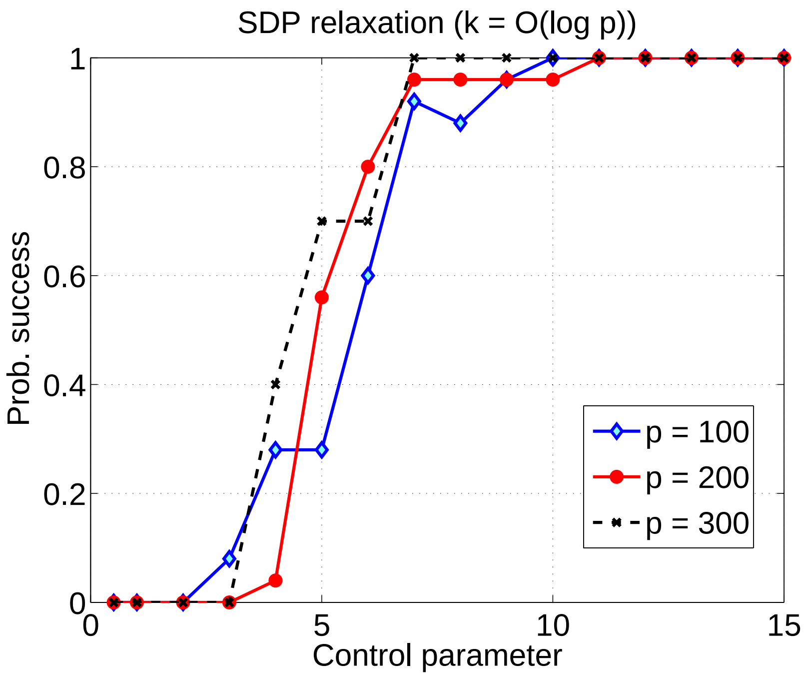
$$\theta_{\text{sdp}}(n, p, k) \quad := \quad \frac{n}{k \log(p-k)} \quad > \quad \theta_{\text{crit}},$$

a rank one solution (when it exists) specifies correct support w.h.p.

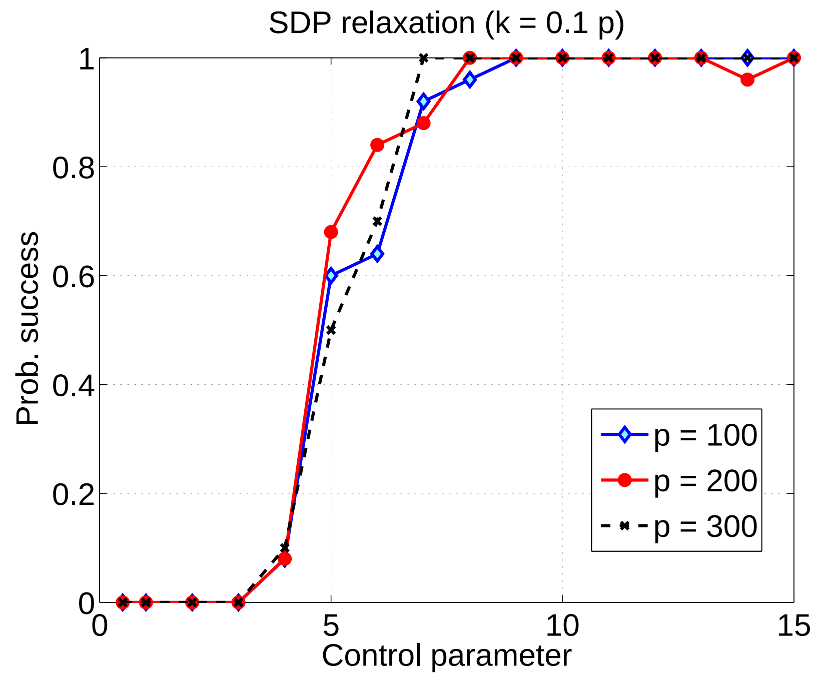
Remarks:

- technical condition $k = \mathcal{O}(\log p)$: likely an artifact

Performance of SDP relaxation



(a) Log. sparsity



(b) Linear sparsity

Probability of success $\mathbb{P}[S(\widehat{\beta}) = S(\beta^*)]$ versus rescaled sample size

$$\theta_{\text{sdp}}(n, p, k) = \frac{n}{k \log(p-k)}.$$

Summary and open directions

1. When does more computation yield greater statistical accuracy?
   - Multivariate regression: second-order cone programming versus quadratic programming (Lasso)
   - Sparse PCA: diagonal thresholding versus SDP relaxation
2. When are polynomial-time algorithms as good as "optimal" algorithms?
   - Multivariate regression: Lasso/SOCP order-optimal for $k = o(p)$
   - Sparse PCA: SDP relaxation order-optimal for $k = \mathcal{O}(\log p)$