APTS High-Dimensional Statistics

Po-Ling Loh[*]

July 17, 2024

Contents

[*] Department of Pure Mathematics and Mathematical Statistics, University of Cambridge. Please send corrections to: pll28@cam.ac.uk

1 Introduction

- High-dimensional statistics: # parameters $p \gg$ # of observations n
- Applications: genetic analysis, health studies, medical imaging, astronomy, climatology ($p \approx 10,000$, $n \approx 100$)
- Key is to leverage low-dimensional structure in data set

Tools: Linear algebra, optimization theory, concentration inequalities

2 The Lasso

Reference: Bühlmann & van de Geer, *Statistics for High-Dimensional Data*, 2011

Linear model:

$$y_i = x_i^T \beta^* + \epsilon_i, \qquad \text{for} \quad 1 \le i \le n.$$

Observations are $\{(x_i, y_i)\}_{i=1}^n$, goal is to estimate $\beta^* \in \mathbb{R}^p$. Assume ϵ_i's are i.i.d. and $\mathbb{E}(\epsilon_i) = 0$, and x_i's fixed or random (in random design case, $x_i \perp\!\!\!\perp \epsilon_i$).

Classical statistics: Ordinary least squares estimator given by

$$\widehat{\beta}_{OLS} \in \arg\min_{\beta} \left\{ \sum_{i=1}^n (y_i - x_i^T \beta)^2 \right\} = (X^T X)^{-1} X^T y.$$

Consistency:

$$\widehat{\beta}_{OLS} \xrightarrow{P} \beta^* \qquad \text{as} \quad n \to \infty.$$

Asymptotic normality:

$$\sqrt{n}(\widehat{\beta}_{OLS} - \beta^*) \xrightarrow{d} N(0, \sigma^2 \Sigma^{-1}) \qquad \text{as} \quad n \to \infty,$$

when $\sigma^2 = Var(\epsilon_i)$ and $x_i \sim N(0, \Sigma)$.

What about $p > n$? Issues:

$$X^T X = \sum_{i=1}^n x_i x_i^T$$

is a sum of n rank-1 matrices, hence

$$\text{rank}(X^T X) \leq n < p,$$

so $(X^T X)^{-1}$ is not defined. In fact, solutions to $\arg\min_\beta \|y - X\beta\|_2^2$ span an entire subspace.

Solution is to enforce/assume sparsity. Suppose $\|\beta^*\|_0 \leq k < n$, and solve the regularized problem

$$\widehat{\beta}_{Lasso} \in \arg\min_\beta \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

a convex relaxation of

$$\widehat{\beta} \in \arg\min_\beta \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right\}.$$

Theorem 1. *Under suitable conditions on X and ϵ, if λ is chosen sufficiently large, then*

$$\|\widehat{\beta}_{Lasso} - \beta^*\|_2 \leq c\lambda\sqrt{k}.$$

Theorem 2. *Under additional suitable conditions, we have*

$$\text{supp}(\widehat{\beta}_{Lasso}) = \text{supp}(\beta^*).$$

How big should n be? In Gaussian case, “suitable conditions” hold w.h.p. when $n \geq Ck \log p$. Information-theoretic arguments show that $n \geq C' k \log\left(\frac{p-k}{k}\right)$ is required for consistent estimation w.h.p.

2.1 Proof of Theorem 1

Basic inequality:

$$\frac{1}{2n}\|y - X\widehat{\beta}\|_2^2 + \lambda\|\widehat{\beta}\|_1 \leq \frac{1}{2n}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1.$$

After some algebra, and using the substitution $y = X\beta^* + \epsilon$:

$$\frac{1}{n}\|X(\widehat{\beta} - \beta^*)\|_2^2 \leq \frac{2\epsilon^T X(\widehat{\beta} - \beta^*)}{n} + 2\lambda\|\beta^*\|_1 - 2\lambda\|\widehat{\beta}\|_1. \tag{1}$$

Next step: We want to lower-bound the LHS by $\alpha\|\widehat{\beta} - \beta^*\|_2^2$, but unfortunately, $\lambda_{\min}\left(\frac{X^T X}{n}\right) = 0$. However, $\widehat{\beta} - \beta^*$ is not an *arbitrary* vector, but is in a special cone set

$$\mathbb{C} = \left\{ v \in \mathbb{R}^p : \|v_{S^c}\|_1 \leq 3\|v_S\|_1 \right\},$$

where $S = \operatorname{supp}(\beta^*)$. Note that vectors supported only on S are always in this cone; the intuition is that regularization encourages $\widehat{\beta}$ to be close to the sparse set, hence in $\mathbb{C}$.

To show that $\widehat{\nu} := \widehat{\beta} - \beta^* \in \mathbb{C}$, argue as follows: lower-bounding the LHS of inequality (1) by 0, we have

$$\begin{aligned} 0 &\leq \frac{2\epsilon^T X \widehat{\nu}}{n} + 2\lambda\|\beta^*\|_1 - 2\lambda\|\widehat{\beta}\|_1 \\ &\leq 2\|\widehat{\nu}\|_1 \left\| \frac{X^T\epsilon}{n} \right\|_\infty + 2\lambda\|\beta^*_S\|_1 - 2\lambda\|\widehat{\beta}_S\|_1 - 2\lambda\|\widehat{\beta}_{S^c}\|_1 \\ &\leq 2\|\widehat{\nu}\|_1 \left\| \frac{X^T\epsilon}{n} \right\|_\infty + 2\lambda\|\widehat{\nu}_S\|_1 - 2\lambda\|\widehat{\nu}_{S^c}\|_1 \end{aligned} \tag{2}$$

using Hölder's inequality in the first inequality and the triangle inequality in the second.

We now impose the assumption that λ is sufficiently large:

$$\lambda \geq 2\left\| \frac{X^T\epsilon}{n} \right\|_\infty .$$

Plugging into the chain of inequalities gives

$$0 \leq 2\|\widehat{\nu}\|_1 \cdot \frac{\lambda}{2} + 2\lambda\|\widehat{\nu}_S\|_1 - 2\lambda\|\widehat{\nu}_{S^c}\|_1 = 3\lambda\|\widehat{\nu}_S\|_1 - \lambda\|\widehat{\nu}_{S^c}\|_1, \tag{3}$$

implying that $\widehat{\nu} \in \mathbb{C}$.

We impose one more condition, the *restricted eigenvalue (RE) condition*:

$$v^T \frac{X^T X}{n} v \geq \alpha\|v\|_2^2, \qquad \forall v \in \mathbb{C}$$

We now apply this condition to lower-bound the LHS of inequality (1). Combined with the upper bounds in inequalities (2) and (3) then gives

$$\alpha\|\widehat{\nu}\|_2^2 \leq 3\lambda\|\widehat{\nu}_S\|_1 - \lambda\|\widehat{\nu}_{S^c}\|_1 \leq 3\lambda\|\widehat{\nu}_S\|_1 \leq 3\lambda\sqrt{k}\|\widehat{\nu}_S\|_2 \leq 3\lambda\sqrt{k}\|\widehat{\nu}\|_2.$$

Hence,

$$\|\widehat{\nu}\|_2 \leq \frac{3\lambda\sqrt{k}}{\alpha}.$$

We can also argue about ℓ_1-error, as follows:

$$\|\widehat{\nu}\|_1 = \|\widehat{\nu}_{S^c}\|_1 + \|\widehat{\nu}_S\|_1 \leq 4\|\widehat{\nu}_S\|_1 \leq 4\sqrt{k}\|\widehat{\nu}_S\|_2 \leq \frac{12\lambda k}{\alpha}.$$

Restating, we have the following result:

Theorem. *Suppose the regularization parameter satisfies* $\lambda \geq 2\left\| \frac{X^T\epsilon}{n} \right\|_\infty$ *and the design matrix satisfies* $v^T \frac{X^T X}{n} v \geq \alpha\|v\|_2^2, \quad \forall v \in \mathbb{C}$. *Then the Lasso solution satisfies the bounds*

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{3\lambda\sqrt{k}}{\alpha}, \quad \textit{and} \quad \|\widehat{\beta} - \beta^*\|_1 \leq \frac{12\lambda k}{\alpha}.$$

2.2 Case study: Gaussians

Recap: "suitable conditions" for Theorem 1 were:

1. RE condition on X
2. $\lambda \geq 2 \left\| \frac{X^T \epsilon}{n} \right\|_\infty$.

RE condition imposes some kind of curvature on objective function, which is important for showing closeness of $\widehat{\beta}$ to β^*: we want that if the objective function values are close, the points where the function is evaluated are close, as well. Furthermore, it is reasonable to need λ sufficiently large, or else we are back to solving the OLS problem.

We also need to verify that the sufficient conditions for the theorem hold in reasonable settings of interest. First consider a deterministic design matrix ($X \in \mathbb{R}^{n \times p}$ is fixed), such that X satisfies the RE condition, while $\epsilon_i \sim N(0, \sigma^2)$. If we further assume that X is column-normalized, so $\max_{1 \leq j \leq p} \frac{\|X_j\|_2}{\sqrt{n}} \leq C$, where X_j is the j^{th} column of X. Then $\left\| \frac{X^T \epsilon}{n} \right\|_\infty$ is the absolute maximum of p zero-mean Gaussians, each with variance at most $\frac{C^2 \sigma^2}{n}$. Hence, from standard Gaussian tail bounds (exercise!), we have

$$\mathbb{P}\left(\left\| \frac{X^T \epsilon}{n} \right\|_\infty \geq C\sigma \left(\sqrt{\frac{2 \log p}{n}} + t \right) \right) \leq 2e^{-nt^2/2}, \qquad \forall t > 0,$$

so we can choose $t \asymp \sqrt{\frac{\log p}{n}}$ and conclude that the choice $\lambda \geq C' \sigma \sqrt{\frac{\log p}{n}}$ is valid, w.h.p.

For the "random design" setting, suppose $x_i \sim N(0, \Sigma)$ and $\epsilon_i \sim N(0, \sigma^2)$. Then we have the following two facts:

- Fact 1: The RE condition holds with high probability $(1 - \exp(-cn))$ when $n \geq Ck \log p$, with $\alpha = \frac{1}{4} \lambda_{\min}(\Sigma_x)$.
- Fact 2: With high probability,

$$\left\| \frac{X^T \epsilon}{n} \right\|_\infty \leq C' \sigma |||\Sigma|||_2^{1/2} \sqrt{\frac{\log p}{n}}.$$

Thus, we can safely choose any $\lambda \geq C' \sigma |||\Sigma|||_2^{1/2} \sqrt{\frac{\log p}{n}}$ for the Lasso theory to hold. For the "optimal" choice of λ, we are guaranteed that

$$\|\widehat{\beta}_{Lasso} - \beta^*\|_2 \leq C'' \sqrt{\frac{k \log p}{n}},$$
$$\|\widehat{\beta}_{Lasso} - \beta^*\|_1 \leq 4C'' k \sqrt{\frac{\log p}{n}}.$$

Details for the Gaussian case may be found in Raskutti, Wainwright & Yu (2010). Extensions to the sub-Gaussian setting exist as well, using concentration results as in the monograph by Vershynin (2012).

2.3 Proof of Theorem 2

We mention some highlights of the proof of the theorem regarding support recovery. More details may be found in Wainwright (2009).

The key idea is a “primal-dual witness” (PDW) construction. The main steps are as follows:

1. Solve the restricted problem, where the support of β is restricted to the true support S. Show that extending the solution to the p-dimensional space by augmenting 0’s yields a local/global optimum $\widehat{\beta}$.
2. Establish strict dual feasibility at $\widehat{\beta}$.
3. Argue that all solutions $\widetilde{\beta}$ fo the extended problem are also supported on S (using RE and other inequalities regarding the proximity of $\widetilde{\beta}$ to $\widehat{\beta}$).

We now elaborate on each of these steps. If $\mathrm{supp}(\widehat{\beta}) = \mathrm{supp}(\beta^*)$, the solution to the Lasso should agree with the solution to the restricted problem

$$\widehat{\beta}_S \in \arg\min_{\beta_S \in \mathbb{R}^S} \left\{ \frac{1}{2n} \|y - X_S \beta_S\|_2^2 + \lambda \|\beta_S\|_1 \right\},$$

where we only regress on the covariates corresponding to the true support set $\mathrm{supp}(\beta^*)$. More precisely, $\widehat{\beta} = (\widehat{\beta}_S, 0_{S^c})^T$ should be a solution of the Lasso.

How do we ensure this? We can examine the (sub)gradient of the Lasso objective, and check whether this vector $\widehat{\beta}$ makes the gradient 0. In other words, we need

$$-\frac{1}{n} X^T (y - X\widehat{\beta}) + \lambda \,\mathrm{sign}(\widehat{\beta}) = 0. \tag{4}$$

The function $\mathrm{sign}(\cdot)$ is defined componentwise, according to

$$\mathrm{sign}(u) = \begin{cases} 1 & \text{if } u > 0 \\ -1 & \text{if } u < 0 \\ \text{anything in } [-1, 1] & \text{if } u = 0. \end{cases}$$

So we can substitute $\mathrm{sign}(\widehat{\beta}) = (\mathrm{sign}(\widehat{\beta}_S), \hat{z}_{S^c})^T$, and solve the system of equations (4) for $\hat{z}_{S^c}$. Rewritten in block matrix form:

$$\frac{1}{n} \begin{pmatrix} X_S^T X_S & X_S^T X_{S^c} \\ X_{S^c}^T X_S & X_{S^c}^T X_{S^c} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_S - \beta_S^* \\ 0 \end{pmatrix} - \frac{1}{n} \begin{pmatrix} X_S^T \epsilon \\ X_{S^c}^T \epsilon \end{pmatrix} + \lambda \begin{pmatrix} \hat{z}_S \\ \hat{z}_{S^c} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Then if we can show that

$$\|\hat{z}_{S^c}\|_\infty \le 1,$$

we will guarantee that $\widehat{\beta} = (\widehat{\beta}_S, 0_{S^c})^T$ is indeed a solution to the Lasso.

Note that after showing $\widehat{\beta}$ is a solution, we still need to show that any other solution is also supported in S. This requires another side argument (see the lemma below), and it turns out that ensuring the slightly stronger condition

$$\|\hat{z}_{S^c}\|_\infty < 1 \tag{5}$$

(together with some other conditions, namely the mutual incoherence condition) will give us the desired uniqueness.

Lemma 3. *If strict dual feasibility (5) holds, then $(\widehat{\beta}_S, 0_{S^c})^T$ is the unique optimal solution.*

Proof. Suppose $\widetilde{\beta}$ is another solution. Then

$$\frac{1}{2n}\|y - X\widehat{\beta}\|_2^2 + \lambda\langle \hat{z}, \widehat{\beta}\rangle = \frac{1}{2n}\|y - X\widetilde{\beta}\|_2^2 + \lambda\langle \tilde{z}, \widetilde{\beta}\rangle,$$

so

$$\mathcal{L}_n(\widehat{\beta}) - \lambda\langle \hat{z}, \widetilde{\beta} - \widehat{\beta}\rangle = \mathcal{L}_n(\widetilde{\beta}) + \lambda\left(\|\widetilde{\beta}\|_1 - \langle \hat{z}, \widetilde{\beta}\rangle\right).$$

By the zero-subgradient condition, we have $\nabla\mathcal{L}_n(\widehat{\beta}) + \lambda\hat{z} = 0$, so

$$\mathcal{L}_n(\widehat{\beta}) + \langle \nabla\mathcal{L}_n(\widehat{\beta}), \widetilde{\beta} - \widehat{\beta}\rangle - \mathcal{L}_n(\widetilde{\beta}) = \lambda\left(\|\widetilde{\beta}\|_1 - \langle \hat{z}, \widetilde{\beta}\rangle\right).$$

Note that the LHS is upper-bounded by 0, by convexity. Therefore,

$$\|\widetilde{\beta}\|_1 \leq \langle \hat{z}, \widetilde{\beta}\rangle \leq \|\hat{z}\|_\infty \|\widetilde{\beta}|_1 \leq \|\widetilde{\beta}\|_1,$$

implying that $\langle \hat{z}, \widetilde{\beta}\rangle = \|\widetilde{\beta}\|_1$. But together with inequality (5), this implies that $\widetilde{\beta}_{S^c} = 0_{S^c}$. □

Theorem. *Suppose* $\operatorname{supp}(\beta^*) = S$ *and* $\epsilon_i \sim N(0, \sigma^2)$*, and* X *satisfies the following assumptions:*

1. $\max_{1 \leq j \leq p} \frac{\|X_j\|_2}{\sqrt{n}} \leq C$*,*
2. $\lambda_{\min}\left(\frac{X_S^T X_S}{n}\right) \geq c_{\min} > 0$*,*
3. $\exists \alpha \in [0, 1)$ *such that* $\max_{j \in S^c} \|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \leq \alpha$*.*

If $\lambda = \frac{2C\sigma}{1-\alpha}\left(\sqrt{\frac{2\log(p-k)}{n}} + \delta\right)$*, for some* $\delta > 0$*, then the Lasso solution* $\widehat{\beta}$ *is unique, with support contained within* S*, and*

$$\|\widehat{\beta} - \beta^*\|_\infty \leq \underbrace{\frac{\sigma}{\sqrt{c_{\min}}}\left(\sqrt{\frac{2\log k}{n}} + \delta\right) + \left|\!\left|\!\left|\left(\frac{X_S^T X_S}{n}\right)^{-1}\right|\!\right|\!\right|_\infty \lambda}_{B(\lambda, X)},$$

with probability at least $1 - 4e^{-n\delta^2/2}$*.*

In particular, if $\min_{j \in S} |\beta_j^*| > B(\lambda, X)$, the Lasso is variable selection consistent.

Remark 1. *Most of the canonical Lasso theory is proven for (sub)-Gaussian design matrices. However, this may not be so useful in compressed sensing, where* $X \in \mathbb{C}^{n \times p}$*. Furthermore, we may have structural constraints on* X *to make multiplication/storage easier (more on that later).*

2.4 Extensions/generalization

Many extensions exist to various problem settings, including the following:

1. *Sparse logistic regression.* We observe $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$ and $y_i \in \{0, 1\}$. The y_i's are related to the x_i's via
$$P(y_i = 1 \mid x_i, \beta^*) = \frac{\exp(x_i^T \beta^*)}{1 + \exp(x_i^T \beta^*)} = \exp\left(x_i^T \beta^* - \log(1 + \exp(x_i^T \beta^*)\right).$$
The penalized log likelihood objective takes the form
$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^n \left(-y_i x_i^T \beta^* + \log(1 + \exp(x_i^T \beta^*)\right) + \lambda \|\beta\|_1 \right\}.$$

2. *Graphical Lasso.* We observe $x_i \sim N(0, \Sigma^*)$, and wish to estimate a sparse matrix $\Theta^* := (\Sigma^*)^{-1}$ (this is useful in graphical model estimation). The penalized log likelihood objective is
$$\widehat{\Theta} \in \arg\min_{\Theta} \left\{ \operatorname{tr}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \lambda \sum_{i \neq j} |\Theta_{ij}| \right\}.$$
(More details on this later.)

3. *Low-rank matrix approximation.* We observe $\{(X_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^{p_1 \times p_2}$ and the y_i's are related to the x_i's via
$$y_i = \operatorname{tr}(X_i^T \Theta^*) + \epsilon_i,$$
and the ϵ_i's are i.i.d. We assume $\operatorname{rank}(\Theta^*)$ is small. Then we use the nuclear-norm penalized objective
$$\widehat{\Theta} \in \arg\min_{\Theta} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(y_i - \operatorname{tr}(X_i^T \Theta)\right)^2 + \lambda \sum_{j=1}^{\min(p_1, p_2)} \sigma_j(\Theta) \right\},$$
where the σ_j's are the singular values. This encourages sparsity of singular values, hence low-rankedness.

All the above problems may be analyzed using a similar procedure as the Lasso for linear regression: Start with a basic inequality, manipulate, and bound the norm of the error. The RE condition is replaced by a restricted strong convexity (RSC) condition. More details may be found in Negahban, Ravikumar, Wainwright & Yu (2012).

2.5 Alternative approach (noiseless case)

Basis pursuit: For matrix $\Phi \in \mathbb{R}^{m \times n}$ and vector $y \in \mathbb{R}^m$, consider the optimization problem
$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \|x\|_0 \\ \text{s.t.} \quad & \Phi x = y. \end{aligned} \tag{6}$$
In general, non-convex program (6) is NP-hard to solve, so we consider the *basis pursuit* program
$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \|x\|_1 \\ \text{s.t.} \quad & \Phi x = y. \end{aligned} \tag{7}$$

Definition: A matrix $\Phi \in \mathbb{R}^{m \times n}$ satisfies the (ϵ, k)-RIP if for all k-sparse vectors x,

$$(1-\epsilon)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\epsilon)\|x\|_2^2. \tag{8}$$

Equivalently, whenever $|S| \leq k$, all eigenvalues of $\Phi_S^T \Phi_S$ are in $[1-\epsilon, 1+\epsilon]$, or

$$\left|\!\left|\!\left| \Phi_S^T \Phi_S - I_k \right|\!\right|\!\right|_2 \leq \epsilon$$

(exercise!).

Theorem 4. *Let $k \leq n$. If Φ satisfies $(\epsilon, 2k)$-RIP with $\epsilon \leq \sqrt{2} - 1$, the solution $\hat{x}$ to the basis pursuit program (7) with $y = \Phi x^*$ satisfies*

$$\|\hat{x} - x^*\|_2 \leq \frac{c}{\sqrt{k}} \|x^*_{tail(k)}\|_1,$$

where $x^*_{tail(k)}$ denotes x^* with top k components zeroed out.

Note: In particular, if $\|x^*\|_0 \leq k$, the BP program (7) is exact, since $\|x^*_{tail(k)}\|_1 = 0$. For a proof, see Candes, “The RIP and its implications for compressed sensing,” 2008.

Connections: Our earlier analysis shows that the Lasso strategy leads to ℓ_2-error bounds under an RE condition on X. Some random matrix ensembles (e.g., spiked identity Gaussian matrices) may be seen to satisfy RE, but not RIP. For more details and comparisons, see:

- Raskutti et al., “RE properties for correlated Gaussian designs,” *JMLR*, 2010.
- Bühlmann & van de Geer, “Statistics for high-dimensional data,” 2011.

3 Graphical models

Let $G = (V, E)$ denote the undirected graphical model associated with the joint distribution, where $V = \{1, \dots, p\}$ and $E \subseteq \{0,1\}^{\binom{p}{2}}$, and $(j,k) \notin E$ if and only if $X_j \perp\!\!\!\perp X_k \mid X_{V \setminus \{j,k\}}$. The graph G is also known as the *conditional independence graph.* For each $j \in V$, let $N(j) = \{k \in V : (j,k) \in E\}$ denote the neighborhood set of j. Let $d = \deg(G)$ denote the degree of G.

3.1 Gaussian graphical models

We begin by discussing edge recovery methods for Gaussian graphical models. Consider the case where $(X_1, \dots, X_p) \sim N(0, \Sigma)$ are joint observations from a multivariate normal distribution, and let $\Theta = \Sigma^{-1}$ denote the inverse covariance matrix. The edge recovery algorithms presented in this section are largely based on a critical observation regarding the relationship between entries of Θ and edges of the conditional independence graph G.

3.1.1 Inverse covariance matrix and edge structure

Recall that the probability density function of the multivariate Gaussian distribution is given by

$$\begin{aligned} q(x_1, \dots, x_p) &= \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} x^T \Theta x\right) \\ &\propto \exp\left(-\frac{1}{2} \sum_{j,k} \Theta_{jk} x_j x_k\right). \end{aligned}$$

It is easy to see (exercise!) that for any $j \neq k$ with $\Theta_{jk} = 0$, we may write

$$q(x_1, \dots, x_p) = q_1(x_j, x_{\backslash\{j,k\}}) q_2(x_k, x_{\backslash\{j,k\}}),$$

for functions $q_1, q_2 > 0$, from which we may deduce that $X_j \perp\!\!\!\perp X_k \mid X_{\backslash\{j,k\}}$. Conversely, if $\Theta_{jk} \neq 0$, such a decomposition is impossible, implying that $X_j \not\perp\!\!\!\perp X_k \mid X_{\backslash\{j,k\}}$. It follows that for all $j \neq k$,

$$(j,k) \in E \iff \Theta_{jk} \neq 0. \tag{9}$$

In other words, the support set $\text{supp}(\Theta)$, discounting diagonals, corresponds precisely to the edge structure of the graph. This is the mainstay of many neighborhood selection algorithms for multivariate Gaussian graphical models.

3.1.2 Edge recovery via matrix estimation

Based on the observations of the previous section, it suffices to devise an estimate of the inverse covariance matrix Θ when we are given a data matrix $X \in \mathbb{R}^{n \times p}$ of n i.i.d. observations from the joint distribution. A simple calculation (exercise!) shows that the maximum likelihood estimator

$$\widehat{\Theta}_{MLE} = \arg\min_{\Theta \succ 0} \left\{ \text{tr}(\widehat{\Sigma}\Theta) - \log\det(\Theta) \right\} \tag{10}$$

is given by $\widehat{\Theta}_{MLE} = \left(\widehat{\Sigma}\right)^{-1}$, provided the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ is invertible. However, when the number of nodes p exceeds the number of dimensions n, the matrix $\widehat{\Sigma}$ is low-rank, hence uninvertible. Various alternative estimators have consequently been proposed that are applicable in high dimensions; we will discuss two such methods below. Both algorithms produce an estimate $\widehat{\Theta}$ of Θ, from which we may in turn estimate the nonzero pattern $\text{supp}(\Theta)$.

Graphical Lasso. The graphical Lasso, first proposed in the literature by Yuan and Lin [25], adds a penalty term to the maximum likelihood expression (10):

$$\widehat{\Theta}_{GLASSO} = \arg\min_{\Theta \succ 0} \left\{ \text{tr}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}| \right\}. \tag{11}$$

The penalty term consists of the regularization parameter λ, multiplied by the ℓ_1-norm of off-diagonal entries of Θ, and encourages a sparse matrix solution. We then define our estimate of the edge set to be $\widehat{E} = \text{supp}(\widehat{\Theta})$, where we abuse notation slightly and disregard the diagonal entries of $\widehat{\Theta}$.

Note that the graphical Lasso is a convex program. Hence, the solution $\widehat{\Theta}_{GLASSO}$ may be obtained efficiently using standard interior point methods [3]. However, generic optimization algorithms may be fairly slow when applied to extremely large data sets, and various authors have proposed more efficient methods specifically designed for solving the graphical Lasso program (11) (e.g., [6, 7, 13, 9]).

3.1.3 Edge recovery via linear regression

We now outline a fundamental relationship between linear regression and estimation of Θ. Note that for each $1 \leq j \leq p$, by properties of Gaussian random variables, we may write

$$X_j = \theta_j^T X_{\setminus\{j\}} + W_j,$$

where

$$\theta_j = \left(\Sigma_{\setminus\{j\},\setminus\{j\}}\right)^{-1} \Sigma_{\setminus\{j\},j} \in \mathbb{R}^{p-1}, \tag{12}$$

$W_j \in \mathbb{R}$ is normally distributed with mean 0, and $W_j \perp\!\!\!\perp X_{\setminus\{j\}}$.

On the other hand, note that by block matrix inversion [8] (exercise!), we have

$$\begin{aligned} \Theta &= \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,\setminus\{1\}} \\ \Sigma_{\setminus\{1\},1} & \Sigma_{\setminus\{1\},\setminus\{1\}} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} a_1 & -a_1 \Sigma_{1,\setminus\{1\}} \left(\Sigma_{\setminus\{1\},\setminus\{1\}}\right)^{-1} \\ -a_1 \left(\Sigma_{\setminus\{1\},\setminus\{1\}}\right)^{-1} \Sigma_{\setminus\{1\},1} & \left(\Sigma_{\setminus\{1\},\setminus\{1\}} - \Sigma_{\setminus\{1\},1}\Sigma_{1,1}^{-1}\Sigma_{1,\setminus\{1\}}\right)^{-1} \end{pmatrix}, \end{aligned} \tag{13}$$

where $a_1 = \left(\Sigma_{1,1} - \Sigma_{1,\setminus\{1\}}\Sigma^{-1}_{\setminus\{1\},\setminus\{1\}}\Sigma_{\setminus\{1\},1}\right)^{-1}$. Hence, the first column of Θ is a constant multiple of the vector $\begin{pmatrix} -1 \\ \theta_1 \end{pmatrix}$, and an analogous statement may be made for each value of j. In particular, recovering $\text{supp}(\theta_j)$, for each $1 \leq j \leq p$, allows us to recover $\text{supp}(\Theta)$; using the equivalence (9), the set $\text{supp}(\theta_j)$ corresponds precisely to the neighborhood set $N(j)$.

Nodewise Lasso. Motivated by the relationship described above, Meinshausen and Bühlmann [15] proposed a method for estimating the neighborhood sets $\{N(j) : 1 \leq j \leq p\}$ via successive linear regressions: In particular, the relation (12) implies that the random variable $\theta_j^T X_{\setminus\{j\}}$ is the best linear predictor of X_j in terms of $X_{\setminus\{j\}}$, so performing a linear regression of the observations corresponding to X_j upon the vector-valued observations $X_{\setminus\{j\}}$ converges to θ_j as $n \to \infty$. In the setting of interest where $p > n$, the estimate of θ_j is given by the solution of the Lasso program

$$\widehat{\theta}_j = \arg\min_{\theta} \left\{ \|X^j - X^{\setminus j}\theta\|_2^2 + \lambda_j \|\theta\|_1 \right\}, \tag{14}$$

where X^j denotes the j^{th} column vector of the $n \times p$ data matrix and $X^{\setminus j}$ denotes the $n \times (p-1)$ block containing the remaining columns. We then define our neighborhood estimates to be

$$\widehat{N(j)} = \text{supp}(\widehat{\theta}_j), \qquad \forall 1 \leq j \leq p,$$

and combine our neighborhood estimates into a global edge estimate $\widehat{E}$ using an AND or OR rule:

Input: Neighborhood estimates $\{N(j)\}_{1 \leq j \leq p}$
Output: Edge set estimate $\widehat{E}$

AND rule: For each $j \neq k$,

$$(j,k) \in \widehat{E} \iff j \in \widehat{N(k)} \text{ AND } k \in \widehat{N(j)}$$

OR rule: For each $j \neq k$,

$$(j,k) \in \widehat{E} \iff j \in \widehat{N(k)} \text{ OR } k \in \widehat{N(j)}.$$

The theoretical results described in the next subsection guarantee that with a sufficiently large sample size and under certain regularity conditions, we have $\widehat{N(j)} = N(j)$, with high probability, for all j. Hence, the estimated edge set $\widehat{E}$ will be a consistent estimate of E, regardless of which rule is applied.

3.1.4 Statistical theory

We now highlight some theoretical results concerning the success of the algorithms described above. At a high level, all the results guarantee that when the number of samples n scales as a power of d times $\log p$, we have $\widehat{E} = E$, with high probability. However, the results differ in the types of conditions imposed on the underlying distribution. For the results of this section, we will use Σ^* and Θ^* to denote the true covariance and inverse covariance matrices, respectively, so our data matrix $X \in \mathbb{R}^{n \times p}$ consists of n i.i.d. draws from the distribution $N(0, \Sigma^*)$. The constants c_i appearing in the statistical results refer to universal constants, the values of which may vary between theorems. We have also suppressed the dependence of the results on minimum eigenvalues of Σ^*.

Theory for graphical Lasso. We begin by discussing the performance of the graphical Lasso (11). Numerous theoretical results have been derived concerning the convergence of $\widehat{\Theta}_{GLASSO}$ to Θ^*, where convergence is measured in various norms (e.g, Frobenius norm [19], spectral norm, and elementwise ℓ_∞-norm [17]). Since the topic of this section is neighborhood selection, we will focus on edge recovery guarantees; i.e., conditions under which $\widehat{E} = E$, with high probability. In the following theorem, we require the *α-incoherence* condition:

$$\max_{e \in S^c} \|\Gamma^*_{eS} \left(\Gamma^*_{SS}\right)^{-1}\|_1 \leq 1 - \alpha \tag{15}$$

where $\Gamma^* = \Sigma^* \otimes \Sigma^* \in \mathbb{R}^{p^2 \times p^2}$ is the tensor product of true covariance matrices, the augmented edge set S is defined according to $S := E \cup \{(j,j) : j \in V\}$, and $S^c := (V \times V) \backslash S$.

Theorem 5 (Ravikumar et al. [17])**.** *Suppose we have the α-incoherence condition* (15) *for some $\alpha \in (0,1]$. Also suppose the regularization parameter is chosen to be $\lambda = \frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}} + \delta$, for some $\delta \in [0,1]$, and suppose the sample size satisfies $n \geq c_2 \left(1 + \frac{8}{\alpha}\right)^2 d^2 \log p$. Then with probability at least $1 - c_3 \exp(-c_4 n \delta^2)$, the estimated edge set $\widehat{E}$ based on the graphical Lasso satisfies $\widehat{E} \subseteq E$. Furthermore, $\widehat{\Theta}_{GLASSO}$ satisfies the elementwise ℓ_∞-norm bound*

$$\|\widehat{\Theta}_{GLASSO} - \Theta^*\|_{\max} \leq c_5 \left(\lambda + \left(1 + \frac{8}{\alpha}\right)\sqrt{\frac{\log p}{n}}\right),$$

so if Θ^* *also satisfies the minimum signal strength condition*

$$\min_{(j,k)\in E} |\Theta^*_{jk}| > c_5 \left(\lambda + \left(1 + \frac{8}{\alpha}\right)\sqrt{\frac{\log p}{n}}\right),$$

we are guaranteed that $\widehat{E} = E$.

Note that the first part of Theorem 5, which stipulates that $\widehat{E} \subseteq E$ with high probability, guarantees that the edge set generated by the graphical Lasso algorithm does not include any false edges.

Theory for nodewise Lasso. The theory for the nodewise regression method may be derived from variable selection guarantees for the Lasso algorithm [27, 22]. Under an incoherence assumption on functions of subblocks of the data matrix X, as well as a minimum signal strength assumption on the true regression coefficients

$$\theta^*_j = \left(\Sigma^*_{\backslash\{j\},\backslash\{j\}}\right)^{-1} \Sigma^*_{\backslash\{j\},j},$$

we may guarantee that the solutions $\{\widehat{\theta}_j\}_{1\le j\le p}$ to the nodewise regression programs (14) satisfy

$$\text{supp}(\widehat{\theta}_j) = \text{supp}(\theta^*_j), \qquad \forall 1 \le j \le p,$$

with high probability. This in turn implies that $\widehat{N(j)} = N(j)$ for all j, so the nodewise regression method succeeds. We summarize in the following theorem. To declutter the theorem statement, we assume that the columns of X have been renormalized so that $\frac{1}{\sqrt{n}} \max_{1\le j\le p} \|X^j\|_2 \le 1$.

Theorem 6. *Suppose there exists a parameter* $\alpha \in (0, 1]$ *such that*

$$\max_{1\le j\le p} \left\{ \max_{k\in(N(j)\cup j)^c} \left\| \Sigma^*_{k,N(j)} \left(\Sigma^*_{N(j),N(j)}\right)^{-1} \right\|_1 \right\} \le 1 - \alpha. \tag{16}$$

Suppose the regularization parameters for the nodewise Lasso are chosen such that $\lambda_j = \frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}} + \delta$, *for some* $\delta \in \left[0, \frac{1}{\alpha^2 d}\right]$, *and suppose the sample size satisfies* $n \ge c_2 d \log p$. *Then with probability at least* $1 - c_3 \exp(-c_4\alpha^2 n\delta^2)$, *the estimated edge set* $\widehat{E}$ *based on nodewise regression satisfies* $\widehat{E} \subseteq E$. *Furthermore,*

$$\max_{1\le j\le p} \|\widehat{\theta}_j - \theta^*_j\|_\infty \le c_5 \left(\lambda + \sqrt{\frac{\log p}{n}}\right),$$

so if we also have the minimum signal strength condition

$$\min_{1\le j\le p} \min_{j,k\in N(j)} |(\theta^*_j)_k| > c_5 \left(\lambda + \sqrt{\frac{\log p}{n}}\right), \tag{17}$$

we are guaranteed that $\widehat{E} = E$.

Note that the condition (16) is the linear regression analog of the α-incoherence condition (15). Furthermore, the minimum signal strength condition (17) may be translated into a minimum signal strength condition on Θ^* via the relation (13).

3.2 Ising models

We now shift our focus from Gaussian to discrete random variables. The probability mass function of an Ising model from statistical physics, parametrized by node potentials $\{\theta_j\}_{1\le j\le p}$ and edge potentials $\{\theta_{jk}\}_{(j,k)\in E}$, with $E \subseteq V \times V$, is given by

$$q(x_1, \dots, x_p) = \frac{1}{Z} \exp\left(\sum_{j=1}^{p} \theta_j x_j + \sum_{(j,k)\in E} \theta_{jk} x_j x_k\right), \qquad \forall x \in \{-1, 1\}^p,$$

where

$$Z = Z(\theta) = \sum_{x\in\{-1,1\}^p} \exp\left(\sum_{j=1}^{p} \theta_j x_j + \sum_{(j,k)\in E} \theta_{jk} x_j x_k\right)$$

is the normalizing constant or *partition function* [10, 1]. Using similar reasoning as in the Gaussian setting (exercise!), we have the relation

$$(j,k) \in E \iff \Theta_{jk} \neq 0, \qquad \forall j \neq k,$$

where $\Theta \in \mathbb{R}^{p\times p}$ is the symmetric matrix with diagonal entries equal to $\{\theta_j\}$ and off-diagonals equal to

$$\Theta_{jk} = \begin{cases} \theta_{jk} & \text{if } (j,k) \in E \\ \theta_{kj} & \text{if } (k,j) \in E \\ 0 & \text{if } (j,k), (k,j) \notin E. \end{cases}$$

Importantly, although the matrix Θ encodes the edges of the graphical model, it no longer corresponds to the inverse covariance matrix of the joint distribution.

3.2.1 Logistic regression

A straightforward calculation shows that the conditional distributions in the Ising model take the form

$$\log q(x_j \mid x_{\setminus\{j\}}) = -f\left(2\theta_j x_j + 2\sum_{k\in N(j)} \theta_{jk} x_j x_k\right), \tag{18}$$

where $f(t) = \frac{1}{1+\exp(t)}$ is the logistic function. This motivates a nodewise neighborhood selection method based on logistic regression. In particular, for high-dimensional graphical models, we optimize the ℓ_1-penalized logistic regression programs

$$\widehat{\theta}^j = \arg\min_{\theta\in\mathbb{R}^p} \left\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} f\left(2\theta_j x_{ij} + 2\sum_{k\in V\setminus\{j\}} \theta_{jk} x_{ij} x_{ik}\right)}_{\mathcal{L}_n(\theta)} + \lambda_j \sum_{k\in V\setminus\{j\}} |\theta_{jk}| \right\}. \tag{19}$$

Here, the minimization is over vectors θ with one coordinate denoted by θ_j and $p-1$ coordinates denoted by $\{\theta_{jk} : k \in V\setminus\{j\}\}$, and the ℓ_1-penalty encourages sparsity. The estimated neighborhood sets are given by

$$\widehat{N(j)} = \text{supp}(\widehat{\theta}^j), \qquad \forall 1 \le j \le p,$$

and we combine our neighborhood estimates into an estimate $\widehat{E}$ of the edge set via an AND or OR rule, just as in the case of the nodewise Lasso for Gaussian graphical models. Note that the programs (19) are all convex, and may be solved efficiently (e.g., see Koh et al. [12] for an interior-point implementation).

3.2.2 Statistical theory

We now describe a result providing statistical guarantees for the nodewise logistic regression algorithm. Let $\{(\theta^j)^\star\}_{1 \le j \le p}$ denote the true parameter vectors, where

$$(\theta^j)^\star = (\theta_j^*; \theta_{jk}^* : k \in V \backslash \{j\}) \in \mathbb{R}^p.$$

The result below assumes an incoherence condition on the Fisher information matrix $J = \nabla^2 \mathcal{L}_n(\theta^*)$:

$$\max_{1 \le j \le p} \left\{ \max_{k \notin N(j)} \left\| J_{k,N(j)} \left(J_{N(j),N(j)}\right)^{-1} \right\|_1 \right\} \le 1 - \alpha, \quad \text{for some } \alpha \in (0,1]. \tag{20}$$

This is the logistic regression analog to the earlier incoherence condition (16) for linear regression. We then have the following result:

Theorem 7 (Ravikumar et al. [18])**.** *Suppose the true parameters of the Ising model satisfy the incoherence condition (20). If the regularization parameters of the nodewise logistic regression program are chosen such that* $\lambda_j = \frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}} + \delta$*, for some* $\delta \in [0,1]$*, and the sample size satisfies* $n \ge c_2 d^3 \log p$*, then with probability at least* $1 - c_3 \exp(-c_4 n \delta^2)$*, the estimated edge set* $\widehat{E}$ *based on nodewise logistic regression satisfies* $\widehat{E} \subseteq E$*. Furthermore,*

$$\max_{1 \le j \le p} \|\widehat{\theta}^j - (\theta^j)^\star\|_\infty \le c_5 \lambda \sqrt{d},$$

so if we also have the minimum signal strength condition

$$\min_{1 \le j \le p} \left\{ \min_{k \in N(j)} |\left(\theta^j\right)_k^\star| \right\} > c_5 \lambda \sqrt{d},$$

we are guaranteed that $\widehat{E} = E$*.*

4 Spectral methods

4.1 Linear algebra "review"

References:

- Horn & Johnson, *Matrix Analysis*
- Fuzhen Zhang, *Matrix Theory*

4.1.1 Eigenvalues and singular values

Definition: For a matrix $A \in \mathbb{C}^{n\times n}$, a vector $0 \neq v \in \mathbb{C}^n$ is an *eigenvector* with associated *eigenvalue* $\lambda \in \mathbb{C}$ if

$$Av = \lambda v.$$

Properties:

- Eigenvectors corresponding to different eigenvalues are linearly independent.
- If $A = A^*$ (i.e., A is Hermitian), all eigenvalues are real; in particular, this happens when A is real and symmetric.
- If $A \succeq 0$, meaning $A = A^*$ and $v^*Av \geq 0$ for all $v \in \mathbb{C}^n$, all eigenvalues of A are nonnegative.

Definition: For a matrix $A \in \mathbb{C}^{m\times n}$, nonnegative square roots of eigenvalues of A^*A are *singular values* of A.

Note: Singular values exist even when A is not a square matrix. Since $A^*A \succeq 0$, it makes sense to talk about square roots of eigenvalues by the last property above.

4.1.2 Matrix decompositions

Definition: A matrix $U \in \mathbb{C}^{n\times n}$ is *unitary* if

$$U^*U = UU^* = I.$$

(If $U \in \mathbb{R}^{n\times n}$, this means U is an *orthogonal* matrix—all rows/columns are orthonormal, and linear transformation corresponding to U consists of rotation/scaling.)

Spectral decomposition: If $A = A^*$, there exists a unitary matrix U such that

$$U^*AU = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

where λ_i's are real-valued eigenvalues of A, and columns $\{u_i\}_{i=1}^n$ of U are orthonormal eigenvectors corresponding to eigenvalues (i.e., A is *diagonalizable* via a unitary transformation). The spectral decomposition is generally not unique, although the (ordered) eigenvalues are unique.

Special case: If $A \in \mathbb{R}^{n\times n}$ and $A = A^T$, can choose $U \in \mathbb{R}^{n\times n}$ to be an orthogonal matrix.

Singular value decomposition: If $A \in \mathbb{C}^{m\times n}$ has nonzero singular values $\{\sigma_1, \ldots, \sigma_r\}$ (with $r \leq \min\{m, n\}$), there exist unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ such that

$$A = U \begin{pmatrix} D & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{pmatrix} V^*,$$

where $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Columns $\{u_i\}_{i=1}^m$ and $\{v_j\}_{j=1}^n$ of U and V are known as *left* and *right* singular vectors of A.

Note that this is the "full SVD" as opposed to the "economy SVD," which has a square diagonal matrix in the center and two rectangular matrices with orthonormal columns.

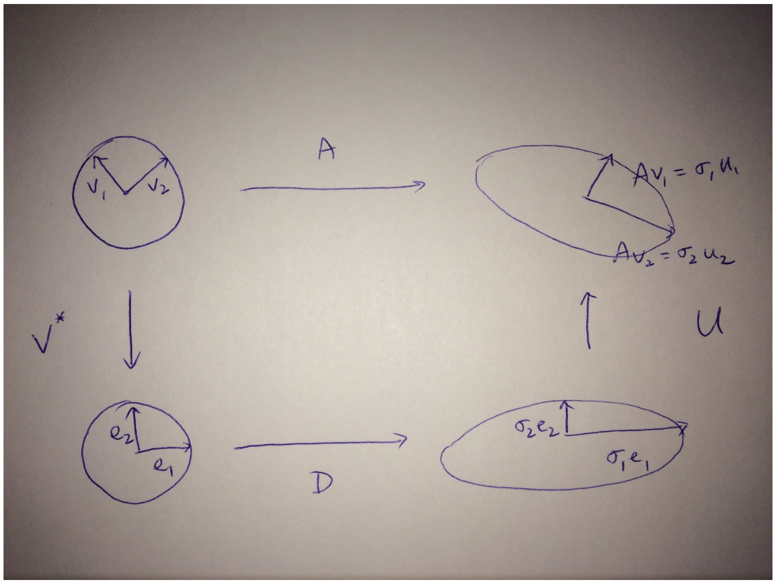
Properties:

- In general, SVD is *not* unique (although we often choose $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, making the matrix D unique).

- We have the rank-r decomposition
$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^*.$$

- If $A \in \mathbb{R}^{m\times n}$, can take $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ to be orthonormal matrices.

- **Geometric interpretation:** Note that the SVD implies that $AV = UD$, so we have $Av_i = \sigma_i u_i$ for all i, and similarly, $u_i^* A = \sigma_i v_i^*$. The sketch shows a geometric interpretation of the SVD via a commutative diagram (the sketch is for two dimensions, with $r = m = n = 2$):



The linear transformation A may be decomposed as an isometry V^*, followed by a scale transformation D, followed by another isometry U.

Lemma 8. *A vector v is an eigenvector of A^*A with eigenvalue λ_1 iff v is a right singular vector of A with singular value $\sigma_1 = \sqrt{\lambda_1}$.*

4.2 Best-fit subspaces

Consider a matrix $A \in \mathbb{R}^{m\times n}$. (From now on, we will work in $\mathbb{R}^n$ unless otherwise stated.) Denote the rows by $\{a_i^T\}_{i=1}^m$, where each $a_i \in \mathbb{R}^n$ may be viewed as a point in n-dimensional space.

4.2.1 One-dimensional case

Question: Find the (unit) vector $v \in \mathbb{R}^n$ that minimizes the sum of squared distances of the points $\{a_i\}$ to the line $\{cv : c \in \mathbb{R}\}$. (Equivalently, find the "best-fit line" through the origin.)

Since $a_i^T v$ corresponds to length of projection of vector a_i on v, distance $d(a_i, v)$ from a_i to line is $\|a_i - (a_i^T v) \cdot v\|_2$. We want to minimize

$$\sum_{i=1}^{m} \|a_i - (a_i^T v) \cdot v\|_2^2.$$

By the Pythagorean theorem, $\|a_i\|_2^2 = \|a_i - (a_i^T v) \cdot v\|_2^2 + \|(a_i^T v) \cdot v\|_2^2$, so we equivalently want to maximize

$$\sum_{i=1}^{m} \|(a_i^T v) \cdot v\|_2^2 = \sum_{i=1}^{m} (a_i^T v)^2 = \|Av\|_2^2$$

over v (note that $\sum_{i=1}^m a_i a_i^T = A^T A$ for the last equality).

Lemma 9. *A vector $v_1 \in \mathbb{R}^n$ is a right singular vector of A corresponding to the first singular value σ_1, if and only if*

$$v_1 \in \arg \max_{\|v\|_2=1} \|Av\|_2^2.$$

4.2.2 Generalization

Theorem 10. *Suppose $\{v_1, \ldots, v_n\}$ are right singular vectors corresponding to the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. For each $1 \leq k \leq n$, let $V_k := \operatorname{span}\{v_1, \ldots, v_k\}$. Then V_k is a best-fit k-dimensional subspace for the points $\{a_i\}_{i=1}^m$:*

$$V_k \in \arg \min_{\dim(V)=k} \sum_{i=1}^{m} d(a_i, V)^2,$$

where we define the distance function

$$d(a, V) = \min_{v \in V} \|a - v\|_2$$

for a vector a and subspace V.

The proof will rely on the following result:

Lemma 11. *Let $A \in \mathbb{R}^{m \times n}$, and define the vectors*

$$\begin{aligned} v_1 &\in \arg \max_{\|v\|_2=1} \|Av\|_2^2 \\ v_2 &\in \arg \max_{\substack{\|v\|_2=1 \\ v^T v_1 = 0}} \|Av\|_2^2 \\ &\vdots \\ v_n &\in \arg \max_{\substack{\|v\|_2=1 \\ v^T v_1 = v^T v_2 = \cdots = v^T v_{n-1} = 0}} \|Av\|_2^2. \end{aligned}$$

Then $v_1, \ldots, v_n$ are right singular vectors of A with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n \geq 0$. Conversely, any vectors $\{v_1, \ldots, v_n\}$ corresponding to the singular values $\{\sigma_1, \ldots, \sigma_n\}$ satisfy the above system.

Proof of Theorem 10. We use induction to prove the theorem, with the base case proved in the previous subsection.

Inductive step: Suppose V_{k-1} is a best-fit $(k-1)$-dimensional subspace. Note that

$$\begin{aligned}\sum_{i=1}^{m} d(a_i, V_{k-1})^2 &= \sum_{i=1}^{m} \left(\|a_i\|_2^2 - \sum_{j=1}^{k-1} |a_i^T v_j|^2 \right) \\ &= \sum_{i=1}^{m} \|a_i\|_2^2 - \sum_{i=1}^{m} \sum_{j=1}^{k-1} |a_i^T v_j|^2 \\ &= \|A\|_F^2 - \left(\|Av_1\|_2^2 + \cdots + \|Av_{k-1}\|_2^2 \right),\end{aligned}$$

so V_{k-1} maximizes $\|Av_1\|_2^2 + \cdots + \|Av_{k-1}\|_2^2$ among all collections of $k-1$ orthonormal vectors. (The first equality above follows from the Pythagorean theorem.)

Let W be any k-dimensional subspace. Then $W \cap V_{k-1}^{\perp} \neq 0$ by a dimensionality argument (since $\dim(W) = k$ and $\dim(V_{k=1}^{\perp}) = n - (k-1)$). Choose an orthonormal basis $\{w_1, \ldots, w_k\}$ of W such that $w_k \in V_{k-1}^{\perp}$. We have

$$\|Aw_1\|_2^2 + \cdots + \|Aw_{k-1}\|_2^2 \leq \|Av_1\|_2^2 + \cdots + \|Av_{k-1}\|_2^2$$

(since V_{k-1} is a maximizer). Furthermore, $\|Aw_k\|_2^2 \leq \|Av_k\|_2^2$, by the maximality of v_k provided in Lemma 11 (since $v_k, w_k \in V_{k-1}^{\perp}$). It follows that

$$\|Aw_1\|_2^2 + \cdots + \|Aw_k\|_2^2 \leq \|Av_1\|_2^2 + \cdots + \|Av_k\|_2^2,$$

so

$$\|A\|_F^2 - \left(\|Av_1\|_2^2 + \cdots + \|Av_{k-1}\|_2^2 \right) \leq \|A\|_F^2 - \left(\|Aw_1\|_2^2 + \cdots + \|Aw_{k-1}\|_2^2 \right),$$

and

$$\sum_{i=1}^{m} d(a_i, V_k)^2 \leq \sum_{i=1}^{m} d(a_i, W)^2,$$

as wanted. (The inequality above holds for any W.) □

4.2.3 PCA

Setting: Consider data points $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^p$. Want to find low-dimensional projection capturing variation in data.

Solution: Find projection onto k-dimensional subspace maximizing sum of squared projection lengths:

$$\max_{\dim(V) \leq k} \sum_{i=1}^{n} \|P_V x_i\|_2^2$$

(equivalently, minimizing sum of squared distances to V). As we have seen, a solution is given by $V = \text{span}\{v_1, \ldots, v_k\}$, span of top k right singular vectors of X.

Statistical interpretation: Note that $\{v_1, \ldots, v_k\}$ are top k eigenvectors of $\frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$, sample covariance of x_i's assuming data are centered. Then v_i's are directions of maximal variance:

$$v_1 \in \arg\max_{\|v\|_2=1} v^T \left(\frac{X^TX}{n}\right) v.$$

(If $\frac{X^TX}{n} \to \Sigma_x$, the v_i's converge to top eigenvectors of Σ_x.)

Computation:

- Computing X^TX takes $O(np^2)$, and eigendecomposition takes $O(p^3)$.
- Computing top k singular values of X only takes $O(npk)$ (more on this in the next subsection).

Source: *Handbook of Linear Algebra*, 2007

Algorithms for efficient SVD computation for large matrices is an active area of research. For moderately-sized matrices, a standard technique is the *power (iteration) method.*

4.3 Computation of SVD/top eigenvectors

4.3.1 Basic idea

Let $B = A^TA$. We have

$$\begin{aligned} B &= \left(\sum_{i=1}^r \sigma_i v_i u_i^T\right)\left(\sum_{j=1}^r \sigma_j u_j v_j^T\right) \\ &= \sum_{i=1}^r \sigma_i^2 v_i u_i^T u_i v_i^T = \sum_{i=1}^r \sigma_i^2 v_i v_i^T. \end{aligned}$$

Furthermore,

$$\begin{aligned} B^2 &= \left(\sum_{i=1}^r \sigma_i^2 v_i v_i^T\right)\left(\sum_{i=1}^r \sigma_i^2 v_i v_i^T\right) \\ &= \sum_{i=1}^r \sigma_i^4 v_i v_i^T. \end{aligned}$$

By induction, we can show that

$$B^k = \sum_{i=1}^r \sigma_i^{2k} v_i v_i^T.$$

In particular, if $\sigma_1 > \sigma_2$, then $\left\|\left|B^k - \sigma_1^{2k} v_1 v_1^T\right|\right\| \to 0$. If we consider $B^k e_1$, the first column of B^k, then normalizing gives an estimate of v_1.

Additional eigenvectors can be calculated (approximately) by applying the power method to $B - \sigma_1 v_1 v_1^T$. To obtain left singular values of A, consider AA^T instead. (Alternatively, we could apply $A\hat{v}_1$ and then rescale to obtain $\hat{u}_1$.)

4.3.2 Faster method

Instead of computing B^k, can save on computation by computing $B^k x$ directly for a (random) vector x. If A is a $m \times n$ matrix, then $B \in \mathbb{R}^{n\times n}$, and computing B^2 alone takes $O(n^3)$ time. However, if we instead compute $B^k x = A^T A \cdots A^T A x$, then we need to perform $2k$ operations that are each $O(mn)$. Note that if we write $x = \sum_{i=1}^n c_i v_i$, we have

$$B^k x \approx \left(\sigma_1^{2k} v_1 v_1^T\right)\left(\sum_{i=1}^n c_i v_i\right) = c_1 \sigma_1^{2k} v_1,$$

so normalizing the resulting vector gives an approximation of v_1. (We use a random vector so as to avoid the case when $c_1 \approx 0$.)

More comprehensive resource: "Computation of the singular value decomposition," *Handbook of Linear Algebra*, 2007

4.4 Best rank-k approximations

If A has nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, define

$$A_k = \sum_{i=1}^k \sigma_i u_i v_i^T,$$

for $1 \leq k \leq r$. (Truncated SVD.) Note that A_k has rank *exactly* k.

4.4.1 Frobenius norm

Theorem 12 (Schmidt Approximation Theorem, Eckart-Young Theorem)**.** *For all* $1 \leq k \leq r$, *we have*

$$\min_{B\in\mathbb{R}^{m\times n}:\operatorname{rank}(B)\leq k} \|A - B\|_F = \|A - A_k\|_F.$$

Proof. Fix k, and let $B \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(B) \leq k$ be a minimizer. Let B' be the matrix obtained from B by projecting each row of A onto the rowspace $\operatorname{row}(B)$. Then

$$\|A - B'\|_F^2 \leq \|A - B\|_F^2$$

(since the Frobenius norm squared is the sum of squared Euclidean norms of rows, and b_j' is defined to be the closest vector to a_j in a subspace containing b_j). Furthermore, since $\operatorname{row}(B') \subseteq \operatorname{row}(B)$, we have $\operatorname{rank}(B') \leq k$. Also by the optimality of B, we have

$$\|A - B\|_F \leq \|A - B'\|_F,$$

implying that $\|A - B\|_F^2 = \|A - B'\|_F^2$, which is the sum of squared distances of the points $\{a_i\}_{i=1}^m$ to $\operatorname{row}(B)$:

$$\|A - B\|_F^2 = \sum_{i=1}^m d(a_i, \operatorname{row}(B))^2.$$

Claim 1. *The squared Frobenius norm $\|A - A_k\|_F^2$ is equal to the sum of squared distances of $\{a_i\}_{i=1}^m$ to $V_k = \operatorname{span}\{v_1, \dots, v_k\}$:*

$$\|A - A_k\|_F^2 = \sum_{i=1}^m d(a_i, V_k)^2.$$

Proof. We show that the projection of a_i onto V_k is equal to the i^{th} row of A_k. The projection may be written as

$$(a_i^T v_1)v_1 + \cdots + (a_i^T v_k)v_k,$$

and the matrix with this vector in the i^{th} row is

$$Av_1v_1^T + \cdots + Av_kv_k^T.$$

Note that

$$\sum_{i=1}^k Av_iv_i^T = \sum_{i=1}^k \sigma_i u_i v_i^T = A_k,$$

so the claim follows. □

By Theorem 10, we have

$$\sum_{i=1}^m d(a_i, V_k)^2 \le \sum_{i=1}^m d(a_i, \operatorname{row}(B))^2.$$

We have shown that the RHS is equal to $\|A - B\|_F^2$, and Claim 1 shows that the LHS is equal to $\|A - A_k\|_F^2$. The desired result follows (since B is a minimizer). □

4.4.2 Spectral norm

Now recall the definition of the spectral norm:

$$\|A\|_2 = \sup_{\|v\|_2=1} \|Av\|_2 = \sigma_1.$$

Theorem 13 (Eckart-Young-Mirsky Theorem)**.** *For all $1 \le k \le r$, we have*

$$\min_{B\in\mathbb{R}^{m\times n}:\operatorname{rank}(B)\le k} \|A - B\|_2 = \|A - A_k\|_2.$$

Note: In fact, Mirsky (1958) proved a generalization of Theorem 12 to all *unitarily invariant* matrix norms. Recall that a unitarily invariant norm satisfies $\|A\| = \|UAV\|$ for all unitary matrices U and V; the Frobenius and spectral norms satisfy this property:

$$\begin{aligned}\|UAV\|_F^2 &= \operatorname{tr}(V^TA^TU^TUAV) = \operatorname{tr}(V^TA^TAV)\\ &= \operatorname{tr}(A^TAVV^T) = \operatorname{tr}(A^TA) = \|A\|_F^2,\end{aligned}$$

and if we have the SVD $A = \bar{U}\bar{D}\bar{V}^T$, then $UAV = U\bar{U}\bar{D}\bar{V}^TV$ clearly has the same set of singular values, so $\|UAV\|_2 = \sigma_1$. We have used the facts that

- $\|A\|_F^2 = \operatorname{tr}(A^TA)$,
- $\operatorname{tr}(ABC) = \operatorname{tr}(BCA)$.

4.5 Johnson-Lindenstrauss Lemma

Idea: Existence of (random) low-dimensional linear projections approximately preserving ℓ_2-distance between N data points. Requires projecting into sufficiently high-dimensional space. First proved in Johnson & Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, 1984.

4.5.1 Main result

Theorem 14 (JL Lemma)**.** *Suppose $0 < \epsilon < \frac{1}{2}$. Let $\{x_1, \ldots, x_N\} \subseteq \mathbb{R}^n$ be a set of data points, and let $m = \frac{C \log N}{\epsilon^2}$. Then there exists a Lipschitz mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ such that for all pairs (i, j),*

$$(1-\epsilon)\|x_i - x_j\|_2^2 \leq \|f(x_i) - f(x_j)\|_2^2 \leq (1+\epsilon)\|x_i - x_j\|_2^2. \tag{21}$$

Definition: A distribution D on $\mathbb{R}^{m\times n}$ satisfies the (ϵ, δ)-*distributional JL property* if for any fixed $x \in \mathbb{R}^n$, a matrix Φ drawn from D satisfies

$$P\left((1-\epsilon)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\epsilon)\|x\|_2^2\right) > 1 - \delta. \tag{22}$$

Lemma 15 (Distributional JL Lemma)**.** *Suppose $\epsilon, \delta < \frac{1}{2}$ and $m = \frac{C' \log(1/\delta)}{\epsilon^2}$. Then there exists a probability distribution D on $\mathbb{R}^{m\times n}$ satisfying the (ϵ, δ)-distributional JL property.*

Proof of Theorem 14. Lemma 15 applied with $\delta < \frac{1}{\binom{N}{2}}$ implies the existence of a distribution D such that

$$P\Bigg((1-\epsilon)\|x_i - x_j\|_2^2 \leq \|\Phi x_i - \Phi x_j\|_2^2 \leq (1+\epsilon)\|x_i - x_j\|_2^2\Bigg) > 1 - \delta, \quad \forall i, j.$$

Taking a union bound over all pairs (i, j), we then have

$$\begin{aligned} P\Bigg((1-\epsilon)\|x_i - x_j\|_2^2 \leq \|\Phi x_i - \Phi x_j\|_2^2 \leq (1+\epsilon)\|x_i - x_j\|_2^2, \quad \forall i, j\Bigg) &> 1 - \binom{N}{2}\delta \\ &> 0, \end{aligned}$$

implying the existence of at least one matrix Φ in the support of D satisfying the desired properties. (This is an illustration of the *probabilistic method*.) We need to choose dimension

$$m = \frac{C' \log(1/\delta)}{\epsilon^2} = \frac{C \log N}{\epsilon^2}$$

in order for successful application of Lemma 15. □

4.5.2 Proof of distributional lemma

Recall the definition of the spectral norm:

$$\|A\|_2 = \sup_{\|v\|_2 = 1} \|Av\|_2 = \sigma_1.$$

Proof of Lemma 15. Different constructions exist, including:

1. projection onto random m-dimensional subspace (original JL proof),
2. choosing entries of Φ to be i.i.d. $N\left(0, \frac{1}{m}\right)$ variables, and
3. assigning entries of Φ to be i.i.d. with values $\pm\frac{1}{\sqrt{m}}$.

We will analyze the last construction.

We have $\Phi_{ij} = \frac{\xi_{ij}}{\sqrt{m}}$, where ξ_{ij}'s are i.i.d. Rademacher variables. We use the following theorem:

Theorem 16 (Hanson-Wright inequality, 1971)**.** *Suppose $A \in \mathbb{R}^{n\times n}$ and $\xi \in \mathbb{R}^n$ is a vector of i.i.d. Rademacher variables. Then for any $\lambda > 0$,*

$$P\left(\left|\xi^T A\xi - \mathbb{E}(\xi^T A\xi)\right| > \lambda\right) \leq \exp\left(-\min\left\{\frac{c_1\lambda^2}{\|A\|_F^2}, \frac{c_2\lambda}{\|A\|_2}\right\}\right).$$

(Versions of Theorem 16 hold for i.i.d. sub-Gaussian components, as well. See Rudelson & Vershynin, "Hanson-Wright inequality and sub-Gaussian concentration," 2013, for more details.)

Note that it suffices to establish the condition (22) for unit vectors. For a fixed unit vector $x \in \mathbb{R}^n$, define the matrix

$$A_x = \frac{1}{\sqrt{m}}\begin{bmatrix} x^T & 0 & \cdots & 0 \\ 0 & x^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x^T \end{bmatrix} \in \mathbb{R}^{m\times mn},$$

and notice that $\|\Phi x\|_2^2 = \|A_x\xi\|_2^2$, where $\xi \in \mathbb{R}^{mn}$ is a vector of Rademacher variables. Also note that

$$\begin{aligned}\mathbb{E}\left(\|A_x\xi\|_2^2\right) &= \mathbb{E}\left[\operatorname{tr}\left(\xi^T A_x^T A_x\xi\right)\right] = \operatorname{tr}\left(A_x^T A_x\mathbb{E}[\xi\xi^T]\right) = \operatorname{tr}(A_x^T A_x) = \|A_x\|_F^2 \\ &= \|x\|_2^2,\end{aligned}$$

where we have used the fact that $\mathbb{E}[\xi\xi^T] = I$, so applying Theorem 16 with $\lambda = \epsilon$ and $A = A_x^T A_x$ gives

$$P\left(\left|\|\Phi x\|_2^2 - \|x\|_2^2\right| > \epsilon\right) \leq \exp\left(-\min\left\{\frac{c_1\epsilon^2}{\|A\|_F^2}, \frac{c_2\epsilon}{\|A\|_2}\right\}\right).$$

Furthermore, A is a block-diagonal matrix with m blocks equal to $\frac{1}{m}xx^T$, so

$$\begin{aligned}\|A\|_F^2 &= m\cdot\frac{1}{m^2}\|xx^T\|_F^2\frac{1}{m}\operatorname{tr}(xx^T\cdot xx^T) = \frac{1}{m}\operatorname{tr}(xx^T) = \frac{1}{m}, \\ \|A\|_2 &= \frac{1}{m}\cdot\sup_{\|u\|_2=1}(u^T x)(x^T u) = \frac{1}{m}.\end{aligned}$$

Hence, probability of error becomes $\exp(-cm\epsilon^2)$, and choice $m = \frac{C\log(1/\delta)}{\epsilon^2}$ drives this less than δ. □

4.5.3 Johnson-Lindenstrauss applications

Recall: Distributional JL lemma shows that *with high probability*, random projection chosen from ensemble satisfies desired distance-preserving property (provided m is chosen large enough relative to N and δ).

k-means clustering: For data points $\{x_i\}_{i=1}^N \subset \mathbb{R}^n$, find a partition $\mathcal{P}$ into k clusters $\{P_1, \dots, P_k\}$ centered at $\{y_1, \dots, y_k\}$, to minimize objective

$$h(\mathcal{P}; X) = \sum_{j=1}^{k} \sum_{i \in P_j} \|x_i - y_j\|_2^2.$$

In general, problem is NP-hard, but efficient algorithms exist providing γ-approximate clusterings, meaning $h(\mathcal{P}_\gamma; X) \leq \gamma h(\mathcal{P}^*; X)$ for $\gamma > 1$. To save computation, idea is to project into m-dimensional space before performing approximate k-means clustering. In order to prove rigorously that this works, derive the following result:

Lemma 17. *If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a JL embedding for $\{x_i\}_{i=1}^N$ with error tolerance ϵ, then any γ-approximate clustering P_γ for $f(X)$ satisfies*

$$h(\mathcal{P}_\gamma; X) \leq \gamma \left(\frac{1+\epsilon}{1-\epsilon} \right) \cdot h(\mathcal{P}^*; X)$$

(clearly, smaller ϵ leads to more accuracy).

Proof. Note that

$$\begin{aligned} (1-\epsilon) h(\mathcal{P}_\gamma; X) &\overset{(a)}{\leq} h(\mathcal{P}_\gamma; f(X)) \\ &\overset{(b)}{\leq} \gamma \cdot h(\mathcal{P}_f^*; f(X)) \\ &\overset{(c)}{\leq} \gamma \cdot h(\mathcal{P}^*; f(X)) \\ &\overset{(d)}{\leq} \gamma (1+\epsilon) \cdot h(\mathcal{P}^*; X), \end{aligned}$$

where (a) follows from the JL property, (b) follows because P_γ is an approximate clustering, (c) follows from the fact that P_f^* is optimal, and (d) follows from the JL property again. Note that we have used the notation

$$h(\mathcal{P}_\gamma; f(X)) = \sum_{j=1}^{k} \sum_{i \in (P_\gamma)_j} \|f(x_i) - f(y_j)\|_2^2,$$

and $\mathcal{P}_f^*$ denotes the optimal clustering for $\{f(x_i)\}_{i=1}^N$, whereas $\mathcal{P}^*$ denotes the optimal clustering for $\{x_i\}_{i=1}^N$. □

On approximate k-means: See Kumar, Sabharwal & Sen, "A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions," 2004. Gives $(1+\epsilon)$-approximate solution, w.h.p., in $O(Nn)$ time (treating k and ϵ as constants). This is in contrast to exact k-means, for which fastest exact algorithm takes $O(N^{kn+1})$ time. Thus, JL reduces the runtime of an approximate algorithm from $O(Nn)$ to $O(N \log N)$ time.

A deterministic $(1+\epsilon)$-approximate algorithm was proposed by Matousek, "On approximate geometric k-clustering," 2000, and runs in $O(N\epsilon^{-2k^2 n} \log^k N)$ time.

The main idea in these papers is to take a random sample of $O(k)$ points, which w.h.p. contains a constant number of points from the largest cluster. Then by trying all subsets of of constant size from the sample, we can obtain an estimate for the centroid of the largest cluster. We would then prune points from the largest cluster to obtain samples from the smaller clusters.

Other examples:

- Approximate nearest neighbor classifier
- Linear classifiers (e.g., SVMs, which are max margin classifiers—first project, then classify)

4.5.4 Equivalences

In fact, we can show that RIP $\iff$ JL property. Recall the RIP definition:

Definition: A matrix $\Phi \in \mathbb{R}^{m \times n}$ satisfies the (ϵ, k)-RIP if for all k-sparse vectors x,

$$(1-\epsilon)\|x\|_2^2 \le \|\Phi x\|_2^2 \le (1+\epsilon)\|x\|_2^2. \tag{23}$$

Equivalently, whenever $|S| \le k$, all eigenvalues of $\Phi_S^T \Phi_S$ are in $[1-\epsilon, 1+\epsilon]$, or

$$\left\|\Phi_S^T \Phi_S - I_k\right\|_2 \le \epsilon.$$

Note: Obvious similarities exist between RIP condition (23) and JL condition (22). However, RIP is a *deterministic* guarantee holding *uniformly* over k-sparse vectors, whereas JL condition is a *probabilistic* guarantee holding *any fixed* n-dimensional vector.

RIP has been studied extensively in compressed sensing and can be used to derive JL embeddings with desirable properties.

References:

- (JL $\implies$ RIP) Baraniuk et al., "A simple proof of the restricted isometry property for random matriices," *Constructive Approximation*, 2007
- (RIP $\implies$ JL) Krahmer & Ward, "New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property," *SIAM Journal on Mathematical Analysis*, 2011

Theorem 18 (JL $\implies$ RIP)**.** *Suppose $\epsilon < 1$, and $m \ge c_1(\epsilon) k \log\left(\frac{n}{k}\right)$. If D satisfies the $\left(\frac{\epsilon}{2}, \delta\right)$-distributional JL property with $\delta = e^{-m\epsilon}$, then with probability at least $1 - e^{-\epsilon m/2}$, a randomly drawn matrix $\Phi \sim D$ satisfies (ϵ, k)-RIP.*

Theorem 19 (RIP $\implies$ JL)**.** *Suppose $\Phi \in \mathbb{R}^{m \times n}$ satisfies $(\epsilon, 2k)$-RIP. Let $D_\xi = \mathrm{diag}(\xi_1, \ldots, \xi_n)$ be a diagonal matrix of i.i.d. Rademacher random variables. Then ΦD_ξ satisfies the $(3\epsilon, 3\exp(-ck))$-distributional JL property.*

4.5.5 Generating RIP matrices

Already studied in compressed sensing literature. Leads to useful families of matrices for distance-preserving projections via Johnson-Lindenstrauss. In particular:

- Subsampled Fourier matrix
- Subsampled Hadamard matrix: $\frac{(-1)^{\langle i,j\rangle}}{\sqrt{n}}$, where $\langle i,j\rangle$ is dot product of binary representations of i and j, or
$$H_k = \frac{1}{\sqrt{2}} \left(\begin{array}{c|c} H_{k/2} & H_{k/2} \\ \hline H_{k/2} & -H_{k/2} \end{array} \right).$$

Take $\Phi = SH$, where H is Fourier/Hadamard and each row of S has all 0's except a single $\sqrt{\frac{n}{m}}$ in a random column (scales and subsamples rows). Then $\Phi D_\xi x$ may be computed quickly (in $O(n \log n)$ time), since we only need to apply random signs to x, apply proper transform, and then rescale. (Fast JL Transform from Ailon & Chazelle,"Approximate nearest neighbors and the fast JL transform," 2006.)

4.6 Spectral clustering I: Data matrix

Reference: Blum, Hopcroft & Kannan, Chapter 7

Setting: Matrix $X \in \mathbb{R}^{n \times d}$ of data vectors. Goal is to partition dataset into K nonoverlapping groups.

4.6.1 Algorithm

1. Compute best rank-K approximation $X_K = \sum_{i=1}^K \sigma_i u_i v_i^T$ of X using SVD.
2. Perform distance-based clustering method (e.g., k-means algorithm) on rows $\{y_j\}_{j=1}^n$ of X_K to obtain clustering $\{C_i^Y\}_{i=1}^K$.
3. Output clusters $\{\tilde{C}_i\}_{i=1}^K$, with $\tilde{C}_i := \{j : y_j \in C_i^Y\}$.

Recall from the theorem from last time that the i^{th} row of X_K is the projection of x_i onto $\text{span}\{v_1, \ldots, v_K\}$, so we are clustering after projecting onto a best-fit subspace.

4.6.2 Some theory

For a clustering $\mathcal{C} = \{C_k\}_{k=1}^K$ of the data matrix X, define c_i to be centroid of cluster containing data point x_i (mean of cluster data points), and define cluster variance

$$\sigma^2(C) = \max_{v \in \mathbb{R}^p : \|v\|_2 = 1} \frac{1}{n} \sum_{i=1}^n \left((x_i - c_i)^T v \right)^2 = \frac{1}{n} \|X - C\|_2^2,$$

where C is matrix with c_i's as rows.

Assume distance-based clustering step is performed in the following manner:

1. Pick a random row y_i of X_K and form a cluster with all rows y_j such that $\|y_i - y_j\|_2 \leq \tau$.

2. Repeat step 1 on remaining rows of X_K until K clusters are formed, and assign any additional rows arbitrarily. (Some clusters are allowed to be empty if τ is too large.)

Theorem 20. *Suppose a clustering $\mathcal{C}$ of the data exists, such that $K \le d$ and*

(i) $|C_k| \ge \frac{6K\sigma(C)n}{\tau}$, *for all* $1 \le k \le K$ *(minimum cluster size),*

(ii)

$$\min_{c_i \neq c_j} \|c_i - c_j\|_2 \ge \frac{5\tau}{2} \qquad \textit{(minimum cluster separation)},$$

and

(iii) $\frac{32\sigma(C)}{3\tau} \le 1$ *(relative size of cluster variance vs. separation).*

Then with probability at least $1 - \frac{32K\sigma(C)}{3\tau}$, *the spectral clustering algorithm outputs a clustering* $\tilde{\mathcal{C}}$ *differing from* $\mathcal{C}$ *in at most* $\frac{32K\sigma^2(C)n}{\tau^2}$ *points.*

Note: Imagine a fixed τ. How far apart do cluster centers need to be in relation to internal spread in order to obtain a low-error clustering? Separation needs to be at least $\frac{5\tau}{2}$, and smaller $\sigma(C)$ implies smaller error. Result holds as long as cluster sizes are not too small, due to random choice of rows in spectral clustering.

Proof. First show that $\|y_i - c_i\|_2 \le \frac{\tau}{2}$ for most i (distance between projected points and centroids). Let $M \subseteq \{1, \ldots, n\}$ be subset of indices such that inequality is not satisfied. Then

$$\|X_K - C\|_F^2 = \sum_{i=1}^{n} \|y_i - c_i\|_2^2 \ge \sum_{i \in M} \|y_i - c_i\|_2^2 > \frac{|M|\tau^2}{4}.$$

Note that C has at most K distinct rows, so $\operatorname{rank}(C) \le K$. Hence,

$$\vert\!\vert\!\vert X_K - C \vert\!\vert\!\vert_2 \le \vert\!\vert\!\vert X_K - X \vert\!\vert\!\vert_2 + \vert\!\vert\!\vert X - C \vert\!\vert\!\vert_2 \le 2 \vert\!\vert\!\vert X - C \vert\!\vert\!\vert_2 \,,$$

by Eckart-Young-Mirsky. Since $\operatorname{rank}(X_K - C) \le \operatorname{rank}(X_K) + \operatorname{rank}(C) \le 2K$ (the left-hand rank is the dimension of $\operatorname{span}\{x_1' - c_1, \ldots, x_n' - c_n\} \subseteq \operatorname{span}\{x_1', \ldots, x_n', c_1, \ldots, c_n\}$), we have

$$\|X_K - C\|_F^2 \le 2K \vert\!\vert\!\vert X_K - C \vert\!\vert\!\vert_2^2$$

(using the fact that $\|A\|_F \le \sqrt{\operatorname{rank}(A)} \vert\!\vert\!\vert A \vert\!\vert\!\vert_2$—this can be proved easily by taking the SVD and recalling that $\|\cdot\|_F$ and $\vert\!\vert\!\vert \cdot \vert\!\vert\!\vert_2$ are unitarily invariant). Hence,

$$\frac{|M|\tau^2}{4} < \|X_K - C\|_F^2 \le 8K \vert\!\vert\!\vert X - C \vert\!\vert\!\vert_2^2 = 8K\sigma^2(C)n,$$

and we conclude that $|M| < \frac{32K\sigma^2(C)n}{\tau^2}$.

Suppose $i, j \notin M$. If x_i and x_j are in the same cluster in $\mathcal{C}$, then

$$\|y_i - y_j\|_2 \le \|y_i - c_i\|_2 + \|c_i - y_j\|_2 \le \tau. \tag{24}$$

On the other hand, if x_i and x_j are in different clusters, we have

$$\|y_i - y_j\|_2 \geq \|c_i - c_j\|_2 - \|c_i - y_i\|_2 - \|c_j - y_j\|_2 \geq \frac{5\tau}{2} - \frac{\tau}{2} - \frac{\tau}{2} = \frac{3\tau}{2}. \tag{25}$$

We will show that with probability at least $1 - \frac{32K\sigma(C)}{3\tau}$, the K row indices chosen in the clustering step lie in M^c. Indeed, if this were the case, the clusters $\tilde{\mathcal{C}} = \{\tilde{C}_k\}_{k=1}^K$ formed from spectral clustering would (after a permutation) be such that

$$\tilde{C_k} = (C_k \backslash M) \cup M_k,$$

where $M_k \subseteq M$. This is because inequality (24) implies that everything in $C_k \backslash M$ must lie in $\tilde{C_k}$ and inequality (25) implies that elements of $C_j \backslash M$, for $j \neq k$, do not lie in $\tilde{C_k}$. In particular, $\tilde{\mathcal{C}}$ and $\mathcal{C}$ would differ by at most

$$|M| < \frac{32K\sigma^2(C)n}{\tau^2}$$

points, as desired. In order to compute the probability that all K selected row indices lie in M^c, we use a union bound. Let E be the event that all selected row indices lie in M^c, and let E_k be the event that the k^{th} selected row index lies in M^c. Then

$$P(E^c) = P(E_1^c) + P(E_1 \cap E_2^c) + \cdots + P(E_1 \cap \cdots \cap E_{K-1} \cap E_K^c).$$

Suppose we are at stage k and all previously selected vertices lay in M^c. Then at least $\frac{6\sigma(C)Kn}{\tau} - \frac{32K\sigma^2(C)n}{\tau^2}$ row indices would remain to be assigned (since at least one cluster would have been untouched, but some vertices with indices in M might have already been assigned), of which at most $\frac{32K\sigma^2(C)n}{\tau^2}$ lie in M. Hence,

$$P(E_1 \cap \cdots \cap E_{k-1} \cap E_k^c) \leq \frac{\frac{32K\sigma^2(C)n}{\tau^2}}{\frac{6K\sigma(C)n}{\tau} - \frac{32K\sigma^2(C)n}{\tau^2}} = \frac{1}{\frac{3\tau}{16\sigma(C)} - 1} \leq \frac{32\sigma(C)}{3\tau},$$

assuming $\frac{32\sigma(C)}{3\tau} \leq 1$. It follows that

$$P(E^c) \leq \frac{32K\sigma(C)}{3\tau},$$

as wanted. □

Note: Other common clustering methods (e.g., k-means, k-medians, k-centers) have also been proven to work when inter-cluster separation is sufficiently large relative to intra-cluster spread. See Awasthi, Blum & Sheffet, "Center-based clustering under perturbation stability," 2012, or Ben-David, "Clustering is easy when ... what?," 2015.

4.7 Spectral clustering II: Graph Laplacians

Reference: von Luxburg, "A tutorial on spectral clustering," 2007.

Setting: Weight matrix $W \in \mathbb{R}^{n \times n}$, where $w_{ij} \geq 0$ measures similarity between data points x_i and x_j in dataset of size n. Goal is to partition dataset into K nonoverlapping groups.

4.7.1 Definitions

Let $d_i = \sum_{j=1}^n w_{ij}$ and $D = \text{diag}(d_1, \ldots, d_n)$.

- Unnormalized Laplacian: $L = D - W$
- Symmetric normalized Laplacian: $L_{sym} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2}$
- Random walk normalized Laplacian: $L_{rw} = D^{-1}L = I - D^{-1}W$

(Note that $D^{-1}W$ is transition matrix of random walk. Can show that eigenvectors of L_{rw} correspond to eigenvectors of L_{sym} left-multiplied by $D^{-1/2}$, with same eigenvalues, so all eigenvectors/eigenvalues are real: $L_{sym}v = \lambda v$ implies $(I - D^{-1/2}WD^{-1/2})v = \lambda v$, so

$$\begin{aligned} L_{rw}(D^{-1/2}v) &= (I - D^{-1}W)(D^{-1/2}v) = D^{-1/2}(v - D^{-1/2}WD^{-1/2}v) \\ &= \lambda(D^{-1/2}v).) \end{aligned}$$

4.7.2 Clustering algorithm

1. Form matrix $L = f(W) \in \mathbb{R}^{n\times n}$.
2. Compute *bottom* K eigenvectors $\{u_i\}_{i=1}^K \subseteq \mathbb{R}^n$ of L, and form matrix $U \in \mathbb{R}^{n\times K}$.
3. Perform distance-based clustering method on rows $\{y_j\}_{j=1}^n$ of U to obtain clustering $\{C_i\}_{i=1}^K$. (In practice, for L_{rw}, first rescale rows of U to unit norm before clustering—see perturbation theory justification.)
4. Output clusters $\{\tilde{C}_i\}_{i=1}^K$, with $\tilde{C}_i := \{j : y_j \in C_i\}$.

Note that eigenvectors of L and W agree when D is a multiple of I (e.g., d-regular, unweighted graphs). Why do we want to cluster eigenvectors of transformed Laplacian matrices instead? We will motivate these methods from the perspective of graph cuts in this lecture, and consistency in Banach spaces in the next lecture.

4.7.3 Connection to graph cuts

Consider an undirected graph $G = (V, E)$ with $V = \{1, \ldots, n\}$ and edge weight matrix $W \in \mathbb{R}^{n\times n}$.

For a partitioning $\{A_i\}_{i=1}^K$ of $\{1, \ldots, n\}$, define

$$\text{cut}(A_1, \ldots, A_K) = \frac{1}{2}\sum_{i=1}^K W(A_i, \bar{A}_i),$$

where $W(A, B) = \sum_{i\in A, j\in B} w_{ij}$ for disjoint sets A and B (sum of weights of edges across clusters). Mincut problem is relatively easy to solve when $K = 2$ (e.g., Karger's algorithm, Stoer-Wagner algorithm). Although polynomial-time algorithms exist for any fixed K, they are not computationally practical for large K. Other objectives of interest, leading to a more balanced cut:

$$\text{RatioCut}(A_1, \ldots, A_K) = \sum_{i=1}^K \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|},$$

$$\mathrm{Ncut}(A_1,\ldots,A_K)=\sum_{i=1}^{K}\frac{\mathrm{cut}(A_i,\bar{A}_i)}{\mathrm{vol}(A_i)},$$

where $\mathrm{vol}(A)=\sum_{i\in A} d_i=\sum_{i\in A, j\in V} w_{ij}$. (Ncut stands for "normalized cut.") With additional weight terms, minimizing these objectives become NP-hard (Wagner & Wagner, "Between min cut and graph bisection," 1993). Spectral clustering corresponds to solving an appropriate relaxation:

- RatioCut $\implies$ clustering with unnormalized Laplacian L
- Ncut $\implies$ clustering with normalized Laplacian L_{rw}

RatioCut (case $K=2$): Optimize

$$\min_{A\subseteq V}\mathrm{RatioCut}(A,\bar{A}).$$

For a fixed subset A, define the vector $v\in\mathbb{R}^n$ with entries

$$v_i=\begin{cases}\sqrt{|\bar{A}|/|A|}, & \text{if } i\in A,\\ -\sqrt{|A|/|\bar{A}|}, & \text{if } i\in\bar{A}.\end{cases}$$

Note that

$$\begin{aligned}
v^T L v &= v^T D v - v^T W v = \sum_{i=1}^{n} d_i v_i^2 - \sum_{i,j=1}^{n} w_{ij} v_i v_j\\
&= \frac{1}{2}\left(\sum_{i=1}^{n} d_i v_i^2 - 2\sum_{i,j=1}^{n} w_{ij} v_i v_j + \sum_{j=1}^{n} d_j v_j^2\right)\\
&= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(v_i-v_j)^2\\
&= \frac{1}{2}\sum_{i\in A, j\in\bar{A}}^{n} w_{ij}\left(\sqrt{\frac{|\bar{A}|}{|A|}}+\sqrt{\frac{|A|}{|\bar{A}|}}\right)^2 + \frac{1}{2}\sum_{i\in\bar{A}, j\in A} w_{ij}\left(-\sqrt{\frac{|A|}{|\bar{A}|}}-\sqrt{\frac{|\bar{A}|}{|A|}}\right)^2\\
&= \left(\frac{|\bar{A}|}{|A|}+\frac{|A|}{|\bar{A}|}+2\right)\cdot\mathrm{cut}(A,\bar{A})\\
&= \left(\frac{|A|+|\bar{A}|}{|A|}+\frac{|A|+|\bar{A}|}{|\bar{A}|}\right)\cdot\mathrm{cut}(A,\bar{A})\\
&= n\cdot\mathrm{RatioCut}(A,\bar{A}).
\end{aligned}$$

Furthermore, v satisfies

$$\sum_{i=1}^{n} v_i=\sum_{i\in A}\sqrt{\frac{|\bar{A}|}{|A|}}-\sum_{i\in\bar{A}}\sqrt{\frac{|A|}{|\bar{A}|}}=|A|\sqrt{\frac{|\bar{A}|}{|A|}}-|\bar{A}|\sqrt{\frac{|A|}{|\bar{A}|}}=0,$$

and

$$\|v\|_2^2=\sum_{i\in A}\frac{|\bar{A}|}{|A|}+\sum_{i\in\bar{A}}\frac{|A|}{|\bar{A}|}=n.$$

This leads to the following relaxation:

$$\begin{aligned}&\min_{v\in\mathbb{R}^n} v^T Lv\\ &\text{s.t. } v^T 1 = 0, \quad \|v\|_2 = \sqrt{n}. \end{aligned} \tag{26}$$

Also note that the smallest eigenvalue of L is 0, with corresponding eigenvector 1 (by the computation above, we see that $L \succeq 0$). Hence, although the optimization problem (26) is highly nonconvex, solution corresponds to (rescaled) eigenvector with second smallest eigenvalue (the constraint $v^T 1 = 0$ imposes orthogonality to the first eigenvector). In order to obtain partition, could define

$$A = \{i : \hat{v}_i \geq 0\}.$$

Another method: Use k-means clustering on the rows of the matrix

$$U = \left(1 \;\middle|\; \hat{v}\right) \in \mathbb{R}^{n\times 2}.$$

This is exactly the spectral clustering algorithm.

RatioCut (General K): For a partition $\{A_i\}_{i=1}^K$ of V, define indicator vectors $h_j = (h_{1,j}, \ldots, h_{n,j})^T$ such that

$$h_{i,j} = \begin{cases} 1/\sqrt{|A_j|}, & \text{if } i \in A_j \\ 0, & \text{otherwise.} \end{cases}$$

Can check that $H^T H = I$, where $H \in \mathbb{R}^{n\times K}$ has columns $\{h_j\}_{j=1}^K$, and

$$(H^T LH)_{ii} = h_i^T Lh_i = \sum_{j,k} w_{jk}(h_{ij} - h_{ik})^2 = \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}.$$

Hence,

$$\text{RatioCut}(A_1, \ldots, A_K) = \sum_{i=1}^K (H^T LH)_{ii} = \text{tr}(H^T LH).$$

Obtain relaxation

$$\begin{aligned}&\min_{H\in\mathbb{R}^{n\times K}} \text{tr}(H^T LH)\\ &\text{s.t. } H^T H = I. \end{aligned} \tag{27}$$

Can be shown (using the Rayleigh-Ritz Theorem) that this is exactly variational characterization for finding bottom K eigenvectors of L, so solution is matrix U with eigenvectors in the columns. Then perform k-means on rows of U to obtain partitioning.

4.7.4 Basics of spectral graph theory

References:

- Spielman, lecture notes on spectral graph theory.
- Chung, "Spectral graph theory," 1997.

Recall: Laplacian matrix of a weighted graph $G = (V, E)$ is given by $L = D - W$.

For $x \in \mathbb{R}^n$, we have

$$x^T Lx = \sum_{(u,v)\in E} w_{u,v}(x_u - x_v)^2 \geq 0,$$

where $w_{u,v} \geq 0$ (calculation done above). Eigenvalues are $0 = \lambda_1 \leq \lambda_2 \leq \dots$. Furthermore, $L1 = 0$, so the all 1's vector has eigenvalue 0.

The following lemma relates the second smallest eigenvalue to graph connectivity:

Lemma 21. *The second eigenvalue of L satisfies $\lambda_2 > 0$ if and only if G is connected.*

In fact, the lemma above generalizes: If $\lambda_k = 0$ and $\lambda_{k+1} \neq 0$, then G has exactly k connected components.

We now consider the spectra of the following (unweighted) graphs:

(i) *The complete graph K_n on n vertices.* The Laplacian L_{K_n} has $n - 1$'s on the diagonal and -1's everywhere else.

(ii) *The star graph S_n on n vertices, with edge set $\{(1, u) : 2 \leq u \leq n\}$.* The Laplacian L_{S_n} has $n - 1$ in the first entry, 1's in all remaining diagonal entries, and -1's in the first row and column.

(iii) *The path graph P_n on n vertices, with edge set $\{(u, u+1) : 1 \leq u < n\}$.* The Laplacian L_{P_n} is tridiagonal, with 1's in the first and last diagonal entries, 2's in all remaining diagonal entries, and -1's in the off-diagonals.

(iv) *The ring graph R_n on n vertices, which has all edges in common with P_n, as well as the edge $(1, n)$.* The Laplacian L_{R_n} is equal to L_{P_n}, except all diagonal entries are 2 and $L_{R_n}(1, n) = L_{R_n}(n, 1) = -1$.

In fact, the following series of lemmas characterize the full spectra:

Lemma 22. *The graph Laplacian L_{K_n} has eigenvalue 0 with multiplicity 1, and eigenvalue n with multiplicity $n - 1$.*

Lemma 23. *The graph Laplacian L_{S_n} has eigenvalue 0 with multiplicity 1, eigenvalue 1 with multiplicity $n - 2$, and eigenvalue n with multiplicity 1.*

Lemma 24. *The graph Laplacian L_{P_n} has eigenvalues $\{2(1-\cos(\pi k/n))\}_{k=0}^{n-1}$, each with multiplicity 1.*

Lemma 25. *The graph Laplacian L_{R_n} has eigenvalues $\{2(1 - \cos(2\pi k/n)\}_{k=0}^{\lfloor n/2 \rfloor}$. When n is even, each eigenvalue has multiplicity 2, except the first and the last; when n is odd, each eigenvalue has multiplicity 2, except eigenvalue 0.*

The following lemma collects some additional statements concerning the spectra of (normalized) graph Laplacians, which we state for unweighted graphs, for simplicity. The proofs are all fairly straightforward (see Chung for details).

Lemma 26. *Suppose $G = (V, E)$ is an unweighted graph with $|V| = n$. Let $0 = \nu_1 \leq \nu_2 \leq \cdots \leq \nu_n$ denote the eigenvalues of the* normalized *graph Laplacian $L' = D^{-1/2}LD^{1/2}$.*

(i) If G is not a complete graph, we have $\nu_2 \leq 1$. If G is complete, then $\nu_2 = \frac{n}{n-1}$.

(ii) We have $\nu_i \leq 2$ for all i, with $\nu_n = 2$ if and only if a connected component of G is bipartite and nontrivial.

(iii) Suppose $n \geq 2$. If G has no isolated vertices, then

$$\nu_n \geq \frac{n}{n-1}.$$

Note: For the examples discussed in the previous subsection, we can show that:

- L'_{K_n} has eigenvalues 0 (with multiplicity 1), and $\frac{n}{n-1}$ (with multiplicity $n-1$).
- L'_{S_n} has eigenvalues 0, 1 (with multiplicity $n-2$), and 2.
- L'_{P_n} has eigenvalues $1 - \cos\left(\frac{\pi k}{n-1}\right)$ for $0 \leq k \leq n-1$.
- L'_{R_n} has eigenvalues $1 - \cos\left(\frac{2\pi k}{n}\right)$ for $0 \leq k \leq \left\lfloor \frac{n}{2} \right\rfloor$.

Can check that $\nu_n = 2$ only for S_n, P_n, and R_n (in case when n is even). Other inequalities may also be verified.

4.8 Stochastic block models

Model: We have a graph with n nodes and K communities. Let $B \in \mathbb{R}^{K \times K}$ be a symmetric matrix of connection probabilities between communities (generally assumed to be unknown), and let $Z = (Z_1, \ldots, Z_n)^T \in \{1, \ldots, K\}^n$ be a vector of community assignments. Then adjacency matrix $A \in \mathbb{R}^{n \times n}$ is generated using independent Bernoulli draws, where $P(A_{ij} = 1) = B_{Z_i, Z_j}$. Want to recover Z based on observing A.

We are interested in the provable performance of vanilla spectral clustering; in order to obtain the sharpest recovery results, one needs to use more sophisticated techniques.

4.8.1 Spectral clustering

References:

- Lei & Rinaldo, "Consistency of spectral clustering in SBMs," 2015.
- Rohe et al., "Spectral clustering and the high-dimensional SBM," 2011. (The main difference is that the expected node degrees are assumed to be linear in n, whereas the work of Lei & Rinaldo can handle expected node degrees as small as $\Omega(\log n)$.)

Theorem 27. *Suppose spectral clustering with 10-approximate k-means produces a partitioning $\tilde{\mathcal{C}}$. Suppose the true partition $\mathcal{C}$ has all communities of size $\frac{n}{K}$, and assume $p_{\max} = \max_{i,j} B_{ij}$ satisfies*

$$\frac{c_1 \log n}{n} \leq p_{\max} \leq \frac{c_2 n \lambda_{\min}^2(B)}{K^3}.$$

With probability at least $1 - \frac{1}{n^3}$, the clustering $\tilde{\mathcal{C}}$ differs from $\mathcal{C}$ in at most $\frac{CK^2 p_{\max}}{\lambda_{\min}^2(B)}$ points.

Interpretation: The quantity $\lambda_{\min}(B)$ measures difference between edge probabilities for within-community vs. between-community connections. Consider case when $B = \alpha_n B_0$ and $B_0 = \lambda I_K + (1-\lambda)1_K 1_K^T$, so within-community connections have probability α_n and between-community connections have probability $(1-\lambda)\alpha_n$. Also, $\lambda_{\min}(B) = \lambda\alpha_n$ since $\lambda_{\min}(B_0) = \lambda$, and $p_{\max} = \alpha_n$, so $\frac{CK^2 p_{\max}}{n\lambda^2_{\min}(B)} = \frac{CK^2}{n\lambda^2\alpha_n}$, which converges to 0, say, when $\alpha_n \asymp \frac{\log n}{n}$. We see that a larger value of λ (corresponding to larger separation between connection probabilities) thus implies a smaller error rate guarantee.

In the condition of the theorem, note that $np_{\max}$ is the maximum expected node degree, and the first inequality in the displayed condition is a statement about the minimum edge density, whereas the second inequality can be viewed as an upper bound on the number of communities K. Recall that the regime $p \geq \frac{c_1 \log n}{n}$ guarantees that Erdös-Renyi graph is connected a.s.

Proof of Theorem 27 (sketch). The main idea is to think of A as a perturbation of underlying probability matrix $P = \mathbb{E}(A) \in \mathbb{R}^{n\times n}$. The main idea is that the eigenvectors of P are well-behaved, and matrix of top K eigenvectors has exactly K distinct rows. This is captured in the following lemma:

Lemma 28. *Suppose B is full-rank. Let $V_0 D_0 V_0^T = P$ be a spectral decomposition, where $V_0 \in \mathbb{R}^{n\times K}$ and $D_0 \in \mathbb{R}^{K\times K}$. Then $V_0 = \Theta X$ (i.e., V_0 has n rows, which are copies of the K rows of X), where $\Theta \in \mathbb{R}^{n\times K}$ is the membership matrix of the communities, and $X \in \mathbb{R}^{K\times K}$, with $\|X_{i,:} - X_{j,:}\|_2 = \sqrt{\frac{2K}{n}}$, for $i \neq j$.*

In particular, V_0 has K distinct (and well-separated) rows, so it makes sense to cluster them. We omit the rest of the proof. □

Remark: In fact, one might wonder why we cannot simply perform approximate k-means clustering on the rows of A, since P can be exactly partitioned into K clusters. It can be shown that $\|A\|_F^2 \approx pn^2$ and $\|P\|_F^2 \leq n$, so we have $\|A-P\|_F^2 \gtrsim n^2$. The proofs of the statements above show that $\|U-V\|_F^2 \precsim \frac{1}{n}$. In fact, although A is not exactly becoming "closer" to P as $n \to \infty$, the matrices of top-K eigenvectors are converging.

4.8.2 Optimization approach

Setting: For simplicity, suppose we have two communities with exactly $\frac{n}{2}$ nodes each, and $B = \begin{pmatrix} a/n & b/n \\ b/n & a/n \end{pmatrix}$.

Eigenvalues of $P = \mathbb{E}[A]$ are $\{\frac{a+b}{n}, \frac{a-b}{n}, 0\}$, where 0 has multiplicity $n-2$. Eigenvectors of first two eigenvalues are $(1,\dots,1)$ and $(1,\dots,1,-1,\dots,-1)$. Communities could be recovered by taking the second eigenvector and assigning vertices to communities based on signs of components.

On the other hand, we observe A rather than P, so extracting the second eigenvector might not give an integral solution. We would have to round, and need $|a-b|$ to be sufficiently large in order for this to work, w.h.p. Turns out that the following method works better:

Optimization problem:

$$\max \quad x^T A x = \sum_{i,j} A_{ij} x_i x_j$$

$$\begin{aligned} \text{s.t.} \quad & x_i = \pm 1, \forall i, \\ & \sum_j x_j = 0. \end{aligned}$$

However, this is very difficult to optimize. (Note that If we relax the condition $x_i = \pm 1$ to a condition such as $\|x\|_2^2 = n$, we extract the second eigenvector—that is computable in $O(n^2)$ time using the power iteration method.)

Instead, using the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we can write $x^T A x = \mathrm{tr}(A x x^T)$, and denoting $X = xx^T$, we can rewrite the optimization problem as

$$\begin{aligned} \max \quad & \mathrm{tr}(AX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall i, \\ & \mathrm{rank}(X) = 1, \\ & X1 = 0. \end{aligned}$$

In particular, the objective function is now *linear* in the lifted variable X. If we relax the constraint $\mathrm{rank}(X) = 1$, we arrive at the semidefinite program

$$\begin{aligned} \max \quad & \mathrm{tr}(AX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall i, \\ & X \succeq 0 \\ & X1 = 0, \end{aligned}$$

which is convex.

For more details, including on conditions under which the second-eigenvector and SDP methods succeed, see the survey by Abbe (2018).

References

[1] R. J. Baxter. *Exactly Solved Models in Statistical Mechanics.* Dover Books on Physics. Dover Publications, 2007.

[2] J. Bento and A. Montanari. Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems*, pages 1303–1311, 2009.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.

[4] T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

[5] T. T. Cai, Z. Ren, and H. H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Statist.*, 10(1):1–59, 2016.

[6] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 30(1):55–66, 2008.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, July 2008.

[8] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1990.

[9] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. QUIC: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15:2911–2947, 2014.

[10] E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925.

[11] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[12] K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale ℓ_1-regularized logistic regression. *Journal of Machine Learning Research*, 8(Jul):1519–1555, 2007.

[13] R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Electron. J. Statist.*, 6:2125–2149, 2012.

[14] N. Meinshausen. A note on the Lasso for Gaussian graphical model selection. *Statistics & Probability Letters*, 78(7):880–884, 2008.

[15] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[16] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

[17] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.

[18] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38:1287, 2010.

[19] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[20] N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

[21] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

[22] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, May 2009.

[23] W. Wang, M. J. Wainwright, and K. Ramchandran. Information-theoretic bounds on model selection for Gaussian Markov random fields. In *ISIT*, pages 1373–1377, 2010.

[24] D. M. Witten, J. H. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

[25] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[26] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014.

[27] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.