CHAPTER 4

Matrix estimation

Over the past decade or so, matrices have entered the picture of high-dimensional statistics for several reasons. Perhaps the simplest explanation is that they are the most natural extension of vectors. While this is true, and we will see examples where the extension from vectors to matrices is straightforward, matrices have a much richer structure than vectors allowing “interaction” between their rows and columns. In particular, while we have been describing simple vectors in terms of their sparsity, here we can measure the complexity of a matrix by its *rank*. This feature was successfully employed in a variety of applications ranging from *multi-task learning* to *collaborative filtering*. This last application was made popular by the NETFLIX prize in particular.

In this chapter, we study several statistical problems where the parameter of interest θ is a matrix rather than a vector. These problems include: multivariate regression, covariance matrix estimation and principal component analysis. Before getting to these topics, we begin by a quick reminder on matrices and linear algebra.

4.1 BASIC FACTS ABOUT MATRICES

Matrices are much more complicated objects than vectors. In particular, while vectors can be identified with linear operators from $\mathbb{R}^d$ to $\mathbb{R}$, matrices can be identified to linear operators from $\mathbb{R}^d$ to $\mathbb{R}^n$ for $n \geq 1$. This seemingly simple fact gives rise to a profusion of notions and properties as illustrated by Bernstein’s book [Ber09] that contains facts about matrices over more than a thousand pages. Fortunately, we will be needing only a small number of such properties, which can be found in the excellent book [GVL96], that has become a standard reference on matrices and numerical linear algebra.

Singular value decomposition

Let $A = \{a_{ij}, 1 \le i \le m, 1 \le j \le n\}$ be a $m \times n$ real matrix of rank $r \le \min(m,n)$. The *Singular Value Decomposition* (SVD) of A is given by

$$A = UDV^\top = \sum_{j=1}^r \lambda_j u_j v_j^\top \,,$$

where D is a $r \times r$ diagonal matrix with positive diagonal entries $\{\lambda_1, \dots, \lambda_r\}$, U is a matrix with columns $\{u_1, \dots, u_r\} \in \mathbb{R}^m$ that are orthonormal and V is a matrix with columns $\{v_1, \dots, v_r\} \in \mathbb{R}^n$ that are also orthonormal. Moreover, it holds that

$$AA^\top u_j = \lambda_j^2 u_j \,, \qquad \text{and} \qquad A^\top A v_j = \lambda_j^2 v_j$$

for $j = 1, \dots, r$. The values $\lambda_j > 0$ are called *singular values* of A and are uniquely defined. If rank $r < \min(n,m)$ then the singular values of A are given by $\lambda = (\lambda_1, \dots, \lambda_r, 0, \dots, 0)\top \in \mathbb{R}^{\min(n,m)}$ where there are $\min(n,m) - r$ zeros. This way, the vector λ of singular values of a $n \times m$ matrix is a vector in $\mathbb{R}^{\min(n,m)}$.

In particular, if A is a $n \times n$ symmetric positive semidefinite (PSD), i.e. $A^\top = A$ and $u^\top A u \ge 0$ for all $u \in \mathbb{R}^n$, then the singular values of A are equal to its eigenvalues.

The largest singular value of A denoted by $\lambda_{\max}(A)$ also satisfies the following variational formulation:

$$\lambda_{\max}(A) = \max_{x \in \mathbb{R}^n} \frac{|Ax|_2}{|x|_2} = \max_{\substack{x \in \mathbb{R}^n \\ y \in \mathbb{R}^m}} \frac{y^\top A x}{|y|_2 |x|_2} = \max_{\substack{x \in \mathcal{S}^{n-1} \\ y \in \mathcal{S}^{m-1}}} y^\top A x \,.$$

In the case of a $n \times n$ PSD matrix A, we have

$$\lambda_{\max}(A) = \max_{x \in \mathcal{S}^{n-1}} x^\top A x \,.$$

Norms and inner product

Let $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ be two real matrices. Their size will be implicit in the following notation.

Vector norms

The simplest way to treat a matrix is to deal with it as if it were a vector. In particular, we can extend ℓ_q norms to matrices:

$$|A|_q = \Big(\sum_{ij} |a_{ij}|^q \Big)^{1/q} \,, \quad q > 0 \,.$$

The cases where $q \in \{0, \infty\}$ can also be extended matrices:

$$|A|_0 = \sum_{ij} \mathbb{1}(a_{ij} \neq 0) \,, \qquad |A|_\infty = \max_{ij} |a_{ij}| \,.$$

The case $q = 2$ plays a particular role for matrices and $|A|_2$ is called the *Frobenius* norm of A and is often denoted by $\|A\|_F$. It is also the Hilbert-Schmidt norm associated to the inner product:

$$\langle A, B\rangle = \mathsf{Tr}(A^\top B) = \mathsf{Tr}(B^\top A)\,.$$

Spectral norms

Let $\lambda = (\lambda_1, \dots, \lambda_r, 0, \dots, 0)$ be the singular values of a matrix A. We can define spectral norms on A as vector norms on the vector λ. In particular, for any $q \in [1, \infty]$,

$$\|A\|_q = |\lambda|_q\,,$$

is called *Schatten q-norm* of A. Here again, special cases have special names:

- $q = 2$: $\|A\|_2 = \|A\|_F$ is the Frobenius norm defined above.
- $q = 1$: $\|A\|_1 = \|A\|_*$ is called the Nuclear norm (or trace norm) of A.
- $q = \infty$: $\|A\|_\infty = \lambda_{\max}(A) = \|A\|_{\mathrm{op}}$ is called the operator norm (or spectral norm) of A.

We are going to employ these norms to assess the proximity to our matrix of interest. While the interpretation of vector norms is clear by extension from the vector case, the meaning of "$\|A-B\|_{\mathrm{op}}$ is small" is not as transparent. The following subsection provides some inequalities (without proofs) that allow a better reading.

Useful matrix inequalities

Let A and B be two $m \times n$ matrices with singular values $\lambda_1(A) \ge \lambda_2(A) \ldots \ge \lambda_{\min(m,n)}(A)$ and $\lambda_1(B) \ge \ldots \ge \lambda_{\min(m,n)}(B)$ respectively. Then the following inequalities hold:

$$\max_k \big|\lambda_k(A) - \lambda_k(B)\big| \le \|A - B\|_{\mathrm{op}}\,, \qquad \text{Weyl (1912)}$$

$$\sum_k \big|\lambda_k(A) - \lambda_k(B)\big|^2 \le \|A - B\|_F^2\,, \qquad \text{Hoffman-Weilandt (1953)}$$

$$\langle A, B\rangle \le \|A\|_q \|B\|_q\,,\ \frac{1}{p} + \frac{1}{q} = 1, p, q \in [1, \infty]\,, \qquad \text{Hölder}$$

4.2 MULTIVARIATE REGRESSION

In the traditional regression setup, the response variable Y is a scalar. In several applications, the goal is not to predict a variable but rather a vector $Y \in \mathbb{R}^T$, still from a covariate $X \in \mathbb{R}^d$. A standard example arises in genomics data where Y contains T physical measurements of a patient and X contains the expression levels for d genes. As a result the regression function in this case $f(x) = \mathbb{E}[Y|X = x]$ is a function from $\mathbb{R}^d$ to $\mathbb{R}^T$. Clearly, f can be estimated independently for each coordinate, using the tools that we have developed in the previous chapter. However, we will see that in several interesting scenarios, some structure is shared across coordinates and this information can be leveraged to yield better prediction bounds.

The model

Throughout this section, we consider the following multivariate linear regression model:

$$\mathbb{Y} = \mathbb{X}\Theta^* + E\,, \tag{4.1}$$

where $\mathbb{Y} \in \mathbb{R}^{n\times T}$ is the matrix of observed responses, $\mathbb{X}$ is the $n \times d$ observed design matrix (as before), $\Theta \in \mathbb{R}^{d\times T}$ is the matrix of unknown parameters and $E \sim \mathsf{subG}_{n\times T}(\sigma^2)$ is the noise matrix. In this chapter, we will focus on the prediction task, which consists in estimating $\mathbb{X}\Theta^*$.

As mentioned in the foreword of this chapter, we can view this problem as T (univariate) linear regression problems $Y^{(j)} = \mathbb{X}\theta^{*,(j)} + \varepsilon^{(j)}, j = 1, \dots, T$, where $Y^{(j)}, \theta^{*,(j)}$ and $\varepsilon^{(j)}$ are the jth column of $\mathbb{Y}, \Theta^*$ and E respectively. In particular, an estimator for $\mathbb{X}\Theta^*$ can be obtained by concatenating the estimators for each of the T problems. This approach is the subject of Problem 4.1.

The columns of Θ^* correspond to T different regression tasks. Consider the following example as a motivation. Assume that the Subway headquarters want to evaluate the effect of d variables (promotions, day of the week, TV ads,…) on their sales. To that end, they ask each of their $T = 40,000$ restaurants to report their sales numbers for the past $n = 200$ days. As a result, franchise j returns to headquarters a vector $\mathbb{Y}^{(j)} \in \mathbb{R}^n$. The d variables for each of the n days are already known to headquarters and are stored in a matrix $\mathbb{X} \in \mathbb{R}^{n\times d}$. In this case, it may be reasonable to assume that the same subset of variables has an impact of the sales for each of the franchise, though the magnitude of this impact may differ from franchise to franchise. As a result, one may assume that the matrix Θ^* has each of its T columns that is row sparse and that they *share the same sparsity pattern*, i.e., Θ^* is of the form:

$$\Theta^* = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix},$$

where $\bullet$ indicates a potentially nonzero entry.

It follows from the result of Problem 4.1 that if each task is performed individually, one may find an estimator $\hat\Theta$ such that

$$\frac{1}{n}\mathbb{E}\|\mathbb{X}\hat\Theta - \mathbb{X}\Theta^*\|_F^2 \lesssim \sigma^2 \frac{kT\log(ed)}{n}\,,$$

where k is the number of nonzero coordinates in each column of Θ^*. We remember that the term $\log(ed)$ corresponds to the additional price to pay for not knowing where the nonzero components are. However, in this case, when the number of tasks grows, this should become easier. This fact was proved in [LPTVDG11]. We will see that we can recover a similar phenomenon when the number of tasks becomes large, though larger than in [LPTVDG11]. Indeed, rather than exploiting sparsity, observe that such a matrix Θ^* has rank k. This is the kind of structure that we will be predominantly using in this chapter.

Rather than assuming that the columns of Θ^* share the same sparsity pattern, it may be more appropriate to assume that the matrix Θ^* is low rank or approximately so. As a result, while the matrix may not be sparse at all, the fact that it is low rank still materializes the idea that some structure is shared across different tasks. In this more general setup, it is assumed that the columns of Θ^* live in a lower dimensional space. Going back to the Subway example this amounts to assuming that while there are 40,000 franchises, there are only a few canonical profiles for these franchises and that all franchises are linear combinations of these profiles.

Sub-Gaussian matrix model

Recall that under the assumption ORT for the design matrix, i.e., $\mathbb{X}^\top\mathbb{X} = nI_d$, then the univariate regression model can be reduced to the sub-Gaussian sequence model. Here we investigate the effect of this assumption on the multivariate regression model (4.1).

Observe that under assumption ORT,

$$\frac{1}{n}\mathbb{X}^\top\mathbb{Y} = \Theta^* + \frac{1}{n}\mathbb{X}^\top E\,.$$

Which can be written as an equation in $\mathbb{R}^{d\times T}$ called the *sub-Gaussian matrix model (sGMM)*:

$$y = \Theta^* + F\,, \tag{4.2}$$

where $y = \frac{1}{n}\mathbb{X}^\top\mathbb{Y}$ and $F = \frac{1}{n}\mathbb{X}^\top E \sim \mathsf{subG}_{d\times T}(\sigma^2/n)$.

Indeed, for any $u \in \mathcal{S}^{d-1}, v \in \mathcal{S}^{T-1}$, it holds

$$u^\top F v = \frac{1}{n}(\mathbb{X}u)^\top E v = \frac{1}{\sqrt{n}} w^\top E v \sim \mathsf{subG}(\sigma^2/n)\,,$$

where $w = \mathbb{X}u/\sqrt{n}$ has unit norm: $|w|_2^2 = u^\top \frac{\mathbb{X}^\top\mathbb{X}}{n} u = |u|_2^2 = 1$.

Akin to the sub-Gaussian sequence model, we have a *direct* observation model where we observe the parameter of interest with additive noise. This enables us to use thresholding methods for estimating Θ^* when $|\Theta^*|_0$ is small. However, this also follows from Problem 4.1. The reduction to the vector case in the sGMM is just as straightforward. The interesting analysis begins when Θ^* is low-rank, which is equivalent to sparsity in its unknown eigenbasis.

Consider the SVD of Θ^*:

$$\Theta^* = \sum_j \lambda_j u_j v_j^\top \,.$$

and recall that $\|\Theta^*\|_0 = |\lambda|_0$. Therefore, if we knew u_j and v_j, we could simply estimate the λ_js by hard thresholding. It turns out that estimating these eigenvectors by the eigenvectors of y is sufficient.

Consider the SVD of the observed matrix y:

$$y = \sum_j \hat\lambda_j \hat u_j \hat v_j^\top \,.$$

Definition 4.1. The **singular value thresholding** estimator with threshold $2\tau \ge 0$ is defined by

$$\hat\Theta^{\mathrm{SVT}} = \sum_j \hat\lambda_j \mathbb{1}(|\hat\lambda_j| > 2\tau) \hat u_j \hat v_j^\top \,.$$

Recall that the threshold for the hard thresholding estimator was chosen to be the level of the noise with high probability. The singular value thresholding estimator obeys the same rule, except that the norm in which the magnitude of the noise is measured is adapted to the matrix case. Specifically, the following lemma will allow us to control the operator norm of the matrix F.

Lemma 4.2. *Let A be a $d \times T$ random matrix such that $A \sim \mathsf{subG}_{d\times T}(\sigma^2)$. Then*

$$\|A\|_{\mathrm{op}} \le 4\sigma\sqrt{\log(12)(d \vee T)} + 2\sigma\sqrt{2\log(1/\delta)}$$

with probability $1-\delta$.

Proof. This proof follows the same steps as Problem 1.4. Let $\mathcal{N}_1$ be a 1/4-net for $\mathcal{S}^{d-1}$ and $\mathcal{N}_2$ be a 1/4-net for $\mathcal{S}^{T-1}$. It follows from Lemma 1.18 that we can always choose $|\mathcal{N}_1| \le 12^d$ and $|\mathcal{N}_2| \le 12^T$. Moreover, for any $u \in \mathcal{S}^{d-1}, v \in \mathcal{S}^{T-1}$, it holds

$$\begin{aligned}
u^\top A v &\le \max_{x\in\mathcal{N}_1} x^\top A v + \frac{1}{4}\max_{u\in\mathcal{S}^{d-1}} u^\top A v \\
&\le \max_{x\in\mathcal{N}_1}\max_{y\in\mathcal{N}_2} x^\top A y + \frac{1}{4}\max_{x\in\mathcal{N}_1}\max_{v\in\mathcal{S}^{T-1}} x^\top A v + \frac{1}{4}\max_{u\in\mathcal{S}^{d-1}} u^\top A v \\
&\le \max_{x\in\mathcal{N}_1}\max_{y\in\mathcal{N}_2} x^\top A y + \frac{1}{2}\max_{u\in\mathcal{S}^{d-1}}\max_{v\in\mathcal{S}^{T-1}} u^\top A v
\end{aligned}$$

It yields

$$\|A\|_{\mathrm{op}} \le 2\max_{x\in\mathcal{N}_1}\max_{y\in\mathcal{N}_2} x^\top A y$$

So that for any $t \geq 0$, by a union bound,

$$\mathbb{P}\big(\|A\|_{\mathrm{op}} > t\big) \leq \sum_{\substack{x \in \mathcal{N}_1 \\ y \in \mathcal{N}_2}} \mathbb{P}\big(x^\top A y > t/2\big)$$

Next, since $A \sim \mathsf{subG}_{d\times T}(\sigma^2)$, it holds that $x^\top A y \sim \mathsf{subG}(\sigma^2)$ for any $x \in \mathcal{N}_1, y \in \mathcal{N}_2$.Together with the above display, it yields

$$\mathbb{P}\big(\|A\|_{\mathrm{op}} > t\big) \leq 12^{d+T} \exp\big(-\frac{t^2}{8\sigma^2}\big) \leq \delta$$

for

$$t \geq 4\sigma\sqrt{\log(12)(d \vee T)} + 2\sigma\sqrt{2\log(1/\delta)}\,.$$

□

The following theorem holds.

Theorem 4.3. *Consider the multivariate linear regression model* (4.1) *under the assumption* ***ORT*** *or, equivalently, the sub-Gaussian matrix model* (4.2)*. Then, the singular value thresholding estimator* $\hat\Theta^{\mathrm{SVT}}$ *with threshold*

$$2\tau = 8\sigma\sqrt{\frac{\log(12)(d \vee T)}{n}} + 4\sigma\sqrt{\frac{2\log(1/\delta)}{n}}\,, \tag{4.3}$$

satisfies

$$\begin{aligned}\frac{1}{n}\|\mathbb{X}\hat\Theta^{\mathrm{SVT}} - \mathbb{X}\Theta^*\|_F^2 = \|\hat\Theta^{\mathrm{SVT}} - \Theta^*\|_F^2 &\leq 144\,\mathrm{rank}(\Theta^*)\tau^2 \\ &\lesssim \frac{\sigma^2\,\mathrm{rank}(\Theta^*)}{n}\Big(d \vee T + \log(1/\delta)\Big)\,.\end{aligned}$$

with probability $1-\delta$.

Proof. Assume without loss of generality that the singular values of Θ^* and y are arranged in a non increasing order: $\lambda_1 \geq \lambda_2 \geq \ldots$ and $\hat\lambda_1 \geq \hat\lambda_2 \geq \ldots$. Define the set $S = \{j \,:\, |\hat\lambda_j| > 2\tau\}$.

Observe first that it follows from Lemma 4.2 that $\|F\|_{\mathrm{op}} \leq \tau$ for τ chosen as in (4.3) on an event $\mathcal{A}$ such that $\mathbb{P}(\mathcal{A}) \geq 1-\delta$. The rest of the proof is on $\mathcal{A}$.

Note that it follows from Weyl's inequality that $|\hat\lambda_j - \lambda_j| \leq \|F\|_{\mathrm{op}} \leq \tau$. It implies that $S \subset \{j \,:\, |\lambda_j| > \tau\}$ and $S^c \subset \{j \,:\, |\lambda_j| \leq 3\tau\}$.

Next define the oracle $\bar\Theta = \sum_{j\in S} \lambda_j u_j v_j^\top$ and note that

$$\|\hat\Theta^{\mathrm{SVT}} - \Theta^*\|_F^2 \leq 2\|\hat\Theta^{\mathrm{SVT}} - \bar\Theta\|_F^2 + 2\|\bar\Theta - \Theta^*\|_F^2 \tag{4.4}$$

Using Cauchy-Schwarz, we control the first term as follows

$$\|\hat\Theta^{\mathrm{SVT}} - \bar\Theta\|_F^2 \leq \mathrm{rank}(\hat\Theta^{\mathrm{SVT}} - \bar\Theta)\|\hat\Theta^{\mathrm{SVT}} - \bar\Theta\|_{\mathrm{op}}^2 \leq 2|S|\|\hat\Theta^{\mathrm{SVT}} - \bar\Theta\|_{\mathrm{op}}^2$$

Moreover,

$$\begin{aligned}\|\hat\Theta^{\rm SVT}-\bar\Theta\|_{\rm op} &\le \|\hat\Theta^{\rm SVT}-y\|_{\rm op}+\|y-\Theta^*\|_{\rm op}+\|\Theta^*-\bar\Theta\|_{\rm op}\\ &\le \max_{j\in S^c}|\hat\lambda_j|+\tau+\max_{j\in S^c}|\lambda_j|\le 6\tau\,.\end{aligned}$$

Therefore,

$$\|\hat\Theta^{\rm SVT}-\bar\Theta\|_F^2\le 72|S|\tau^2=72\sum_{j\in S}\tau^2\,.$$

The second term in (4.4) can be written as

$$\|\bar\Theta-\Theta^*\|_F^2=\sum_{j\in S^c}|\lambda_j|^2\,.$$

Plugging the above two displays in (4.4), we get

$$\|\hat\Theta^{\rm SVT}-\Theta^*\|_F^2\le 144\sum_{j\in S}\tau^2+\sum_{j\in S^c}|\lambda_j|^2$$

Since on S, $\tau^2=\min(\tau^2,|\lambda_j|^2)$ and on S^c, $|\lambda_j|^2\le 3\min(\tau^2,|\lambda_j|^2)$, it yields,

$$\begin{aligned}\|\hat\Theta^{\rm SVT}-\Theta^*\|_F^2 &\le 432\sum_j\min(\tau^2,|\lambda_j|^2)\\ &\le 432\sum_{j=1}^{\mathrm{rank}(\Theta^*)}\tau^2\\ &= 432\,\mathrm{rank}(\Theta^*)\tau^2\,.\end{aligned}$$

□

In the next subsection, we extend our analysis to the case where $\mathbb{X}$ does not necessarily satisfy the assumption ORT.

Penalization by rank

The estimator from this section is the counterpart of the BIC estimator in the spectral domain. However, we will see that unlike BIC, it can be computed efficiently.

Let $\hat\Theta^{\rm RK}$ be any solution to the following minimization problem:

$$\min_{\Theta\in\mathbb{R}^{d\times T}}\Big\{\frac{1}{n}\|\mathbb{Y}-\mathbb{X}\Theta\|_F^2+2\tau^2\,\mathrm{rank}(\Theta)\Big\}\,.$$

This estimator is called *estimator by rank penalization with regularization parameter* τ^2. It enjoys the following property.

Theorem 4.4. *Consider the multivariate linear regression model* (4.1)*. Then, the estimator by rank penalization* $\hat\Theta^{\text{RK}}$ *with regularization parameter* τ^2*, where* τ *is defined in* (4.3) *satisfies*

$$\frac{1}{n}\|\mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^*\|_F^2 \le 8\,\text{rank}(\Theta^*)\tau^2 \lesssim \frac{\sigma^2\,\text{rank}(\Theta^*)}{n}\Big(d \vee T + \log(1/\delta)\Big)\,.$$

with probability $1-\delta$*.*

Proof. We begin as usual by noting that

$$\|\mathbb{Y} - \mathbb{X}\hat\Theta^{\text{RK}}\|_F^2 + 2n\tau^2\,\text{rank}(\hat\Theta^{\text{RK}}) \le \|\mathbb{Y} - \mathbb{X}\Theta^*\|_F^2 + 2n\tau^2\,\text{rank}(\Theta^*)\,,$$

which is equivalent to

$$\|\mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^*\|_F^2 \le 2\langle E, \mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^*\rangle - 2n\tau^2\,\text{rank}(\hat\Theta^{\text{RK}}) + 2n\tau^2\,\text{rank}(\Theta^*)\,.$$

Next, by Young's inequality, we have

$$2\langle E, \mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^*\rangle = 2\langle E, U\rangle^2 + \frac{1}{2}\|\mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^*\|_F^2\,,$$

where

$$U = \frac{\mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^*}{\|\mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^*\|_F}\,.$$

Write

$$\mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^* = \Phi N\,,$$

where Φ is a $n \times r, r \le d$ matrix whose columns form orthonormal basis of the column span of $\mathbb{X}$. The matrix Φ can come from the SVD of $\mathbb{X}$ for example: $\mathbb{X} = \Phi\Lambda\Psi^\top$. It yields

$$U = \frac{\Phi N}{\|N\|_F}$$

and

$$\|\mathbb{X}\hat\Theta^{\text{RK}} - \mathbb{X}\Theta^*\|_F^2 \le 4\langle \Phi^\top E, N/\|N\|_F\rangle^2 - 4n\tau^2\,\text{rank}(\hat\Theta^{\text{RK}}) + 4n\tau^2\,\text{rank}(\Theta^*)\,. \tag{4.5}$$

Note that $\text{rank}(N) \le \text{rank}(\hat\Theta^{\text{RK}}) + \text{rank}(\Theta^*)$. Therefore, by Hölder's inequality, we get

$$\begin{aligned}
\langle E, U\rangle^2 &= \langle \Phi^\top E, N/\|N\|_F\rangle^2 \\
&\le \|\Phi^\top E\|_{\text{op}}^2 \frac{\|N\|_1^2}{\|N\|_F^2} \\
&\le \text{rank}(N)\|\Phi^\top E\|_{\text{op}}^2 \\
&\le \|\Phi^\top E\|_{\text{op}}^2\big[\text{rank}(\hat\Theta^{\text{RK}}) + \text{rank}(\Theta^*)\big]\,.
\end{aligned}$$

Next, note that Lemma 4.2 yields $\|\Phi^\top E\|_{\text{op}}^2 \le n\tau^2$ so that

$$\langle E, U\rangle^2 \le n\tau^2\big[\text{rank}(\hat\Theta^{\text{RK}}) + \text{rank}(\Theta^*)\big]\,.$$

Together with (4.5), this completes the proof. □

It follows from Theorem 4.4 that the estimator by rank penalization enjoys the same properties as the singular value thresholding estimator even when $\mathbb{X}$ does not satisfies the ORT condition. This is reminiscent of the BIC estimator which enjoys the same properties as the hard thresholding estimator. However this analogy does not extend to computational questions. Indeed, while the rank penalty, just like the sparsity penalty, is not convex, it turns out that $\mathbb{X}\hat{\Theta}^{\text{RK}}$ can be computed efficiently.

Note first that

$$\min_{\Theta \in \mathbb{R}^{d\times T}} \frac{1}{n}\|\mathbb{Y}-\mathbb{X}\Theta\|_F^2 + 2\tau^2 \operatorname{rank}(\Theta) = \min_k \Big\{ \frac{1}{n} \min_{\substack{\Theta \in \mathbb{R}^{d\times T} \\ \operatorname{rank}(\Theta)\le k}} \|\mathbb{Y}-\mathbb{X}\Theta\|_F^2 + 2\tau^2 k \Big\} .$$

Therefore, it remains to show that

$$\min_{\substack{\Theta \in \mathbb{R}^{d\times T} \\ \operatorname{rank}(\Theta)\le k}} \|\mathbb{Y}-\mathbb{X}\Theta\|_F^2$$

can be solved efficiently. To that end, let $\bar{\mathbb{Y}} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top\mathbb{Y}$ denote the orthogonal projection of $\mathbb{Y}$ onto the image space of $\mathbb{X}$: this is a linear operator from $\mathbb{R}^{d\times T}$ into $\mathbb{R}^{n\times T}$. By the Pythagorean theorem, we get for any $\Theta \in \mathbb{R}^{d\times T}$,

$$\|\mathbb{Y}-\mathbb{X}\Theta\|_F^2 = \|\mathbb{Y}-\bar{\mathbb{Y}}\|_F^2 + \|\bar{\mathbb{Y}}-\mathbb{X}\Theta\|_F^2 .$$

Next consider the SVD of $\bar{\mathbb{Y}}$:

$$\bar{\mathbb{Y}} = \sum_j \lambda_j u_j v_j^\top$$

where $\lambda_1 \ge \lambda_2 \ge \dots$. The claim is that if we define $\tilde{\mathbb{Y}}$ by

$$\tilde{\mathbb{Y}} = \sum_{j=1}^k \lambda_j u_j v_j^\top$$

which is clearly of rank at most k, then it satisfies

$$\|\bar{\mathbb{Y}}-\tilde{\mathbb{Y}}\|_F^2 = \min_{Z:\operatorname{rank}(Z)\le k} \|\bar{\mathbb{Y}}-Z\|_F^2 .$$

Indeed,

$$\|\bar{\mathbb{Y}}-\tilde{\mathbb{Y}}\|_F^2 = \sum_{j>k} \lambda_j^2 ,$$

and for any matrix Z such that $\operatorname{rank}(Z) \le k$ with SVD

$$Z = \sum_{j=1}^k \mu_j x_j y_j^\top ,$$

where $\mu_1 \ge \mu_2 \ge \dots$, we have by Hoffman-Weilandt

$$\|Z-\bar{Y}\|_F^2 \ge \sum_{j\ge 1} |\lambda_j - \mu_j|^2 \ge \sum_{j>k} \lambda_j^2 .$$

Therefore, any minimizer of $\mathbb{X}\Theta \mapsto \|\mathbb{Y} - \mathbb{X}\Theta\|_F^2$ over matrices of rank at most k can be obtained by truncating the SVD of $\mathbb{Y}$ at order k.

Once $\mathbb{X}\hat{\Theta}^{\text{RK}}$ has been found, one may obtain a corresponding $\hat{\Theta}^{\text{RK}}$ by least squares but this is not necessary for our results.

Remark 4.5. While the rank penalized estimator can be computed efficiently, it is worth pointing out that a convex relaxation for the rank penalty can also be used. The estimator by nuclear norm penalization $\hat{\Theta}$ is defined to be any solution to the minimization problem

$$\min_{\Theta \in \mathbb{R}^{d \times T}} \Big\{ \frac{1}{n} \|\mathbb{Y} - \mathbb{X}\Theta\|_F^2 + \tau \|\Theta\|_1 \Big\}$$

Clearly this criterion is convex and it can actually be implemented efficiently using semi-definite programming. It has been popularized by matrix completion problems. Let $\mathbb{X}$ have the following SVD:

$$\mathbb{X} = \sum_{j=1}^{r} \lambda_j u_j v_j^\top ,$$

with $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_r > 0$. It can be shown that for some appropriate choice of τ, it holds

$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \frac{\lambda_1}{\lambda_r} \frac{\sigma^2 \operatorname{rank}(\Theta^*)}{n} d \vee T$$

with probability .99. However, the proof of this result is far more involved than a simple adaption of the proof for the Lasso estimator to the matrix case (the readers are invited to see that for themselves). For one thing, there is no assumption on the design matrix (such as INC for example). This result can be found in [KLT11].

4.3 COVARIANCE MATRIX ESTIMATION

Empirical covariance matrix

Let $X_1, \ldots, X_n$ be n i.i.d. copies of a random vector $X \in \mathbb{R}^d$ such that $\mathbb{E}\big[XX^\top\big] = \Sigma$ for some unknown matrix $\Sigma \succ 0$ called *covariance matrix*. This matrix contains information about the moments of order 2 of the random vector X. A natural candidate to estimate Σ is the *empirical covariance matrix* $\hat{\Sigma}$ defined by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top .$$

Using the tools of Chapter 1, we can prove the following result.

Theorem 4.6. *Let $X_1, \ldots, X_n$ be n i.i.d. sub-Gaussian random vectors such that $\mathbb{E}[XX^\top] = \Sigma$ and $X \sim \mathsf{subG}_d(\|\Sigma\|_{\text{op}})$. Then*

$$\|\hat{\Sigma} - \Sigma\|_{\text{op}} \lesssim \|\Sigma\|_{\text{op}} \Big(\sqrt{\frac{d + \log(1/\delta)}{n}} \vee \frac{d + \log(1/\delta)}{n} \Big) ,$$

with probability $1-\delta$.

Proof. Observe first that without loss of generality we can assume that $\Sigma = I_d$. Indeed, note that since $\mathbb{E}\big[XX^\top\big] = \Sigma \succ 0$, then $X \sim \mathsf{subG}_d(\|\Sigma\|_{\mathrm{op}})$. Moreover, $Y = \Sigma^{-1/2}X \sim \mathsf{subG}_d(1)$ and $\mathbb{E}[YY^\top] = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_d$. Therefore,

$$\begin{aligned}
\frac{\|\hat\Sigma - \Sigma\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} &= \frac{\|\frac{1}{n}\sum_{i=1}^n X_iX_i^\top - \Sigma\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}}\\
&\le \frac{\|\Sigma^{1/2}\|_{\mathrm{op}}\|\frac{1}{n}\sum_{i=1}^n Y_iY_i^\top - I_d\|_{\mathrm{op}}\|\Sigma^{1/2}\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}}\\
&= \|\frac{1}{n}\sum_{i=1}^n Y_iY_i^\top - I_d\|_{\mathrm{op}}\,.
\end{aligned}$$

Let $\mathcal{N}$ be a 1/4-net for $\mathcal{S}^{d-1}$ such that $|\mathcal{N}| \le 12^d$. It follows from the proof of Lemma 4.2 that

$$\|\hat\Sigma - I_d\|_{\mathrm{op}} \le 2\max_{x,y\in\mathcal{N}} x^\top(\hat\Sigma - I_d)y$$

So that for any $t \ge 0$, by a union bound,

$$\mathbb{P}\big(\|\hat\Sigma - I_d\|_{\mathrm{op}} > t\big) \le \sum_{x,y\in\mathcal{N}} \mathbb{P}\big(x^\top(\hat\Sigma - I_d)y > t/2\big)\,. \tag{4.6}$$

It holds,

$$x^\top(\hat\Sigma - I_d)y = \frac{1}{n}\sum_{i=1}^n \big\{(X_i^\top x)(X_i^\top y) - \mathbb{E}\big[(X_i^\top x)(X_i^\top y)\big]\big\}\,.$$

Using polarization, we also have

$$(X_i^\top x)(X_i^\top y) = \frac{Z_+^2 - Z_-^2}{4}\,,$$

here $Z_+ = X_i^\top(x+y)$ and $Z_- = X_i^\top(x-y)$. It yields

$$\begin{aligned}
&\mathbb{E}\big[\exp\big(s\big((X_i^\top x)(X_i^\top y) - \mathbb{E}\big[(X_i^\top x)(X_i^\top y)\big]\big)\big)\big]\\
&\quad= \mathbb{E}\big[\exp\big(\frac{s}{4}\big(Z_+^2 - \mathbb{E}[Z_+^2]\big) - \frac{s}{4}\big(Z_-^2 - \mathbb{E}[Z_-^2]\big)\big)\big]\\
&\quad\le \Big(\mathbb{E}\big[\exp\big(\frac{s}{2}\big(Z_+^2 - \mathbb{E}[Z_+^2]\big)\big)\big]\mathbb{E}\big[\exp\big(-\frac{s}{2}\big(Z_-^2 - \mathbb{E}[Z_-^2]\big)\big)\big]\Big)^{1/2}\,,
\end{aligned}$$

where in the last inequality, we used Cauchy-Schwarz. Next, since $X \sim \mathsf{subG}_d(1)$, we have $Z_+, Z_- \sim \mathsf{subG}(2)$, and it follows from Lemma 1.12 that

$$Z_+^2 - \mathbb{E}[Z_+^2] \sim \mathsf{subE}(32)\,, \qquad \text{and} \qquad Z_-^2 - \mathbb{E}[Z_-^2] \sim \mathsf{subE}(32)$$

Therefore for any $s \le 1/16$, we have for any $Z \in \{Z_+, Z_-\}$, we have

$$\mathbb{E}\big[\exp\big(\frac{s}{2}\big(Z^2 - \mathbb{E}[Z^2]\big)\big)\big] \le e^{128s^2}\,,$$

It yields that

$$(X_i^\top x)(X_i^\top y) - \mathbb{E}\big[(X_i^\top x)(X_i^\top y)\big] \sim \mathsf{subE}(16)\,.$$

Applying now Bernstein's inequality (Theorem 1.13), we get

$$\mathbb{P}\big(x^\top(\hat\Sigma - I_d)y > t/2\big) \le \exp\Big[-\frac{n}{2}(\Big(\frac{t}{32}\Big)^2 \wedge \frac{t}{32})\Big]\,.$$

Together with (4.6), this yields

$$\mathbb{P}\big(\|\hat\Sigma - I_d\|_{\mathrm{op}} > t\big) \le 144^d \exp\Big[-\frac{n}{2}(\Big(\frac{t}{32}\Big)^2 \wedge \frac{t}{32})\Big]\,. \tag{4.7}$$

In particular, the right hand side of the above inequality is at most $\delta \in (0,1)$ if

$$\frac{t}{32} \ge \Big(\frac{2d}{n}\log(144) + \frac{2}{n}\log(1/\delta)\Big) \vee \Big(\frac{2d}{n}\log(144) + \frac{2}{n}\log(1/\delta)\Big)^{1/2}$$

This concludes our proof. □

Theorem 4.6 indicates that for fixed d, the empirical covariance matrix is a consistent estimator of Σ (in any norm as they are all equivalent in finite dimension). However, the bound that we got is not satisfactory in high-dimensions when $d \gg n$. To overcome this limitation, we can introduce sparsity as we have done in the case of regression. The most obvious way to do so is to assume that few of the entries of Σ are non zero and it turns out that in this case thresholding is optimal. There is a long line of work on this subject (see for example [CZZ10] and [CZ12]).

Once we have a good estimator of Σ, what can we do with it? The key insight is that Σ contains information about the projection of the vector X onto *any* direction $u \in \mathcal{S}^{d-1}$. Indeed, we have that $\mathrm{var}(X^\top u) = u^\top \Sigma u$, which can be readily estimated by $\widehat{\mathrm{Var}}(X^\top u) = u^\top \hat\Sigma u$. Observe that it follows from Theorem 4.6

$$\begin{aligned}\big|\widehat{\mathrm{Var}}(X^\top u) - \mathrm{Var}(X^\top u)\big| &= \big|u^\top(\hat\Sigma - \Sigma)u\big|\\ &\le \|\hat\Sigma - \Sigma\|_{\mathrm{op}}\\ &\lesssim \|\Sigma\|_{\mathrm{op}}\Big(\sqrt{\frac{d+\log(1/\delta)}{n}} \vee \frac{d+\log(1/\delta)}{n}\Big)\end{aligned}$$

with probability $1-\delta$.

The above fact is useful in the Markowitz theory of portfolio section for example [Mar52], where a portfolio of assets is a vector $u \in \mathbb{R}^d$ such that $|u|_1 = 1$ and the risk of a portfolio is given by the variance $\mathrm{Var}(X^\top u)$. The goal is then to maximize reward subject to risk constraints. In most instances, the empirical covariance matrix is plugged into the formula in place of Σ.

4.4 PRINCIPAL COMPONENT ANALYSIS

Spiked covariance model

Estimating the variance in all directions is also useful for *dimension reduction*. In *Principal Component Analysis (PCA)*, the goal is to find one (or more) directions onto which the data $X_1, \dots, X_n$ can be projected without loosing much of its properties. There are several goals for doing this but perhaps the most prominent ones are data visualization (in few dimensions, one can plot and visualize the cloud of n points) and clustering (clustering is a hard computational problem and it is therefore preferable to carry it out in lower dimensions). An example of the output of a principal component analysis is given in Figure 4.1. In this figure, the data has been projected onto two orthogonal directions PC1 and PC2, that were estimated to have the most variance (among all such orthogonal pairs). The idea is that when projected onto such directions, points will remain far apart and a clustering pattern will still emerge. This is the case in Figure 4.1 where the original data is given by $d = 500,000$ gene expression levels measured on $n \simeq 1,387$ people. Depicted are the projections of these $1,387$ points in two dimension. This image has become quite popular as it shows that gene expression levels can recover the structure induced by geographic clustering. How is it possible to "compress" half a million dimensions into only two? The answer is that the data is intrinsically low dimensional. In this case, a plausible assumption is that all the $1,387$ points live close to a two-dimensional linear subspace. To see how this assumption (in one dimension instead of two for simplicity) translates into the structure of the covariance matrix Σ, assume that $X_1, \dots, X_n$ are Gaussian random variables generated as follows. Fix a direction $v \in \mathcal{S}^{d-1}$ and let $Y_1, \dots, Y_n \sim \mathcal{N}_d(0, I_d)$ so that $v^\top Y_i$ are i.i.d. $\mathcal{N}(0,1)$. In particular, the vectors $(v^\top Y_1)v, \dots, (v^\top Y_n)v$ live in the one-dimensional space spanned by v. If one would observe such data the problem would be easy as only two observations would suffice to recover v. Instead, we observe $X_1, \dots, X_n \in \mathbb{R}^d$ where $X_i = (v^\top Y_i)v + Z_i$, and $Z_i \sim \mathcal{N}_d(0, \sigma^2 I_d)$ are i.i.d. and independent of the Y_is, that is we add a isotropic noise to every point. If the σ is small enough, we can hope to recover the direction v (See Figure 4.2). The covariance matrix of X_i generated as such is given by

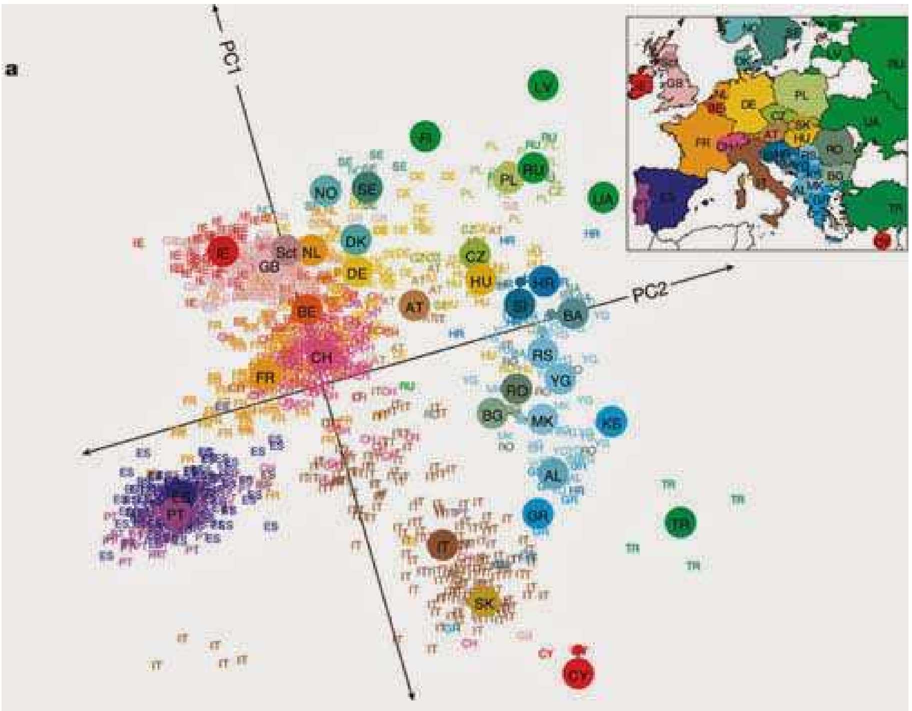
$$\Sigma = \mathbb{E}\big[XX^\top\big] = \mathbb{E}\big[((v^\top Y)v + Z)((v^\top Y)v + Z)^\top\big] = vv^\top + \sigma^2 I_d\,.$$

This model is often called the *spiked covariance model*. By a simple rescaling, it is equivalent to the following definition.

Definition 4.7. A covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is said to satisfy the spiked covariance model if it is of the form

$$\Sigma = \theta vv^\top + I_d\,,$$

where $\theta > 0$ and $v \in \mathcal{S}^{d-1}$. The vector v is called the *spike*.



Courtesy of Macmillan Publishers Ltd. Used with permission.

Figure 4.1. Projection onto two dimensions of $1,387$ points from gene expression data. Source: Gene expression blog.

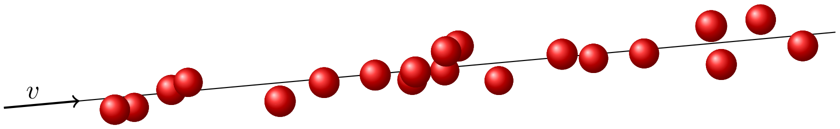


Figure 4.2. Points are close to a one dimensional space space by v.

This model can be extended to more than one spike but this extension is beyond the scope of these notes.

Clearly, under the spiked covariance model, v is the eigenvector of the matrix Σ that is associated to its largest eigenvalue $1 + \theta$. We will refer to this vector simply as *largest eigenvector*. To estimate it, a natural candidate is the largest eigenvector $\hat{v}$ of $\tilde{\Sigma}$, where $\tilde{\Sigma}$ is any estimator of Σ. There is a caveat: by symmetry, if u is an eigenvector, of a symmetric matrix, then $-u$ is also an eigenvector associated to the same eigenvalue. Therefore, we may only estimate v up to a sign flip. To overcome this limitation, it is often useful to describe proximity between two vectors u and v in terms of the principal angle between their linear span. Let us recall that for two unit vectors the principal angle between their linear spans is denoted by $\angle(u, v)$ and defined as

$$\angle(u, v) = \arccos(|u^\top v|)\,.$$

The following result form perturbation theory is known as the Davis-Kahan $\sin(\theta)$ theorem as it bounds the sin of the principal angle between eigenspaces. This theorem exists in much more general versions that extend beyond one-dimensional eigenspaces.

Theorem 4.8 (Davis-Kahan $\sin(\theta)$ theorem)**.** *Let Σ satisfy the spiked covariance model and let $\tilde\Sigma$ be any PSD estimator of Σ. Let $\tilde v$ denote the largest eigenvector of $\tilde\Sigma$. Then we have*

$$\min_{\varepsilon\in\{\pm1\}} |\varepsilon\tilde v - v|_2^2 \le 2\sin^2\big(\angle(\tilde v, v)\big) \le \frac{8}{\theta^2}\|\tilde\Sigma - \Sigma\|_{\mathrm{op}}^2\,.$$

Proof. Note that for any $u \in \mathcal{S}^{d-1}$, it holds under the spiked covariance model that

$$u^\top \Sigma u = 1 + \theta(v^\top u)^2 = 1 + \theta\cos^2(\angle(u, v))\,.$$

Therefore,

$$v^\top\Sigma v - \tilde v^\top \Sigma \tilde v = \theta[1 - \cos^2(\angle(\tilde v, v))] = \theta\sin^2(\angle(\tilde v, v))\,.$$

Next, observe that

$$\begin{aligned}
v^\top\Sigma v - \tilde v^\top\Sigma\tilde v &= v^\top\tilde\Sigma v - \tilde v^\top\Sigma\tilde v - v^\top\big(\tilde\Sigma - \Sigma\big)v \\
&\le \tilde v^\top\tilde\Sigma\tilde v - \tilde v^\top\Sigma\tilde v - v^\top\big(\tilde\Sigma - \Sigma\big)v \\
&= \langle \hat\Sigma - \Sigma, \tilde v\tilde v^\top - vv^\top\rangle && (4.8)\\
&\le \|\tilde\Sigma - \Sigma\|_{\mathrm{op}}\|\tilde v\tilde v^\top - vv^\top\|_1 && \text{(Hölder)}\\
&\le \sqrt{2}\|\tilde\Sigma - \Sigma\|_{\mathrm{op}}\|\tilde v\tilde v^\top - vv^\top\|_F && \text{(Cauchy-Schwarz)}\,.
\end{aligned}$$

where in the first inequality, we used the fact that $\tilde v$ is the largest eigenvector of $\tilde\Sigma$ and in the last one, we used the fact that the matrix $\tilde v\tilde v^\top - vv^\top$ has rank at most 2.

Next, we have that

$$\|\tilde v\tilde v^\top - vv^\top\|_F^2 = 2(1 - (v^\top\tilde v)^2) = 2\sin^2(\angle(\tilde v, v))\,.$$

Therefore, we have proved that

$$\theta\sin^2(\angle(\tilde v, v)) \le 2\|\tilde\Sigma - \Sigma\|_{\mathrm{op}}\sin(\angle(\tilde v, v))\,,$$

so that

$$\sin(\angle(\tilde v, v)) \le \frac{2}{\theta}\|\tilde\Sigma - \Sigma\|_{\mathrm{op}}\,.$$

To conclude the proof, it remains to check that

$$\min_{\varepsilon\in\{\pm1\}} |\varepsilon\tilde v - v|_2^2 = 2 - 2|\tilde v^\top v| \le 2 - 2(\tilde v^\top v)^2 = 2\sin^2(\angle(\tilde v, v))\,.$$

□

Combined with Theorem 4.6, we immediately get the following corollary.

Corollary 4.9. *Let $X_1, \dots, X_n$ be n i.i.d. copies of a sub-Gaussian random vector $X \in \mathbb{R}^d$ such that $\mathbb{E}\big[XX^\top\big] = \Sigma$ and $X \sim \mathsf{subG}_d(\|\Sigma\|_{\mathrm{op}})$. Assume further that $\Sigma = \theta vv^\top + I_d$ satisfies the spiked covariance model. Then, the largest eigenvector $\hat v$ of the empirical covariance matrix $\hat\Sigma$ satisfies,*

$$\min_{\varepsilon \in \{\pm 1\}} |\varepsilon \hat v - v|_2 \lesssim \frac{1+\theta}{\theta}\Big(\sqrt{\frac{d + \log(1/\delta)}{n}} \vee \frac{d + \log(1/\delta)}{n}\Big)$$

with probability $1-\delta$.

This result justifies the use of the empirical covariance matrix $\hat\Sigma$ as a replacement for the true covariance matrix Σ when performing PCA in low dimensions, that is when $d \ll n$. In the high-dimensional case, where $d \gg n$, the above result is uninformative. As before, we resort to sparsity to overcome this limitation.

Sparse PCA

In the example of Figure 4.1, it may be desirable to interpret the meaning of the two directions denoted by $\mathsf{PC1}$ and $\mathsf{PC2}$. We know that they are linear combinations of the original 500,000 gene expression levels. A natural question to ask is whether only a subset of these genes could suffice to obtain similar results. Such a discovery could have potential interesting scientific applications as it would point to a few genes responsible for disparities between European populations.

In the case of the spiked covariance model this amounts to have v to be sparse. Beyond interpretability as we just discussed, sparsity should also lead to statistical stability as in the case of sparse linear regression for example. To enforce sparsity, we will assume that v in the spiked covariance model is k-sparse: $|v|_0 = k$. Therefore, a natural candidate to estimate v is given by $\hat v$ defined by

$$\hat v^\top \hat\Sigma \hat v = \max_{\substack{u \in \mathcal{S}^{d-1} \\ |u|_0 = k}} u^\top \hat\Sigma u\,.$$

It is easy to check that $\lambda^k_{\max}(\hat\Sigma) = \hat v^\top \hat\Sigma \hat v$ is the largest of all leading eigenvalues among all $k \times k$ sub-matrices of $\hat\Sigma$ so that the maximum is indeed attained, though there my be several maximizers. We call $\lambda^k_{\max}(\hat\Sigma)$ the k-sparse leading eigenvalue of $\hat\Sigma$ and $\hat v$ a k-sparse leading eigenvector.

Theorem 4.10. *Let $X_1, \dots, X_n$ be n i.i.d. copies of a sub-Gaussian random vector $X \in \mathbb{R}^d$ such that $\mathbb{E}\big[XX^\top\big] = \Sigma$ and $X \sim \mathsf{subG}_d(\|\Sigma\|_{\mathrm{op}})$. Assume further that $\Sigma = \theta vv^\top + I_d$ satisfies the spiked covariance model for v such that $|v|_0 = k \le d/2$. Then, the k-sparse largest eigenvector $\hat v$ of the empirical covariance matrix satisfies,*

$$\min_{\varepsilon \in \{\pm 1\}} |\varepsilon \hat v - v|_2 \lesssim \frac{1+\theta}{\theta}\Big(\sqrt{\frac{k\log(ed/k) + \log(1/\delta)}{n}} \vee \frac{k\log(ed/k) + \log(1/\delta)}{n}\Big)\,.$$

with probability $1-\delta$.

Proof. We begin by obtaining an intermediate result of the Davis-Kahan $\sin(\theta)$ theorem (Theorem 4.8). Note that we get from (4.8) that

$$v^\top \Sigma v - \hat v^\top \Sigma \hat v \le \langle \hat\Sigma - \Sigma, \hat v \hat v^\top - vv^\top\rangle$$

Since both $\hat v$ and v are k sparse, there exists a (random) set $S \subset \{1, \dots, d\}$ such that $|S| \le 2k$ and $\{\hat v \hat v^\top - vv^\top\}_{ij} = 0$ if $(i,j) \notin S^2$. It yields

$$\langle \hat\Sigma - \Sigma, \hat v \hat v^\top - vv^\top\rangle = \langle \hat\Sigma(S) - \Sigma(S), \hat v(S)\hat v(S)^\top - v(S)v(S)^\top\rangle$$

Where for any $d \times d$ matrix M, we defined the matrix $M(S)$ to be the $|S| \times |S|$ sub-matrix of M with rows and columns indexed by S and for any vector $x \in \mathbb{R}^d$, $x(S) \in \mathbb{R}^{|S|}$ denotes the sub-vector of x with coordinates indexed by S. It yields, by Hölder's inequality that

$$v^\top \Sigma v - \hat v^\top \Sigma \hat v \le \|\hat\Sigma(S) - \Sigma(S)\|_{\rm op} \|\hat v(S)\hat v(S)^\top - v(S)v(S)^\top\|_1 \,.$$

Following the same steps as in the proof of Theorem 4.8, we get now that

$$\min_{\varepsilon \in \{\pm 1\}} |\varepsilon \hat v - v|_2^2 \le 2\sin^2\big(\angle(\hat v, v)\big) \le \frac{8}{\theta^2} \sup_{S\,:\,|S|=2k} \|\hat\Sigma(S) - \Sigma(S)\|_{\rm op} \,.$$

To conclude the proof, it remains to control $\sup_{S\,:\,|S|=2k} \|\hat\Sigma(S) - \Sigma(S)\|_{\rm op}$. To that end, observe that

$$\begin{aligned}
\mathbb{P}\Big[\sup_{S\,:\,|S|=2k} &\|\hat\Sigma(S) - \Sigma(S)\|_{\rm op} > t\|\Sigma\|_{\rm op}\Big] \\
&\le \sum_{S\,:\,|S|=2k} \mathbb{P}\Big[\sup_{S\,:\,|S|=2k} \|\hat\Sigma(S) - \Sigma(S)\|_{\rm op} > t\|\Sigma(S)\|_{\rm op}\Big] \\
&\le \binom{d}{2k} 144^{2k} \exp\Big[-\frac{n}{2}\big(\Big(\frac{t}{32}\Big)^2 \wedge \frac{t}{32}\big)\Big] \,.
\end{aligned}$$

where we used (4.7) in the second inequality. Using Lemma 2.7, we get that the right-hand side above is further bounded by

$$\exp\Big[-\frac{n}{2}\big(\Big(\frac{t}{32}\Big)^2 \wedge \frac{t}{32}\big) + 2k\log(144) + k\log\big(\frac{ed}{2k}\big)\Big]$$

Choosing now t such that

$$t \ge C\sqrt{\frac{k\log(ed/k) + \log(1/\delta)}{n}} \vee \frac{k\log(ed/k) + \log(1/\delta)}{n} \,,$$

for large enough C ensures that the desired bound holds with probability at least $1-\delta$. □

4.5 PROBLEM SET

Problem 4.1. Using the results of Chapter 2, show that the following holds for the multivariate regression model (4.1).

1. There exists an estimator $\hat{\Theta} \in \mathbb{R}^{d\times T}$ such that
$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \sigma^2 \frac{rT}{n}$$
with probability .99, where r denotes the rank of $\mathbb{X}$.
2. There exists an estimator $\hat{\Theta} \in \mathbb{R}^{d\times T}$ such that
$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \sigma^2 \frac{|\Theta^*|_0 \log(ed)}{n}\,.$$
with probability .99.

Problem 4.2. Consider the multivariate regression model (4.1) where $\mathbb{Y}$ has SVD:
$$\mathbb{Y} = \sum_j \hat{\lambda}_j \hat{u}_j \hat{v}_j^\top\,.$$
Let M be defined by
$$\hat{M} = \sum_j \hat{\lambda}_j \mathbb{1}(|\hat{\lambda}_j| > 2\tau)\hat{u}_j \hat{v}_j^\top\,, \tau > 0\,.$$

1. Show that there exists a choice of τ such that
$$\frac{1}{n}\|\hat{M} - \mathbb{X}\Theta^*\|_F^2 \lesssim \frac{\sigma^2 \operatorname{rank}(\Theta^*)}{n}(d \vee T)$$
with probability .99.
2. Show that there exists a matrix $n \times n$ matrix P such that $P\hat{M} = \mathbb{X}\hat{\Theta}$ for some estimator $\hat{\Theta}$ and
$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \frac{\sigma^2 \operatorname{rank}(\Theta^*)}{n}(d \vee T)$$
with probability .99.
3. Comment on the above results in light of the results obtain in Section 4.2.

Problem 4.3. Consider the multivariate regression model (4.1) and define $\hat{\Theta}$ be the any solution to the minimization problem
$$\min_{\Theta\in\mathbb{R}^{d\times T}} \Big\{\frac{1}{n}\|\mathbb{Y} - \mathbb{X}\Theta\|_F^2 + \tau\|\mathbb{X}\Theta\|_1\Big\}$$

1. Show that there exists a choice of τ such that

$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \frac{\sigma^2 \operatorname{rank}(\Theta^*)}{n}(d \vee T)$$

with probability .99.

`[Hint:Consider the matrix`

$$\sum_j \frac{\hat{\lambda}_j + \lambda_j^*}{2} \hat{u}_j \hat{v}_j^\top$$

`where` $\lambda_1^* \geq \lambda_2^* \geq \dots$ `and` $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots$ `are the singular values of` $\mathbb{X}\Theta^*$ `and` $\mathbb{Y}$ `respectively and the SVD of` $\mathbb{Y}$ `is given by`

$$\mathbb{Y} = \sum_j \hat{\lambda}_j \hat{u}_j \hat{v}_j^\top$$

2. Find a closed form for $\mathbb{X}\hat{\Theta}$.

18.S997 High-dimensional Statistics
Spring 2015