EPFL, Autumn 2024
Yoav Zemel
Guillaume Obozinski

Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 10

We will use several times in this exercise sheet *Von Neumann's inequality*. We will prove it in a last exercise. *Von Neuman's inequality* says that if we let $A, B \in \mathbb{R}^{d\times K}$, with $K \leq d$, two matrices whose singular values are respectively $\sigma_1(A) \geq \ldots \geq \sigma_K(A)$ and $\sigma_1(B) \geq \ldots \geq \sigma_K(B)$, then

$$|\mathrm{tr}(A^\top B)| \leq \sum_{k=1}^{K} \sigma_k(A)\sigma_k(B).$$

Exercise 10.1 Proving part of Eckart-Young's theorem...

(a) Use Von Neumann's inequality to show that if A and B are as above then

$$\|A - B\|_F^2 \geq \sum_{k=1}^{K} (\sigma_k(A) - \sigma_k(B))^2.$$

If $UDV^\top$ is the SVD of A then we have

$$\|A\|_F^2 = \mathrm{tr}(UDV^\top VDU^\top) = \mathrm{tr}(U^\top UDV^\top VD) = \|D\|_F^2 = \sum_{k=1}^{K} \sigma_k(A)^2.$$

and so, using Von Neumann's inequality,

$$\|A-B\|_F^2 = \|A\|_F^2 - 2\mathrm{tr}(A^\top B) + \|B\|_F^2 \geq \sum_{k=1}^{K} \sigma_k(A)^2 - 2\sum_{k=1}^{K} \sigma_k(A)\sigma_k(B) + \sum_{k=1}^{K} \sigma_k(B)^2.$$

(b) Solve the problem

$$\min_B \sum_{k=1}^{K} (\sigma_k(A) - \sigma_k(B))^2 \quad \text{s.t.} \quad \mathrm{rank}(B) \leq r.$$

Let $\alpha_k = \sigma_k(A)$ and $\beta_k = \sigma_k(B)$. We can rewrite the previous problem as

$$\min_\beta \|\alpha - \beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_0 \leq r.$$

But since the coefficients of α are sorted in decreasing order, the smallest error is obtained for the vector β^* such that $\beta_k^* = \alpha_k 1_{\{k\leq r\}}$. So any matrix B with singular values $\alpha_1, \ldots, \alpha_r, 0, \ldots, 0$ is a minimum.

(c) For a fixed matrix A, show that there exists a matrix B of rank r such that

$$\|A - B\|_F^2 = \sum_{k=r+1}^{K} \sigma_k(A)^2$$

We can use B the matrix of projections of the rows of A on the r first principal directions of A. More precisely, if $UDV^\top$ is the SVD of A, and $U_{[r]} \in \mathbb{R}^{d\times r}$, $S_{[r]} \in \mathbb{R}^{r\times r}$, and $V_{[r]} \in \mathbb{R}^{d\times r}$) denote respectively the matrices formed of the r first column of U, S and V, then ,if we let $B = U_{[r]}S_{[r]}V_{[r]}^\top$, we have $\|A - B\|_F^2 = \|\sum_{k=r+1}^{K} s_k u_k v_k^\top\|_F^2 = \sum_{k=r+1}^{K} s_k^2$, which proves the result since $s_k = \sigma_k(A)$. Note that if several singular values are equal, then the SVD is not unique: the right and left subspaces associated with a given singular value is of dimension larger than one and therefore admits several bases. However, any of these bases would attain the proposed value. In that case, there are several matrices B that attain the considered value.

(d) If $A = USV^\top$, let $U_{[r]} \in \mathbb{R}^{d\times r}$, $S_{[r]} \in \mathbb{R}^{r\times r}$, and $V_{[r]} \in \mathbb{R}^{d\times r}$) denote respectively the matrices formed of the r first column of U, S and V. Use the previous result to show that $B^* = U_{[r]}S_{[r]}V_{[r]}^\top$ minimizes:

$$\min_B \|A - B\|_F^2 \quad \text{s.t.} \quad \text{rank}(B) \leq r,$$

By questions (a) and (b) we know that $\|A-B\|_F^2 \geq \sum_{k=r+1}^{K} \sigma_k(A)^2$ for any matrix B of rank K. By question (c), we know that this value can be attained by $B^* = U_{[r]}S_{[r]}V_{[r]}^\top$.

Exercise 10.2 Probabilistic version of PCA. Let $\mathbf{D} \in \mathbb{R}^{p\times K}$ be a fixed full column rank matrix (thus with $K \leq p$). We consider the following generative model:

$$\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_K), \quad \mathbf{x} \mid \mathbf{z} \sim \mathcal{N}(\mathbf{D}\mathbf{z}, \sigma^2\mathbf{I}_p)$$

(a) Use that with previous model we equivalently have that $\mathbf{x} = \mathbf{D}\mathbf{z} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_p)$ where ε and $\mathbf{z}$ are independent, to obtain the marginal distribution of $\mathbf{x}$; in particular compute its mean and it covariance.

Solution: $\mathbf{x}$ is a linear combination of independent Gaussian r.v.s, and so it is Gaussian as well. Its mean is

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{D}\mathbf{z} + \varepsilon] = \mathbf{D}\mathbb{E}[\mathbf{z}] + \mathbb{E}[\varepsilon] = 0.$$

And so, given that all variables are centered, its variance is

$$\text{Cov}(\mathbf{x}) = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbb{E}[\mathbf{D}\mathbf{z}\mathbf{z}^\top\mathbf{D}^\top] + \mathbb{E}[\varepsilon\varepsilon^\top] = \mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p,$$

where we used that, by independence, $\mathbb{E}[\mathbf{D}\mathbf{z}\varepsilon^\top] = \mathbb{E}[\mathbf{D}\mathbf{z}]\mathbb{E}[\varepsilon]^\top = 0$.

(b) Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is an i.i.d sample from the model above. Express its log-likelihood as a function of $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$.

$$\begin{aligned}
\ell(\mathbf{D}, \sigma^2) = \sum_{i=1}^n \log p(\mathbf{x}_i) &= -\tfrac{1}{2}\sum_{i=1}^n \mathbf{x}_i^\top[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p]^{-1}\mathbf{x}_i - \tfrac{n}{2}\log\det[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p] \\
&= -\tfrac{n}{2}\sum_{i=1}^n \text{tr}(\mathbf{x}_i\mathbf{x}_i^\top[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p]^{-1}) - \tfrac{n}{2}\log\det[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p] \\
&= -\tfrac{n}{2}\text{tr}(\widehat{\Sigma}[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p]^{-1}) - \tfrac{n}{2}\log\det[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p].
\end{aligned}$$

(c) Verify that $[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p]^{-1} = \sigma^{-2}\mathbf{I}_p - \sigma^{-2}\mathbf{D}[\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top$.

We can check that

$$\sigma^{-2}\Big(\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p\Big)\Big(\mathbf{I}_p - \mathbf{D}[\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top\Big) = \mathbf{I}_p.$$

Indeed, the LHS expression above multiplied by σ^2 is equal to

$$\begin{aligned}&\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p - \mathbf{D}\mathbf{D}^\top\mathbf{D}[\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top - \mathbf{D}\sigma^2\mathbf{I}_p[\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top\\ =&\sigma^2\mathbf{I}_p + \mathbf{D}\Big([\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}][\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1} - \mathbf{D}^\top\mathbf{D}[\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1} - \sigma^2\mathbf{I}_p[\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1}\Big)\mathbf{D}^\top\end{aligned}$$

and we see that what is inside of the parentheses is equal to zero.

(d) Show that when $\sigma^2 \to 0$, then $\sigma^2\ell(\mathbf{D}, \sigma^2)$ converges to $-\frac{n}{2}\mathrm{tr}(\widehat{\Sigma}(\mathbf{I} - \mathbf{H}))$ with $\mathbf{H} = \mathbf{D}[\mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top$.

We can write $\sigma^2\ell(\mathbf{D}, \sigma^2) = -\frac{n}{2}[A_1(\sigma^2) + A_2(\sigma^2)]$ with

$$A_1(\sigma^2) := \sigma^2\mathrm{tr}(\widehat{\Sigma}[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p]^{-1}) \quad \text{and} \quad A_2(\sigma^2) := \sigma^2 \log\det[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p].$$

First note that using question (c) we can rewrite $A_1(\sigma^2)$ as

$$A_1(\sigma^2) = \mathrm{tr}(\widehat{\Sigma}) - \mathrm{tr}(\mathbf{D}[\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top\widehat{\Sigma}),$$

so that when $\sigma^2 \to 0$ we have $A_1(\sigma^2) \to \mathrm{tr}(\widehat{\Sigma}) - \mathrm{tr}(\mathbf{H}\widehat{\Sigma})$.

Then note that $A_2(\sigma^2) = \sigma^2\sum_{j=1}^{K}\log(s_j + \sigma^2) + \sigma^2(p - K)\log\sigma^2$, where $s_1 \geq \ldots \geq s_K$ are the eigenvalues of $\mathbf{D}\mathbf{D}^\top$, so that when $\sigma^2 \to 0$, then $A_1(\sigma^2) \to 0$ because $\sigma^2\log\sigma^2 \to 0$.

(e) Using Von Neumann's inequality, prove that the projector on the subspace spanned by the K top eigenvectors of $\widehat{\Sigma}$ maximizes $\mathrm{tr}(\widehat{\Sigma}\mathbf{H})$.

Indeed by VN's inequality $\mathrm{tr}(\widehat{\Sigma}\mathbf{H}) \leq \sum_{k=1}^{K} s_k$ and we obtain an equality for $\mathbf{H} = U_{[k]}U_{[k]}^\top$ where $U_{[k]} \in \mathbb{R}^{d\times K}$ is the matrix whose columns are the k top eigenvectors of $\widehat{\Sigma}$. We have not proven here that $U_{[k]}U_{[k]}^\top$ is the unique maximizer, but, if the eigenvalues of $\widehat{\Sigma}$ are distinct, a closer look at equality cases in VN's inequality would allow us to prove that it is.

(f) Explain why when σ^2 is small, the maximum likelihood estimator for D can be expected to be a matrix whose columns span the kth right principal subspace of $\mathbf{X}$ and whose singular values are the top singular values of $\mathbf{X}$ where $\mathbf{X}$ is the design matrix of the data.

When σ^2 is small we proved in the previous questions that the log-likelihood is dominated by the term $\frac{1}{\sigma^2}\frac{n}{2}\mathrm{tr}(\widehat{\Sigma}(\mathbf{H} - \mathbf{I}))$ which entails that, when the eigenvalues of $\widehat{\Sigma}$ are distinct, the matrix $\mathbf{H}$ maximizing the likelihood should be close to $U_{[k]}U_{[k]}^\top$.

Of course in practice we maximize with respect to D, but the space spanned by the columns of D is the same as the space spanned by the columns of $\mathbf{H}$. By analyzing more closely the maximum likelihood estimator for D we could show that $D^\top D$ estimates the covariance structure of the data in the basis of the subspace given by the columns of $U_{[k]}$.

(g) In which sense is the probabilistic model introduced at the beginning of this exercise a probabilistic counterpart of PCA?

Intuitively first, the model is constructed as a generative model where a datapoint is generated in the subspace spanned by the columns of D with a covariance structure given by $DD^\top$ and the observed data is then obtained by adding a noise of variance σ^2. It therefore makes sense that this model would estimate a principal subspace of the data. Then, we gave some argument towards proving that the maximum likelihood estimator would estimate approximately the same subspace as PCA when σ^2 is small.

Practical Exercise

Exercise 10.3 (PCA and Dimensionality Reduction) Import the file `data.csv` using the `read.csv` function in `R`. It contains a list of 10 dimensional vectors with their class.

(a) Using the `svd` function, compute the principal components of the given data set (exclude the class).

(b) How many principal components do you need to explain more than 99% of the variance?

Solution: The singular values s_j drop suddenly after $j = 2$ and thus the first two principal components capture most of the variance. Although the data is of dimension 10, it appears that it is essentially of dimension 2.

(c) Plot the first two principal components and describe the shape of the data. Would it be a good idea to use linear classifiers to classify this data set? If not, why not?

Solution: Thus projecting high-dimensional data to a low-dimensional subspace can reveal important information. Linear classifiers can not be used because the resulting data does not appear linearly separable.

(d) Construct a function $f : \mathbb{R}^2 \to \mathbb{R}^3$ to transform the data by mapping the first two principal components so as to render the data easily classifiable using linear classifiers. *Note: You may choose the function by inspection or trial and error.*

Solution: Pick $f(x, y) = \exp(ax^2 + by^2)$ and fit $a, b > 0$ by inspection.

Bonus Exercise

Exercise 10.4 (von Neumann's Inequality). The goal of this problem is to establish *Von Neumann's inequality*. Let $A, B \in \mathbb{R}^{d\times d}$ two matrices whose singular values are respectively $\sigma_1(A) \geq \ldots \geq \sigma_d(A)$ and $\sigma_1(B) \geq \ldots \geq \sigma_d(B)$. *Von Neuman's inequality* says that

$$|\mathrm{tr}(A^\top B)| \leq \sum_{k=1}^{d} \sigma_k(A)\sigma_k(B).$$

(a) Why can we assume without loss of generality that A is a diagonal matrix?[*Hint: inject the SVD of A.*]

Let $U_A D V_A^\top$ be the singular value of A. Then we consider $\tilde{A} = D$ and $\tilde{B} = U_A^\top B V_A$. Then $\mathrm{tr}(A^\top B) = \mathrm{tr}(V_A D U_A^\top B) = \mathrm{tr}(D U_A^\top B V_A) = \mathrm{tr}(\tilde{A}^\top \tilde{B})$ and we have $\sigma_k(A) = \sigma_k(D) = \sigma_k(\tilde{A})$ and $\sigma_k(B) = \sigma_k(\tilde{B})$ because if $USV^\top$ is the SVD of B then $U_A^\top U$ and $V_A^\top V$ are still orthogonal matrices. So if we show the result for the pair $(\tilde{A}, \tilde{B})$, the result will be true for the pair (A, B).

(b) If $A = D$ is diagonal, prove that Von Neumann's inequality is equivalent to the inequality $|\mathrm{tr}(DUSV^\top)| \leq \mathrm{tr}(DS)$, where $USV^\top$ is the SVD of B.

This is just because D and S contain precisely the singular values of A and B.

(c) Let $P_k = \mathrm{Diag}(\underbrace{1,\ldots,1}_{k \text{ ones}}, \underbrace{0,\ldots,0}_{d-k \text{ zeros}})$. Let $\sigma_{d+1}(A) = \sigma_{d+1}(B) = 0$ by convention, and let $a_k = \sigma_k(A) - \sigma_{k+1}(A)$ and $b_k = \sigma_k(B) - \sigma_{k+1}(B)$, so that we have

$$D = \sum_{k=1}^{d} a_k P_k \quad \text{and} \quad S = \sum_{l=1}^{d} b_l P_l.$$

Show that Von Neumann's inequality can equivalently written as

$$|\sum_{k=1}^{d}\sum_{l=1}^{d} a_k\, b_l\, \mathrm{tr}(P_k U P_l V^\top)| \leq \sum_{l=1}^{d} a_k\, b_l\, \mathrm{tr}(P_k P_l).$$

This is an immediate consequence of the linearity of the trace.

(d) Deduce from the previous question that it is sufficient to prove

$$|\,\mathrm{tr}(P_k U P_l V^\top)| \leq \mathrm{tr}(P_k P_l),$$

which is actually exactly Von Neumann's inequality but for a particular kind of matrix.

By construction $a_k \geq 0$ and $b_l \geq 0$ so that

$$|\sum_{k=1}^{d}\sum_{l=1}^{d} a_k\, b_l\, \mathrm{tr}(P_k U P_l V^\top)| \leq \sum_{k=1}^{d}\sum_{l=1}^{d} a_k\, b_l\, |\mathrm{tr}(P_k U P_l V^\top)|,$$

so indeed the previous inequality would allow to conclude.

(e) Let $\mathbf{u}_k$ denote the kth column of U and $\mathbf{v}_l$ denote the lth column of V. Show that $\operatorname{tr}(P_k U P_l V^\top) = \sum_{i=1}^{l} \langle P_k \mathbf{u}_i, \mathbf{v}_i \rangle \leq l$ and deduce that in fact $\operatorname{tr}(P_k U P_l V^\top) \leq \min(k, l)$.
Let $U_{[l]}$ (resp. $V_{[l]}$) be the matrix formed of the l first columns of U (resp. of V), then $UP_lV^\top = UP_lP_lV^\top = U_{[l]}V_{[l]}^\top$. So that

$$\operatorname{tr}(P_k U P_l V^\top) = \operatorname{tr}(P_k U_{[l]} V_{[l]}^\top) = \operatorname{tr}(V_{[l]}^\top P_k U_{[l]}) = \sum_{i=1}^{l} \langle \mathbf{v}_i, P_k \mathbf{u}_i \rangle.$$

Permuting under the trace we have $\operatorname{tr}(P_k U P_l V^\top) = \operatorname{tr}(P_l V^\top P_k U)$ and so by exchanging the roles of U and $V^\top$ we get the symmetric inequality (with this time the rows of U and V).

(f) Use this last result to prove Von Neumann's inequality.

Finally we proved $\operatorname{tr}(P_k U P_l V^\top) = \min(k, l)$ but clearly $\operatorname{tr}(P_k P_l) = \min(k, l)$, so we proved the inequality $|\operatorname{tr}(P_k U P_l V^\top)| \leq \operatorname{tr}(P_k P_l)$, and therefore the original inequality.

(g) Assume now that $A, B \in \mathbb{R}^{d \times K}$ with $K < d$, why is the inequality still true?

We can always add zero columns to A and B to turn them into square matrices. This adds only singular values equal to zero to both matrices.