Problems

December 11, 2024

Problems marked with $^{\$}$ may be somewhat challenging—thus their solutions (or original solutions) may (perhaps) earn some extra credit(s). Problems marked with * deserve special attention; their solutions are substantially complementing the main exposition of the methodology. Even if a student does not solve such a problem independently, it is recommended that they familiarize at least with the facts these problems convey, if not with their solutions. Especially in case when the problem is marked also by $^{\$}$, its solution is usually not required for the understanding of the main exposition—but its statement ususally is.

Problems designated as "R problems" are (typically) not solved in the traditional mathematical way, but via reading the documentation, experimenting with the software, etc., until arriving to a guess for the solution. This guess should be then validated in some way in the software—mere speculations do not constitute a valid solution. Once the problem is solved, make some record of your session: if it does not pose difficulties for you, print a transcript of it (transcript is always required if you are submitting solutions over email in lieu of a normal parrticipation in class), otherwise at least write down some results (although the demonstration in class may not necessarily reproduce all of that).

Problems set in blue have been already sufficiently discussed in class, and their solution no longer earns a credit.

1*. Matrices generating quadratic forms can be considered symmetric without loss of generality. Give a formal justification of this claim.

2*. Suppose that (not necessarily square) matrices $\mathbf{A}$ and $\mathbf{B}$ are such that both $\mathbf{AB}$ and $\mathbf{BA}$ are square matrices. Show that matrices $\mathbf{AB}$ and $\mathbf{BA}$ have the same *nonzero* eigenvalues.

3*. Let $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$, where $\mathbf{Q}$ is an orthogonal and $\boldsymbol{\Lambda}$ a diagonal matrix. Show that the diagonal of $\boldsymbol{\Lambda}$ consists of eigenvalues and the columns of $\mathbf{Q}$ are the corresponding eigenvectors.

4*. Let $\mathbf{A} = \mathbf{ULV}^\top$ be a singular decomposition of matrix $\mathbf{A}$. Figure out from that the eigenvalue decomposition of $\mathbf{A}^\top\mathbf{A}$.

5. Give the form of a linear transformation in $\mathbb{R}^2$ that is the rotation counterclockwise by an angle φ. Is that an orthogonal transformation? (Orthogonal transformation = linear transformation whose matrix is orthogonal). In the applications, software, etc., orthogonal transformations are often mentioned as "rotations". Comment on this terminology, for simplicity considering only transformations in $\mathbb{R}^2$.

6*. Let $\mathbf{A}$ be a symmetric $\mathrm{p} \times \mathrm{p}$ nonnegative definite matrix. Find a $\mathrm{p} \times \mathrm{m}$ matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\top$ has the minimal Euclidean distance from $\mathbf{A}$.

7*$. Give a proof of the theorem of Eckart and Young, as formulated in the notes. (The crux of this problem is in showing that the desired approximation of a *diagonal* matrix is diagonal itself: once this is established, the rest of the proof is straightforward, as indicated in the notes.)

8*. Suppose that $\mathbf{A}$ is a symmetric and nonnegative definite matrix, $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ where $\mathbf{Q}$ is an orthogonal and $\mathbf{\Lambda}$ a diagonal matrix. Find $\mathbf{x}$ for which $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ is maximal, under the condition that $\|\mathbf{x}\| = 1$. Is such an $\mathbf{x}$ unique?

9*. Suppose that $\mathbf{A}$ is symmetric and nonnegative definite and $\mathbf{B}$ is symmetric and positive definite. The maximum of

$$\frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{B}\mathbf{x}}$$

for $\mathbf{x} \neq 0$ is the largest eigenvalue of $\mathbf{B}^{-1}\mathbf{A}$ and is attained for the corresponding eigenvector x.

10*. Suppose that $\mathbf{A}$ is symmetric and nonnegative definite and $\mathbf{B}$ and $\mathbf{C}$ are symmetric and positive definite. The maximum of

$$\frac{(\mathbf{x}^\top\mathbf{A}\mathbf{y})^2}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})(\mathbf{y}^\top\mathbf{C}\mathbf{y})}$$

for $\mathbf{x} \neq 0$ and $\mathbf{y} \neq 0$ is the largest eigenvalue of both $\mathbf{B}^{-1}\mathbf{A}\mathbf{C}^{-1}\mathbf{A}^\top$ and $\mathbf{C}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top$, and is attained for the corresponding eigenvectors $\mathbf{x}$ and $\mathbf{y}$, respectively.

11*. Prove that $\dfrac{\partial \mathbf{a}^\top\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$, $\dfrac{\partial \mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$, $\dfrac{\partial \mathbf{a}^\top\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^\top$.

12*$. Prove that $\dfrac{\partial \log\det(\mathbf{X})}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^\top$.

13*. For the stochastic version of the variance-covariance matrix, we have the transformation formula, for any nonrandom $\mathbf{A}$

$$\mathrm{Var}(\mathbf{A}\mathbf{y}) = \mathbf{A}\,\mathrm{Var}(\mathbf{y})\mathbf{A}^\top.$$

Construct the transformation theory, including formulation and proof of the analog of the above formula, for the sample variance-covariance matrices, working with the (non-random) data matrix $\mathbf{Y}$ instead of the random vector $\mathbf{y}$, and considering *the same* (non-random) $\mathbf{A}$.

14. Describe what conclusion for the data can you derive from the fact that their sample variance-covariance matrix is singular.

15*. Show that the (sample) variance-covariance matrix computed from the scaled data is the (sample) correlation matrix.

16. Let (X, Y) be a random vector. Assuming that all required moments exist, find (non-random) $\mathfrak{a}$ and $\mathfrak{b}$ such that $\mathsf{E}(Y - \mathfrak{a} - \mathfrak{b}X)^2$ is minimal.

17*. Consider the extension of the previous problem: let $\mathbf{f}$ and $\mathbf{z}$ be random vectors with respectively $\mathfrak{m}$ and $\mathfrak{p}$ components, such that both $\mathsf{E}(\mathbf{f}) = 0$ and $\mathsf{E}(\mathbf{z}) = 0$. Show that an $\mathfrak{m} \times \mathfrak{p}$ matrix $\mathbf{U}$ minimizing $\mathsf{E}\,\|\mathbf{f} - \mathbf{U}\mathbf{z}\|^2$ has the form $\mathbf{U} = \mathsf{Cov}(\mathbf{f}, \mathbf{z})[\mathsf{Var}(\mathbf{z})]^{-1}$. (It is assumed that all moments exist.)

18. (An R problem) Functions `prcomp()` and `princomp()` both compute principal components; if they compute them directly from the data matrix (not from the variance-covariance or correlation matrix), they give slightly different results. Figure out why—and indicate how they can be reconciled.

19. (Another R problem) Wanting to demonstrate how `predict.prcomp()` method for `prcomp()`, I executed the following code (results are abbreviated)

```
> trackmen.pcs <- prcomp(trackmen,scale=T)
> t(as.matrix(trackmen) %*% trackmen.pcs$rot)[1,]
...
      usa      ussr    wsamoa
 86.07319  87.32818 104.55099
```

This is what I believed should be the result of the `predict.prcomp()` method for `prcomp()`. Thus I executed

```
> predict(trackmen.pcs)[,1]
...
  switzerl     taipei   thailand     turkey        usa       ussr     wsamoa
-1.6389715  0.9505025  2.7618174  0.2660800 -3.4305560 -2.6268513  7.2312164
```

Apparently, the result is different... What am I doing wrong?

20*. Let the singular decomposition of the *centered* data matrix $\mathbf{Y}$ is $\mathbf{Y} = \mathbf{ULV}^\top$. Show how this decomposition can be used for computing the principal components.

21*. Prove that principal components are uncorrelated (their sample correlation is zero).

22*. Show that if the data matrix $\mathbf{Y}$ can be viewed as a matrix whose rows are independent random vectors that have all distribution with mean $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$, then the sample variance-covariance matrix $\mathbf{S_Y}$ (as defined in the lectures), is an unbiased estimator of $\boldsymbol{\Sigma}$: that is, $\mathrm{E}(\mathbf{S_Y}) = \boldsymbol{\Sigma}$.

23. Given that all results of factor analysis are equivalent under a rotation by an orthogonal matrix $\mathbf{A}$: is the order of the resulting factors essential?

24*. Verify that all stochastic assumptions of the predictive factor model (transparency entitled "Factor model: predictive form") of factor analysis are preserved by a rotation by any orthogonal matrix $\mathbf{A}$.

25*. Assuming that all stochastic assumptions of the predictive factor model (*orthogonal factor model*) are satisfied, calculate $\mathrm{Cov}(\mathbf{y}, \mathbf{f})$.

26*. Suppose that $\boldsymbol{\Sigma}$ is a symmetric nonnegative definite matrix, and $\mathbf{LL}^\top$ is its low rank (rank $\mathfrak{m}$) approximation (as in the transparency entitled: "Low-rank approximation"). Let $\boldsymbol{\Psi} = \boldsymbol{\Sigma} - \mathbf{LL}^\top$. Show that $\boldsymbol{\Psi}$ is nonnegative definite - and in particular, that its diagonal elements are nonnegative.

27. Let $x_1, x_2, \ldots, x_m$, $y_1, y_2, \ldots, y_n$ are two samples arising as results of independent random variables, all of them with the normal distribution with the same variance; the mean of the x_i's is μ_x, the mean of the y_i's is μ_y. You can test the equality $\mu_x = \mu_y$ either (i) by the two sample t-test (function `t.test()` in R) or (ii) by the F-test of the equality of all means in the one-way ANOVA layout. Compare both approaches and summarize the result.

28*. Suppose that random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ has (multivariate) normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Show that $\boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$ has (multivariate) normal distribution $\mathrm{N}(\mathbf{0}, \mathbf{I})$, the normal distribution with mean zero and variance-covariance matrix equal to the identity matrix $\mathbf{I}$.

29*. Show that with normal distribution, orthogonal transformation preserves iid property: if $X_1, X_2, \ldots, X_n$ are independent random variables, each with the same normal distribution with mean 0, then so are the components of the random vector $\mathbf{AX}$, where $\mathbf{X}^\top = (X_1, X_2, \ldots, X_n)$, for any orthogonal matrix $\mathbf{A}$.

30. Suppose that we have two random vectors, $\mathbf{X} = (X_1, X_2)^\top$ and $\mathbf{Y} = (Y_1, Y_2)^\top$. These vectors are independent – that means, each component of $\mathbf{X}$ is independent of each component of $\mathbf{Y}$ – and each of them has multivariate normal distribution:

$$\mathbf{X} \quad \text{with mean} \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and variance-covariance matrix} \quad \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

$$\mathbf{Y} \quad \text{with mean} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and variance-covariance matrix} \quad \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}.$$

What is the distribution of $Z_1 = X_1 - X_2 + X_3 - X_4$? What is the distribution of $Z_2 = X_1 + X_2 - X_3 - X_4$? Are they the same? Are Z_1 and Z_2 independent?

31*. Suppose that $\mathbf{Y}$ is a random matrix with lines $\mathbf{y}_i^\top$, where $\mathbf{y}_i$ are iid random vectors, vectors that are independent between themselves (but not necessarily their components are independent) and have the same normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Derive the putative form of maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, via the solutions of the corresponding likelihood equations.

32$^\$$. Show that given a $p \times p$ symmetric positive definite matrix $\mathbf{B}$ and a $b > 0$, we have for every positive definite $p \times p$ matrix $\boldsymbol{\Sigma}$,

$$\frac{1}{(\det(\boldsymbol{\Sigma}))^b} e^{-\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B})/2} \leqslant \frac{1}{(\det(\mathbf{B}))^b} (2b)^{pb} e^{-pb},$$

with equality holding only for $\boldsymbol{\Sigma} = \frac{1}{2b}\mathbf{B}$.

(This can be used to prove that the maximum likelihood estimators, as derived in the previous problem, are really maximizing the likelihood.)

33*. Suppose that $\mathbf{Y}$ is a random matrix with lines $\mathbf{y}_i^\top$, where $\mathbf{y}_i$ are iid random vectors. Show that if $\mathbf{A}$ and $\mathbf{B}$ are (non-random) matrices such that $\mathbf{A}\mathbf{B}^\top = \mathbf{O}$, then the elements of $\mathbf{AY}$ and $\mathbf{BY}$ are uncorrelated. Use that to show that if the (same) distribution of all $\mathbf{y}_i$ is normal, then $\bar{\mathbf{y}}$, the random vector of columnwise sample means of $\mathbf{Y}$, and $\mathbf{S}_\mathbf{Y}$, the (random) sample variance-covariance matrix calculated out of $\mathbf{Y}$, are independent.

34. (An R problem) Lecture notes say (transparency entitled "Remarks", preceding the transparency entitled "Stochastic underpinning?") that canonical variates are usually scaled so that the variance of them is one. Is it true for the R function `cancor()`? How is it done there? You are not to provide a proof by examining the source code, but verify your answer on some dataset.

35*. (A genuinely R problem) When I checked the connection of correspondence analysis to canonical correlations, my R calculations went smoothly:

```
> haireye=apply(HairEyeColor,1:2,sum)
> library(MASS)
> corresp(haireye,nf=3)
First canonical correlation(s): 0.45691646 0.14908593 0.05097489
...
> hair=as.data.frame(as.table(haireye))
> hairdat=hair[rep(row.names(hair),hair$Freq),1:2]
> X=model.matrix(~Hair-1,data=hairdat)
> Y=model.matrix(~Eye-1,data=hairdat)
> cancor(X,Y)
$cor
[1] 0.45691646 0.14908593 0.05097489
...
```

However, from the mathematical point of view this should not be possible! Note first that the 0-1 matrices $\mathbf{X}$ and $\mathbf{Y}$ sum rowwise to the vectors of 1's; thus, taking $\mathbf{a} = \mathbf{b} = (1, 1, 1, 1)^\top$, and consider the correlation of $\mathbf{Xa}$ and $\mathbf{Yb}$, gets me screwed

```
> as.vector(apply(X,1,sum))
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
...
> as.vector(apply(Y,1,sum))
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
...
> cor(apply(X,1,sum),apply(Y,1,sum))
[1] NA
Warning message:
In cor(apply(X, 1, sum), apply(Y, 1, sum)) : the standard deviation is zero
```

If I want to calculate the canonical correlations via the method on ... it bombs[1]

```
> solve(var(X))
Error in solve.default(var(X)) :
  system is computationally singular: reciprocal condition number =
  1.38702e-17
> solve(var(Y))
Error in solve.default(var(Y)) :
  system is computationally singular: reciprocal condition number = 7.41607e-18
```

[1] some versions of R behave differently, however

36*. Prove the three properties stated on the transparency with the title "Wishart distribution: first properties".

37*. Prove the property on the transparency with the title "Wishart distribution: the important property".

38*. (Statistical/R problem) Consider two-way layout *saturated model* in the (univariate) ANOVA, a linear model with two factors, each with two levels: the mean μ_{ij}, of every observation whose first factor is set at i and second factor is set at j, is modeled as

$$\mu_{ij} = \nu + \alpha_i + \beta_j + \gamma_{ij}, \qquad i = 1, 2, \quad j = 1, 2.$$

where ν is customarily referred to as *intercept*, α_i and β_j is called the *effect* of the first and second factor, respectively, and γ_{ij} models their *interaction*. If $\gamma_{ij} = 0$ for all i and j, the model is called *additive*; to test additivity, one can test the hypothesis

$$H_0 : \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0.$$

Under certain conditions, this is equivalent to testing

$$H_0' : \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0.$$

These "certain conditions" are those typically adopted to ensure that the saturated model is identified (that is, uniquely determined when there are enough data). Give one set of such conditions, and show, in particular, what of them is needed to ensure the above equivalence. Are your conditions used by R?

39. Let $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_n$ be a random sample from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the sample mean $\bar{\mathbf{y}}$ and the sample variance-covariance matrix $\mathbf{S}$. Consider one-dimensional projections of this random sample: for given $\mathbf{a}$, the one-dimensional random sample is $\mathbf{a}^\top\mathbf{x}_1$, $\mathbf{a}^\top\mathbf{x}_2$, ..., $\mathbf{a}^\top\mathbf{x}_n$. Hotelling's one-sample statistic T_a^2 for such a projected sample is nothing else than the square of one-sample t-statistic, where the appropriate mean, sample mean and sample standard deviation depend on $\mathbf{a}$ and respectively on $\boldsymbol{\mu}$, $\bar{\mathbf{x}}$ and $\mathbf{S}$. Show that the Hotelling's one-sample statistic T^2 for the original (unprojected, p-dimensional sample) is equal to the maximum of all projected statistics $T_{\mathbf{a}}^2$, over all $\mathbf{a} \neq \mathbf{0}$; that is, show that $T^2 = \max_{\mathbf{a}\neq\mathbf{0}} T_{\mathbf{a}}^2$.

40$^\$$. Is the Canberra metric (as given in the transparencies) or some of its modifications really a metric? (Prove or disprove.)

41. Verify all claims stated on the transparency entitled "Recovering inner products" (currently page 250 of the 2nd set).

42. Let **C** is a similarity matrix with elements c_{ij}, and let **D** be a dissimilarity matrix with elements $d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{1/2}$. Show that if **C** is nonnegative definite, then **D** is Euclidean, that is, induced by some inner product.

43*. Refer to the transparency entitled "Dendrogram" and prove the claim stated there: show that the tree distance between objects and/or clusters read out of a dendrogram - in a way described in the text of the transparency - is an ultrametric.

44*. Suppose that the original dissimilarity used in clustering is an ultrametric, and an agglomerative method with single linkage is used. Prove or disprove: the tree distance in the resulting dendrogram is an extension of the original dissimilarity.

45$^{\$(c)}$. Suppose that the clusters in $\mathbb{R}^2$ arise as a mixture of distribution: as two samples of size n (the same size is assumed just for simplicity) from two bivariate normal distributions with expected values $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ - for simplicity, assume that their variance-covariance matrix is the same, $\boldsymbol{\Sigma}$, and that $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = 10$. If n grows to ∞, what is the limit of the distance of two clusters that arise this way (a) in the single linkage (b) complete linkage (c) average linkage? Give just intuitive justification of your claims, no formal probabilistic reasoning is required here.

46. For a collection of n data points in $\mathbb{R}^2$, consider the coordinatewise mean and the coordinatewise median. Show that the mean is equivariant (that is, transforms accordingly: mean of transformed data is their original mean transformed by the same transformation) with respect to any orthogonal transformation (rotation, say). Show that the coordinatewise median does not have this property.

47. Consider a task of classification with two classes, based on X: we assume that the distribution of X is either P_1 or P_2, with densities respectively $f_1(x)$, and $f_2(x)$. There are, however, three possible decision outcomes: outcomes 1 and 2 correspond into classifying an item into class 1 or 2, respectively, the outcome 3 means "undecided". In the decision-theoretic setting, all this is expressed by a loss function that posits $\mathcal{L}(1, P_1) = \mathcal{L}(2, P_2) = 0$, $\mathcal{L}(2, P_1) = \mathcal{L}(1, P_2) = 1$, and $\mathcal{L}(3, P_1) = \mathcal{L}(3, P_2) = q$, with $0 < q < 1$. Given the general prior probabilities π_1 and π_2, derive the optimal Bayes classification rule in this case.

48. Suppose that a supervised classification method classifying into two classes, 1 and 2, enables you to predict (that is, to estimate/determine somehow) the posterior probabilities *for some* given prior probabilities π_1 and $\pi_2 = 1 - \pi_1$. (In view of the fact that $\pi_2 = 1 - \pi_1$, one can consider the posterior probabilities to be parametrized by π_1 alone – and without loss of generality assume $\pi_1 = 1/2$.)

Given the formulas for the posterior probabilities for given π_g and true f_g, one can naturally posit that analogous formulas should be satisfied by the *estimates* of the posterior probabilities and the *estimates* $\widehat{f}_g$ of f_g. So, let us assume that we can obtain $\hat{q}_1(\mathbf{x}, 1/2)$ and $\hat{q}_2(\mathbf{x}, 1/2)$ for any $\mathbf{x}$; can we recover from these the predictions $\hat{q}_1(\mathbf{x}, \pi_1)$ and $\hat{q}_2(\mathbf{x}, \pi_1)$ for any given π_1? We cannot recover in general recover the density estimates $\widehat{f_1}(\mathbf{x})$ and $\widehat{f_0}(\mathbf{x})$, but perhaps posterior probabilities may be possible – show how, and then indicate how this could be applied for incorporating prior probabilities into the method of k nearest neighbors.

49. Consider a general classification rule for two classes as in the transparency entitled "The special case of two classes continued": the rule that based on $\mathbf{x}$ classifies to class 1 if $\text{rule}(\mathbf{x}) \geqslant C$ and otherwise classifies to class 2. Suppose now threshold $\in [0, 1]$, and $\text{rule}(\mathbf{x}) = Y$, where Y is a random number uniformly distributed in the interval $[0, 1]$. How does the ROC curve look like for this classification rule?

50. Give a detailed derivation of the scores given on the transparency entitled "Classification scores, LDA and QDA".

51. Using the theory developed in the lectures about MANOVA, show that the rank of the matrix $\mathbf{B}$ defined on the transparency entitled "LDA another way: Fisher's linear discriminants" is $K - 1$.

52. Prove the equivalence to LDA when classification is done using *all* linear discriminants, as stated on the transparency "And the classification rule based on them".

53. Prove the equivalence of the least squares regression to the LDA, as stated in the second paragraph of the transparency entitled "Regression interpretation".

54. Verify the claim stated in the transparency entitled "The connection to the LDA": show that in the LDA situation, when f_1 and f_2 are multivariate normal with the same variance-covariance matrix, the posterior probabilities have the form shown in the transparency.

55*. Prove the property stated in the first paragraph of the transparency entitled "Duality to principal components" (currently page 250 of the 2nd set).

56. Show that the solution for ridge regression estimation prescription, the vector β minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{\top}\boldsymbol{\beta}$$

is $\boldsymbol{\beta} = (\mathbf{X}^{\top}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^{\top}\mathbf{y}$, regardless of the rank of $\mathbf{X}$.