Problems

December 16, 2023

In the applications, software, etc., orthogonal transformations are often mentioned as "rotations". Comment on this terminology, for simplicity considering only transformations in $\mathbb{R}^2$.

1 Give a form of the linear transformation that is a rotation by an angle φ. Is that an orthogonal transformation?

2. Do all orthogonal transformations have the form considered in Problem 1?

3. Show (in general,. not only in $\mathbb{R}^2$) that every orthogonal transformation preserves distances and angles.

4. Is singular decomposition of a matrix unique?

5. Let $\mathbf{A}$ be a square matrix and let $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}^\top$ be its singular decomposition. Give a characterization of the invertibility of $\mathbf{A}$ in terms of this singular decomposition.

6. "Every symmetric matrix $\mathbf{A}$ is similar to a diagonal matrix" - how does this relate to its singular decomposition $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}^\top$?

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

8. Let $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}^\top$ be a singular decomposition of matrix $\mathbf{A}$. Show that the matrix $\mathbf{A}^\top\mathbf{A}$ can be diagonalized and demonstrate how.

9. Suppose that $\mathbf{A}$ is a symmetric 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 $\|\mathbf{x}\| = 1$.

10. Find the compact and general matrix form for the (sample) variance-covariance matrix $\mathbf{S_Y}$. Show with the help of this form that every $\mathbf{S_Y}$ is nonnegative definite.

11. Find the formula for $\mathbf{S_{YA}}$.

12. Suppose that $\mathbf{y}$ is a random vector now, and $\mathrm{Var}(\mathbf{y})$ is its variance-covariance matrix. Find $\mathrm{Var}(\mathbf{Ay})$.

13. Describe what does it imply for the data if the variance-covariance matrix is singular.

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

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

16. 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.

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

18. Suppose that $\mathbf{A}$ and $\mathbf{B}$ are $\mathrm{p} \times \mathrm{q}$ and $\mathrm{q} \times \mathrm{p}$ matrices, respectively. Show $\mathbf{AB}$ and $\mathbf{BA}$ have the same nonzero eigenvalues.

19. Suppose that $\mathbf{A}$ is symmetric and nonnegative definite and $\mathbf{B}$ is 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.

20. Suppose that $\mathbf{A}$ is symmetric and nonnegative definite and $\mathbf{B}$ and $\mathbf{C}$ are 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.

21. Verify that all stochastic assumptions (regression form) of the *orthogonal* factor model (factors assumed uncorrelated) are preserved by a rotation by any orthogonal matrix $\mathbf{A}$.

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

23. Assuming that all stochastic assumptions (regression form) of the orthogonal factor model (including the assumption of uncorrelated common factors) are satisfied, calculate $\mathrm{Cov}(\mathbf{y}, \mathbf{f})$.

24. If $\mathbf{L}$ is a $p \times m$ matrix, find $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\top$ has the minimal Hilbert-Schmidt distance from the sample variance-covariance matrix $S_{\mathbf{Y}} = \mathrm{var}(\mathbf{Y})$.

25. Let $\mathbf{f}$ and $\mathbf{z}$ be random vectors with respectively m and p components, such that both $\mathrm{E}(\mathbf{f}) = 0$ and $\mathrm{E}(\mathbf{z}) = 0$. Show that an $m \times p$ matrix $\mathbf{U}$ minimizing $\mathrm{E}\,\|\mathbf{f} - \mathbf{U}\mathbf{z}\|^2$ has the form $\mathbf{U} = \mathrm{Cov}(\mathbf{f}, \mathbf{z})[\mathrm{Var}(\mathbf{z})]^{-1}$.

Problems 26 and 27, and possibly also Problem 28 are to be solved by experimentation in R. Once you arrive to the solution, make some record of your session: if it does not pose difficulties for you, print a transcript of the session, otherwise at least write down some results.

26. Lecture notes say (page 113, "Remarks") 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 at least on two datasets.

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, on the basis of experimentation with at least (and rather also at most) datasets.

28. Verify the approach to computing the canonical correlations via SVD, as outlined on page 121 of Lecture Notes. You can do it either computationally in R (one dataset being sufficient for this task), or mathematically.

The following two problems are to be solved in a strictly mathematical way.

29. 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 $\mathbf{I}$.

30. Consider two possible situations: (i) random variables $Y_1, Y_2, \ldots, Y_n$ have each (one-dimensional) normal distribution, and nothing else (in particular, independence) is assumed (ii) random vector $(Y_1, Y_2, \ldots, Y_n)^\top$ consisting of (the coresponding) random variables has (multivariate) normal distribution. Comment on a relationship of (i) and (ii): if there is some implication (one implies another or vice versa), then prove it; if in general an implication does not hold, show a counterexample.

31. 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, $\mathrm{E}(\mathbf{S_Y}) = \Sigma$.

32. 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 $\mathbf{X}^\top = (X_1, X_2, \ldots, X_n)$.

33. Prove the three properties stated on the transparency with the title "Wishart distribution: first properties" (page 156).

34. Prove the property on the transparency with the title "Wishart distribution: the most important property" (page 157).

35. 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 $\boldsymbol{\Sigma}$,

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

with equality holding only for $\boldsymbol{\Sigma} = \dfrac{1}{2b}\mathbf{B}$. (This proves that the maximum likelihood estimators, as derived in the lectures, are really maximizing the likelihood.)

36. 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 $\mathbf{AB}^\top = \mathbf{O}$, 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 Y, and $\mathbf{S_Y}$, the (random) sample variance-covariance matrix calculated out of $\mathbf{Y}$, are independent.

37. Consider two-way layout modeling in (univariate) ANOVA, 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

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

To have the model identified, we adopt the restrictions

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i \gamma_{ij} = 0 \text{ for } j = 1, 2, \quad \sum_j \gamma_{ij} = 0 \text{ for } i = 1, 2.$$

Show that in this model with these restrictions,

$$\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0 \quad \text{is equivalent to} \quad \gamma_{ij} = 0 \text{ for all } i \text{ and } j.$$

38. In the exposition of repeated measures, we never used in inferences about contrasts more than $p - 1$ contrasts simultaneously. Explain why - briefly but thoroughly, with an eye on the methodology we used.

39. Refer to the transparencies entitled "Paired T^2" and "And two-sample T^2", momentarily on pages 208 and 209. Apparently, the methods are not equivalent, as the p-values are different. Explain what is going on: what are the methods used, what are their assumptions, etc.

40. Refer to the (corrected) transparency entitled "Some insights" (momentarily page 226). Prove all statements after "We have that".

41. Provide necessary detailed explanation for the transparency entitled "Likelihood ratio motivation for Wilks' Λ" (momentarily page 211 of the second set). In particular, verify the formula for maximized likelihood under the model and submodel, and also demonstrate the equivalence to the ratio of RSS to RSS_H in the univariate case.

42. Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ 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 a, the one-dimensional random sample is $\mathbf{a}^\top \mathbf{x}_1, \mathbf{a}^\top \mathbf{x}_2, \ldots, \mathbf{a}^\top \mathbf{x}_n$. 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 $T^2 = \max_{\mathbf{a} \neq \mathbf{0}} T_{\mathbf{a}}^2$.

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

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

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

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

47. Show that the tree distance between objects and/or clusters read out of a dendrogram is an ultrametric.

48. 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.

49. 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?

50. 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.

51. Suppose that the data come from two classes, with prior probabilities π_1 and $\pi_2 = 1 - \pi_1$, and respective densities of classifiers $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$. The classification based on $\mathbf{x}$ has three possible results: 1, 2, and 3, the last corresponding to "undecided". The corresponding cost functions satisfy $c(1|1) = c(2|2) = 0$, $c(1|2) = c(2|1) = 1$, and $c(1|3) = c(2|3) = c$. Derive the optimal Bayes classification rule in this case. What does it reduce to when $c = 1$ and $\pi_1 = \pi_2 = 0.5$?

52. Consider supervised classification classifying into two classes, 1 and 2, on the basis of the value $\mathbf{x}$ of classifiers considered realizations of random elements $\mathbf{X}$. Let π_1 and π_2 are prior probabilities for classes 1 and 2, and $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are the respective densities of X (for simplicity, assume they are with respect to a Lebesgue measure on $\mathbb{R}^p$). Derive the formulas for the posterior probabilities, the conditional probabilities of the item belonging to classes 1 and 2 given $\mathbf{X} = \mathbf{x}$, as shown in the transparency entitled "Special cases".

53. Show that in the special case when f_1 and f_2 are multivariate normal, with the same variance-covariance matrix, then the posterior probabilities have the form shown on the transparency entitled "The connection to the LDA".

54. 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 $\widehat{q}_1(\mathbf{x}, 1/2)$ and $\widehat{q}_2(\mathbf{x}, 1/2)$ for any $\mathbf{x}$; can we recover from these the predictions $\widehat{q}_1(\mathbf{x}, \pi_1)$ and $\widehat{q}_2(\mathbf{x}, \pi_1)$ 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.

55. Show that the rank of the matrix $\mathbf{B}$ defined on the transparency entitled "LDA another way: Fisher's linear discriminants" is $K - 1$ (as stated on the next transparancy "Fisher linear discriminants").

56. 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".

57. 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 $\mathbf{X}$.

58. Prove the equivalence of the least squares regression to the LDA, as stated on the transparency entitled "LDA as regression".

59. Suppose that $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\top}$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_q)^{\top}$ are random vectors composed of indicator random variables, random variables that attain only values 0 and 1. Let p_{ij} be the probability that *both* X_i and Y_j are equal to 1; the marginal probabilities of X_i and Y_j being equal to 1 are then respectively

$$p_{i\cdot} = \sum_j p_{ij} \qquad \text{and} \qquad p_{\cdot j} = \sum_i p_{ij}$$

Let the variance covariance matrix of $(\mathbf{X}^{\top}, \mathbf{Y}^{\top})^{\top}$ be

$$\mathrm{Var}\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}}\\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}}\end{pmatrix}$$

Show that the elements of $\boldsymbol{\Sigma}_{\mathbf{XY}}$ are $p_{ij} - p_{i\cdot}p_{\cdot j}$.

60. Verify mathematically that the matrix $\mathbf{R}^{-1/2}\mathbf{E}\mathbf{C}^{-1/2}$, as defined on the transparency entitled "Correspondence analysis" (currently page 151 of the first set) has 1 among its eigenvalues.

61. Verify mathematically that the elements of the matrix $\sqrt{n}\mathbf{R}^{-1/2}\mathbf{E}\mathbf{C}^{-1/2}$ are "Pearson residuals" (as claimed in the transparency entitled "Interpretation I", currently page 152 of the first set): their squares are the summands in the Pearson χ^2 statistics for testing independence in the contingency table.

62. Demonstrate the fact that when the points with one label and points with another label are separated by a hyperplane, maximum likelihood estimation of logistic regression collapses. For simplicity, consider only one-dimensional situation, with one classifier x, when for some c, all points with one label have $x_i < c$ and all points with another label have $x_i > c$.

63. Demonstrate that maximum likelihood estimates in logistic regression transforms accordingly when 0 is relabeled to 1 and 1 is relabeled to 0.

1. The following is a part of a specific output in software environment R of the results for linear regression evaluated on 32 automobile models. The predicted variable `mpg`, as usually summarized in a vector $\mathbf{y}$, records consumption in miles per gallon; the predictors, usually summarized in a matrix $\mathbf{X}$, are various other characteristics: number of cylinders, displacement, etc.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) What method was used to obtain the estimates? Give a concise definition in the form of a minimization prescription; introduce additional notation if necessary.

(b) What are the necessary assumptions required for obtaining these results? Illustrate, using matrix formalism, on the method used to obtain the results.

(c) How does the variable denoted as `(Intercept)` enter the matrix $\mathbf{X}$?

(d) The last column gives p-values for the estimates: what *stochastic* assumptions are necessary to ensure validity of these p-values?

2. Suppose that instead of the original responses y_i, we use $cy_i + d$ (such a situation may occur, for instance, when measurement units are changed). How do the estimates change? Give a short justification of your answer.