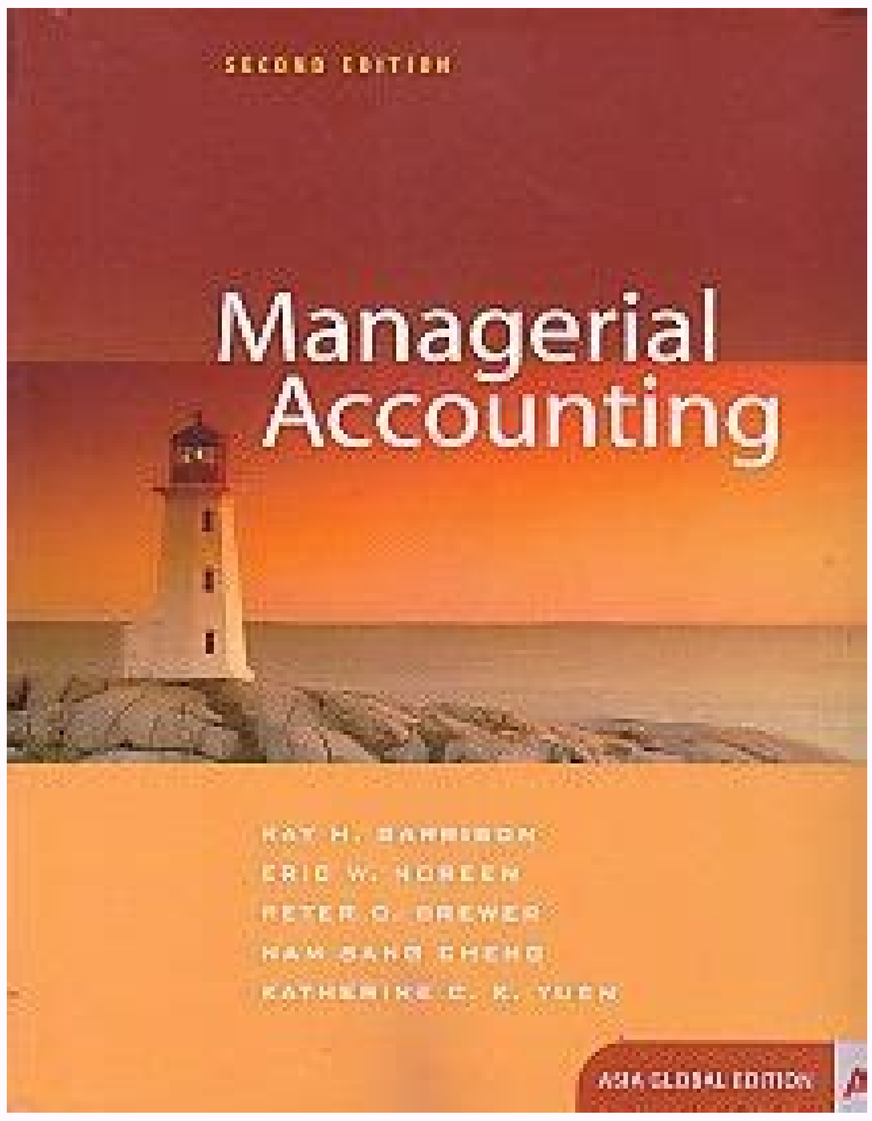
I'm not robot
reCAPTCHA
Continue

10483080.173913 3204001.9189189 50871731910 16995732144 9100431.4591837 51793721315 654049.41538462 53581363755 64163410976 87959128650 122885002.42857 28325779287 13448864.304348 6625018.9 9506460.5945946 20961888.156627 28730016.516129 87368441286 30882949178 107544194.5 105924483060 21066577.641026

Applied multivariate statistical analysis solution manual 5th edition



Decision analysis: an analytic and systematic approach to the study of decision-making.

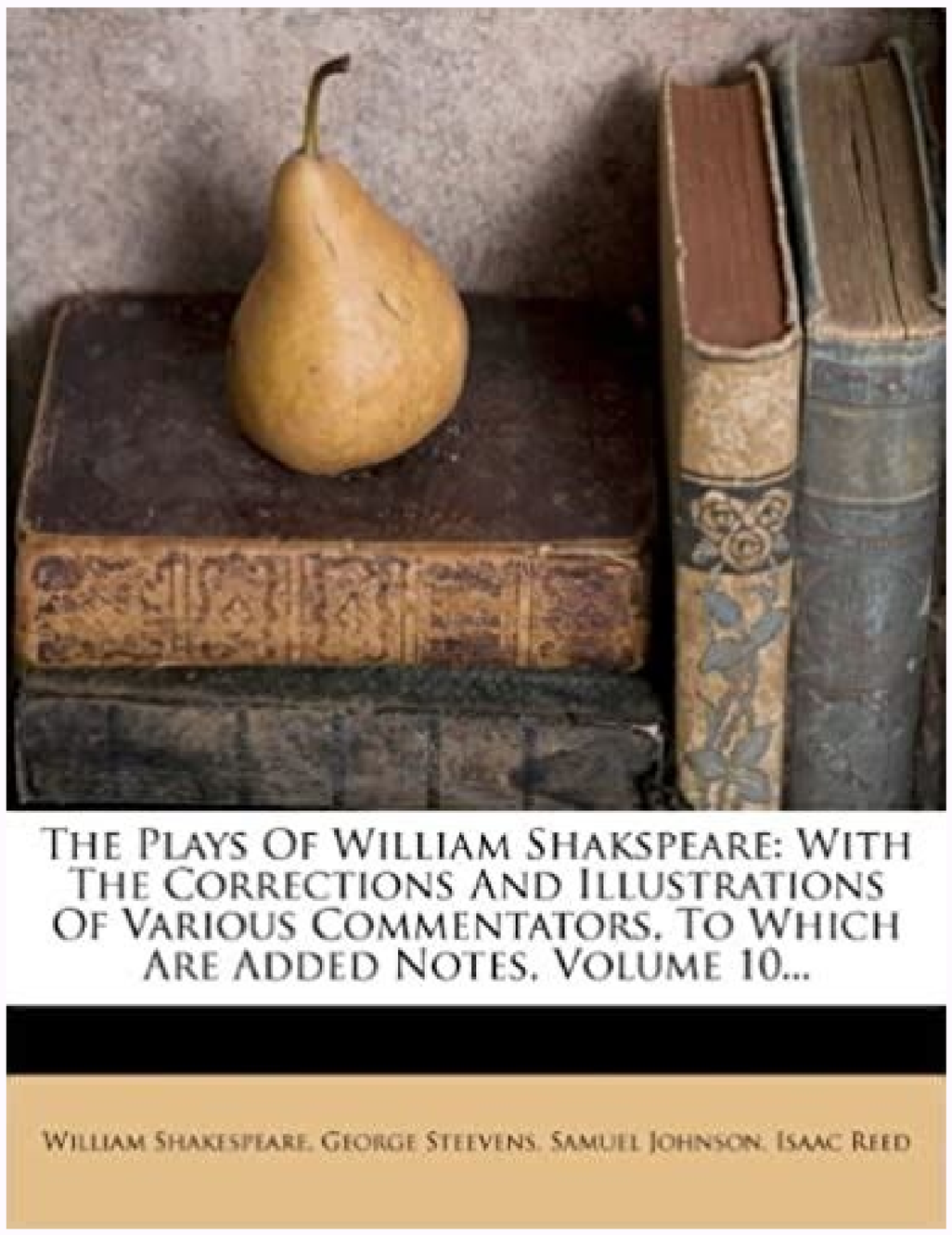
The five steps of decision-making:
1. Clearly define the problem at hand
2. List all possible alternatives: a decision alternative is defined as the course of action that is available for the decision maker.
3. Identify the possible future outcomes for each decision alternative: outcomes are also known as states of nature.
4. Identify the payoff for each combination of alternatives and outcomes: outputs are called payoffs and can also be nonmonetary. They are also conditional in a sense that when something happens demand also plays a role. The easiest way to present payoff values to is by constructing a payoff table, or decision table
5. Select one of the decision analysis modeling techniques. Apply the decision model and make your decision.

Types of decision-making environments:
There are three decision-making environments, which depend on how much knowledge, or information people have.

1. Decision making under certainty: decision makers know for sure the payoff for every decision alternative. Typically there is only one outcome for each alternative.
2. Decision making under uncertainty: decision makers have no information at all about the various outcomes. They do not know the probability that something will occur.
3. Decision making under risk: decision makers have some knowledge regarding the probability of occurrence for each outcome. Decision makers attempt to identify the alternative that optimizes the expected payoff.

Decision making under uncertainty:
There are 5 different decision-making criteria to handle uncertainty situations:

1. Maximax: it selects the decision alternative that maximizes the maximum payoff over all alternatives. It takes an extremely rosy view of the future and locates the alternative with the overall highest possible payoff (optimistic criterion).
2. Maximin: it takes an extremely conservative view of the future (pessimistic criterion). It finds the alternative that maximizes the minimum payoff over all decision alternatives.
3. Realism: offers a compromise between optimistic and pessimistic decisions. We use a parameter called the coefficient of realism to measure the decision maker's level of optimism regarding the future. It is denoted by α. Realism payoff for alternative = α x (maximum payoff for alternative) + (1- α) x (minimum payoff for alternative). It is also called the weighted average criterion.
4. Equally likely: finds the decision alternative that has the highest average payoff.



1 90 Chapter 4 The M u ltiva r i ate Norm a l Distri bution • • •

xj2 j= l = ◆ (42 + 52 + 48 + 58) = ◆ (4 + 5 + 4 + 3) = = 50 4 The sample variances and covariances are sl l 4 = = s22 = = s1 2 = = ◆ ◆ (xj l - x1) 2 j= l ◆ ((42 - 50) 2 + (52 - 50) 2 + (48 - 50) 2 + (58 - 50) 2) = 34 4 ◆ ◆ (xj2 - x2) 2 j= l ◆ ((4 - 4) 2 + (5 - 4) 2 + (4 - 4) 2 + (3 - 4) 2) = .5 4 ◆ ◆ (xj l - xl) (xj2 - x2) j= l ◆ ((42 - 50) (4 - 4) + (52 - 50) (5 - 4) + (48 - 50) (4 - 4) + (58 - 50) (3 - 4)) = - 1 .5 and sn = [34 - 1 .5 .5 - 1 .5] Section 1 .3 The Organ ization of Data 11 The sample correlation is r1 2 = ◆Vs; = r2 1 = r1 2 so R = G raphical Techniques - 1 .5 v'34v3 = - . 2 2 1 This follows because each C has the largest possible number, q 1 , of linearly independent rows, all perpendicular to the vector 1. 29. Let A be a k X k (square) matrix. 1 1 , I Sn l = I [(n - 1)/n]I S I = I [(n - 1)/n]Ip i i S I = [(n - 1)/n]PI S I . 4.34. Are there any suspect observations? 56 Chapter 2 Matrix Algebra and Ra ndom Vectors The transpose operation A' of a matrix changes the columns into rows, so that the first column of A becomes the first row of A' , the second column becomes the second row, and so forth. ◆◆ ◆ ◆◆ OJ n 1 0 0 0 0 0 0 1 0 0 0 Let Xr, . We shall be manipulating these quantities shortly using the algebra of vectors discussed in Chapter 2. Table 1 .2 shows the measured values of x 1 = density (grams/ cubic centimeter) x2 = strength (pounds) in the machine direction x3 = strength (pounds) in the cross direction A novel graphic presentation of these data appears in Figure 1 . - r and n m 7.11 H0 ni + n(i l - i) - n - r - 1 _ l: 2 (m - r + q + 1) (Z'Z) - (7-26) A2 > > Aq+ l > 0 Z 1 (Z1Z 1) - 1 Z1 q + 1 Z 1b 1. Consequently, large x 1 coordinates (in absolute value) are not as unexpected as large x2 coordinates. As we have already noted in this book, f(x) will often be the multivariate normal density function. .·· ' . This data mining exercise is implemented using the general data mining process de scribed earlier and SAS Enterprise Miner software. 1 25 -. It will become increasingly clear that many multivariate methods are based upon an underlying probability model known as the multivariate normal distribu tion. Of course, this cause-effect relationship cannot be substantiat ed, because the experiment did not include a random assignment of payrolls. 4.23. 48.9 1 04 21 80.3 135 25 (xi) , • J • •. Chapter 8 TABLE 8.6 Exercises 473 NATI ONAL TRACK RECORDS FOR M E N Country Argentina Australia Austria Belgium Bermuda Brazil Burma Canada Chile China Colombia Cook Islands Costa Rica Czechoslovakia Denmark Dominican Republic Finland France German Democratic Republic Federal Republic of Germany Great Britain and Northern Ireland Greece Guatemala Hungary India Indonesia Ireland Israel Italy Japan Kenya Korea Democratic People ' s Republic of Korea Luxembourg Malaysia Mauritius Mexico Netherlands lOO m (s) 10.39 10.31 10.44 10.34 10.28 10.22 10.64 10.17 10.34 10.51 10.43 12.18 10.94 10.35 10.56 10.14 10.43 10.11 10.12 200 m (s) 20.81 20.06 20.81 20.68 20.58 20.43 21 .52 20.22 20.80 21 .04 21.05 23.20 21.90 20.65 20.52 20.65 20.69 20.38 20.33 10.16 20.37 44.50 1.73 3.53 13.21 27.61 132.23 10.11 20.21 44.93 1.70 3.51 13.01 27.51 129.13 10.22 10.98 10.26 10.60 10.59 10.61 10.71 10.01 10.34 10.46 10.34 10.91 20.71 21 .82 20.62 21.42 21.49 20.96 21 .00 19.72 20.81 20.66 20.89 21.94 46.56 48.40 46.02 45.73 47.80 46.30 47.80 45.26 45.86 44.92 46.90 47.30 1 .78 1.89 1.77 1 .76 1.84 1.79 1 .77 1.73 1.79 1.73 1.79 1.85 3.64 3.80 3.62 3.73 3.92 3.56 3.72 3.60 3.64 3.55 3.77 3.77 14.59 14.16 13.49 13.77 14.73 13.32 13.66 13.23 13.41 13.10 13.96 14.13 28.45 30.11 28.44 28.81 30.79 27.81 28.93 27.52 27.72 27.38 29.23 29.67 134.60 139.33 132.58 131.98 148.83 132.35 137.55 131.08 128.63 129.75 136.25 130.87 10.35 10.40 11.19 10.42 10.52 20.77 20.92 22.45 21 .30 20.95 47.40 46.30 47.70 46.10 45.10 1.82 1.82 1 .88 1.80 1.74 3.67 3.80 3.83 3.65 3.62 13.64 14.64 15.06 13.46 13.36 29.08 31.01 31 .77 27.95 27.61 141 .27 154.10 152.23 129.20 129.02 400 m (s) 46.84 44.84 46.82 45.04 45.91 45.21 48.30 45.68 46.20 47.30 46.10 52.94 48.66 45.64 45.89 46.80 45.49 45.28 44.87 800 m 1500 m 5000 m 10,000 m Marathon (min) (min) (min) (min) (min) 3.70 14.04 1.81 29.36 137.72 3.57 1.74 13.28 27.66 128.30 3.60 13.26 135.90 1.79 27.72 3.60 13.22 1 .73 27.45 129.95 14.68 3.75 1.80 30.55 146.62 3.66 13.62 1.73 28.62 133.13 3.85 14.45 1.80 30.28 139.95 3.63 1 .76 13.55 28.09 130.15 3.71 1.79 13.61 29.30 134.03 3.73 1.81 13.90 29.13 133.53 3.74 1.82 13.49 27.88 131.35 4.24 16.70 2.02 35.38 164.70 3.84 1.87 14.03 28.81 136.58 3.58 13.42 1 .76 28.19 134.32 3.61 13.50 1 .78 28.11 130.78 14.91 3.82 1.82 154.12 31 .45 3.61 1.74 13.27 27.52 130.87 3.57 1.73 132.30 13.34 27.97 3.56 13.17 1.73 27.42 129.92 (continues on next page) 474 Chapter 8 TABLE 8.6 Principa l Components (continued) Country New Zealand Norway Papua New Guinea Philippines Poland Portugal Rumania Singapore Spain Sweden Switzerland Taipei Thailand Turkey USA USSR Western Samoa lOO m (s) 10.51 10.55 10.96 10.78 10.16 10.53 10.41 10.38 10.42 10.25 10.37 10.59 10.39 10.71 9.93 10.07 10.82 200 m (s) 20.88 21 .16 21.78 21 .64 20.24 21.17 20.98 21 .28 20.77 20.61 20.46 21 .29 21 .09 21 .43 19.75 20.00 21.86 400 m (s) 46.10 46.71 47.90 46.24 45.36 46.70 45.87 47.40 45.98 45.63 45.78 46.80 47.91 47.60 43.86 44.60 49.00 800 m 1500 m 5000 m 10,000 m Marathon (min) (min) (min) (min) (min) 13.21 27.70 1.74 3.54 128 . This common practice of creating new variables that are sums of the original variables and then including them in the data set has caused enough lost time that we emphasize the necessity of being alert to avoid these consequences. I f a xi is the largest eigenvalue, then the major axis lies in the direction e1 . Therefore, Next, and Lx = \1£ = v14 = 3 .742 cos (O) so 0 x' y Ly = VfY = v'6 = 2.449 = = LxLy 3 . 4 , we obt a i n ed t h e canoni c al cor r e l a t i o ns bet w een t h e t w o head relattihoentmatwo lreixg variables for white leghorn fowl. ' J - 1 . Suppose 3 1 s = 3 9 2 (3 X 3) 1 2 1 1 38 Chapter 3 = = = R = [i ! !] Sa m p l e Geometry and Ra ndom Sa m p l i ng Then s1 1 4, s22 9, and s3 3 = = 1. (a) 1 46 Chapter 3 Sa m p l e Geometry and Random Sam p l i n g 3.2. Given the data matrix Graph the scatter plot in p = 2 dimensions, and locate the sample mean on your diagram. As s u mi n g a common covar i a nce ma t i s of i n t e r e s t t o s e e whet h er t h e pr o f i l e s of mal e s and f e mal e s ar e t h e same.A sample of 30 males and 30 females gave the sample mean vectors 6.7.803333 7.6.060033 3.4.790067 ' 4.4.050033 and pooled covariance matrix .606 .262 .066 .161 ..206266 ..167337 ..817310 ..012943 . - . In our attempt to make the study of multivariate analysis appealing to a large audience of both practitioners and theoreticians, we have had to sacrifice a consistency XV xvi Preface of level. ◆ Section 4.6 Assessi n g the Ass u m ption of Norma l ity 1 77 We summarize the major conclusions of this section as follows: In the next three sections, we consider ways of verifying the assumption of nor mality and methods for transforming nonnormal observations into observations that are approximately normal. 48.9 . From a business perspective, the nuggets of knowledge represent actionable information that can be exploited for a competitive advantage. [See (4-31).] Set the significance level at a = .10. j=l Examples of determinants (evaluated using Definition 2A.24) are 1 3 = 1 4 (-1) 2 + 3 6 1 1 1 1 (-1) 3 = 1 (4) + 3(6) (-1) = -14 6 4 In general, 3 1 6 7 4 5 = 3 4 51 (-1) 2 + 1 72 51 (- 1) 3 + 6 27 4 (-1) 4 -7 -7 2 -7 1 = 3(39) - 1 (-3) + 6(- 5 7) = -222 1 0 0 1 0 (-1) 2 0 0 3 + 0 0 1 (-1) 4 = 1(1) = 1 0 1 0 =1 + 0 (-1) 0 1 0 1 0 0 0 0 1 Supplement 2A Vectors and Matrices: Basic Concepts 95 If I is the k X k identity matrix, I I I = 1. Data Mining Techniques. 1 821 - .1821 .9833 A = 43.3748 0.0000] 0.0000 14.9103 v = = [8.547 - 1 .0000 - . Example 1 .5 {M u lti ple scatter plots for paper strength measu rements) Paper is manufactured in continuous sheets several feet wide. Result 4.2. If X is distributed as Np(JL, I) , then any linear combination of + variables a' X == + .· + is distributed as N(a' JL, a' Ia). Chapter 8 Exercises 47 1 component analysis of the data using both the covariance matrix S and the cor relation matrix R. This is no coincidence, as the follow ing result indicates. We concl u de t h at bot h t h e additiTheve manner. On the other hand, if one of the samplers had waited at a campsite far in the in terior of the area and interviewed only canoeists who reached that spot, suc cessive measurements would not be independent. 12.5. Consider the matrix of distances 1 2 3 4 0 1 0 11 2 0 5 3 4 0 1 2 3 4 Cluster the four items using each of the following procedures. The weekly rates of return are defined as (current Friday closing price - previous Friday closing price)/(previous Friday closing price), adjusted for stock splits and div idends. The eigenvalues of A are 3 and 1 . ' ·; ·· . .·◆ • • • •• • • • • . A counterclockwiselike rotation of the axes in Figure 1 . Now, (8-24) defines a hyperellipsoid that is centered at x and whose axes are given by the eigenvectors of s- 1 or, equivalently, of S. But Res u l t pr o vi d es an i n terpretation of a squared statistical distance. a k (X 1) - IL (1) = ak 1 i 1) - J.L P) + a k 2 (X◆1) - J.L◆1)) + + ak p + ··· ··· 1 p - r11.p( )) + a k p vr=a-PP ◆ 1 1 (x p( ) - J.Lp( )) r=- v a-PP ◆) wher e Ther e f o r e , t h e canoni c al coef f i c i e nt s f o r t h e X ?)> :x· TABLE 6 F-D I STRI B UT I O N P E RC E NTAG E PO I NTS (a == .01) .0 1 v2 Fv vl 1 1 2 4052. 484 9.3 Chapter 9 Factor Ana lysis and I nference for Structu red Cova ria nce Matrices M ETHODS OF ESTI MATIO N Given observations x 1 , x 2 , . Find the maximum and minimum values o f the ratio x' Ax/x' x for any nonzero vectors x' = [x 1 , x2 , x3] if A= 2.39. A second group of variables, representing physical-fitness scores, if available, might correspond to another factor. (b) Sketch the n = 3-dimensional representation of the data, and plot the de viation vectors y1 - x1 1 and y2 - x2 1. The measurements of the p variables in a single trial, such as Xj [Xi 1 , Xi 2 , . Linoff. Agai{3 andn, usiTheing (7r-mean tors ande covar 6), wevecget Cov ([-◆]) = Cov (= u{-(◆◆◆r◆-◆-i=-z-(-i-, z)-◆i-z, J Sidependent nce Cov (. 488 Chapter 9 Factor Ana lysis and I nference for Structured Cova ria nce Matrices " " The first two eigenvalues, A 1 = 2.85 and A2 = 1.81, of R are the only eigenvalues greater than unity. Let e2 V= e r 1 ◆ Then V is normal with mean vect{ e'or 21eandk - 2e'ek - 2 = Cov(V; , Vk) = o:u u u ' otherwise 2)-1and by (7-10), That is, thena2 ar=ee'inedependent N(O , oZ V ' Z J (Z = ' Z) = i V◆ V◆-r-1 . (See Exercise 7.8.) For a data point with high leverage, hjj approaches 1 and the prediction at Zj is almost solely determined by yj , the rest of the data having little to say about the mat ter. 1 , A.2 , . Figure 3.3 = where xil is perpendicular to Yi - xil. 5 In addition, special purpose programs such as CART have been used successfully in data mining applications. 0 1) 8 5981. (pXl) (pXp) (pXl) 2 Proof. 0 3 [ ]] [2 2 .05 -.05 V2 = Y.I2 v'6 .8 - 1 - V2 -- V2 v'6 1 - 1 - V2 1 - V2 v'6 . Kennedy Birthplace (region of United States) Midwest South Midwest West South East Elected first term? Z 7A.) generalized inverse r1 q+1 5 Technically, both n (Z 1) . In the scatter diagram of Figure 1 . _ = x-x , _1 = _ x-x 0 implies that treatment 2 is larger, on aver age, than treatment 1 . That is, the eigenval ues after A2 are all relatively small and about the same size.) The scatter diagram indicates the orientation of the points, and their co ordinates can be used to calculate the sample covariance s 1 2 . Chernoff [6] suggested representing p-dimensional observations as a two-dimensional face whose characteristics (face shape, mouth curvature, nose length, eye size, pupil position, and so forth) are determined by the measurements on the p variables. = = of determinants (see Result 2A.11), By the properties and I v-112 1 1 s n I I v -112 1 I v -112 sn v-112 1 = Consequently, i I I v-112 1 I ii./ + q, I I v-112 1 I sn I I v-1/2 1 I sn I I v-1/2 1 I v-112 Li ' v-1;2 + v- 1;2 q, y-112 l I y -1/2 s n y-1/2 1 I L Z L ◆ + 'lI z I IRI I = " " (9-41) " = ---- by (9-30). In particular, (12-40) where e l and e 2 are the first two eigenvectors of s or, equivalently, of X◆ X C = (n - 1) S. 0000 1 5 . That is, calculate the observed values of b'X and c'X, and then use the sample mean, variance, and covariance formulas. 10.) Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing so ciological theories. The reduced rank K > 1 approximation to P - rc ' K . 59 q) =) = -- -- -- - U'1 0\ N CAN O N I CAL VAR I ATE CO E F F IC I E NTS AN D CAN O N I CAL CO R R E LATI O N S Z 1(a 1' ·· 1) Standardized variables Z 2(1) Z3(1) Z4(1) Z5(1) .42 .21 .17 - .02 .44 a2·' a3' ·· - .30 .65 .85 - .29 - . (b) If we j oin clusters { 1 } and { 2 } , the new cluster { 12} has ESS 1 = ◆ (xj - x) 2 = (2 - 1 .5) 2 + (1 - 1.5) 2 = .5 and the ESS associated with the grouping { 12 } , { 3 } , {4} is ESS = .5 + 0 + 0 = .5. 2 pr o vi d es a test for paral el profiles. Of course, it is imperative to examine the adequacy of the model before the estimated function becomes a permanent part of the decision-making apparatus. Figure 1 .7 gives the three-dimensional scatter plot for the standardized variables. In this case, the maximum of the likelihood function [see (9-25) with jL = x and i = LL ' + q, , where L and q, are the maximum likelihood estimates of L and '\}I, respectively] is proportional to I 1-n/2 i exp (- ◆ [(◆ (xi tr I - 1 = - I LL ' + X) (xi - x) ')]l q, , -n/2 exp (- ◆ n tr [(LL + q,) -l S n J) (9-35) Using Result 5.2, (9-34), and (9-35), we find that the likelihood ratio statistic for testing H0 is maximized likelihood under H0 2 ln A - 2 ln . A more extreme percentile must serve to determine obser vations that do not fit the pattern of the remaining data. 0 8 00 .08 0 .05 -.05 0 [-0.0.21 -0.1] 0.1 -0.1 [- .. 357 - . Using the residual mean square s 2 as an estimate of a-2 , we have ---(7-20) Var (ej) = s 2 (1 - hjj) , j = 1, 2, . ., .., . Then replacing the ith diagonal element of R by hj2 = 1 - l/Ji , we obtain a "reduced" sample correlation matrix R r = hi2 r1 2 r1 2 h i2 ri p r2 p ri p r2 p h p*2 Section 9.3 Methods of Estimation 491 Now, apart from sampling variation, all of the elements of the reduced sample cor relation matrix R should be accounted for by the m common factors. A strategy for on-line multivariate exploratory graphical analysis, moti vated by the need for a routine procedure for searching for structure in multivariate data, is given in [32]. Do the scatter plots of pairs of observations on different characteristics give the elliptical appearance expected from normal populations? A similar result holds for many other important univariate statistics. Thus, Ae 1 = Ae 1 gives [2] [ ] [ ] 13 -4 -4 13 -2 2 -2 10 or e1 1 e1 1 e2 1 = 9 e2 1 e3 1 e3 1 13e 1 1 - 4e2 1 + 2e3 1 = 9e 1 1 -4e 1 1 + 13e2 1 - 2e3 1 = 9e2 1 2e 1 1 2e2 1 + 10e3 1 = 9e3 1 - Moving the terms on the right of the equals sign to the left yields three homogeneous equations in three unknowns, but two of the equations are redundant. The sec ond equality for I A I follows by considering 4.11. , _ \10 De- 1/2 - dtag - (12-25) (12-26) (12-27) for scaling purposes. see €= 1 Zb = g Z (Z'Z)-1Z' r+ 1 Z (Z' Z) - 1 Z' = L g g . List any assumptions required in carrying out this test. Then the sample principal comp o nents, .Yz == ei(x - x) are realizations of population principal compone n ts li == ej (X - p.. Prthanediescttiomnatofinagnewthe observation, suofch as YQ,Accorat ding to the regres ioisnormodele unceror tain or (new response (expected value of at (new error) 2 wher e i s di s t r i b ut e d as and i s i n dependent of and, hence, of and s . appr o pr i a t e l y modi f i e d a. (d) Given the results in Parts a-c, do you feel that the stock rates-of-return data can be summarized in fewer than five dimensions? ·· · . Here AA ' has eigenvalue-eigenvector pairs (Ar , ui) , so AA ' u l· = A?u l l. (x i P - Xp) (xi 1 - X1 / + · · · + j=l j=l Again, the divisors n cancel in the statistic, so S may be used. We call this factor the market fac tor, as we did in the principal component solution. 6 For more information on data mining in general and data mining tools in particular, see the ref erences at the end of this chapter. Also, check that the length of 3x is three times the length of x. In addition, as we discuss later, a component associated with an eigenvalue near zero and, hence, deemed unimportant, may indicate an unsuspected linear dependency in the data . (d) Prepare a table of the correlation coefficients between each principal com ponent you decide to retain and the original variables. Van Nos trand, 1958. (See Exercise 7.3.) In Figure 7.2(d), the residuals form a horizontal band. United Parcel Service (UPS) has a package level detail database of about 17 terabytes to track its shipments. The following exercises may require a computer. Additionally, more mathematics is required to derive multivariate statistical techniques for making inferences than in a univari ate setting. [1] 1. 1 43 . 6 [◆] [0 8 -. Zero covariance implies that the corresponding components are independently distributed. Consequently, many statisticians prefer graphical diagnostics based on stu dentized residuals. 92 1 .84 1 .75 1 . e). • • • • • • : ·- ' • • • • ' • • • 7 • • • • • •

2) - JL (2)) ' , we get (Tl p O"l,q+l O"l,q+2 (2-39) O"q , q +l O"q , q +2 O"qp which gives all the covariances, O"ij ' i = 1, 2, . No tice that the variables and their three-digit ranges are indicated in the boxes along the SW-NE diagonal. 82 1 .77 1 .72 1 .67 1 . 87 1 .85 1 . - - -

components of mean vectors, 225, 232, 235 for contrasts, 280 development, 223-26 for differences in mean vectors, 287 , 290, 291 for paired comparisons, 275 as projections, 258 for regression coefficients, 367 for treatment effects, 305-06, 312 Single linkage (see Cluster analysis) Singular matrix, 96 Singular-value decomposition, 101 , 714, 721
Special causes (of variation) , 239 Specific variance, 480 Spectral decomposition, 61 -62, 100 SStress, 701 Standard deviation: population, 73 sample, 7 Standard deviation matrix: population, 73 sample, 140 Standardized observations, 8, 445 Standardized variables, 432 Stars, 25 Strategy for multivariate comparisons, 332 Stress, 701 Structural
equation models, 524-29 Studentized residuals, 377 Sufficient statistics, 173 Sums of squares and cross products matrices: between, 298 total, 298 within, 298, 299 Time dependence (in multivariate observations), 256-57, 410-14 T2-sta tis tic: definition of, 211-12 Subject I index distribution of, 212 invariance property of, 215-16 in quality control, 243,
248, 250, 251 , 459 in profile analysis, 319 for repeated measures designs, 279 single-sample, 212 two-sample, 285 Trace of a matrix, 98 Transformations of data, 194-202 Variables: canonical, 545-46, 556-57 dummy, 357 endogenous, 525 exogenous, 525 latent, 525 predictor, 354 response, 354 standardized, 432 Variance: definition, 69 generalized,
124, 136 767 geometrical interpretation of, 119 total sample, 138, 438, 447, 567 Varimax rotation criterion, 504 Vectors: addition, 52, 84 angle between, 53 , 86 basis, 86 definition of, 50, 84 inner product, 53, 87 length of, 52, 86 linearly dependent, 54, 85 linearly independent, 54, 85 linear span, 85 perpendicular (orthogonal), 54, 87 proj ection of,
55, 88 random, 67 scalar multiplication, 84 unit, 52 vector space, 85 Wilks's lambda, 217, 299, 395 Wishart distribution, 174 It appears from the plot that the data as a whole are not normally distributed. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1987. G&E X1 -1 Bost. 1 05223590 S i m p l e Statistics X2 4.477573765 0.080 1 04466 X3
3.703 1 85794 0.08229677 1 X1 X2 X3 0.0080 1 9 1 4 1 9 0. In addition, it is possible to link algebraically the calculation of x and S directly to X using matrix operations. Plot the minimum stress(q) versus q and interpret the graph. Loadings obtained by a principal component factor analy sis are related to the principal components, which have, by
design, a variance op timizing property. 2 Example 4.9 {Co nstructi ng a Q-Q plot) A sample of n = 10 observations gives the values in the following table: Ordered observations (j -))/n Probability levels Standard normal quantiles q(j) .05 .15 .25 .35 .45 .55 .65 .75 .85 .95 -1.645 - 1.036 -.674 -.385 -.125 .125 .385 .674 1.036 1.645 xu) - 1.00 -.10 .16
.41 .62 .80 1.26 1.54 1.71 2.30 Here, for example, P[Z < .385] = 1.385 00 1 e -z212 dz 2'1T " ◆ v = .65. 5. 1 From our discussion in Section 2.3 with A = I- 1 , we can write • . The eigenvectors associated with these eigenvalues can be determined by solving the following equations: 1 00 Chapter 2 Matrix Algebra and Ra ndom Vectors From the first
expression, xl = xl x 1 + 3x2 = x2 or x 1 = -2x2 There are many solutions for x 1 and x2 • Setting x2 = 1 (arbitrarily) gives x 1 = - 2, and hence, is an eigenvector corresponding to the eigenvalue 1. Also, (7-33) confirming that }he predicted values Y(i) are perpendicular to all residual vectors e(k) . In practice, it is typically necessary to repeat one of
more of these steps sever al times until a satisfactory solution is achieved. , p - 1) (2-52) 82 Chapter 2 Matrix Algebra and Ra ndom Vectors Proof. REFERENCES 1. , n (7-21) We expect the studentized residuals to look, approximately, like independent draw ings from an N (0, 1) distribution. 5 12.6 12.7 12 . b ' c = [7 -3 6] [) [= [- 13]] The product b
' c is a 1 X 1 vector or a single number, here -13. Clustering Algorithms. For example, if X contains p = 12 variables, and the fac tor model in (9-4) with m = 2 is appropriate, then the p (p + 1)/2 = 12 (1 3)/2 = 78 elements of I are described in terms of the mp + p = 12(2) + 12 = 36 parameters eij and l/Ji of the factor model. Ar e meas u r e s of j o b
s a t i s f a ct i o n as s o ci a t e d wi t h j o b characteristics? (See Exercise 3.16.) Consequently, E(Xi Xj) = I + p.. If the eigenvalues are not all distinct, the eigenvectors corresponding to common eigen values may be chosen to be orthogonal. 12.12. ◆ ◆ TABLE 1 2. Ward, Jr. , J. , m satisfy E(i (i)))=I =0 and M u ltivariate M u ltiple Reg ress ion
Section 7.7 Proof. 2 M u ltivariate Li near Reg ress ion Models SAS ANALYSIS FOR EXAMPLE 7.8 U S I N G PROC. PROCRU STES ANALYS IS: A M ETHOD FOR CO M PARING CON F I G U RATI O N S Starting with a given n X n matrix of distances D , or similarities S, that relate n ob jects, two or more configurations can be obtained using different
techniques. a• a •••• a • •• a. In par!_icular, the CS?mmunalities hf ef1 + . But • x is a nonzero vector, and 0 # x == E' y implies that y # 0. Hypothesis construction and testing. Multivariate Analysis in Behavioural Research. ng - ··· + ng ei w-1 w -1 =A - rimin Exer c i s e out l i n es t h e der i v at i o n of t h e Fi s h er di s c ant s . In the general
multiresponse situation, p variables are simultaneously record ed on n items. Fl F2 F3 .16 .30 .19 .35 .13 - .090 .065 - .139 .156 .376 .341 .433 .990 .406 .245 .830 .595 .000 .336 .671 - .169 .275 .000 .445 - . : . 0 1 4832 444 Chapter 8 Principal Components the first principal component may be viewed as the ln (volume) of a box with adjusted
dimensions. The geometrical interpretation of the sample principal components is illustrat ed in Figure 8.4 f2r p ◆ 2. , m, j = 1, 2, . ' .... The ith column, p consisting of all n measurements on the ith variable, determines the ith point. 89 1 . Although the signs of the sample correlation and the sample covariance are the same, the correlation is
ordinarily easier to interpret because its magnitude is bound ed. 17, 2 . Normality is sus pect if the points deviate from a straight line. Also, (3-14) If we compare (3-14) with (3-13), we see that Assuming now that I S I = (n - 1) - (p - 1) (volume) 2 holds for the volume gener ated in n space by the p - 1 deviation vectors d 1 , d 2 , . Squared distances
(see Chapter 1) and the multivariate normal density can be ex pressed in terms of matrix products called quadratic forms (see Chapter 4). From another point of view, if X is distributed as Np (/L, I) , the contours of constant density are ellipsoids whose axes already lie in the directions of maximum variation. However, Cov (X) Cov (W). ( ) 1 2 X have
covar i a nce t h ey ar e not i n dependent . 486 Chapter 9 Factor Ana lysis and I nference for Structu red Covaria n ce Matrices The representation in (9-13), when applied to the sample covariance matrix S or the sample correlation matrix R, is known as the principal component solution. These eigenvectors are the coefficient vectors for the first two
sample principal components. Certain conclusions can be reached concerning the sampling distributions of X and s n without making further assumptions regarding the form of the underlying joint distribution of the variables. Since= F ==1.34 = F3,(1n4(o.0i5)nte=ract3.3io4,nweeffedoctsnot). Let A, U = [u 1 , . As we shall see in Chapters 8 and 12,
there are several techniques that seek to represent p-dimensional observations in few dimensions such that the original dis tances (or similarities) between pairs of observations are (nearly) preserved. The communalities (.98, .88, .98, .89, .93) indicate that the two factors account for a large percentage of the sample vari ance of each variable. ==
== == Result 8.2 says that Total population variance o-1 1 + o-22 + · · · + a-P P A1 + A2 + · · · + AP (8-6) and consequently, the proportion of total variance due to (explained by) the kth prin cipal component is == == Proportion of total population variance due to kth principal component k A 1 + A2 + · · · + A p == 1, 2, .
iahnatooduresatlimeasmatoursreposofsvares iaanbility when the parent popul Let be t h e maxiwhichmiums a fluiknctleiihooodn of estThen imatotrhofe and consider estimating thofe parameter is given by (See and For example, The maximum likeliarhoode these timaximatomr umof JL'lilkel-1ijLhoodis [L'es:it-i1m[L,atwherors ofe andX andI,
rTheespectmaxiivelmyum. Thus, forany Var(a'Y) == a-Var(2(a a-' Za*/3 a*)a'e')(a=-Var(a* a'ea*)) = a' Io2a 2 ' (a - a*)-'1(a - a*) a*' a*] = asa'inZce- (a*'a -Za*)= 'ca' *-=c'(a=- a*)Becaus Z (Z ' Ze)a* cis fixedfrandom (thae-condia*) ' (taio-n a*)(1a -is a*)pos'iZtiv=e unles a= a*, Var(a'Y) is minimized by the choice a*' Y = c' (Z' Zf Z'Y = c' {3 . Kourti, T.,
and J. Cross .. In matrix notation, thbee destheignesmatponsriexs, and let ej be the er " 13 (1) Z2 7.7 M u ltivariate M u ltiple Reg ression Z1 Z 2 ◆ 0). As in Chapter 1, if n observations have been obtained, the entire data set can be placed in an n X p array (matrix) : 112 Section 3.2 X= The Geometry of the Sample X1 1 X12 X21 X22 Xl p X2 p Xn l
Xn2 Xn p X (n xp) X1 1 X 1 2 X2 1 X22 Xl p X2 p x1 x2 Xn l Xn2 Xn p x'n 113 Each row of represents a multivariate observation. There is almost a continuum of sites distinguished by the different species of trees. Since rQ > .9351, we do not reject the hypothesis of normality. Let Ak = {ai j } and B = {bi j } (mX) (m x k) be two matrices of equal
dimension. 3.12. The general theory of likelihood ratio tests is beyond the scope of this book. , x k - 1 · • For example, to construct perpendicular vectors from 3 4 1 0 and 0 0 2 -1 Supp lement 2A Vectors and Matrices: Basic Concepts 89 we take 4 0 0 2 so u◆ u 1 = 42 + 02 + 02 + 22 = 20 and x; u 1 = 3(4) + 1 (0) + 0(0) - 1 (2) = 10 Thus, 3 1 u2 = 0
-1 4 10 0 20 0 2 1 1 0 -2 4 1 0 and z 1 - -V20 0 2 ' 1 1 1 z2 = v'6 0 -2 Matrices Definition 2A.l3. Examine the data on speeds for mar ginal and multivariate normality. + c' (x n - x) (xn - x)'c n- 1 (x 1 - x) (x 1 - x)' + (x2 - x) (x 2 - x)' + . 1 0 , we notice that some points in, for example, the (x2 , x3) scatter plot seem to be disconnected from the others. 94 1
.89 23 4.28 3 . (a) Using the data in the table, calculate the Euclidean distances between pairs of cereal brands. 1 , e 1) , (A.2 , e 2) , . Ordinarily, it is preferable to find regions of IL values that are plausible in light of the observed data. e. Statistical Ecology-a Primer on Methods and Com puting, New York: John Wiley, 1988. Wish. , u p] and V = [v1
, v2 , . , p dimensions. Much of the information contained in the data can be assessed by calculating certain summary numbers, known as descriptive statistics. Check that the Calnonzer o ei g enval u es ar e t h e same as t h os e i n par t a. Consequently, Equation (2-51) says that the largest eigenvalue, A 1 , is the maximum value of the quadratic form
x' Bx for all points x whose distance from the origin is unity. If n is the total of the frequencies in the data matrix X, we first construct a ma trix of proportions P = {Pi j } by dividing each element of X by n. we have n (xj - Qyj - b) ' (xj - Qyj - b) L j= l (xj - Qyj) ' (xj - Qyj) + n b ' b " The last term is nonnegative, so the best fit occurs for b = 0. From
Example 9.5, we determine I LzL ◆ + 'l'z I IRI " " " 1 .000 .572 1.000 .513 .602 1.000 .411 .393 .405 1.000 .458 .322 .430 .523 1.000 1.000 .577 1.000 .509 .599 1.000 .387 .389 .436 1.000 .462 .322 .426 .523 1.000 = .194414 .193163 = 1"0065 Using Bartlett ' s correction, we evaluate the test statistic in (9-39): LL ' + 'II I - 1 - (2p + 4m + 5)/6 J ln I I
Sn I " [n = [" " 100 - 1 -] (10 + 8 + 5) ln (1.0065) 6 = .62 Section 9.4 Factor Rotation 501 Since ◆ [(p - m) 2 - p - m] == ◆ [(5 - 2) 2 - 5 - 2] == 1, the 5% critical value xi(.05) == 3.84 is not exceeded, and we fail to reject H0 • We conclude that the data do not contradict a two-factor model. Consider the census-tract data listed in Table 8.5. Suppose
the observations on X5 = median value home were recorded in thousands, rather than ten thou sands, of dollars; that is, multiply all the numbers listed in the sixth column of the table by 10. Con sider the vector yi [x l i , x2 i, . A small donation will help us alot. Section 2 . Scatter plots for the lumber stiffness measurements are given in Figure 4.11 on
page 193. In these circumstances, the specific factors play the domi nant role, whereas the major aim of factor analysis is to determine a few important common factors. A Standard izing the Sample Pri nci pal Components Sample principal components are, in general, not invariant with respect to changes in scale. (See bk == I2il2 fk ak == I1F2 ek a
b I1i i 1 2 I2ii 2 1 a == p * 2 a I2i i2 1 I1 i i 1 2 b == p * 2 b I1i i 1 2 I2ii 2 1 1 0.3 I2ii 2 1 I1i i 1 2 I NTERPRETI NG TH E POPU LATION CAN O N I CAL VARIABLES Canoni c al var i a bl e s ar e , i n gener a l , ar t i f i c i a l . Consequently, - ◆] 8 3 ' [-◆] -2 = The deviation vectors d1 3s1 2 • The concepts of length, angle, and projection have
provided us with a geomet rical interpretation of the sample. From this example, the chicken-bone measurements for {xP) sskkulunll brengteadh h { ftiebmuria lelngtength h iJ == A z (1) V == B z (2) B z == BD i◆2 . (The i nteger s in the plot i n d icate the n u mber of poi nts occ u pyi n g the same locatio n . , (A.p , e p) where > A = e2X, . -.1 ,
(XZP-ar1L)e has a x◆-disstt1raindarbutiodn.normal2 variables, and we con For Par t b, we not e t h at P[(X c] i s t h e pr o babi l i t y as (X JL) JL)' I 1 2 sfrigonedmPartot a,thPe [el(Xlip-soijL)d (' XI-1-(X◆-t-) ' IJL)- (X -x◆(JL)a)] c by a,andPar the densittybholNp(JLds,. Outliers are best detected visually whenever this is possible. From the definitions
of y1 and y2 , we have ' > or y == E X (2 Xl) (2 X2)(2Xl) Now E is an orthogonal matrix and hence has inverse E ' . - = 432 Chapter 8 Principa l Components where e1 x, e2 x, . The biplot is shown in Figure 12.27. (a) "' . 0 03 -. Construct an alter nate control chart, based on the sum of squares d◆ j , to monitor the unexplained variation in the original
observations summarized by the additional principal components. Variable 1 . (See Exercise 1 .6.) Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores. However, for m 2, LL' produces numbers that are, in general, larger than the sample
correlations. A square matrix kAk is nonsingular if kAk kx 0 (x) (x)(x) (k X1) implies that kx 0 . Also, by independence, Cov (e, F) = E (e, F') = 0. (4-22), n(I 1 - i) Wp, r - q (I) n(I 1 - i) ni , 0.) New York: McGraw-Hill, 1970. ·· '. Specifically, there exist r positive constants A 1 , A2 , . (There were a few missing values.) 472 Chapter 8 Principa l
Com ponents 8.17. We emphasize the applications of multivariate methods and, consequently, have attempted to make the mathematics as palatable as possible. Finally, n-r- 1 n -r- 1 e' e = e' [I - Z (Z' Z) - 1 Z'] e = ◆ e' e e e ee = ◆ Ve V € €= 1 €= 1 which has the wp , n - r - 1 (I) distribution, by (4-22). 2 .057 v2 -2.648zi2) 2.475z◆2) wherand 2,e rz?
es)p,ectiv1,ely2. = c ((p - l) X p) n2 n1 Cx csp ooled c ' . 4.18. This implies that does not have an inverse. (11-57) A ( ) == 3 150 == .02 2%, a Section 1 1 .6 627 Cl assification with Several Pop u l ations I f we adopt t h e hol d out es t i m at e of t h e expect e d as our cr i t e r i o n, we find for the data on irises: Single variable Misclas ification rate
AER .253 .480 .053 .040 Pairs of variables Misclas ification rate xl , x2 xl , x3 Xl , X4 X2 , X3 x2 , x4 X3 , X4 .207 .040 .040 .047 .040 .040 Wetinguiseshtihnatg ththeethsirnegleespvareciieasblofe iris. 4 3 2 Figure 3 . Figure 1 .25 depicts a cluster of points whose center of gravity (sample mean) is indicated by the point Q. and R.A. Johnson "A New
Family of Power Transformations to Improve Nor mality or Symmetry." Biometrika, 87, no. (a) Comment on the pattern of correlation within the centrarchid family x1 through x 4 • Does the walleye appear to group with the other fish? Then PR 2 where A = diag (A1 , n ◆ = j= l xjxj + n ◆ y;·yl j=l - p 2 ◆ Ai i =l = tr [X*X* ' J + tr [Y*Y* '] - 2 tr [A J
(12-46) A.2 , . Let x = ◆measured radiation with door closed and 1 x2 = ◆ measured radiation with door open 222 Chapter 5 n I nferences a bout a Mean Vector 42 [.564] [.0144 .0117] .603203.018 -163..0117391] .0146 ' = [-163.391 200.228 ..0026,02, [[.-704,.710, .710].704 95% 42[.564 - I-Ll ' .603 - [-163.203.039118 -163200..232891] [
..560364 - i-Ll] 2(4041) F2,4o(.05) 3. The generalization of the distance formulas of (1-19) and (1-20) to p dimen sions is straightforward. rIfperforismthance,e numberhow ofmatmie sofclaths eifexpect ied holeddoutactuobsal ererrvoatriroanste,iEn tAER he ith ,grisoprup,ovided by then an esti AER (11-48) (11-52) (i) g == n◆Z) 1, 2, . The first
subscript on an entry in arrays S n and R indicates the row; the second subscript indicates the column. 1 x I B I is obta ed Similarly, expanding the determinant O last row gives = 1 A I· l o' I 1 by the (b) ◆ ◆ ◆◆ ◆ ◆ !I I :, K:-C1 I · I I A I cl = by the last row gives 0, Chapter 4 Exe rcises But expanding the determinant 205 1 :, A:C I = 1. The good ones
try all subsets: z 1 alone, z2 alone, , z 1 and z2 , The best choice is decided by . However, because of the way the stars are constructed, each variable gets equal weight in the visual impression. Since, by Result 4.9, " is the sum of its eigenvalues, this trace will equal its minimum value, zero, if This choice is unique because is of full rank " implies that
0, in which case and *,... Model Assessment In the model development stage of data mining, several models may be examined si multaneously. Young, "The Approximation of One Matrix by Another of Lower Rank." Psychometrika, 1 (1936) , 21 1-218. , Zj r] Yj = [lj 1 , lj 2 , . Let A be an m X k matrix of real numbers. Use whatever methodology,
including transformations, you feel is appropriate. 1 8 Chernoff faces over time. 89 1 .84 25 4.24 3 .39 2.99 2.76 2.60 2.49 2.40 2.34 2.28 2.24 2.16 2.09 2.01 1 .96 1 .92 1 . 1 , the best two-dimensional approximation to the data matrix X approximates the jth observation xj in terms of the sample values of the first two principal components. 277 6 == .
29 . (See Exercise 1 .10.) Contours of constant distances computed from (1 -22) and (1-23) are hyperellipsoids. The standardized measurements are 1 92 Chapter 4 TABLE 4.4 The M u ltiva riate Norm a l D istri bution FOU R M EASU R E M E NTS OF STI FFN ESS WITH STAN DARDIZED VALUES xl x2 x3 x4 Observation no. L j= l A V' (p X p) (p x p)
(p x p) ◆) B e cause the configurations are centered to have zero means xi = 0 and n = U Yi = 0 . l Cov (I: , Yk) = e;Iek = 0 i#k (8- 5) If some Ai are equal, the choices of the corresponding coefficient vectors, e z , an d hence I: , are not unique. Given the original data displayed in Exercise 1.4, do you feel that it is better to determine principal
components from the sample covariance matrix or sample correlation matrix?) Fortu nately, because of the following relation between and A , we do not need the dis tribution of the latter to carry out the test. · + c'xn) Sample mean == ----- n 1 = c' (x 1 + x2 + · · · + x n) = c' x n (3-31) . n If y Exa m p l e 6 . 149 9.158 12.132 6.978 6.890 73.0 73.0
77.0 61.5 66.5 79.5 74.0 59.5 68.0 75.0 66.5 63.0 136.5 135.5 139.0 1 1 8.0 133.5 150.0 137.0 116.0 123 .0 141.0 1 17.0 1 17.0 Source: Data courtesy of Kevin E. Then, be any ei g enval u e-ei g envect o r pai r f o r I Z (Z ' Z) 1 1 2 by (7-6), [I - Z (Z'Z-1) Z'] = [I - Z (Z ' Z-)1- Z2'] so Ae = [I - Z(Z ' Z) Z'] e = [I - Z(Z' Z) -1Z'] e = A[I - Z(Z ' Z)-1Z'] e =
ThatResultis7., 2A),=andf0 orromResNow,ult t4.r 9[I,t-r [IZ-(ZZ(' ZZ) ' ZZ)'-] 1Z= '] =-A1 -+ A2(s+·ee· th+e Aproof e , wher n 1 A1 - A-2 val· · uesAofn arAieequaltheeione,genvalanduesof[Z ' . 24 .36 .38 .19 - .08 . B. 5 37 Distance We note that the distances in (1-22) and (1-23) are completely determined by the coefficients (weights) a i k ' i = 1 , 2, .
For example, p-variate normality is indicated if 1. We have chosen to provide explanations based upon algebraic concepts and to avoid the derivations of statistical results that require the calculus of many variables. Alternative ly, if both variables are standardized, their subsequent magnitudes will be of the same order, and X2 (or Z2) will play a
larger role in the construction of the principal com ponents. simSeeultAppendi aneous confx, Tablidence 1.) sThetatement95% ssimweulrtea neous confidence statement is T1 3 - T3 3 belongs to T13 - T33 t513 (.00208))(n11 n13) nw33 -.-.004343 2..0818,7(.0or0614)(-.061, Weownedconclnurudesingthhomes at the averis higaherge maiby
n.0t25enanceto .061andhourlaborpercospatt ifeontr gover n ment day t h an privately ownedT1nur3 -sinT2g3homes . ISBN 0-13-092553-5 1. • • • • • • • .. Compare the results in (a) and (b). , OJ . 17 . 76400000 4.265 50000 Mean Sq u a re 0.83383333 0. 30050000 0. (a) Determine the power transformation A 1 that makes the x 1 values approxi-
mately normal. A., and G. Convert the women ' s track records in Table 1.9 to speeds measured in meters per second. A rotation of the factors often reveals a simple structure and aids interpretation. Let X 1 , X 2 , X 3 , X 4 , and X 5 be independent and identically distributed randon1 vectors with mean vector IL and covariance matrix I. The sample
covariance s;k = ◆ 1 n (xi; - X;) (xik - Xk) n i = 1 , 2 , . the question of Cons i d er t h e popul a t i o n means r e pr e s e nt i n g t h e ernectageedrebyspsonstraeigshttolifnoesur, itsreshatownmentins Fifogrurthee 6.fi4r.stThigrosup.broAken-plloitneofgrthaphesteimeans , s t h e for populProfaitleiosncan1. Consideration of bias motivates a slightly
modified definition of the sample variance-covariance matrix. school class, x3 = percentage of applicants accepted, x4 = student -faculty ratio, X5 = estimated annual expenses and X6 = graduation rate (%) . Theyiancearmate apprricesopristiasteeriifousnorlymvialoitlaytappear s t o hol d , but t h e as s u mpt i o n of e d. Refer to Exercise 8.18.
likelihood estimator of is ◆, where 2 j= 1 is the maximum likelihood estimator of Var I A I L(jL , i) = 6. In the second case, the trapezoid has very little height above the plane. Young, F. Plots like those in Figure 1 . Dun & Bradstreet is the largest firm in terms of num ber of employees, but is "typical" in terms of profits per employee. (a) Graph the
two vectors. Examples of vectors are • • • a = [◆J x= 1 -1 b= 1 ' -1 Vectors are said to be equal if their corresponding entries are the same. Moreover, from the results in Section 4.4, we know that X is exactly normally distributed when the un derlying population is normal. The sum of the matrices A and B is an m X k matrix C, written C == A + B,
such that the arbitrary element of C is given by == == C·l 1. This has no effect on the sample covariance matrix S and gives the ith principal component i = 1 , 2, . 2 (1964) , 211-252. By adding a scale to the curve, we would obtain a one-dimensional ordination. Cormack, R. lWeity can thenthseelegrctoupsto ofmaxiY-mvalizueesthriesIraattiivoe.
Specifically, p(i) e (i) = (Z' Z) -1 Z' e(i) 1 f3 u) = [I - Z (Z' Z) - Z' J eu) ' + i = 1, 2, . The eigenvector e 1 is determined from - (A. (See [1] .) (In1 our.3cas819e,) (16 - 3 + 1)/2 F1 = -.3819 (1 1 - 31 + 1)/2 =7.55 = (1 -.5230.5230) ((1116 -- 313 ++ 1)1)//22 = 4·26 and 1 - 31 + 1 = 3 = (16 - 3 + 1) = 14 = 1 andSimitlhaFrrerloyem,fobefre=,
weo4.re2r,6eFj3e, 1c4tF3(H0:.0, 15)4(.0=15)3.=34.3=3We4, and(haveno fweaFct1ore=rj1e7.ctef55fH0:ects)F3at, 1=4t(h.0e5)5%== le3.vel(3n4,o. + clp xp z2 = Cz l Xl + Czz X2 + . The sampling distributions of Ai and ei are difficult to derive and beyond the scope of this book. Although there are many choices for initial estimates of specific
variances, the most popular choice, when one is working with a correlation matrix, is l/1[= 1/ r i i , where r ii is the ith diagonal element of R- 1 . Since, by definition, and cos (O) = cos (02 - 0 1) = cos (02) cos (0 1) + sin (02) sin (0 1) the angle 0 between the two vectors x ' = [x 1 , x2] and cos (0) = cos (02 - 0 1) = y' = [y1 , y2] is specified
by Y) ( )) ((◆) Ly Lx Ly Lx ◆ We find it convenient to introduce the dimensions, the inner product of x and y is 2 + x2 = XlYl + XzYz LxLy (2- 3) inner product of two vectors. In this case I = LL' + 'I' , and testing the adequacy of the m common factor model is equivalent to testing (pIX p) (9-33) (pLx m) (mL'x p) + (pwx p) versus H1 : I any other
positive definite matrix. Consequently, with a large sample size n, we expect the observed proportion Pi I of the observations lying in the interval (xi - ◆' xi + ◆) to be about .683. That is, with ref erence to the x l and x2 axes, we define the distance from the point p = (x l ' x2) to the origin 0 = (0, 0) as d(O, P) (1-17) = where sl l and s22 denote the
sample variances computed with the xl and x2 measurements. In particular, ei k is pro portional to the correlation coefficient between }j and Xk . New York: John Wiley, 1 984. OJ [Examples of symmetric matrices are 1 0 I = 0 1 0 , (3 X 3) Q Q 1 a c e f c b g d B = e g c a (4 X 4) f d a d 92 Chapter 2 Matrix Algebra and Random Vectors Definition
2A.22. 39 1 . Also, find a 95% prediction interval for a new facility ' s CPU requirement corresponding to the same z0 • A computer program provides the estimated regression function = y = (Z ' Z)-1 = [8.42 + 1.08z 1 + .42z2 8.17969 - .06411 .00052 .08831 -.00107 .01440 and s = 1.204. Note that the ith specific factor si is associated only with the
ith response Xi . 1 Individual 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 5 SWEAT DATA xl (Sweat rate) x2 (Sodium) x3 (Potassium) 3.7 5.7 3.8 3.2 3.1 4.6 2.4 7.2 6.7 5.4 3.9 4.5 3.5 4.5 1.5 8.5 4.5 6.5 4.1 5.5 48.5 65.1 47.2 53.2 55.5 36.1 24.8 33.1 47.4 54.1 36.9 58.8 27.8 40.2 13.5 56.4 71.6 52.8 44.1 40.9 9.3 8.0 10.9 12.0 9.7 7.9 14.0 7.6
8.5 11.3 12.7 12.3 9.8 8.4 10.1 7.1 8.2 10.9 11.2 9.4 Source: Courtesy of Dr. Gerald Bargman. , J. (c) (d) 2.8. 2.9. A- 1 . uAlltso, VarUnder (z0f3) thezf0uCovrther(f3as)szu0mptz0io(nZ't'Z)hat z0ia-s norsinmcealCovly di(sf3tr)ib uted, Result asserts that p is Nr+1 (f3, a-2(Z'Z)-1) independently of s2/a-2 , which = [1, + + ··· (1) (2) + Y Esti mati ng the Reg
ression Fu nction at z0 (7-3), = [1, = + + ··· = Zo /3 + " Result 7.7. Yo " (7-3), o = 100(1 - a)% e (7 - 1 8) = n - r-1 (a/2) IJ ± IS 100(a/2 Proof. In the radiotherapy data listed in Table 1 .7 (see also the radiotherapy data on the CD-ROM), the n = 98 observations on p = 6 variables represent patients ' reactions to radiotherapy.) ' is zero because the
entry is the co variance between a component of Xi and a component of X e , and these are inde pendent. - im; .) ± tv {£;;2 a 6-59) = X - Xm . Are they the same? popul I t i s not sary to assume that the populations are multivariate normal. 72 2.64 2.55 22 7.95 5.72 4.82 4.31 3.99 3.76 3.59 3.45 3.35 3.26 3.12 2.98 2.83 2.73 2.67 2.58 2.50 23 7 .88 5.66
4.76 4.26 3.94 3.71 3 .54 3.41 3 .30 3.21 3 .07 2.93 2.78 2.69 2.62 2.54 2.45 24 7.82 5.61 4.72 4.22 3.90 3.67 3.50 3.36 3.26 3.17 3.03 2.89 2.74 2.64 2.58 2.49 2.40 25 7.77 5.57 4.68 4.18 3.85 3.63 3.46 3.32 3 .22 3.13 2.99 2.85 2.70 2.60 2.54 2.45 2.36 26 7.72 5.53 4.64 4.14 3.82 3.59 3.42 3.29 3.18 3.09 2.96 2.81 2.66 2.57 2.50 2.42 2.33 2.38 2.29
27 7.68 5.49 4.60 4.1 1 3 .78 3.56 3.39 3.26 3.15 3.06 2.93 2.78 2.63 2.54 2.47 28 7.64 5 .45 4.57 4.07 3 .75 3.53 3.36 3.23 3.12 3.03 2.90 2.75 2.60 2.51 2.44 2.35 2.26 29 7.60 5 .42 4.54 4.04 3 .73 3.50 3.33 3.20 3.09 3.00 2.87 2.73 2.57 2.48 2.41 2.33 2.23 30 7.56 5.39 4.51 4.02 3.70 3.47 3.30 3.17 3.07 2.98 2.84 2.70 2.55 2.45 2.39 2.30 2.21 40
7.31 5.18 4.31 3.83 3.51 3.29 3.12 2.99 2.89 2.80 2.66 2.52 2.37 2.27 2.20 2.11 2.02 60 7 . (See [12] for further discussion.) ((1 80 Chapter 4 The M u ltivariate Norma l D istribution • • • • • A 0-0 plot for the data in Exa m p l e 4.9. Figure 4.5 the normal quantiles q (j) , is shown in Figure 4.5. The pairs of points (q (j) , x u)) lie very nearly along
a straight line, and we would not reject the notion that these data are normally distributed-particularly with a sample size as small as n = 10. 12.10. Hernandez, F. King, B. Consider now the problem of determining whether a given p X 1 vector JL o is a plausible value for the mean of a multivariate normal distribution. Construct a 95% confidence
interval for the mean CPU time, E (Yo I z0) {30 + {3 1 Zo1 + {3 2 z0 2 at z0 = [1, 130, 7.5]. /3k ; - /3 q ; 1 Xm 1 • f3ki - /3q1 are (i. The remainin ◆ eigen':alues are small and about equal, although A4 is somewhat smaller than A2 and A3 • Thus, there is some evidence that the corresponding population correlation matrix p may be of the "equal-
correlation" form of (8-15). Given the air-pollution data in Table 1 .5, examine the pairs X5 = N0 2 and x6 = 0 3 for bivariate normality. , X n] ' is the data matrix. 96 1 . , x n 1 be n measurements on the first variable. Using Result 8.4, we obtain PY 1 , z 1 == e1 1 VA;- == .707 vT.4 == .837 and PY 1· z2 == e2 1 VA;- == .707 vT.4 == .837 (') ( ) ◆ In
this case, the first principal component explains a proportion A1 1.4 == == .7 p 2 of the total (standardized) population variance. Although it is impossible to simultaneously plot all the measurements made on several variables and study the configurations, plots of individual variables and plots of pairs of variables can still be very informative.
Similarly, AP is the smallest value of the qua dratic form for all points x one unit from the origin. In addition, the resulting credit ratings ("Good" or "Bad") for each applicant was recorded. The n observations of the p variables can also be . e upper)th percentile of a t-distribution with We f o recas t by whi c h es t i m at e s By Res u l t has and The f o r
e cas t er r o r i s t h en Thus , so the predictor is unbiased. qi is the ith component of We have cons i d er e d t h e mul t i v ar i a t e t w ow ay model wi t h catnatiioonsn of. , • I I • ... If X1 , X2 , , Xn denote a random sample from a normal population, the appropriate test statistic is (X - JLo) 1 ◆ 1 2 2 £.J Xj and s = £.J (Xj - X) where X = ◆ t= Vn n n 1
sj j =l j =l • • • - ' 21 0 - - _ Section 5.2 The P l a u s i b i l ity of f.Lo as a Va l ue for a Normal Pop u l ation Mean 21 1 This test statistic has a student ' s !-distribution with n - 1 degrees of freedom (d.f.). , e P and A be the diagonal matrix with eigenvalues A 1 , A2 , . ' p 1, 2, . These deviation vectors lie in the direction of d i , but have a squared length
of n - 1 . N. The key proper ties, which we shall soon discuss in some mathematical detail, can be stated rather simply. 17 39.15 9 . The answer may have implications for job design. The ith column of the data matrix, or variable, is represented as an arrow from the origin to the point with co ordinates (e l i' e2 J' the entries in the ith column of the
second matrix [e l ' e 2] ' in the approximation (12-43). In Chapter 1, we described scatter plots for pairs of characteristics. 32 2.71 2.30 2.08 1 . Data mining requires a rich collection of tools and algorithms used by a skilled analyst with sound subj ect matter knowledge (or working with someone with sound subj ect matter knowledge) to produce
acceptable results. Time Warner has a "typical" number of employees, but comparatively small (negative) profits per employee. In addition, the coordi nates of all points P = (x1 , x2) that are a constant squared distance c 2 from Q satisfy a 1 1 (x 1 - Yl) 2 + 2a 1 2 (X1 - Yl) (x2 - Y2) + a 22 (x2 - Y2) 2 = C2 (1-21) By definition, this is the equation of
an ellipse centered at Q. , Xi p] , will usually be correlated. s o t h at ProveConsthat everider tyheie gdefenvalinitiuoenofof aan eigenvalposuie,tivwhere defeinite matAe.rixMulitsipposly onitivte.he lConseft byide'er sthoethseattse'of pointsAe' e. Repeat Exercise 12.16 using the national track records data for men given in Table 8.6. Compare the results
with those of Exercise 12.16. 54450000 7.92 3.73 3.32 Pr > F X 2 Mean 9 . k ; - i. Thus ca 1 P ca 2 P ca l l ca l 2 ca 2 1 ca 22 cA = (n X p) ca n l ca n 2 ca np Two matrices A and B of the same dimensions can be added. 1/2 - _ JL , -x S n(x - JL)' S-1 (x - JL) [_ , __ 8. 4 The sample mean vector and sample correlation matrix were, respectively, 49.95] x '
= [39.88, 48.11, 45.08, and 1 .000 .7501 .6329 .6363 .6925 .7386 .7501 1 .000 R = .6625 .6329 .6925 1.000 .6363 .7386 .6625 1.000 The eigenvalues of this matrix are A1 = 3.085, A2 = .382, A3 = .342, and A4 = .217 We note that the first eigenvalue is nearly equal to 1 + (p - 1) r = 1 + (4 - 1) (.6854) = 3.056, where r is the arithmetic average of
the off diagonal e}ements of R. Then it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations. title 'Principal Component Ana lysis'; data t u rt l e; i nfile ' E 8-4.d at'; i n put l e n gth width heig ht; x

data? These scatter plot arrays are further pursued in our discussion of new soft• ware graphics in the next section. 7 383 Thatto thosise, of is (athbiatasise,d estimator ofIf imporunltaents vartheiacolbleusmnsare miofs inarg feroperm pthendie modelcular, the least squares estimates may be misleading. If X is distributed as Np (IL , I) with I I I # 0, show that the joint density can be written as the product of marginal densities for o if I 1 2 = X2 X1 and X (q Xl) (q (p - q)) ((p - q) Xl) Hint: Show by block multiplication that I 1i 0 ◆ , Is t he I. 1 1 (a) , but then the brush could be moved to provide a sequence of highlighted points. , ek X X where the symbol j_ (attained when x = e k +l , k is read "is perpendicular to. ....,... (b) The variance is not constant. 657 2.616 2.575 2.536 2.498 63 .657 9.925 5.841 4.604 4.032 3.707 3.499 3.355 3.250 3. Then A is said to be symmetric if A = A' . 00000000 2 . (This is true, for example, in business, economics, ecology, geology, and sociology.) You should consult [7] and 1 2 Chapter 1 Aspects of M u ltiva riate Ana lysis [8] for detailed accounts of design principles that, fortunately, also apply to multi variate situations. Attention is usually centered on contrasts when comparing treatments. Moreover, this first principal component explains a proportion A1 == 100.16 == .992 101 A1 + A2 of the total population variance. (See [1 4] . As originally designed, Chernoff faces can handle up to 18 variables. (S ee, f o r exampl e , [1 1] . Consequently, in terms of the singular values and vectors from D ◆ 112 PD◆ 1 12 , the reduced rank K < J approximation to the row profiles D◆1 P is K P* (12-38) l1c ' + L Ak D◆ 112 uk (D◆I2 vk) ' k=2 In terms of the singular values and vectors A k , uk and vk obtained from o -1 12 (P - rc') D -1 12 we can write K-1 P * - 11 c' L Ak D◆ 112 uk (D◆I2 vk) ' k=1 (Row profiles for the archaeological data in Table 12.8 are shown in Figure 12.20.) = · r c ' · Inertia Total inertia is a measure of the variation in the count data and is defined as the weighted sum of squares where the Ak are the singular values obtained from the singular-value decomposition 3 of D ◆ 1 12 (P - rc') D◆ 1/2 (see the proof of Result 12. If we set s ee also (2-22)] i =l p 1 " - e .e◆ B- 1/2 = ◆ i =l Y J\i ,.. Definition 2A.31. A naturalist for the Alaska Fish and Game Department studies grizzly bears with the goal of maintaining a healthy population. 80 3.79 3.24 3.21 3.19 3.17 3.16 3.14 6 3.78 3 .46 3 .29 3.18 3.11 3 .05 3.01 2.98 2.96 2.94 2.90 2.87 2.84 2.81 2.80 2.78 2.76 7 3.59 3.26 3.07 2.96 2.88 2.83 2.78 2.75 2.72 2.70 2.67 2.63 2.59 2.57 2.56 2.54 2.51 8 3 .46 3.11 2.92 2.81 2.73 2.67 2.62 2.59 2.56 2.54 2.50 2.46 2.42 2.40 2.38 2.36 2.34 9 3.36 3.01 2.81 2.69 2.61 2.55 2.51 2.47 2.44 2.42 2.38 2.34 2.30 2.27 2.25 2.23 2.21 10 3 .29 2.92 2.73 2.61 2.52 2.46 2.41 2.38 2.35 2.32 2.28 2.24 2.20 2.17 2.16 2.13 2.11 11 3 .23 2.86 2.66 2 . As indicated, each facial feature represents a single financial indicator, and the longitudinal changes in these indicators are thus evident at a glance. Then c1 y = 5 [◆] [◆] = -2 1 - 10 and c2 y = (- 1.2) [◆] [=◆:◆] -2 = 2.4 Definition 2A.3 (Vector Addition). The linear combination of the original data, using a, is a constant. We conclude this brief overview of multivariate analysis with a quotation from F. , ck , not all zero, such that . Moreover, r+ 1 1 r+1 1 Z (Z'Z) Z' = i◆= 1 Aj Zei ejZ' ◆ i =1 qi q; -- == l == l l1 • ' r + 1 Qk Section 7 . , and M. (n - JL = (X ------ (p + · · · + (u Zn _ 1) 2 . 1 1 .8299 .8801 .9126 .9269 .9410 .9479 .9538 .9599 .9632 .9671 .9695 .9720 .9771 .9822 .9879 .9905 .9935 .8788 .9198 .9389 .9508 .9591 .9652 .9682 .9726 .9749 .9768 .9787 .9801 .9838 .9873 .9913 .9931 .9953 .9032 .9351 .9503 .9604 .9665 .9715 .9740 .9771 .9792 .9809 .9822 .9836 .9866 .9895 .9928 .9942 .9960 (A correlati on coefficient test fo r norma lity) Let us calculate the correlation coefficient rQ from the Q-Q plot of Example 4.9 (see Figure 4.5) and test for normality. 1 2 + (n l n2 + -2 X n2) at 0 (A profi le analysis of l ove and ma rriage data) Asveyedparadult of atslawirgterh rsetsupdyectoftolotveheiandr marmarriargeiage,"contE. This scale may not be compatible with that of the principal components, so an arbitrary multiplier can be introduced that adjusts all of the vec tors by the same amount. Hint: For c = 0, evaluate Cov (X1 , X2) = E[X1 (X1)] For c very large, evaluate Cov (X1 , X2) . This region is determined by the data, and for the moment, we shall denote it by R(X) , where X = [X , X 2 , . Therefore, its expected value is i= l i= l For any random vector V with E (V) = #Lv and Cov (V) = I v , we have E (VV') = I v + #Lv#L'v . 7] 9. .40 .30 .20 3 .10 .00 2 - 2.0 3 • 5 - 1 .0 e3 2 9 .0 2 3 • • 1 .0 2.0 3.0 q (j) Figure 4.6 A 0-0 plot of the radiation data (door closed) from Exa m p l e 4. From the second expression, x 1 = 3x 1 x 1 + 3x2 = 3x2 implies that x 1 = 0 and x2 = 1 (arbitrarily), and hence, [◆] X= is an eigenvector corresponding to the eigenvalue 3. 1 3 (Co nstructi ng a chi-square pl ot) Let us construct a chi-square plot of the generalized distances given in Exam ple 4.12. It is conve nient to denote normalized eigenvectors by e, and we do so in what follows. In general, the statistical distance of the point P = (x 1 , x2) from the fixed point (y1 , y2) for situations in which the variables are correlated has the general form d(P , Q) = v'a 1 1 (x 1 - Y1) 2 + 2a 1 2 (X1 - Y1) (x2 - Y2) + a 22 (x2 - Y2) 2 (1-20) Q = and can always be computed once a 1 1 , a 1 2 , and a 22 are known. 0, r + 1. , yP = e◆ x, we have and this equation defines an ellipsoid (since A1 , A2 , . Multivariate Analysis.)> x· TABLE 5 F-D I STR I B UT I O N PE RCE NTAG E POI NTS (a == .05) .05 F If, 1 ' v 2 (.05) v2 vl 1 2 1 161 .5 18.51 2 3 1 99.5 215.7 1 9 .00 19.16 4 224.6 1 9.25 5 230.2 19.30 6 234.0 19.33 7 236.8 1 9.35 8 238 . The resulting plot is called a chi-square plot or gamma plot, because the chi-square distribution is a special case of the more gener al gamma distribution.) = (X?J - J.L?))/ VO::Z , = Uk , V Z (2) . X - JL = • • • With so many unobservable quantities, a direct verification of the factor model from observations on X1 , X2 , . The B such that AB = BA = I is called the inverse of A and is denoted by A- 1 . 6 1 877 1 88 1 . , c2 P]. , p, we have a 1 1 a 12 0 < (dist ance) 2 = [x 1 , x2, . Each observation can be expressed as a linear combination xj = (xj e l) e l + (xj e 2) e 2 + . The multivariate normal distribution is completely specified once the mean vector IL and variance-covariance matrix I are given (see Chapter 4) , so it is not surprising that these quantities play an important role in many multivariate procedures. k The matrix ◆ \IA; eiei = PA 112 P' is called the square root of A and is denoted by A112 . U'1 U'1 ._,... When the num ber of observations n is large, dot plots are not feasible. (7-10), (7-6), 4.9, X + ··· + + ··· + + + 2A . has probability .5. Calcu late the value of the angle between them. Yet, if linear combinations of the columns of exist that are nearly 0, the cal culation of is numerically unstable. The remain ing minor axes lie in the directions defined by e 2 , , e P . Find the ar rays x, S n , and R. For example, let [3 -1 2 4 0 5 (2X 3) A Then [43 where -1 2 0 5 (2 X 3) e1 1 e1 2 c2 1 e22 = = = =]] [◆ -◆] 4 (3 X 2) (3) (3) (3) (4) (4) (3) (4) (4) + + + + 3 and _ [◆ -◆] B = (3 X 2) 4 3 [3211 2031] [ee21 11 ee221 2] _ (2 X 2) (-1) (6) + (2) (4) = 11 (-1) (- 2) + (2) (3) = 20 (0) (6) + (5) (4) = 32 (0) (-2) + (5) (3) = 31 As an additional example, consider the product of two vectors. So lving Data Mining Pro blems Through Pattern Recognition. For this reason, an orthogonal transformation of the factor loadings, as well as the implied orthogonal transformation of the factors, is called factor rotation. 20 Chapter 1 Aspects of M u ltivariate Ana lysis Li n ki ng M u lti ple Two-D imensional Scatter Plots One of the more exciting new graphical procedures involves electronically connect ing many two-dimensional scatter plots. • A final maximization result will provide us with an interpretation of eigenvalues. Machine . The constant of proportionality is the product of the variances, which, in turn, is pro portional to the product of the squares of the lengths (n - 1)sii of the di . How ever, a negative value of c creates a vector with a direction opposite that of x. For any two deviation vectors di and d k , n (3-6) d i d k == ◆ (xji - xi) (xj k - xk) j=l Let fJi k denote the angle formed by the vectors di and d k . = (e) I AB I = I A I I B I (f) I c A I = c k I A I , where c is a scalar. We first consider a scaled version B = D◆ 1 12 PD ◆ 1 12 of the correspon dence matrix P . Xjp Xj p - Xp Xp have the same sample covariance matrix S as the original observations. a'X + (a' X+ Example 4.4 X + ... Johnson.--5th ed.) Repeat the princi pal component analysis outlined in Exercise 8.18 for the men. Exercises V 112 pV 112 = I. 1 I NTRODUCTION A generalization of the familiar bell-shaped normal density to several dimensions plays a fundamental role in multivariate analysis. From Example 3.9, 3 3 and S = X= -◆ OJ [◆ iI [◆] Consequently, using (3-36), we find that the two sample means for the derived observations are Sample mean of b ' X = b' X = [2 2 - 1] Sample mean of c' X = c' X = [1 - 1 3] Using (3-36), we also have Sample variance of b'X = b ' Sb = [2 2 - 1] [◆] [◆] = 17 (check) (check) [◆] ◆] [-l [l] [i ◆] [-◆] -n [= [2 2 - 1] Sample variance of c' X = c' Sc =3 3 -2 1 1 2 =3 (check) = 13 (check) = [1 -1 3] - = [1 -1 3] Chapter 3 Sample covariance of b ' X and c ' X = b ' Sc Exe rcises 1 45 [-◆ ◆] [-◆] [-!] = 3 = [2 2 - 1] = [2 2 - 1] -2 1 1 2 ◆ (check) As indicated, these last results check with the corresponding sample quan tities computed directly from the observations on the linear combinations. The diagonal matrix A h as entries A. 8 671 and 1 The normalized eigenvectors and are obtained by1, 2solving and scal_1ing the results su[ch1.0t7143hat - 2.8671 1. 97 1 (x i) Density . U1 ._,... 1 + 2 i=l has rank r 1 + 1 and generates the unique projection o f y on the space spanned by the linearly independent columns of Z. Then the expression in (9-2) can be written X - where JL = LF + L* = E = LTT'F + LT and F* E = L*F* + = T'F = 0 E (9-7) Since E(F*) and Cov (F*) = = T' E(F) T' Cov (F)T = T'T = (m XI m) it is impossible, on the basis of observations on X, to distinguish the loadings L from the loadings L* . (See [13] for a description of this test and tables of critical values.) Example 7.7 {Residual plots) Three residual plots for the computer data discussed in Example 7.6 are shown in Figure 7.3. The sample size n = 7 is really too small to allow definitive judg ments; however, it appears as if the regression assumptions are tenable. 16. If you are author or own the copyright of this book, please report to us by using this DMCA report form. 2 The Geometry of t h e Sample 117 and We note that .X 1 1 and d 1 = [◆] y1 - .X1 1 are perpendicular, because (:Xl l)'(yl - :Xl l) = [2 2 2] A similar result holds for x2 1 and d 2 = - = 4-6+2 = 0 y2 - x2 1. Suppose, then, that the data have not yet been observed, but we intend to col lect n sets of measurements on p variables. Let 0 = (0, 0, . Supplement 1 2A Data M i n ing 733 7. In cases where it is possible to capture the essence of the data in three dimensions, these representations can actually be graphed. A drift like this would be very rare if the year-to-year values were independent observations from the same distribution. Should they be the same? , e m] = 0 = I. For the standardized variables, [1..40 1..40] [1..20 1..20] and [.5 ..46] With p 1, [.86,.28] [.54, .74] Vv = 1 X i)) = a-}!'f . 64 23 2.94 2.55 2.34 2.21 2. Bear 4 Bear 3 Bear 2 Bear 1 - ...c= ◆ gf 11) ◆ 1 80 - 1 80 - {j ◆ 1 60 OJ) 1 60 ◆ 11) 1 40 - 140 1 2 3 4 I 1 5 Year r I 2 Bear 5 I 3 I 4 1 80 - {j ◆ 1 60 OJ) ◆ 11) I ◆ ◆ I 2 I 3 Year Year Bear 6 Bear 7 1 80 - 1 80 1 60 140 - I 1 5 / I 4 I 1 80 - {j ◆ 1 60 OJ) ◆ 11) 1 40 - 5 I 1 / I 2 I 3 I 4 {j gf ◆ 11) 1 60 ◆ 11) 140 - 140 1 2 3 Year 1.15 - ◆ 1 60 - OJ) I 5 Year Figure {j 4 5 I 1 r I 2 I 3 I 1 I / 2 I 3 Year - 1 80 OJ) - 1 40 - - {j - I I 4 5 Year I n d ivi d u a l g rowth cu rves for l e n gth for fem a l e g rizzly bears. We first define the overal mean pXp = I 2 = · · · = I g = I. The distinction between • divisors can be ignored with principal component solutions. The dual substitution of S and the approximate maximum likelihood estimator into the test statistic of (9-39) does not affect its large sample properties. The only require ments are that the parent population, whatever its form, have a mean IL and a finite covariance I. regarded as p points in n-dimensional space. If an arbitrary matrix A has the same number of rows and columns, then A is called a square matrix. Section 2.2 Example 2.2 Some Basics of Matrix and Vector Al gebra 55 (Identifying l i nearly i ndependent vecto rs) Consider the set of vectors Setting implies that c1 + c2 + c3 = 0 2c1 - 2c3 = 0 c 1 - c2 + c3 = 0 with the unique solution c 1 = c2 = c3 = 0. Of the 75 applicants, 33 were classi fied as Good credit risks (with negative expected costs) . [ ] change i n r a t e 1.3005 -..74205395 1 Factor 1: of extrusi.on . That is, Yj l e l + Yj 2 e 2 + . From the model in (9-4) , (X - ◆-L) (X - IL) ' = (LF + e) (LF + e) ' = (LF + e) ((LF) ' + e') = LF(LF) ' + e (LF) ' + LFe ' + ee ' so that I = Cov (X) = E(X - IL) IL)' (X - = LE (FF') L' + E (eF') L' + L E (Fe ') + E (ee ') = LL' + 'II according to (9-3). File loading please wait... It also plays a basic role Let be diisstdiribstutriebdutasedNasp(x , wherwiteh x denot0.eThen s t h e chi s q uar e diThestriNbutp(ion withdistrdegributeioesnofasfsriegedom. Gulledge, Jr. "Use of the Correlation Coefficient with Normal Probability Pilots." The American Statistician, 39, no. Consequently, the soft ware package may not run unless the Heywood case option is selected. Arguments over the psychological interpretations of several early studies and the lack of powerful com puting facilities impeded its initial development as a statistical method. 8425 6291 6778 4969 5922 4825 7307 6019 8.25. g "" n· X · ..£.J i =1 x=g-- = L ni i l l nt g 2: xi j i =1 jL =1 g ni L i= 1 =1 whitrainchinigs stehte. Finally, for 5 -4 0 = (A - 5) 2 - (-4) 2 the eigenvalues satisfy s= ' = (A - 9) (A - 1) -4 5 [ ] and we determine the eigenvalue-eigenvector pairs A 1 = 9, e1 = [1/ v2, - 1/ vl] and A 2 = 1, e 2 = [1/vl, 1/vl] . A second drawback is that the (automatic) selection methods are not capable of indicating when trans formations of variables are useful. At first glance, none of these utilities appears to be similar to any other. Let X and Y be random matrices of the same dimension, and let A and B be conformable matrices of constants. Even so, we suggest first visually inspecting the data whenever possible. Deleting the outlier and the cases corresponding to the older paper and ad j usting the ranges of the remaining observations leads to the scatter plots in Figure 1 . fForrom thobse ervations from whiandchThesobshowsappar act u ervations from the confusion matrix has the form Predicted membership Actmemberual ship where number of items ◆orrectly clas ified as items number of i t e ms mi s c l a s i f i e d as i t e ms number of i t e ms ◆or r e ct l y cl a s i f i e d number of i t e ms mi s c l a s i f i e d The apparent error rate is then APER which is recognized as the of items in the training set that are misclas ified. Given the plot in Figure 6.5, this As s u m i n g t h at t h e pr o f i l e s ar e par a l e l , we can t e s t f o r pr o f i l e s . Fish caught by the same fisherman live alongside of each other, so the data should provide some evidence on how the fish group. We adopt this terminology because the bulk of this book is devoted to proce dures designed for analyzing samples of measurements from larger collections. New York: John Wiley, 1973. 2 N (z'o /3 , o- z0 (Z' Z) - 1 z 0) (z'o P - z'o{J)j \1o-2 z'o (Z ' Z) - 1 z0 v?J;? 85 1 . + (x n - x) (x n - x)' = c' c n-1 [ ] 1 42 Chapter 3 Sa m p l e Geometry and Ra ndom Sa m p l i ng or (3-33) Sample variance of c ' X = c ' S c Equations (3-32) and (3-33) are sample analogs of (2-43). Radiation 31 32 33 34 35 36 37 38 39 40 41 42 .10 .20 .11 .30 .02 .20 .20 .30 .30 .40 .30 .05 Source: Data courtesy o f J. We do not assume the reader is familiar with matrix algebra. 864 2.837 2.813 2.793 2.775 2.759 2.744 2.732 2.720 2.710 2.700 2.692 2.684 2.676 2.669 2.663 2. Now use the result in Part a. U'1 w ._,... An example of an orthogonal matrix is 1 1 21 2 1 2 A= 1 - 21 21 2 1 2 2 1 21 2 1 -2 1 2 1 2 1 21 21 2 Clearly, A = A', so AA' = A' A = AA. (See [10] and [22].) Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimen sions. 1 A plot of the data matrix X as n = 3 poi nts i n p = 2 space. The covariance matrix of X (l) is I 1 1 , that of X (2) is I22 , and that of elements from x (l) and X (2) is I 1 2 (or I 2 1). Ray. If we assume that the coordinate variables vary independently of one another, the distance from P to Q is given by d(P , Q) = I (x i - YI) 2 'I S1 1 + (x2 - Y2) 2 S22 (1-15) The extension of this statistical distance to more than two dimensions is straight-forward. •I ... The first principal component, which explains 96% of the total variance, has an interesting subject-matter interpretation. , d ◆ should behave like a chi-square random variable. Given the outcomes Y and the values of the preftictor variables Z with full col umn rank, we determine the least squares estimates ƒJ (i) exclusively from the obser vations Y(i) on the ith response. In the spirit of the principal component solution, consideration of the estimat ed eigenvalues Ai, Ai, . What wil happen as isLetntacrnteasdiesst?ances and ecomment wher e (T heA/sand t h e ; i e o r s of t h e mat r i x e;'s arShowe thePreiogperenvaltieuses(1)and-(4)asofsotciheatseqduarnore-mroalotizmated eirigxenvect i n (2 2 2). , ZP . If the ob servations were generated from a multivariate normal distribution, each bivariate distribution would be normal, and the contours of constant density would be ellipses. We shall simply summarize the pertinent large sam ple results. , ,u p , example, in the p-dimensional multivariate normal case, lT 1 b lT 1 p , lT2 2 ' . 94 1 .85 1 .77 1 .72 1 . In two dimensions, we can often sketch the axes of the mean-centered el lipse by eye. The value of rik remains unchanged if the measurements of the ith variable are changed to yjl = a xj i + b, j = 1, 2, . Consider the air-pollution data listed in Table 1 .5. In general, the ith point yi = [x 1 i , x2 i, . In traditional statistical applications, sample sizes are relatively small, data are carefully collected, sample results provide a basis for inference, anomalies are treat ed but are often not of immediate interest, and models are frequently highly 5 SAS Institute's data mining program is currently called Enterprise Miner. Consider the Euclidean distances from the point Q to the point P and the origin 0. The icons in the figure represent various activities in the data mining process. Do the data seem to be normally distributed? 2.0 2.5 3.0 3.5 Year 4.0 4.5 5.0 Figure 1 . (See Figure 2.5 .) = L x l cos (O) (2-8) I (2-9) Matrices A matrix is any rectangular array of real numbers. The ordered distances and the corresponding chi-square percentiles for p = 2 and n = 10 are listed in the following table: 1 1 2 3 4 5 6 7 8 9 10 d Jn 1 qc ,2 1Q .59 .81 .83 .97 1.01 1.02 1.20 1.88 4.34 5.33 .10 .33 .58 .86 1 .20 1.60 2.10 2.77 3.79 5.99 c 1 - 2) 1 86 Chapter 4 The M u ltiva r i ate Norm a l D istri bution • • • • • • • • • • 1..-..Lr-----L--'---'---◆ 0 1 Figure 4.7 2 3 4 5 6 qc,2 ((j- !) / 1 0) A ch i-sq u a re plot of the ordered d i stances in Exa m p l e 4. 095445 1 2 0 .4472 1 3 60 P r > ITI 0.4286 0.0208 DF 1 3 4 Pr > F 0. The two specimens (9 and 16) with large squared distances stand out as clearly different from the rest of the pattern in Figure 4.9. Once these two points are removed, the remaining pattern conforms to the expected straight-line relation. This may be done to validate assumptions or to reinforce prior convictions. Although the coeffi cients and the correlations can lead to different rankings as measures of the impor tance of the variables to a given component, it is our experience that these ranking s are often not appreciably different. Also, by Result 4.8, Ve and Vk have covariance matrix (e{;ek) I = (0) I = 0 if e ◆ k. I. , k, and eiej = 0 for i ◆ j. Can the associations in the data be well represented in one dimension? Then the ex pected value of X, denoted by E(X) , is the n X p matrix of numbers (if they exist) E(X) (2-23) where, for each element of the matrix, 2 2 If you are unfamiliar with calculus, you should concentrate on the interpretation of the expected value and, eventually, variance. The interested reader will find a proof in [6] , Chapter 8 . London: Chapman and Hall, 1996. On the other hand, very large samples in variably produce statistically significant lack of fit. The fundamental scalar relation about the existence of an inverse number a - 1 such that a -1 a = aa - 1 = 1 if a ◆ 0 has the following matrix algebra extension: If there exists a matrix B such that · • B A = A B = I (k X k) (k X k) (k X k) (k X k) (k X k) then B is called the inverse of A and is denoted by A- 1 . Here a large value of (xj - x) ' S - 1 (xj - x) will suggest an unusual observation, even though it cannot be seen visually. rHatibutfiioens"ld, aandsoci"outologicsomest, sur" andmaletshandeir lefevelmals eofs wer"pasesiasonatkedet"oandresp"compani oenatfoleo"wilonve.g quesRecenttionsl,yusmaringried ond t o t h 8-point scale in the figure below. Given the foregoing component coefficients, the first principal component appears to be essentially a weighted average of the first four variables. Our purpose now is to develop a "statistical" distance that accounts for differ ences in variation and, in due course, the presence of correlation. J [ ] = 2 8-9 3 [-1, 6 5 Before the sample is selected, T has the distribution of a (3 - 1)2 ,3_ = 4F2, l (3 2) F2 2 random variable. 4 The deviation (column) vectors are d1 = [-2, 1, 1] , d2 = [1 , 0, - 1] , and d3 = [0, 1, -1]. Our partial proof of Result 4.2 indicates how the linear combination definition of a normal density re lates to the multivariate density in (4-4) . Let P = (x1 , x2 , . The results are given in Table 12. z (1) z (2) U1 . The pattern of residuals may be funnel shaped, as in Figure 7.2(b), so that there is large variability for large y and small variability for small y . p = (V 1 12) - 1 I (V 112) - r , V 1 12 p X I p Chapter 2 2.24. Example 1 .7 (Looking for g roup structu re i n th ree d i mensions) Referring to Example 1 .6, it is interesting to see if male and female lizards oc cupy different parts of the three dimensional space containing the size data. Cluster the countries into groups using several values of K. The second factor contrasts the chemical stocks with the oil stocks. A is square, A I = I A22l l A 11 - A 12A2i A21 I for I A22l # 0 = IA 11 I A 22 - A 21 A !i A 12 I for IA 11 I # 0 Hint: Partition A and verify that [OI, -A 12I A2◆] [AA2111 AA2212] [-A22!1 A21 OI J [Au - A1,2A2"◆A2 1 A22 J Take determinants on both sides of this equality. Filliben, J. 176 - . Corresponding interpretations hold for the other scatter plots in the figure. Thus, for this example with m = 1, it is possible to get a unique numeri cal solution to the equations I = LL' + '1'. (See Exercise 1.15.) Track records from many nations were used to develop an index of performance for both male and female athletes. We would also like to give special thanks to Wai Kwong Cheang for his help with the calculations for many of the examples. 9 into a K-means clustering program. • A method for calculating the A ' s and e ' s is described in Supplement 2A. Examine these standardized values for large or small values. • • • • • • . = o1 2 2 However, partitioning X and I as I [I 1 1 2 X= [1!-l I = [- - - - -;- - J- = (I1X212) : (IlX2l)] we(X s,eXe th) atandX1X=3 ar[e◆in:]dependent Therefore, haveby Rescovarulitance matThisriixmIpl1i2es=X[3◆isJin. Berthold, M . • • • 0E8 pling distribution of the likelihood-ratio test statistic A . • • • Definition 2A.4. The space of all real m-tuples, with scalar multiplication and vector addition as just defined, is called a vector space. The data that determine distance will, however, remain fixed. It is hoped that this will make interpretation easier. The cor r e l a t i o ns and c a n hel p s u ppl y meani n gs f o r t h e canoni c al vartionsiatbetes. , x n 1] are the n measure ments on the first variable. - - - - - - - - - - - - - - .73 ◆· l == ---- - --- Specific variances Estimated factor loadings Specific variances Estimated factor loadings 1 - h? (If the hy pothesis H0 : B == 0 were not rej ected, then all simultaneous intervals would in clude zero.) The B onferroni simultaneous intervals also cover zero. (New York: NYU Press, 1978). 4 5 4 3 • Considering data in the form of arrays facilitates the exposition of the subject matter and allows numerical calculations to be performed in an orderly and efficient manner : The efficiency is twofold, as gains are attained in both (1) describing nu merical calculations as operations on arrays and (2) the implementation of the cal culations on computers, which now use many languages and statistical packages to perform array operations. (See Chapter 8.) For the principal component solution, the estimated loadings for a given factor do not change as the number of factors is increased. A ppendix Table 1 Standard Normal Probabilities Table 2 Student ' s t-Distribution Percentage Points Table 3 x2 Distribution Percentage Points 748 Table 4 F-Distribution Percentage Points (a = .10) Table 5 F-Distribution Percentage Points (a = .05) Table 6 F-Distribution Percentage Points (a = .01) 749 Appendix TABLE 1 STAN DARD N O RMAL PRO BAB I LITI ES P[Z ◆ z] 0 z z .00 .01 .02 .03 .04 .05 .06 .07 .08 .09 .0 .1 .2 .3 .4 .5 .6 .7 .8 .9 .5000 .5398 .5793 .6179 .6554 .6915 .7257 .7580 .7881 .8159 .5040 .5438 .5832 .6217 .6591 .6950 .7291 .761 1 .7910 .8186 .5080 .5478 .5871 .6255 .6628 .6985 .7324 .7642 .7939 .8212 .5120 .5517 .5910 .6293 .6664 .701 9 .7357 .7673 .7967 .8238 .5160 .5557 .5948 .6331 .6700 .7054 .7389 .7703 .7995 .8264 .5199 .5596 .5987 .6368 .6736 .7088 .7422 .7734 .8023 .8289 .5239 .5636 .6026 .6406 .6772 .7123 .7454 .7764 .8051 .8315 .5279 .5675 .6064 .6443 .6808 .7157 .7486 .7794 .8078 .8340 .5319 .5714 .6103 .6480 .6844 .7190 .7517 .7823 .8106 .8365 .5359 .5753 .6141 .6517 .6879 .7224 .7549 .7852 .8133 .8389 1 .0 1.1 1 .2 1.3 1 .4 1 .5 1.6 1 .7 1.8 1 .9 .8413 .8643 .8849 .9032 .9192 .9332 .9452 .9554 .9641 .9713 .8438 .8665 .8869 .9049 .9207 .9345 .9463 .9564 .9649 .9719 .8461 .8686 .8888 .9066 .9222 .9357 .9474 .9573 .9656 .9726 .8485 .8708 .8907 .9082 .9236 .9370 .9484 .9582 .9664 .9732 .8508 .8729 .8925 .9099 .9251 .9382 .9495 .9591 .9671 .9738 .8531 .8749 .8944 .9115 .9265 .9394 .9505 .9599 .9678 .9744 .8554 .8770 .8962 .9131 .9279 .9406 .9515 .9608 .9686 .9750 .8577 .8790 .8980 .9147 .9292 .941 8 .9525 .9616 .9693 .9756 .8599 .8810 .8997 .9162 .9306 .9429 .9535 .9625 .9699 .9761 .8621 .8830 .9015 .9177 .9319 .9441 .9545 .9633 .9706 .9767 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 .9772 .9821 .9861 .9893 .9918 .9938 .9953 .9965 .9974 .9981 .9778 .9826 .9864 .9896 .9920 .9940 .9955 .9966 .9975 .9982 .9783 .9830 .9868 .9898 .9922 .9941 .9956 .9967 .9976 .9982 .9788 .9834 .9871 .9901 .9925 .9943 .9957 .9968 .9977 .9983 .9793 .9838 .9875 .9904 .9927 .9945 .9959 .9969 .9977 .9984 .9798 .9842 .9878 .9906 .9929 .9946 .9960 .9970 .9978 .9984 .9803 .9846 .9881 .9909 .9931 .9948 .9961 . x x The following summarizes these concepts: 1 54 Chapter 4 The M u ltivariate Norm a l Distri bution A contour of constant density for a bivariate normal distribution with o-1 1 = o-2 2 is obtained in the following example. Simple, yet elegant and effective, methods for displaying data are available in [29] . Let A and B be k X k square matrices. Using this normal distribution, we obtain P[l Ai - Ai I z(aj2)Ai V2711 J = 1 - a. 81 - .86 .47 - .1 9 - .49 . ◆ 1 00 - ◆ 1 50 - ◆ ◆ B ear 2 I n divid u a l g rowth cu rves for weight for fe male g rizzly bears. Another important new graphical technique uses software that allows the data analyst to view high-dimensional data as slices of various three-dimensional per spectives. 88 1 .85 1 . 1 N (O , 1) . I - \ ... 29832868 9. Using the inner product, we have the natural extension of length and angle to vectors of n components: Lx = length of x =◆ (2- 5) 54 Chapter 2 M atrix Algebra and Ra ndom Vectors Since, again, cos (0) x' y = 0. + = • = e = = A A = = A A = A A + A = 0, an A A A = + = = (7-39) + = 0 + = e Result 7.10. Cryer . 1 4000000 0.560 1 587 1 0.42804465 9.49000000 - - - - - - - - - X3 - - - - - - - - SD Mean 1 . The array x is a single col umn with p rows. When the number of char acteristics p is large, the large number of scatter plots p(p - 1)/2 may prevent viewing them all. mXn ( ) ( ) ( ) = Definition 2A.24. 468 Chapter 8 Principa l Components 8.8. Use the results in Example 8.5. (a) Compute the correlations ryn zk for i = 1 , 2 and k = 1, 2, . obtain four or five clusters. Consider, then, the following seven pairs of measurements on two variables: (x1) : Variable 2 (x2) : Variable 1 3 5 4 5.5 2 6 4 7 8 10 2 5 5 7.5 These data are plotted as seven points in two dimensions (each axis represent ing a variable) in Figure 1.1. The coordinates of the points are determined by the paired measurements: (3 , 5) , (4, 5.5) , . The two configurations appear to be quite similar, but a quan titative measure would be useful. 00000000 Statistic Wil ks' La m bda P i l l a i's Tra ce Hotel l i n g-Lawley Trace Roy's G reatest Root Pr > F 0.0625 0.0625 0.0625 0.0625 Den D F 2 2 2 2 Num DF 2 2 2 2 F 1 5 . 9 8 13.34 27.69 1 .76 3.62 131 .48 14.72 1.90 31 .36 4.01 148.22 14.74 30.64 3.83 1.81 145.27 13.29 27.89 1 .76 3.60 131.58 13.13 27.38 1.79 3.62 128.65 13.25 27.67 1 .76 3.64 132.50 15.11 31.32 1.88 3.89 157.77 27.73 13.31 1 .76 3.55 131.57 13.29 27.94 1.77 3.61 130 .63 13.22 27.91 1 .78 3.55 131 .20 14.07 30.07 1.79 3.77 139.27 32.65 15.23 3.84 1.83 149.90 13.56 28.58 1.79 3.67 131 .50 13.20 27.43 3.53 1 .73 128.22 13.20 27.53 1.75 3.59 130.55 16.28 34.71 4.24 2.02 161.83 Source: IAAF/ATFS Track and Field Statistics Handbook for the 1 984 Los Angeles Olympics. Moreover, eiej = 0 for i ◆ j. 1 At this point, "independently" means that the x2 measurements cannot be predicted with any ac curacy from the x 1 measurements, and vice versa. Bellman, R. Then there exist an m X m orthogonal matrix U and a k X k orthog onal matrix V such that A = UAV' where the m X k matrix A has (i, i) entry Ai 0 for i = 1, 2, . ' e p of S: 1. Proving that a' X is normally distributed if X is multivariate normal is more difficult. Knowing the position on a line along the maj or axes of the cloud of points would be almost as good as know ing the three measurements Mass, SVL, and HLS. (a) Construct the sample covariance matrix S for the census-tract data when X5 = median value home is recorded in thousands of dollars. [9 -2 6] -2 Chapter 2 2.7. LetDetA beermasingievtenheineigExerenvalciusees2.and6. 8 6 weis lefsailthtoanre.j0e5ctththereeadequac y of,tshoemequadrevidaencetic fittathat the .quadr 01. Let n For t h e model i n wi t h conf i dence at l e as t = - a) , belongs to 5.4) Tk . 7 81zP 1 .631 .060zl2) .944z?) and 1)) . P. The test data set is ordinarily used once at the end of the modeling process to ensure an unbiased assessment of model per formance. h o ee s s ,r s e i ,p .p line having slope 1 and that passes through the origin. Fortunately, efficient computer programs now exist that en able one to get these estimates rather easily. (Proof. + == X + Nq(AJL, AIA ') . In spite of these shortcomings, covariance and correlation coefficients are routinely calculated and analyzed. Additional Problems in Li near Reg ression We shall briefly discuss several important aspects of regression that deserve and re ceive extensive treatments in texts devoted to regression analysis. Exxon 5. 8.6. Data on x 1 = sales and x2 = profits for the 10 largest U.S. industrial corpora tions were listed in Exercise 1.4 of Chapter 1. 158 2. 5 89 [..558962] [..556289] . - 1) - p I + 1 and F - (1 -.7771.7771) ((12(12()1(4)- -313 ++1)1)/22 - 1.34 ((21 1(2)(1)(4)- 31- 3++1)1)===3 14 = SSPres = I SSPint SSPres = (S 1) V1 = v2 == v1 = e v2 == Section 6 . • A A Large sample variances and covariances for the maximum likelihood estimates e ij ' l/Ji have been derived when these estimates have been determined from the sam ple covariance matrix S . Groupings of items will manifest themselves in this representation. X1 ◆ X2 + ◆ X3 • X1 , X2 , X2 (c1 1 x1 + c 1 2 x2 + . Starting with the sample cor 1..0505 1..0505 ..452269 ..460267 [J ..560269 ..442267 1..0926 1..0926 r S1 2 = S2 1 r S22 . 289.6 289.0 288.2 288.0 288.0 288.0 290.0 289.0 289.0 289.0 287.0 287.6 288.4 290.2 286.0 287.0 287.0 290.0 287.0 289.1 288.3 289.0 290.0 288.3 288.7 290.0 288.6 288.2 286.0 290.0 288.7 287.0 288.0 288.6 288.0 289

) (X2 - JL2) (Xp - JLp) (Xp - JLp) 2 (Xp - JLp) (XI - JLI) (Xp - JLp) (X2 - JL2) E(XI - JL I) 2 E(X1 - JL 1) (X2 - JL2) E(X2 - JL2) 2 E(X2 - JL2) (XI - JL 1) E(X1 - JL 1) (Xp - JLp) E(X2 - JL2) (Xp - JLp) E(Xp - JLp) (XI - JL I) E(Xp - JLp) (X2 - JL2) E(Xp - JLp) 2 Section 2.6 Mean Vectors and Cova riance Matrices 71 or I = Cov (X) = Example 2. Show that, if X and z are independent, then each component of X is 3.16. (k X k) = = = = • Result 2A.l0. X3 0 Cent. WIC H E RN Texas A&M University Prentice Hall ------ PRENTICE HALL, Upper Saddle River, New Jersey 07458 Library of Congress Cataloging-in-Publication Data Johnson, Richard Arnold. , n Xj q The name rep ea te d measures stems from the fact that all treatments are administered to each unit. To illustrate scalar multiplication, take c1 = 5 and c2 = - 1 .2. Example 2. These coefficients can be set out in the rectangular array . ma (8-25)], Z = [Z{ll i z (zJ] 1 z (1) ' z (21) ' : = 1 Zn( ) ' Zn(2) ' and the sample canonical variates become z z and wher e The s a mpl e canoni c al cor r e l a t i o ns are fchanged ected byandthemaystandarbe caldiczulataitoen.d, fTheor stacorndarreldaitziedonsobsdiesprvlaatyedionsi,nby substiurteimaing p un obsBzervfatorB,ionsand. = (Z ' Z)-1Z'Y E(P) = P = a-2(Z' Z)-1 e E(e) = 2 (e) = a-2[I - Z (Z' Z)-1Z'] = a-2[I - H] Also, E(e'e) = ... In this sense, S captures the orientation and size of the pattern of scatter. The additional uncertainty in forecasting Y0 , 1which is represented by the extra term s 2 in the expressi o n s 2 (1 + z0(Z ' Z) - z 0) , comes from the presence of the unknown error term s0 • Example 7 . Thesthe feorcmorerecaslatieonsis diarreecustlyualrellyataleldctloostehetosmzeralole. , eP . Getting Started ◆ Chapters 1-4 ◆ Inference About Means Classification and Grouping Chapters 5-7 Chapters 11 and 12 Analysis of Covariance Analysis of Covariance Structure Structure Chapters 8-10 Chapters 8-10 I I For most students, we would suggest a quick pass through the first four chap ters (concentrating primarily on the material in Chapter 1 ; Sections 2.1, 2 . Thus, • • • . If the data are naturally chronological, a plot of the residuals ver sus time may reveal a systematic pattern. ·: : ... 3 Meth ods of Esti mation 495 component solution. Furthermore, when data can be treated as a random sample, statistical inferences are based on a solid foundation. canoni Many tcimalesvarthiiastiedsentandificthateioorn iigsiainaldedvarbyiacom- bles. I LL' I i LL' Comment. The information contained in the single-variable dot diagrams can be used to calculate the sample means .X1 and .X2 and the sample variances s1 1 and s22 . (d) Using the values for the first two principal components, plot the data in a two-dimensional space with 5\ along the vertical axis and y2 along the hor izontal axis. Now, c' = Z (Z ' Z) c . 588 S = 1 . Draw the den drograms and compare the results. 158 - . 68 3.29 16 4.49 3 .63 3 .24 3 .26 3.11 3 .00 3.18 3 . If the data arise from a normal population, the pairs (q(j) , x u)) will be approximately linearly related, since o-q (j) + J.L is nearly the ex pected sample quantile. (c) Construct a chi-square plot of the ordered distances in Part a. The straight-line distance between two arbitrary points P and Q with coordi nates P = (x 1 , x2 , . Regardless of their origin, multivariate techniques must, invariably, be im plemented on a computer. T 2 = 3[8 - 9, 6 - 5] _ _ • The next example illustrates a test of the hypothesis H0 : IL = ILo using data col lected as part of a search for new diagnostic techniques at the University of Wiscon sin Medical School. Section 9.2 The O rthogonal Factor Model 479 and that F and e are independent, so Cov (e, F) = E (eF') = 0 (pXm) These assumptions and the relation in (9-2) constitute the orthogonal factor mode/ . In this case, lift for the first percentile group can be calculated as the ratio of the num ber of Good predicted by the model to the number of Good from a random as signment or Lift = ◆◆ = 1 .33 The lift value indicates the model assigns 10/299 = .033 or 3.3% more Good Supp lement 1 2A Data M i n ing 737 Figure 12.29 C u m u l ative l i ft chart for the log istic reg ression classifier. Detdeteerrmmiinnee the sqanduares-hroowot matthatrix using the matrix in Exercise 2.3. Also, (See Result 2A.15) Using the matrix A' A y = Ax Hint: 2.17. 1 (1994), 237-257. With the x'y notation, we may express the length of a vector and the cosine of the angle between two vectors as Lx = length of x = Vxi + x◆ + + x◆ = ◆ x'y cos (B) = ◆:- ◆: v x' x v y'y · · · " " Definition 2A.ll. The points indicated by the cir cled locations in the figure are outliers-values that are too large relative to the rest of the observations. " Goin ◆ in the other direction, given the estimated loadings L z and specific variances '\}I z obtained from R, we find that the resulting maximum likelihood estimates for a factor analysis of the covariance matrix [(n - 1) /n J S are L = y 1/2 iz and q, = y 1/2 q, z y 1/2' or where a-ii is the sample variance computed with divisor n. The data for the pair of variables x 1 = employees (j obs) and x2 = profits per employee (productivity) are graphed in Figure 1 . You should be aware that there are linear dependencies among the residuals from a linear regression analysis, so the last eigenvalues will be zero, within rounding error. , ei p] also merits inspection. Since this is analogous to testing H0 : IL = /L o versus H1 : IL # IL o [see (5-7)], we see that the confidence region of (5-18) consists of all p0 vectors for which the T2 -test would not reject H0 in favor of H1 at significance level a. toLetp treatments for populationsand1 and 2, respectively. Departures from the assumptions of the model are typically indicated by two types of phenomena: (a) A dependence of the residuals on the predicted value. ' p PY P Xk ◆ V lTkk are the correlation coefficients between the components }j and the variables Xk . See [1]. The univariate central limit theorem also tells us that the sampling distribution of the sample mean,X, for a large sample size is nearly normal, whatever the form of the underlying population distribution. [ ] Result 3.5. The linear combinations · ·· · b'X = b1 X1 + b2 X2 + + bP XP c'X = c1 X1 + c2 X2 + · · + cP XP have sample means, variances, and covariances that are related to x and S by Sample mean of b' X = b' x Sample mean of c' X = c' x Sample variance of b ' X = b ' Sb (3-36) Sample variance of c' X = c' Sc • Sample covariance of b'X and c ' X = b ' S c Section 3.6 Example 3.1 3 Sa m p l e Va l ues of Linear Com b i n ations of Va riables 1 43 {Means and covaria nces fo r l i near combi nations) We shall consider two linear combinations and their derived values for the n = 3 observations given in Example 3.9 as [◆:] [◆:] Consider the two linear combinations b 'X = [2 2 - 1] and c'X = [1 - 1 3] = 2Xl + 2X2 - X3 = X1 - X2 + 3X3 The means, variances, and covariance will first be evaluated directly and then be evaluated by (3-36). 5 Tract 1 2 3 4 5 6 7 8 9 10 11 12 13 14 C E N S US-TRACT DATA Total population (thousands) 5.935 1.523 2.599 4.009 4.687 8.044 2.766 6.538 6.451 3.314 3.777 1 .530 2.768 6.585 Median school years 14.2 13.1 12.7 15.2 14.7 15.6 13.3 17.0 12.9 12.2 13.0 13.8 13.6 14.9 Total employment (thousands) 2.265 .597 1 .237 1.649 2.312 3.641 1.244 2.618 3.147 1.606 2.119 .798 1.336 2.763 Health services employment (hundreds) 2.27 .75 1.11 .81 2.50 4.51 1.03 2.39 5.52 2.18 2.83 .84 1 .75 1.91 Median value home ($10,000s) 2.91 2.62 1 .72 3.02 2.22 2.36 1 .97 1.85 2.01 1 .82 1.80 4.25 2.64 3.17 Note: Observations from adj acent census tracts are likely t o b e correlated. Boston Edison Co. (2) Arizona Public Service (1) 2 8 5 5 Central Louisiana Electric Co. (3) Commonwealth Edison Co. (4) 5 5 Consolidated Edison Co. (NY) (5) 1 5 Figure 1 . In this geometrical representation, we depict yr , . 30 1 23 1 55 4.43000000 N 10 10 To test for interaction,Iwe computI e 275.7098 A* + I 354.7906 = _7771 For (g - 1) (b - 1)(1 1,- A*) (gb(n - 1) - p + 1)/2 F = A* (l (g - 1)(b - 1) - pl + 1)/2 hasgb(nan- exact- pF-+dis1trd.ibf.utioen wi[1]t.h) For ourI (gexampl - 1)(eb. Even those readers with a good knowledge of matrix algebra or those willing to ac cept the mathematical results on faith should, at the very least, peruse Chapter 3, "Sample Geometry," and Chapter 4, "Multivariate Normal Distribution." Our approach in the methodological chapters is to keep the discussion direct and uncluttered. Suppose the data, in tabular form, are Variable 1 (dollar sales) : Variable 2 (number of books) : 42 52 48 58 4 5 4 3 Using the notation just introduced, we have x1 1 = 42 x2 1 = 52 x3 1 = 48 x4 1 = 58 x 1 2 = 4 x2 2 = 5 x3 2 = 4 x4 2 = 3 and the data array X is 42 52 X= 48 58 with four rows and two columns. Are there any "wild" observations that should be checked for accuracy? We notice in Example 8.3, however, that the correlation coefficients displayed in the table confirm the interpretation provid ed by the component coefficients. 1 2 (Effective classification with fewer variables) on discsroimn inantonftuhncreteiosnspeci, Fiesshofer iris prfloewersents.edSeean analTablye sInis hiofsExerdatpioaneerccoliseliencgteword byk Ander Let the clas es be defined as The fol owing four variablseespalwerleengtmeash, ured frosmepal wipldantths of each species. = = o) /3 (4.9 ' = 7.2, = = = = = = 7.2. n 1) r e'e n r I. The first component is a roughly equally weighted sum, or "index," of the five stocks. We can write X= (n Xp) X ◆ 1st (multivariate) observation (3-1) ◆ nth (multivariate) observation The row vector xj , representing the jth observation, contains the coordinates of a point. It is then not possible to represent the data well in fewer than p dimensions. 4 5625 . , g. Let A and B be k X k matrices and c be a scalar. a ni X p If not, we let P = [e 1 , , eq] be the eigenvectors of I corresponding to nonzero eigenvalues , Aq] · Then we replace X by P ' X which has a full rank covariance matrix P ' I.P. . As examples, consider measurements of variables representing consumption and income or variables representing personality traits and physical characteristics. ,consYm andiderathsienprgleoblseetmofofprmodel i n g t h e r e l a t i o ns h i p bet w een r e e d i c t o r var i a ble s z , , Eac h r e 1 sponse is assumed to fol ow its own regres ion. This situation is illustrated in Figure 7.2(c). + (xp - Yp) 2 sPP (1-16) All points P that are a constant squared distance from Q lie on a hyperellipsoid centered at Q whose major and minor axes are parallel to the coordinate axes. 4] -1 - [1..027143 w-1 - 1401 [242 1429 2. Result 9.1. Let X 1 , X 2 , . Perform a correspondence analysis of the data on forests listed in Table 12.10, and verify Figure 12.26 given in Example 12.21 . 155 - . 88 1 . [The first potential reassignment will be based on the distances d 2 (D, (AB)) and d 2 (D, (CD)) .] Compare your solution with the solution in the example. We set andand notparetittihoatnedwitash this assignment, X, JL, and I can respectively be rearranged 24 J.LJ.L24 a-2 a-24 i 1 2 a-23 a-25 X= 3 JL = J.LJ.L3l I = a-a-2123 a-a-3144 i a-1 3 a-a-31 3 a-a-31 55 X5 J.L5 a-25 a-45 ! a-1 5 a-35 a-5 or IL x 1 l X= x21) JL = IL2 Thus, from Result for we have the distribution Itexpris clesearedfbyromsimthplisyexampl egtthhatetapprhe noropmrialatedimeases stributandion fcovaror anyiancessubseftrocanm tbehe s e l e ct i n oressigairnyal. Kent, and J. 17. (See [24] .) In the figure, the faces are used to track the financial well-being of a company over time. 1 . For instance, the dot diagram •• • • •

1 2 Finally, for the linear combination aX1 + bX2 , we have E(aX1 + bX2) = aE(X1) + bE(X2) = aJL1 + bJL2 Var (aX1 + bX2) = E[(aX1 + bX2) - (aJL1 + bJL2)] 2 I I I I = E[a(X1 - JL1) + b(X2 - JL2)] 2 = E[a2 (X1 - JL1) 2 + b2 (X2 - JL2) 2 + 2ab(X1 - JL 1) (X2 - JL2)] = a 2 Var (X1) + b2 Var (X2) + 2abCov (X1 , X2) = a 2 a- 1 1 + b2 a-22 + 2aba- 1 2 (2-41) With c' = [a, b], aX1 + bX2 can be written as [a b] [ ]: = c' X Similarly, E(aX1 + bX2) = aJL1 + b JL2 can be expressed as [a b] If we let [::] = c' p be the variance-covariance matrix of X, Equation (2-41) becomes Var (aX1 + bX2) = Var (c' X) = c' I c since (2-42) Section 2.6 77 Mean Vectors and Covariance Matrices The preceding results can be extended to a linear combination of p random variables: In general, consider the q linear combinations of the p random variables X1 , . 3 Methods of Estimation 497 In this case, the two solution methods produced very different results. Example 3.4 (Calcu lati ng Sn and R fro m deviatio n vectors) Given the deviation vectors in Example 3.3, let us compute the sample variance-covariance matrix Sn and sample correlation matrix R using the geo metrical concepts just introduced. Then • • • • • • a'Xj has positive variance for each constant vector a # 0, then, provided that p < n, S has full rank with probability 1 and I S I > 0. W. ' p (8-21) Yi -- e i (x - x) ' for any observation vector x. Does the choice A = ◆ or A = 0 make much difference in this case? There is clearly one dominant principal component. Let B be positive definite and d be a given vec(p x p) (p xl) tor. The constant of proportionality is c. , n1 = . Monitor the model performance. Recall that jL and i are those choices for ◆ and I that best explain the observed values of the random sample. That is, the axes are reversed. They lie along the axes of the ellipse in the perpendicular directions of maximum ◆ampl◆ varial!_ce. Refer to Exercise 8.20. - X-) n A ll - l1 a-ii = l (Xi) · h(6), Section 4.4 The Sa m p l i n g Distri bution of X and 1 73 5 Sufficient Statistics Frx1 o, xm2,expr. , (5, 7.5) . Perform the decomposition of y1 into .X1 1 and y1 - .X1 1 using the first column of the data matrix in Example 3.9. Use the six observations on the variable X1 , in units of millions, from Table 1.1. (a) Find the projection on 1 ' = [1 , 1, 1, 1, 1, 1] . 36 . Lawley, D. Because Y = Y + e , (or Y'Y Y'Y e'e ◆) resi ual (error) ◆urn total sum of squares = predicted sum of squares 0 square s an + and cross products and cross products cross pro d ucts (7-34) The residual sum of squares and cross products can also be written as e ◆ e = Y' Y - Y'Y = Y'Y - jJ' z ' z jJ (7-35) (Example 7.8 =) ( ) + (Fitti ng a mu ltivariate straig ht-l i n e reg ression model) To illustrate the calculations of /3 ' Y , and e ' we fit a straight-line regression model (see Panel 7.2 on page 386 for SAS output), " " lf1 = f3o l + f3 1 1 Zj1 + Bj 1 lj 2 = f3o 2 + f3 I 2 Zj l + Bj 2 ' j = 1, 2, . 12.17. 2 12.3 12.4 12 . Therefore, useful factor patterns are often not revealed until the factors are rotated (see Section 9.4) . The quantities sik and rik do not, in general, convey all there is to know about the association between two variables. Example 2.6 {Some typical p roducts and their d i mensions) Let A = [1 -2 3 2 4 -1] b = [-!] [_;] c = d = [◆] Then Ab, be ' , b ' c , and d ' Ab are typical products. Compare the results with those in Examples 12.4 and 12.6. 12.9. The vocabulary "richness" of a text can be quantitatively described by counting the words used once, the words used twice, and so forth. 7.7, z'o/3 1 2 E (z'o{J) = z'o{J Var (z'o{J) = z'o (Z ' Z) - z0o- • E (Yo - z o /3) = E (eo) + Yo - z'o/3 = Zo /J + eo - z o /3 = eo + Zo (fJ - p) . They correspond to sub stituting the sample quantities x and S for the "population" quantities IL and I, re spectively, in (2-43). Definition 2A.32. 0001 - .0001 1.0000]] According to Result 12.2, to better align these two solutions, we multiply the nonmetric scaling solution by the orthogonal matrix " Q ◆ .£..; i= l V·U· 1 l ' -[.9833 .1821 - U - - . r 0 (9-24) which is equal to the square of the multiple correlation coefficient between Xi and the other p - 1 variables. 1 and 1 .2, we find that the marginal dot diagrams are the same, but that the scatter diagrams are decidedly different. In this case some of the observations will be negative. The variable t 2 in (5-1) is the square of the distance from the sample mean X to the test value JLo · The units of distance are expressed in terms of sj Vn , or es timated standard deviations of X. 2 xl Scatter plot a n d dot d i a g rams for rea rranged data . 7388 H EAD2 -0. These eigenvalues then provide information on the variability in all direc tions in the p-space representation of the data. "Invariance of Maximum Likelihood Estimators." Annals of Mathematical Sta tistics, 37 , no. Finally, S 1 2 == S2 1 are the matrices of sample co variances computed from observations on pairs of treatment 1 and treatment 2 variables. 34 1.30 1 .24 00 "'0 "'0 (1) :::::l a. 6 1 2 50000 0. Successive choices of ai maximize (8-19) subj ect to 0 = ajSe k = aji\k e k , or ai perpendicular to e k . The principal components of Z may be obtained from the eigenvectors of the correlation matrix p of X. Identification of Outliers. A three-dimensional scatter plot can often reveal group structure. All subsets of the components of X have a (multivariate) normal distribution. 33333333 c.v. T for HO: Parameter = 0 -1 . We have discussed some rather simple techniques for checking the multivariate normality assumption. 758 . tDetheierersmsioncieattheedmajlengtorhands. , V(i 1- 1)i ' V- ((ii -- 11))i , O, . These pictures are valu- able aids in understanding data and often prevent many false starts and subsequent inferential problems. Unfortu nately, this rank is not always well determined from R, and some judgment is necessary. Plot the minimum stress(q) versus q and interpret the graph. (X) 90 Matrix Algebra a n d Ra ndom Vectors Chapter 2 An m X k matrix, say, A, of arbitrary constants can be written A (mX k) a 12 a 22 a1 1 a 21 == a ml a m2 or more compactly as Ak == { ai j }, where the index i refers to the row and the index (mX) j refers to the column. Shapiro, S. When n = 100 and p = 5, there are Section 4.7 Detecting Outl iers and Clea n i n g Data 191 500 values. (a) (b) AA' A' A (c) 2.23. • " " " " " C omment. Section 7.4 I nfe rences About the Reg ression M odel 365 of c' f3 has the smallest posa'sYibl=e varianceaamong al l l i n ear es t i m at o r s of t h e f o r m Y 2 2 that are unbiased for c' f3. (x2) Machine ' 1 04 . 57 1.51 1 .46 1 . p sample Cov It Matrices of E rro rs of Approxi mations .A- 1 S1 1 S22 = = = B S22 B' = V) = = p; 0 A- 1 U = 0 = b (i) Bx (2) :B - 1 V (qXq) (qX 1) (iJ) I , 0 = = (qxq) 0 0 (A- 1) (A- 1) ' (B- 1) (:8- 1) ' V As 1 2 B ', Pi 0 A- 1 0 Pi x(1) 8(i) B (10-32), iJ = Ax (1) (l) = A - 1 iJ x(2) x (p X 1) (p X p) (p X1) (qX 1) :B- 1 , (V) s1 2 A (:8- 1) ' == (10-36) = As 1 1 .N = I (p X p) pf 8 (1)b(1) ' + pf8(2)b(2) 1 + . If the canonical correlations are also moderately vl = P + F + + + o + + + Pi Vi Vi Vi SAM PLE CORRELATIONS B ETWEEN ORIGI NAL VARIABLES AN D CANO N I CAL VARIABLES x (l) ul vl vl ul . This corresponds to the assumption of independence of the rows and columns. Let B be (pXp) a positive definite matrix with eigenvalues A 1 A2 AP 0 and associated normalized eigenvectors e 1 , e 2 , . , p ek +l i e k +l fek +l ek + l = ek + l i e k +l = Var (Yk +l) - 1, But ek + 1 (Ie k + l) = "-k + l ek +l e k +l = "- k +l so Var (Yk + l) = "-k + l · It remains to show that e i perpendicular to e k (that is, e;e k = 0, i =I= k) gives Cov (I: , Yk) = 0. Methods for Statistical Data Analysis of Multivariate Observations. Singular-Value De composition. (See [2] .) The means and covariances of the p X 1 random vector X can be set out as matrices. The predictor variable that explains the largest significant proportion of the variation in Y (the variable that has the largest correlation with the response) is the first vari able to enter the regression function . Given these distances, determine the coordinates of the sites in q == 3 , 4, and 5 dimensions using multidimensional scaling. Data Preparation for Data Mining. In cases where the units of the variables are not commensurate, it is usually de sirable to work with the standardized variables (xj l - xl) ◆ 1 z. (c) Graph (to scale) the triangle formed by y1 , .X 1 1 , and y1 - x1 1 . , yp) be a specified fixed point. By the invariance property of maximum likelihood estimates (see Section 4.3), functions of L and 'I' are estimated by the same functions of i and .q,. Plot the residuals ej against the predicted values Yj = For ffi o + ffi l Zj l + . Using the data on bone mineral content in Table 1.8, perform a principal com ponent analysis of S. 1y, s 1 1 1s th e samp l e covari-. And the average of the squares of the distances of all of the numbers from the mean provides a measure of the spread, or variation, in the numbers. • Example 3.6 {A nonrandom sample) Because of concerns with future solid-waste disposal, a study was conducted of the gross weight of solid waste generated per year in the United States ("Char- acteristics of Municipal Solid Wastes in the United States, 1960-2000," Franklin Associates, Ltd.). Hint: X + Y has Xi j + Yi j as its (i, j) th element. (10-3 5) (q X 1) (pX 1) (10-34) D1i12 == D2◆/2 == I unaf Az for A, R Example 1 0.4 (Canon ical correlati on ana lysis of the ch icken-bone data) I (p x p) 9.14, 1) == X◆ == Xi2) == X ◆2) == (qxq) Section 1 0.4 The Sample Ca nonical Va riates and Sa m p l e Ca nonical Corre lations have the sample correlation matrix 559 1.0 .505 : .569 .602 [i-!◆- 1- :◆-:J - --:.◆6-◆02-◆- - ◆.%42-672--h-◆.◆926?:?:_- -1.- ◆0◆i◆ lation canalal corysirseoflatitohnse headand corandrelsepgondisetsnofg paivarrisaofblevars usiainblges prA ocanoni duces ctalhecortworecanoni 1)) . Using the archaeological data in Table 12.13, determine the two-dimensional metric and nonmetric multidimensional scaling plots. , 0) denote the origin, and let Q = (y1 , y2 , . (b) Outliers masked. dIfbeonechecked i n t h e or d er mul t i v ar i a t e nor m al i t y and t h en or bot h of t h es e as s u mpt i o ns i s vi o l a t e d, i m pr o ved cl a s i f i cationThemayquadrbe posaticsibrlueleifs tarhee datan ala arterenfatirisvtesutoitaclblays itfriacnsatfioonrmwied.th linear discriminant fequalunctiocovarns. Then Vn (X - IL) has an approximate Np (0, I) distribution for large sample sizes. Once X and s 2 are observed, the test becomes: Reject H0 in favor of H1 , at significance level a, if (5-2) where tn_ 1 (aj 2) denotes the upper 100(a/2)th percentile of the !-distribution with n - 1 d.f. If H0 is not rejected, we conclude that JLo is a plausible value for the normal population mean. The inequality is obvious if either b = 0 or d = 0. ' n -r-1 1 (7-13), [I - Z(Z' Z) Z'] = ◆ e e e € , where e{;ek = 0, e # k, and e € e e = 1. , rown (X - lx') ' . 00000000 Mean Sq u a re 40.00000000 T for r HO: Parameter = 0 0.91 4.47 Sou rce Model E r ror Corrected Tota I Sou rce Z1 F Va l u e 20.00 Y 1 Mean 5 .00000000 F Va l u e 20.00 S u m of Sq u a res 1 0.00000000 4.00000000 1 4.00000000 R-Sq u a re 0 . Because 4 and 1 are scalars, premultiplication and postmultiplication of A by x' and x, respectively, where x' == [x 1 , x2] is any nonzero vector, give x' A x == 4x e 1 e i x + x' e2 e2 x (1 X2) (2 X 2) (2Xl) (1 X 2) (2 Xl) (1 X 2)(2 Xl) (1 X 2) (2 Xl) (1 X 2) (2 Xl) == 4yy + y◆ 0 with y1 == x' e 1 == e1 x and y2 == x' e2 == e2 x We now show that y1 and y2 are not both zero and, consequently, that x' Ax == 4yy + y◆ > 0, or A is positive definite. In cases where the weights are the same, it is convenient to ignore the com mon divisor and use the usual Euclidean distance formula. 1 (Calculati ng the po pulation principal components) [-◆ -◆ ◆] Suppose the random variables X1 , X2 and X3 have the covariance matrix I= It may be verified that the eigenvalue-eigenvector pairs are A1 = 5.83, A2 = 2.00, A3 = 0.17, e1 = [.383, -.924, OJ e2 = [0, 0, 1 J e3 = [.924, .383, OJ Therefore, the principal components become Yi = e1X = .383X1 - .924X2 Y2 = e2X = X3 Y3 = e3X = .924X1 + .383X2 The variable X3 is one of the principal components, because it is uncorrelated with the other two variables. • 30 Chapter 1 Aspects of M u ltiva riate Ana lysis Liquidity ◆ Profitability 1 975 1 976 1 978 1 977 1 979 ------◆ Time Figure 1 . We shall return to this issue in Section 5.4. _ - (Eval uati ng T2) Example 5 . Compare the results. (a) tr (c A) = c tr (A) (b) tr (A ± B) = tr (A) ± tr (B) (c) tr (AB) = tr (BA) (d) tr (B- 1 AB) = tr (A) k k • (e) tr (AA') = ◆ ◆ a rj i=1 j=1 Definition 2A.29. nX Hint: A. Here the degrees of freedom are v = (dimension of (dimension of 8 o) · e) - - v0 • Statistical tests are compared on the basis of their power, which is defined as the curve or surface whose height is P[test rejects evaluated at each parameter vector Power measures the ability of a test to reject when it is not true. 803 2 . Northwestern lies in the middle • of the biplot . 92 1 .88 1 .86 1.81 1 . Result 2A.l3. iBoxris. (See [4] .) Hypotheses testing • • • Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends. Examine the marginal normality of the observations on variables X1 , X2 , 4.32. + e◆ j (9-28) s11 + s22 + · · · + sP P Proof. (See Exercise 7 . \ (a) . The cumulative charts for all three classifiers are shown in Figure 12.30. If the ob = i The equivalent statistic served value of this likelihood ratio is too small, the hypothesis H0 : IL = ILo is unlikely to be true and is, therefore, rejected. Of course, the experimenter should randomize the order in which the treatments are presented to each subject. Other methods are ad hoc in nature and are justified by logical or commonsense arguments. not diagonal gives the oblique factor model. ' Ap] and eigenvectors e l , . Let b and d be any two p X 1 vectors. However, with some additional as sumptions about the random vectors F and e, the model in (9-2) implies certain co variance relationships, which can be checked. .-=- ., . Let A = { a i j } be a k X k square matrix. Is the model likely to be useful? 5 5. 6 1 8 1 4 1 62 1 . Its modern beginnings lie in the early 20th-century attempts of Karl Pearson, Charles Spearman, and others to define and measure intelligence. Now, or s1 1 == 134 • Also, Section 3 . Let a = .05 and use the entry corre sponding to n = 40 in Table 4.2. 4.29. 33 .18 . This value is obtained by maxi mizing L(JL o , I) with respect to I. 96 1.91 1 .83 1 .75 1 . With this choice, 0 vi Qu i v i VU ' ui = = 0 1 0 [0, . 731 732 Chapter Clustering, Distance Methods, and Ord i nation 12 structured. From (2-2) it is clear that x is expanded if I c I > 1 and contracted if 0 < I c I < 1. Chernoff faces appear to be most useful for verifying (1) an initial grouping suggested by subj ect-matter knowledge and intuition or (2) final groupings produced by .clustering algorithms. The last column in Table 4.4 reveals that specimen 16 is a multivariate outlier, since x◆(.005) = 14.86; yet all of the individual measurements are well within their respective univariate scatters. NBecaus q(O, Ie1 X-1 I-1 21LI12i-i2I1)1.2IGi2iv(enX2 t-hatILX2)2 andx2X, 2IL-1 1LI2 1ar2Ie 2iin(dependent tdiiosnaltribdiutsitornibofutiXon1 of- X1L11 -- IIL1l2I-2Ii(1X2I2 2-i(x1L22)-. Although the likelihood in (9-25) is appropriate for S, not R, surprisingly, this practice is equivalent to obtaining the maximum likelihood estimates L and q, based on the sample covariance matrix S, setting L z = v - 112 L and q, z = y- 1/2 q, y-112 . 27 23 64 44 30 75 26 124 54 30 14 State lab of hygiene x 2 j 1 (BOD) X 2 j 2 (SS) 25 28 36 35 15 44 42 54 34 29 39 15 13 22 29 31 64 30 64 56 20 21 276 Chapter 6 Com parisons of Severa l M u ltivariate Means Do the two laboratories' chemical analyses agree? trNow,ibutioInl of-IX11 2giIv2eni I2th1 at o-whi1 c-h wea-i2/obto-2ainedo-1by(lan-inPIdi2r)eandct metI1 2hIod.2i o-12/o-2 , agreeing with Result = == + = == 4.6◆ • Section 4.2 The M u ltivariate Norma l Dens ity and Its Properties 1 63 For thAle mull conditivartioianalte nordistmrialbutsiitounsatiaron,e i(tmisulwortivarthiaemphas i z i n g t h e f o l o wi n g: t e) nor m al . Adriaans, P. 5 565 Add ition a l Sa mple Descri ptive Measures If only the first canonical pairs are used, so that for instance, r x(1) [8(1) = " u1 " u r 8( ) J 2 " ur 8(2) and X' (2) [b(1) = (10-38) " " v b(r) J 2 Vi b (2) " vr (x (1) , x (2)) . These data, multiplied by 100, are 20.6, 18.7, 14.2, - 15.7, 19.0, 7.7, -11.6, 8.8, 9.8, and 18.2. Use these 10 observa tions to complete the following. + C1 p xp , c2 1 x1 + c22 x2 + . 07 3.00 2.92 2.83 18 8.29 6.01 5 .09 4.58 4.25 4.01 3.84 3.71 3.60 3.51 3.37 3 .23 3 .08 2.98 2.92 2.84 2.75 19 8.18 5.93 5.01 4.50 4.17 3.94 3.77 3.63 3.52 3 .43 3.30 3.15 3.00 2.91 2.84 2.76 2.67 20 8.10 5.85 4.94 4.43 4.10 3.87 3.70 3.56 3.46 3.37 3.23 3.09 2.94 2.84 2.78 2.69 2.61 21 8.02 5 .78 4.87 4.37 4.04 3.81 3.64 3.51 3.40 3.31 3.17 3 . Prediction. The natural logarithms of the dimensions of 24 male turtles have sample mean vector x ' = [4.725, 4.478, 3.703 J and covariance matrix 3 Scree is the rock debris at the bottom of a cliff. • • • ..... Graybill, F. On the other hand, many valuable insights can be obtained from the data by constructing plots with paper and pencil. It turns out, however, that nor mal distributions are useful in practice for two reasons: First, the normal distribution serves as a bona fide population model in some instances; second, the sampling dis tributions of many multivariate statistics are approximately normal, regardless of the form of the parent population, because of a central limit effect. This example provides a case where it seems sensible to retain a compoIl nent (j/2) associated with an eigenvalue less than unity. 64 Chapter 2 Matrix Al gebra a n d Ra ndom Vectors Example 2.1 1 (A positive defi nite matrix and q uad ratic fo rm) Show that the matrix for the following quadratic form is positive definite: 3x i + 2x◆ - 2V2 x 1 x2 To illustrate the general approach, we first write the quadratic form in matrix notation as By Definition 2A.30, the eigenvalues of A are the solutions of the equa tion I A - AI I == 0, or (3 - A) (2 - A) - 2 == 0. The point B == 0 falls outside the 95% confidence region for B (see Exercise 6.1), and this result is consistent with the T 2 -test. Section 1 .5 35 Distance What is a meaningful measure of distance when the variability in the x 1 direc tion is different from the variability in the x2 direction and the variables x 1 and x2 are correlated? The solid dots in these figures correspond to specimens Section 4.7 1 500 I - I I I - - x1 0 ◆ ('.! - 0 ◆ - 0 ◆0 ◆ ◆ c:P oo 0 - - 0 0 • 0 00 0 0 C"l ◆ 000 c:Pco 0 ({) ,◆• c:P • 0 Oo 0 0 • 0 J I •oca ◆0 o o 0oc?CSJ • 0 • 0 0 0 0 I I • I • 0 0 ◆0 < 1] = .3174. Typically, the di◆gonal entries of will be large. Allowing for specific factors, we find that the approximation becomes I - LL' + 'It VA;- e◆ ---------- = [VA;- e 1 VI;: e 2 VA: e m] VI;: e2 ---------- ---------- + VX: e ◆ l/1 1 0 0 t/12 0 0 0 0 l/Jp (9-13) m where l/Ji = (Ti i - ◆ erj for i = 1 , 2, . Result 4.12 (Law of large numbers). It is usual practice to deter mine an eigenvector so that it has length unity. Therefore, upon division by the standard deviations, we have the "standardized" coordinates x i = x 1j◆ and xi = x2j'\l's; . I). Incorrect readings on an individual variable are quickly detected by locating a large leading digit for the standardized value. 83 1.81 1 .78 1 .75 18 3 .01 2.62 2.42 2 .29 2.20 2.13 2.08 2.04 2.00 1 .98 1 .93 1 . - 500 Chapter 9 Factor Ana lysis and I nfere nce fo r Structu red Cova ria nce Matrices The test statistic in (9-39) is based on the ratio of generalized variances I i I I LL ' + q, I I sn I I sn I Let v - 1;2 be the diagonal matrix such that v - 112 S n v-1;2 R . 8.24. (Note that the first component is essentially a normalized unit vector and might measure the athletic excellence of a given nation. physical phenomenon, selects a number p The values of these variables are all recorded for each distinct item, individual, or > experimental unit. (See [23].) Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants. 569 .422 1 . 286826 1 4 P i l l a i's Trace Hotel l i n g - Lawley Trace Roy's G reatest Root Level of FACTO R 1 0 3 3 3 1 .3385 1 . In sum, distance is determined from a positive definite quadratic form x' Ax. Conversely, a positive definite quadratic form can be interpreted as a squared distance. Riccardi Kathleen Schiaparelli Senior Managing Editor: Linda Mihatov Behrens Assistant Managing Editor: Bayani DeLeon Production Editor: Steven S. If is of full rank, > Z. We verify that (12-31) holds for these choices. For any correspondence matrix, P , the common term in every expansion is r-.J n ;|2 u v 1 D ◆I2 1 = D r 11fiDc = rc ' 714 Chapter 1 2 Cl usteri ng, Distance Methods, and Ord i nation Therefore, we have established the first approximation and (12-30) can always be ex pressed as p I rc' = + ◆ Ak (n ;|2 uk) (DY2 vk) ' k=2 Because of the common term, the problem can be rephrased in terms of P - rc' and its scaled version D◆ 112 (P - rc') D◆ 1 /2 • B y the orthogonality of the singular vec tors of n-1 12 PD -1 12 we have u'k (D 112 11) = 0 and vk' (D 1 12 1J) = 0 ' for k > 1 ' so r c r ' c n-1 12 (P - rc') n-1;2 r c I = " ◆ k =2 Ak uk vk' is the singular-value decomposition of D◆ 1 12 (P - rc') D◆ 1/2 in terms of the singular val ues and vectors obtained from D ◆ 1/2 PD ◆ 1/2 • Converting to singular values and vectors Ak , uk , and vk from D ◆ 1 /2 (P - rc ') D ◆ 1 /2 only amounts to changing k to k - 1 so A k = Ak + 1 , uk = uk + 1 ' and vk = vk + 1 for k = 1 , . Special plots called Q-Q plots can be used to assess the assumption of normality. We knowwhertehatZ1 ,x 2, is ,definarede inasdependent the distributio1)n raofndomthe varsumi ables. , e◆ x are recognized as the principal components of x . Also, each has length unity. 3 (1971), 321-367. roeascedursounablmmare cle ayandssuifchiseismasiplltstethaechoirgoneetcobjesinbasethctiess.dexampl on theeualtliomwsatethcreiitnevesria tofi Our di s c us s i o n has t e nded t o emphas i z e t h e l i n ear di s c r i m i n ant r u l e of orthe linear diandscrimmanyinantcommer c i a l comput e r pr o gr a ms ar e bas e d upon i t . This rule does not have a great deal of theoretical support, however, and it should not be applied blindly. Consider the m X k matrix A with arbitrary elements ai j ' i = 1, 2, . = 1 + - n (nm (nm - n + ··· + nm - n I X. i yp] Xn p Then the coordinates of the first point y1 = [x1 1 , x2 1 , . 8.14. Then c can be selected to pro duce a test with a specified significance level a. X Find The ei g enval u es and ei g envect o r s of (Letc) Thehaveeigenvalcovariuesanceandmateigrenvect o r s of ix p (a) I p1 3 . E (z0 (fJ - fJ"')) = 0 eo fJ "' 1 2 2 Var (Y0 - z0{J) = Var (eo) + Var (z0f3) = a- + z0 (Z ' Z) - z0o- = a-2 (1 + z'o (Z' Z) -1 z 0) . 1 1 (a) . 96 1 .93 1 . Thus I X- . (2, 2) (a nd B) ar e near l y l i n ear l y dependent . (2) a' (xj - x) = 0 for all j (3) a' xj = c for all j (c = a' x) The linear combination of the mean corrected data, using a, is zero. Conse quently, replacing I by S in the approximating normal distribution for X will have a negligible effect on subsequent probability calculations. == == The maximum likelihood estimates of the communalities are for i so (Proportion of total sample variance due to jth factor) = fL + == 1, 2, . Consequently, it is sometimes useful to scale all the deviation vec tors so that they have the same length. (4-9) where the {3's are defined by f3f32l,,qq++l f3f32l,,qq++22 f3f32l,,pp /3q,q+l /3q,q+2 /3q,p Thevalue(condis) ofttihoenalcondicovartioniiannce,g variable(s).2 2 , does not depend upon the We concl u de t h i s s e ct i o n by pr e s e nt i n g t w o f i n al pr o per t i e s of mul t i v ar i a t e nor malconsrtaandom vect o r s . Chapters in this text are divided into sections according to inference about treatment means, inference about covariance structure, and techniques for sorting or grouping. , uk with the same linear span. In matrix algebra, however, the product of two nonz ero ma trices may be the zero matrix. We put the data mining material in Chapter 12 since much of data mining, as it is now applied in business, has a classification and/or grouping obj ective. == a l- 1. 12 allow one to identify readily observations that do not conform to the rest of the data and that may heavily influence inferences based on standard data-generating models. McGregor, "Multivariate SPC Methods for Process and Product Mon itoring," Journal of Quality Technology, 28 (1996), 409-428. , (q (n) , x (n)) , and exam ine the "straightness" of the outcome. Figure 1 . This test is called a sphericity test, because the constant den sity contours are spheres when I = 0"2 1. Verify (2-24): E(X + e k which is the (i, j)th element of AE(X) B . On the other hand, a two-dimensional plot of the sampling units can be obtained by graphing the first two principal com ponents, as in Section 8.4. The idea behind biplots is to add the information about the variables to the principal component graph. The multivariate multiple regression model poses no new computational problems. I I (6- 5 7) I = 0 > xJb- 1) p (a) (1 == 6- 58) xfb- 1) p (can Result we f3 q f3 k T e ; - Tm; = (n m (ie ;. 404 (1988), 1192-1 197. + (x n - x) (x n - x) ' = b' c n - 1 or Sample covariance of b'X and c' X = b ' S c (3-35) I n sum, we have the following result. • • (2, 3) • (1 , 3) (1) 0 • (1 , 2, 3) (1 , 2) Numbers in parentheses correspond to predicator variables CP p l ot fo r com p uter data from Exa m p l e 7.6 with th ree pred icto r va riables (z 1 = orders, z2 = add- delete cou nt, z3 = n u m ber of items; see the exa m p l e a n d orig i n a l sou rce) . Yet the departure from the specified distribution may be very small and technically unimportant to the infer ential conclusions. If the factor model p = LL' + 'I' is correctly specified, the m common factors should account for the off-diagonal elements of p, as well as the communality portions of the diagonal elements . 32 00 I "'0 "'0 (1) :::::l a. 4.6 ASS ESSING TH E ASSU M PTION OF N ORMALITY As we have pointed out, most of the statistical techniques discussed in subsequent chapters assume that each vector observation Xi comes from a multivariate normal distribution. The N u m ber of Principal Com ponents There is always the question of how many components to retain. 7 9 1.00 It is clear from the circled entries in the correlation matrix that variables 1 and 3 and variables 2 and 5 form groups. 3 06 . = = · = = Section 2 . 3 If p > rank (S1 2) = p 1 , the nonzero sample canonical correlations are pf, . Similarly, a random matrix is a matrix whose elements are random variables. We shall discuss and illustrate several methods for displaying multivariate data in two dimensions. Hint: (2-43) , Z1 - E(Z1) = c1 1 (X1 - JL 1) + · · · + c1 p (Xp - JLp) Z2 - E(Z2) = c2 1 (X1 - JL 1) + · · · + c2 p (Xp - JLp) · (Z1 , Z2) = E[(Z1 - E(Z1)) (Z2 - E(Z2))] = E [(c1 1 (X1 - JL 1) + . For p > 4, we cannot graph the joint confidence region for p. Pollatschek) : 1 1 2 3 4 5 Lamentations chapter 0 .76 2.97 4.88 3.86 Lamentations chapter 2 4 3 0 .80 4.17 1 . 4919 .0647 1 .4108 .2493 .2635 .3127 R= 1 - .2249 .2293 .3506 .4108 .4653 1 - .2144 - .2277 - .1917 .0647 - .2249 1 .2293 - .2144 .2045 .2493 .0652 is based on a sample of about 120. The matrix P = rc' can be shown to be the best rank 1 approximation to P. , A p) and V is an orthogo11al matrix whose columns are the == ' e , . 1 • (0) • (3) (2) . It turns out that certain multivariate statistics, like X and S, have large-sample properties analogous to their univariate counterparts. 1 1 after a discussion of factor rotation. 78 17 3 . Theie. , p. 1 06z◆ 2 .057 V; -2.648zl2) 2.475z◆2) Herand e2,z?re)s,pectiv1,el2y.andThez◆pr2) ,ecedi1,ng2rareseulltthsewerstaendartakendizedfromdattahevalSASues sftoartissetitcsal1 sinoalftwvarareiaoutblesputwitshhownthe canoni in Panel10. That is, we can write (9-11) (pIX p) (pLX p) (pL'X p) + (p0X p) LL' Apart from the scale factor ◆ ' the factor loadings on the jth principal component of the population. (3-32) Since (c' xj - c' x) 2 = (c' (xj - x)) 2 = c' (xj - x) (xj - x)'c, we have (c'x 1 - c' x) 2 + (c'x2 - c' x) 2 + . in a (nm - n)p (nm - n - p + 1) Ellipse Format Chart. " canoni may alcsalo varbe ihelatepsfuagail inntsht itshreier cthoempenont var i a bl e s as an ai d i n s u bj e ct m a t t e r i n t e r p r e t a t i o n. The ot h er di s c us s e s t h e di s t r i b ut i o n of anot h er f o r m of l i n ear com bi2natiTheons. However , t h e as s u mpt i o n of nor malas toitythseeemsapprtooprbeiatmorenesescrofitaiclalinfearor quadr aticartuiclersutleh,anbotlihnearrulersucanles. Because the residuals e have covariance matrix a-2 [I - H], the variances of the sq uan vary greatly if the diagonal elements of H, the leverages h j j , are substantially different. Figure 2 • • • :· .01 .03 .05 8.6 Scatter plot of the principal components y1 and y2 of the data o n male turtles. , xp] # [0, 0, . Once established, any suc cessful data mining effort is an ongoing exercise. (b) Treating the distances in (a) as measures of (dis)similarity, cluster the coun tries using the single linkage and complete linkage hierarchical procedures. The contribution to the sample variance sii from the first common factor is e[l . 97 24 4.26 3 .40 3.01 2.78 2.62 2.51 2.42 2.36 2.30 2.25 2.18 2.11 2.03 1 . Thus, x' Bx x' B lf2 B lf2 x x ' P A 112 P' P A 112 P' x y' Ay [= x' x = -- y'y y'y x' PP' x I pXp ( ) '-.r--1 (2-53) Setting x = e 1 gives y = P' e 1 1 0 = 0 since k 1 k i= 1 = For this choice of x, we have y' Ay/y' y = A 1/1 = A 1 , or (2-54) A similar argument produces the second part of (2-51). Then and I == arvaleutsehe, and ofare andcalleId, treshepectively. (a) {- == { (b) ◆ ◆ ◆◆ l A B 1 l A OI I I I B 1 Hint: (a) 0 0 = I A I I B I for I A I * 0 . At the a level of significance, we reject H0 in favor of H1 if the observed (n - 1)p 2 1 (5-7) T - n (X - IL o) s (X - IL o) > F (a (n p) p, n - p) It is informative to discuss the nature of the T2 -distribution briefly and its cente spondence with the univariate test statistic. Mastering Data Mining. 95 1 .92 1 .89 1 .84 1 . 1 4 Distance 33 {Calculating a statistica l d istance) A set of paired measurements (x 1 , x2) on two variables yields x1 = x2 = 0, s1 1 = 4, and s22 = 1. (See [9] and [20] .) Measurements on several accounting and financial variables were used to de velop a method for identifying potentially insolvent property-liability insurers. Since e Y - Z/J satisfies = (Y - Z/J) ' (Y - Z/J) = [Y - zp z (/J - fJ) [Y - zp z (fJ - fJ) J = (Y - Z/J) ' (Y - Z/J) (/J - /J) ' Z'Z (/J - /J) = e' e (/J /J) ' z' z (/J - /3) 1( ) I) = IT L(I I j= 1 + = I = I Z (/3 - /J) I -1 (/J - /J) ' Z' A' A, 2 1 A = I - 1 (/J - /J) ' Z', tr [Z(/3 - /J)I -1 (/J - /J) ' Z'] fJ = fJ . Specify each of the following completely. k \Ill l' k In this case, (A. The region R(X) is said to be a 100 (1 a)% confidence region if, before the sample is selected, P[R(X) will cover the true = 1 a (5-17) This probability is calculated under the true, but unknown, value of The confidence region for the mean of a p-dimensional normal variable (5-6) . o- 1 1 + o-22 + + o-PP = L Var (XJ = "- 1 + "-2 + + AP = L Var (I:) i =l 1 . (x3) . Then, k= 1 the best approximation to P r - rc' s P - rc' - K◆ Ak (D ;I2 uk) (D◆I2 vk) ' k=1 (12-34) • Remark. The same steps wi t h mean Cons i d er t h e ar b i t r ar y r a ndom vect o r X' vector = Partition(lX into [-XX-(2-)) J X = where X(1l = [◆J and X(2l = [◆;] Letthe covar(lbe itahncee covarmatriaicncees ofXmat(r(l2i)xandof XXwi(2)tandh generthe acovarl elementiance matriParx oftiantionelementinto ofXYou ar) eandgivanenelthemente randomofX vect). [ ]]] [[- .0308 .20 .50 [.5 .5] -..0058 -..0058 2 Total 2 24 2 4 4 4 The correspondence matrix is p with marginal totals root matrices are D◆ 1 12 Then P ' ' .24 = and = diag (vlj = rc c .12 vlj r ' .32, = The negative square D◆ 1/2 vl) = diag (v2 , vl) .12 .24 = - .32 = The scaled version of this matrix is A = D◆ 1 12 (P v2 0 0 . sp p) I R I so (3-22) [The proof of (3-21) is left to the reader as Exercise 3.12. wn _ 1 (I I) = (·) I l (n- p - 2) /2 e -tr [AI -1] /2 2 p (n- 1) /2 1Tp (p - 1) /4 1 I l (n- 1) /2 (P II i =l [1].) (n - i)) (4-25) Section 4.5 4. 85 20.5 2 1 .0 2 1 .5 22.0 22.5 Voltage 23.0 23.5 24.0 Figure 5 . , d P _ 1 , we can establish the following general result for p deviation vectors by induction (see [1], p. Clearly most of the variation is scatter about a one-dimensional straight line. The obj ective is to develop a credit scoring rule that can be used to determine if a new applicant is a good credit risk or a

sample splits, from the two laboratories. F. "Market and Industry Factors in Stock Price Behavior." Journal of Business, 39 (1966), 139-1 90. The coordinate axes can then be visually rotated through an angle-call it > ··· 1 , 2, . 492 Factor Ana lys is and I nference for Structu red Cova ria nce Matrices Chapter 9 The Maxi m u m Likeli hood Method If the common factors F and the specific factors e can be assumed to be normally distributed, then maximum likelihood estimates of the factor loadings and specific variances may be obtained. The method followed was to select persons randomly (perhaps using a random number table) from all those who entered the wilderness area during a particular week. 16, the error of approximation is 2: A ̄. + Yj, q - l e q - 1 differs from Xj by Yj q e q + + Yj p e p , the square of whose length is YJq + · · · + YJp · Suspect observations will often be such that at least one of the coordinates Yj q ' . , xP , respectively. 7 63 1 Fisher's Method for D iscri m i nating among Severa l Pop u l ations te becaus. The proj ection of Yi on the unit vector (1/Yn)1 is, by (2-8) , = = ,( ) 1 1 1 1 X1 i + X2 i + Yz. n Vn Vn _ ··· ··· + Xn i 1 - Xz·1 _ _ (3-3) That is, the sample mean xi (xli + x2 i + + x n J/n yi1/n corresponds to the multiple of 1 required to give the projection of Yi onto the line determined by 1. , o] - (p - 1) 1 1 . A generalized variance of zero is indicative of extreme degeneracy, in the sense that at least one column of the matrix of deviations, Section 3.4 Genera l ized Va ria nce x1 - x' x2 - x' X 11 - X1 X112 - X2 X21 - X 1 X22 - X2 X 1 p - X p X2 p - X p x ̄ - x' Xnl - X1 Xn2 - X2 X - 1 x' (n Xp) (l Xp) Xn p - Xp 131 (3-18) can be expressed as a linear combination of the other columns. Obt a i n t h e s i n gul a r v al u e decompos i t i o n of Verthe ify the rpopul elatioanstiohnipcovar s iance matrixand[Equation (2-32)] , pis the wherepopis deviulatiaotniocorn matrelartiixon[Ematquatrixio[nE(quat2-35)io]n. For example, ̄ ̄ ̄J [! ̄ ̄ n [̄ - ̄ ̄ ̄ J [̄ 1 + A + c B Definition 2A.17 (Scalar Multiplication). A likelihood ratio test of E rejects in favor of ft. be ' = 7 -3 [5 8 -4] = 6 35 56 -28 - 15 -24 12 30 48 -24 The product be ' is a matrix whose row dimension equals the dimension of b and whose column dimension equals that of c. Mosdoflotwht eofandqualthevolivartytacontige,ablsrehosownlarelelipiinnsecontFis g(eurlroeipl.seHowever r e veal s t h at cas e i s out ofvarcontiate rol charandtthfoisr lisn(duegas fltoow)an,unusin Fiugurralley largeshvolowsumethatofthgasis poiflonw.t isTheoutuniside the tAlhrletehseigotmaherlimuniitsv. • • • . Simply stated, random sampling implies that (1) measurements taken on different items (or trials) are unrelated to one an other and (2) the j oint distribution of all p variables remains the same for all items. That is, the existence of A- 1 is equivalent to (2-12) (See Result 2A.9 in Supplement 2A.) Example 2.8 (The existence of a matrix i nverse) For A= you may verify that [so .4 - .2 .8 - .6][ ] [3 2 = 4 1 [! ̄ ̄] (- .2)3 + (.4)4 (- .2)2 + (.4) 1 (.8)3 + (6) 4 (8)2 + (-.6)1 [- - .2 .4 .8 - .6] . In Figure 7.4, we have circled the point corresponding to the "best" subset of predictor variables. If L is the p X m matrix of estimated factor loadings obtained by any method (principal component, maximum likelihood, and so forth) then A A L* == LT, where TT' == T'T == I (9-42) is a p X m matrix of "rotated" loadings. Does this ranking correspond with your intiuitive notion of athletic excel lence for the various countries? The expected value and variance of a ' X follow from (2-43) . In particular, qc , p ((j - ̄)In) = x ̄(n - j + ̄)In). Marriott [19], page 89. (a) Calculate statistical distances (xj - x) ' S- 1 (xj - x) , j = 1 , 2, . The scatter plot in Figure 1 .4 supports the claim that a championship team can be bought. So, for instance, 1, 2, 1 1 + 2 [pj (a + d) - 1] This equation relates coefficients 3 and 1 . Figure .07 y2 The diagnostics involving principal components apply equally well to the check ing of assumptions for a multivariate multiple regression model. This format will allow easy interface with existing statistical software packages and provide more convenient hands-on data analysis opportunities. Thus, multiplication by projects a vector onto the space spanned by the columns of Similarly, J is the matrix for the proj ection of y on the plane perpendicular to the plane spanned by the columns of Z. ̄ .. 12(c) ; one of the outliers (16) is now hidden. We prefer rQ because it corresponds directly to the points in the normal-scores plot. cm. 1 0) 8 9 58.91 59.44 59.86 7 12 15 60 . nonical variate ofthe other set. Example 8. " R esult 12.1. The term rc' is common to the approximation P whatever the J I X correspondence matrix P. Database packages with embedded data min ing capabilities, such as SQL Server 2000, represent one evolutionary direction. Start with the initial groups (AB) and (CD) . 1 2 (Uti l ity data as Chernoff faces) From the data in Table 12.5, the 22 public utility companies were represented as Chernoff faces. Large values for the variables SAT, TopIO, and Grad are associated with the private school group. 3.10. Moreover, the estimated covariance (or cor relation) matrix remains unchanged, since A A A A A LL' + W == LTT' L + W == L * L * ' + W (9-43) Equation (9-43) indicates that the residual matrix, S n - LL' - .q, == s n - L * L * ' - .q, , remainAs unchanged. The inertia associated with the best reduced rank K < J approximation to the K centered matrix P - rc' (the K-dimensional solution) has inertia L A ̄. 93 3.79 3.68 3.59 3.46 3.31 3.16 3 . Sett ing Y1 = e1 x, y2 = e2 x, . Several computer programs are now available for this purpose. = (xj 2 - x2) Vs; j = 1 , 2, ... X2(g-1) (b-1)p (XJg-1) (b-1) p (a) (T1 T e # 0. 1 1 for an additional generalized sum of squares property.) ,... 4 (2000), 954-959. A single numerical measure of variability is provided by the determinant of the sample variance-covariance matrix. - 5.2 TH E PLAU SIBI LITY OF POPU LATI ON M EAN 1-Lo AS A VALU E FOR A NORMAL Let us start by recalling the univariate theory for determining whether a specific value JLo is a plausible value for the population mean JL · From the point of view of hypothe sis testing, this problem can be formulated as a test of the competing hypotheses Here H0 is the null hypothesis and H1 is the (two-sided) alternative hypothesis. 1 20 3.3 Chapter 3 Sa m p l e Geometry and Random Sa m p l i n g RAN D O M SAM PLES AN D THE EXPECTED VALU ES O F TH E SAM PLE M EAN AN D COVARIAN CE MATRIX In order to study the sampling variability of statistics like x and Sn with the ultimate aim of making inferences, we need to make assumptions about the variables whose observed values constitute the data set X. 84 3.69 3.58 3.50 3 .44 3.39 3.35 3 .28 3.22 3.15 3.11 3 .08 3.04 3.01 9 5.12 4.26 3 . Mosimann. For example, with m = 2, the vector x may be regarded as representing the point in the plane with coordinates x 1 and x2 • Vectors have the geometrical properties of length and direction. Cross (x3) . Clean the data and perform any variable transformation that seems appropriate. A • • • Li ke l i hood Ratio Tests for Regression Parameters The multiresponse analog of (7-15), the hypothesis that the responses do not depend on Zq+ 1 , Zq+2 , , Zr , becomes • • • H0 : /3 (2) = 0 where {3 = /3 ( ) ((q+ 1)1X m) /3 (2) ((r - q) X m) (7-43) Setting Z] (n XZ(r2- q)) [(n XZ(q+l)) 1 = /3 (2) = 0, Y = 393 'we can write the general model as Under the quantities involved in the H0 : M u ltivari ate M u ltiple Reg ression Section 7.7 Z 1 /3(l) + e and the likelihood ratio test of is based on H0 extra sum of squares and cross products = (Y - Z 1 fJ (l)) ' (Y - Z 1 /3 (l)) - (Y - Z {J)' (Y - Z /3) = n(1 {J (l J = (Z!Z 1 r1 Z lY i 1 = n- 1 (Y - z l {J (l l) ' (Y - z l {J (l J) · I - I) " " " " whereFrom the likelihandood ratio, can be expres ed in terms of generalized variances: (Equivalently, I@I I can be used. The volume generated by deviation vectors of the standardized variables is il lustrated in Figure 3.10 for the two sets of deviation vectors graphed in Figure 3.6. A comparison of Figures 3.10 and 3.6 reveals that the influence of the d 2 vector (large variability in x2) on the squared volume I S I is much greater than its influence on the squared volume I R I · The quantities I S I and I R I are connected by the relationship (3-21) l S I = (s1 1 s22 · . You may verify that e2 = [1/V'I8, - 1/VIS , -4/ V'I8] is also an eigenvector for 9 = A2 , and e3 = [2/3, -2/3, 1/3] is the normalized eigenvector correspond ing to the eigenvalue A3 = 18. + IJL2m i = 1, 2 , . In this way we hope to make the book accessible to a wide audience. In the example to follow, we briefly discuss the results of applying lo gistic regression, decision tree methodology, and a neural network to the problem of credit scoring (determining good credit risks) using a publicly available data set known as the German Credit data. S. ' · ' . We illustrate the profile approximation method using the row profiles; however, an analogous solution re sults if we were to begin with the column profiles. Then (BC 2) ' (BC 2 SC2 B ')-1(BC2) = C2B ' (B ')-1 (C 2 SC2) -1 B-1BC 2 = C2(C 2 SC2) - 1 C 2 , so T2 computed with C 2 or C 1 BC 2 gives the same result. Figure 4.10 on page 190 as shows a situation with two unusual observations. ' 25 . That is, suppose the true model has = with rank r + 1 and ! Z [Z 1 Z2] f3 1 ( ) + q X (Y(nX 1) [(nx(Zq1+1) (nXZ(r2-q) J (rf3q()2X) 11)) Ze)1 13 (1)a21.Z2 P (2) e E(e) (Y Z ' (Y Z) /3 f3 1 1 1 1 ·) ((1 (Z1 ZE1()p Z1 1)Y. tFrhatomRes is, I A uI l=t 2A.11(e), II AP' II =PII·PApplAP' yI =ExerI PcI isAeP'I2.1 1.= I P I A I P' I = I A I I I , since I I I = I P' P I = mat r i x Show2 tI hQatQI'QI =I = 1Alorso-1, froifmRes . The Principal Com po nent (and Pri nci pal Factor) Method The spectral decomposition of (2-20) provides us with one factoring of the covariance ma trix I . p (9-31) 0 _ 1 2 '(, ' · the maximum likelihood estimates of the communalities, and we evaluate the im portance of the factors on the basis of Proportion of total (standardized) = fL + e@ j + . uTherementdatas werare die rseppeatlayededin Tabl5et6.im4es. 140 1 1 . Can we regard the observations here as being normally distributed? At this point, we settle for delineating some simple conditions for S to be of full rank or of reduced rank. -1 (6-42) pg(g g ̄ nk . • The calculations required for Q-Q plots are easily programmed for electronic computers. Their linear combinations span the space of all linear combinations of the columns of Z. It is desirable to make L well defined by imposing the computationally convenient uniqueness condition == a diagonal matrix (9-26) " " The maximum likelihood estimates L and '\}I must be obtained by numerical maximization of (9-25). in itahnceeir riensonepectisveet ofsetvars. , O J = 1 0 Therefore, n -2 max ̄ xjQyj Q j= l Finally, we verify that QQ ' matrix, as required. Thus, x = E' y. Wilk. = - = 280 Chapter 6 Comparisons of Seve ra l M u ltivariate Means , A confidence region for contrasts CJL , with IL the mean of a normal populatio n, is determined by the set of all CJL such that (n - 1) (q - 1) , 1 (6 -17) n(C x - C p) (CSC) - (C x - Cp) 1 (a) l (n q +) Fq- l, n-q+ _ _ _ x§(.01) 20.09. 1 1 Any pair of contrast matrices C and C must be related by C 1 BC , with B nonsingular. In turn, an analysis of the data gathered by experi mentation or observation will usually suggest a modified explanation of the phe nomenon. Relate these quan tities to Sn and R. a1 1 a1 2 a1 3 a 2 1 a 22 a 2 3 a3 1 a 3 2 a33 = a 1 1 a 22 a 33 + a 1 2 a 2 3a3 1 + a 2 1 a 3 2 a 1 3 - a 3 1 a 22 a 1 3 - a 2 1 a 1 2 a 33 - a 3 2 a 2 3 a 1 1 The determinant of any 3 X 3 matrix can be computed by summing the products of elements along the solid lines and subtracting the products along the dashed lines in the following diagram. 4.9. Refer to Exercise 4.8, but modify the construction by replacing the break point 1 by c so that -X1 if -c < X1 < c X2 = xl elsewhere Show that c can be chosen so that Cov (X1 , X2) = 0, but that the two random variables are not independent. 970 - . (See [12] .) In practice, however, the two frequently produce comparable factor loadings if the num ber of variables is large and the number of common factors is small. • 7 • • • • • • • l • • • x • (c) Figure 3.7 Scatter plots with th ree d iffe rent orie ntations. Result 4.7 can be used to show that n (X - IL) ' I - 1 (X - IL) has a x distribution (̄ ̄) ̄ ̄ when X is distributed as NP /L, I or, equivalently, when Vn (X - IL) has an Np (O, I) distribution. The following are true for a random vector X having a multivariate normal distribution: 1. Using the least squares estimates f3 , we can form the matrices of 1 Predicted values: Y = z{J = Z (Z' Z)- Z' Y e = Y - Y = [I - Z (Z ' Z) - 1 Z '] Y Residuals: (7- 3 1) I I M u ltiva riate M u ltiple Reg ress ion Section 7 . The distance of P to the origin 0 is obtained by setting y1 == y2 == · · · == Yp == 0 in (1-16). Example 1 .8 (Li nked scatter plots and brushi ng) To illustrate linked two dimensional scatter plots, we refer to the paper-quality data in Table 1 .2. The di agonal elements are zero, and if the other elements are also small, we may subjectively take the m factor model to be appropriate. + a pp (xp - Yp) 2 + 2a 1 2 (X1 - Yl) (x2 - Y2) + 2a 1 3 (x 1 - Y1) (X3 - Y3) + .) This suggests that we calculate the sample variances using the x1 and x2 coordinates and measure distance as in Equation (1-13). 4.19. The solution from either method can be rotated in order to simplify the interpretation of factors, as described in Section 9.4. It is al ways prudent to try more than one method of solution; if the factor model is appro priate for the problem at hand, the solutions should be consistent with one another. 82 1 .77 1 .72 1 .68 1 . The di mension of a matrix is frequently indicated in parentheses below the letter repre senting the matrix. R. , A; > 0 = A; + 1 , A; + 2 , . k =s + 1 Whatever the correspondence matrix P , the term rc' always provides a (the best) rank one approximation. 0 700 1. Y e' e f3 , 1 V2iT 1 {3, -c' c 1 c- c- j 2 a2 (y - z{3) (y - z{3) I {J f3 . 2.21. A direct approach to obtaining a biplot starts from the singular value decompo sition (see Result 2A.15), which first expresses the n X p mean corrected matrix X c as Xc (n xp) == U A V' (12-41) (n xp) (pXp) (pXp) where A == diag (A 1 , A2 , . 54 2.45 2.39 2.34 2.30 2.27 2.25 2.21 2.17 2.12 2.10 2.08 2.05 2.03 " 12 3.18 2.81 2.61 2.48 2.39 2.33 2.28 2.24 2.21 2.19 2.15 2.10 2.06 2.03 2.01 1 . Nonetheless, some mathematical sophistication and a desire to think quan titatively will be required. [Zjo , Zj r, . If we take into account the vari ability of the points in the cluster and measure distance by the statistical distance in (1-20) , then Q will be closer to P than to 0. Discuss your answer. Similarly, let the discrete random variable X2 have the probability function 0 1 .8 .2 Then E(X2) = Thus, 2: x2 p2 (x2) all x2 E(X) = = (0) (.8) + (1) (.2) E(X1)] [· 1] [E(X .2 2) = = .2. The sample canonical c@rrelations Pk are, therefore, the maximum likelihood estimates of Pk and v'n/ (n - 1) ab v'n/ (n - 1) bk are the maximum likelihood estimates of ak and bk , respectively. ,d i - 1 ' di + 1 , · · · , dp . Supplement 1 2A Data M i n i ng 735 Partition Figure 1 2.28 Neura l S AMPS l O . Then the product ex is a vector with ith entry cxi . Can you interpret the principal components? 4.28. @ 1 00 - I Year / I 3 I ...c= / +-l Year @ 0- I 150 - . These measurements (commonly called data) must frequently be arranged and displayed in various ways. 4.38. U'1 N "'0 "'0 (1) ::::::l a.. @ (x3) Cross • J • •. Consequently, we first calculate the stan dardized values, Zj k = (xj k - xk)/ @ , so the variables contribute equally to 18 Chapter 1 Aspects of M u ltiva riate Ana lysis 15 5 HLS 90 SVL 3 Figure 1 .6 3 D scatter plot of l i zard data from Ta ble 1 . From (4-7), these axes are given by the eigenvalues and eigenvectors of I. The probability that the interval contains JL is 1 - a; among large numbers of such independent intervals, approximately 100 (1 - a)% of them will contain JL . That is, for any prescribed accuracy s > -s < Y J.L < s J approaches unity as n @ oo . Sample variance = == 3 3 _1 In a similar manner, the n c x 1 c'x2 c'x3 == 3 observations on c'X are 1x 1 1 - 1x 12 + 3x 1 3 == 1 (1) - 1 (2) + 3 (5) = 14 = 1 (4) - 1 (1) + 3 (6) == 21 == 1 (4) - 1 (0) + 3(4) == 16 == and (1 4 + 2 1 + 16) Samp1e mean == - _ _ _ - 17 3 2 + (16 - 17) 2 (14--- 17) 2 + (21---17) -----Sample variance == _ 3 1 == 13 1 44 Chapter 3 Sa m p l e Geometry and Ra ndom Sa m p l i ng Moreover, the sample covariance, computed from the pairs of observations (b'x 1 , c' x 1) , (b'x 2 , c'x 2) , (b'x3 , c ' x3) , is Sample covariance (1 - 3) (14 - 17) + (4 - 3) (21 - 17) + (4 - 3) (16 - 17) 9 2 3-1 Alternatively, we use the sample mean vector x and sample covariance matrix S derived from the original data matrix X to calculate the sample means, variances, and covariances for the linear combinations. 8 6 i . Summary numbers, which quantitatively portray certain features of the data, are also necessary to any description. The matrices I, I, and E given after Definition 2A.13 are square matrices. We can us e Res u l t 6. TABLE 1 2 . • 7.6 M O D E L CH ECKI NG AND OTH ER ASPECTS OF REGRESSION Does the Model Fit? Chapter 1 2 TABLE 1 2. 2.38. Build the model on the training set. If not, the normality assumption is suspect. Construct a scree plot to aid your determination. , p W. The eigenvector e1 fol ows from the vec [.4217871 .1096] (.5458) Thus, [.8947, .4466] and - [..28776561]] [.8561] [.4054436] [We must scale so that Z) == Z == [Z (l) , Z (2) ' P2 1 l P22 p-1 11 /2 - -- -- == p*1 2 , p2* 2 _ .2 - .2 3 - . title 'MAN OVA'; data film; i n fi l e 'T6-4.dat'; i n put x1 x2 x3 facto r 1 facto r2; proc g l m d ata = fi l m ; cl ass fact o r 1 facto r 2; m odel x1 x2 x3 = facto r 1 factor2 facto r 1 * factor2 jss3; man ova h = fact o r 1 factor2 factor1 * factor2 j pri nte; mea ns fact or 1 fact or2; PROGRAM COM MAN DS General Linear Models Proced u re Cl ass Leve l I nformation Leve ls Cl ass 2 FACTOR 1 2 FACTOR2 N u m ber of observations i n So u r ce Mod el E rro r Cor rected Tot a I So u rce So u rce Mod e l E rro r Cor rected Tot a I So u rce OUTPUT Va l u es 0 1 0 1 data set = 2 0 Pr > F 0.0023 DF 3 16 19 S u m o f Sq u a res 2 . (b) Perform a principal component analysis using only x 1 through x4 • Int er pret your results. Using (1-13), we see that all points which have coordinates (x 1 , x2) and are a constant squared distance c 2 from the origin must satisfy 2 ̄ S1 1 + 2 ̄ = c2 S22 (1-14) Equation (1 -14) is the equation of an ellipse centered at the origin whose major and minor axes coincide with the coordinate axes. This generalization of variance is an integral part of the comparison of multivariate means. The Q-Q plots constructed from the marginal distributions of xl ' x2 ' and x3 all ap proximate straight lines. 92 1 . + sPP for a factor analysis of S (9-20) " Aj for a factor analysis of R p Criterion (9-20) is frequently used as a heuristic device for determining the appropriate number of common factors. By (12-30) , D@ 1 12 PD@ 1/2 has the singular-value decomposition I D @ 1;2 PD@ 1;2 == @ A'k u k v k k=1 (12-36) 718 Chapter 1 2 Cl usteri ng, D i sta nce Methods, a n d Ord i nation The best rank K approximation is obtained by using the first K terms of this expan sion. Consider the q linear combinations ai l xl + ai 2 x2 + 0 0 0 + ai p xp , i = 1, 2, 0 0 0 ' (3-37) q These can be expressed in matrix notation as a l l xl a 2l xl + + a l2 x2 a 22 x2 + a q l xl + a q 2 X2 + + 0 0 0 0 0 0 0 0 0 + a l p xp a 2 p xp + a q p xp + = a1 1 a1 2 a 2 1 a 22 a lp a2 p xl x2 aq l aq 2 aq p xp = AX (3-38) Taking the ith row of A, ai , to be b' and the kth row of A, ak , to be c' , we see that Equations (3-36) imply that the ith row of AX has sample mean ai x and the ith and kth rows of AX have sample covariance aiSa k . (c) Since x = 3 and y = 1, graph [5 - 3, 1 - 3, 3 - 3] = [2 -2, OJ and [-1 - 1, 3 - 1, 1 - 1] = [-2, 2, O J . 1 3 Com b i n ed g rowth cu rves fo r we ight fo r seven fem a l e g rizzly bears. 2.6. Let] [A= (a) Is A symmetric? Thus, the inclusion of the third variable, which is linearly related to the first two, has led to the case of a zero generalized variance. 00000000 46.00000000 Mean Sq u a re 40. Researchers in the biological, physical, and social sciences frequently collect measurements on several variables. Example 3.2 (Data as p vectors in n d i mensions) Plot the following data as p = 2 vectors in n = 3 space: Here y1 = [4, - 1 , 3 J and y2 = [1 , 3, 5] . , xp) and Q = (y1 , y2 , . 55 1 .49 1 .42 1 . Set LCL = and take UCL = Poisuggesnts tabovethat ththeepruppercesscontin quesrol ttiimonit srheoulpredsebent potexamientniaedl sptoecidetalecausrmineevarwhetiatihoern andim mediate corrective action is warranted. if max A = OEeo < c (5-16) max · · · , · · , 8 . We have it that xi = (x l i . Expanding the determinant 0' B 0' 0' 0' first row (see Definition 2A.24) gives 1 times a determinant of the same form, with the order of I reduced by one. Verrill, S., and R. "Transformations of Multivariate Data." Biometrics, 27 , no. Definition 2A.12. It is not difficult to show (see Exercise 8.5) that the p eigenvalues of the corre lation matrix (8- 1 5) can be divid-ed into two groups. We reject H0 , that JLo is a plausible value of JL , if the observed I t I exceeds a specified percentage point of a !-distribution with n - 1 d.f. Rejecting H0 when I t I is large is equivalent to rejecting H0 if its square, 2 2 -1 2t = (X - JLo) = n (X) (s) (X - JLo) (5-1) JLo s 2jn is large. 8 80 (8 H0 I 8], H0 8 81 H0 H0 I 8 H0: 00 • 5.4 CO N F I D E N CE REG I O N S AN D S I M U LTAN EOUS CO M PARISONS OF COMPO N E NT M EANS To obtain our primary method for making inferences from a sample, we need to ex tend the concept of a univariate confidence interval to a multivariate confidence re gion. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size. The walleye is the most popular fish to eat. However , t h e ar e no l o nger gi v en by (6 6 7) and t h e ex e compl (6-68). 19. New York: John Wiley, 1984. (b) Show that A is positive definite. Section 8.3 (S u m marizing Sample Va riation by Principal Components) 447 Using (8-29), we see that the proportion of the total sample variance explained by the ith sample principal component is Proportion of (standardized) sample variance due to ith sample principal component == A. 1 0 Genera l ized Va riance 1 33 (Creati ng new variables that lead to a zero genera l i zed varia nce) Consider the data matrix 1 9 10 4 12 16 X = 2 10 12 5 8 13 3 1 1 14 where the third column is the sum of first two columns. 1 4 = 30. [See (4-31).] Let the significance level be a = .1 0. What interpretation, if any, can you give to the first principal component? Additional insights can sometimes be gleaned from visual inspection of the slowly spinning data. Order the original observations to get x (l) , X (z) , - • . Includes bibliographical references and index. In this situation, the columns are said to be colinear. What is important is the appearance of the cross product term 2a 1 2 x 1 x2 necessitated by the nonzero correlation r1 2 . the choi-1ce f3 = (c' - a' Z) '. Chapter 4 4.31. > > Z Z' Z and e 1 , e 2 , . 122 .022 0 .014 - .001 - .017 0 .009 .067 - .129 - .215 0 .036 .041 - .254 - .005 -.022 .062 - .109 -.112 .076 .020 - .091 0 498 Chapter 9 Factor Ana lysis and I nference for Structu red Covaria nce Matrices Maximum likelihood: R - LL' - W " 0 .000 .000 .012 .000 -.012 .004 .000 - .018 .000 = 0 .000 .002 - .002 .006 - .025 - .009 - .000 .000 0 .000 0 .000 - .033 - .000 .001 .028 0 .036 - .000 - .034 -.002 .006 .008 - .012 - .000 - .000 - .045 .052 - .013 .000 .000 .000 0 .043 .016 .000 0 .091 .000 0 .000 0 • A Large Sample Test for the N u m ber of Co mmon Facto rs The assumption of a normal population leads directly to a test of the adequacy of the model. (c) There exists a matrix A- 1 such that AA- 1 A- 1 A I . • Example 1 .4 (A scatter plot for baseba l l data) In a July 17, 1978, article on money in sports, Sports Illustrated magazine pro vided data on x 1 = player payroll for National Baseball teams. The essential purpose of factor analysis is to describe, if possible, the covari ance relationships among many variables in terms of a few underlying, but unob servable, random quantities called factors. Most of the variation can be explained by • a single variable determined by a line through the cloud of points. I . 1)a- , 2s = e e+ Y'[I - Z(- Z '-Z)-1Z'] Y Y'[I - H]Y E(s2) = a-2 p e Y = zp + e p = (Z ' Z) - 1 Z ' Y = (Z ' Z) - 1 Z ' (Z p + e) = p + (Z ' Z) - 1 Z ' e e == [[II -- Z(Z(ZZ''ZZ))--11ZZ'']] Y[Zp + e] = [I - Z(Z ' Z)-1Z'] e estimator has P The residuals and Cov ( ) have the properties 0 and Cov (n - n - so defining r - (r n 1) r n-r-1 1 we have Moreover, Now, Proof. 7627 Section 6.9 Perspectives and a Strategy for Analyz i n g M u ltivariate Models Since, with .01, 327 a = g)) ln A* - (31 - @ (4 - 2 2)) 1n .7627 9. This component might be called a general stock-market component, or simply a market component. Example 4.6 4014 Oo) • (The eq u ivalence of zero cova ria nce and i ndependence fo r normal variabl es) 4 1 0 1 O 3 o 0 2 1, 4 1 0 1 lO 3 !o 0 i2 1 _(@-:>:-@@i-@@-:>:-@ L i 4050 • 0 0 4-6)] 0 405 10 Result 4.6. > 00 Section 4.2 The M u ltivari ate Normal Density and Its Properties 1 61 and Notvariaeblltheat. The scatter plot of n points in p-dimensional space provides information on the locations and variability of the points. , p, . 436 Chapter 8 Principa l Components Standardization does not substantially alter the situation for the I n (8- 1 3). 98 6 5859. The gender, by row, for the lizard data in Table 1.3 are f m f f m f m f m f m f m m m m f m m m f f m f f Figure 1.8 repeats the scatter plot for the original variables but with males marked by solid circles and females by open circles. Using (8-7) with Z in place of X, we find that the proportion of total variance explained by the kth princi pal component of Z is (Proportion of (standardized) population variance due to kth principal component where the Ak ' s are the eigenvalues of p . (To simplify the notation, the subscript n will be dropped when the dimension of the vector 1n is clear from the context.) The vector 1 forms equal angles with each of the n coordi nate axes, so the vector (1/Yn)1 has unit length in the equal-angle direction. When p is near 1, the last p - 1 components collectively contribute very little to the total variance and can often be neglected. Since Ie k = Ak e k , premultiplication by e; gives Cov (I: , Yk) = e;Ie k = eiA. 08000000 0.3 224903 1 N 10 10 Level of FACTOR 1 0 Leve l of FACTOR2 0 E = E rror SS&CP Matrix - - - - - - - - - X1 - - - - - - - - - Mean SD 6 . Linden reports that the distributions of standard scores were normal or approximately normal for each of the ten decathlon events. 16 "' * P1 .55 . Figure 12.22 gives an example of a biplot for the public utilities data in Table 12.5. You can see how the companies group together and which variables contribute to their positioning within this representation. The question naturally arises as to how well the solutions coincide. Generalized variance is easier to inter pret when the two or more samples (patterns) being compared have nearly the same orientations. El. X8 2 3 A b i p l ot of the data on p u b l i c util ities. They are, in effect, plots of the sample quantile versus the quantile one would expect to observe if the observations actually were normally distributed. In this case, we compute l S I = (252.04) (123.67) - (-68.43) (-68.43) = 26,487 • The generalized sample variance provides one way of writing the information on all variances and covariances as a single number. To match this formulation more closely, we subtract $1 from the entries in the first row of the cost matrix to obtain the "realistic" cost matrix: 736 Chapter 1 2 Clustering, Dista nce Methods, and Ord i nation Predicted (Decision) Actual Good (Accept) Bad (Rej ect) -$1 $5 0 0 Good Bad This matrix yields the same decisions as the original cost matrix, but the results are easier to interpret relative to the expected cost obj ective function. • Data mining procedures and software continue to evolve, and it is difficult to predict what the future might bring. For example, clicking on Score executes the process up to and includ ing the Score icon. Data analysis, while interesting with one variable, becomes truly fasci nating and challenging when several variables are involved. m k :L :L i = l j=l (a i j - bi j) 2 = tr [(A - B) (A - B) '] Result 2A.16. Hartigan, J. Since (n - 1) s = (X - lx ') ' (X - lx ') (pX n) (pX p) (n Xp) the kth column of S, col k (S) , can be written as a linear combination of the rows of (X - lx') ' . Because x =I= 0 and B is positive definite, x ' B x > 0. Expected cost scores for these applicants were created using the neural network model. , d P = (yp - xp 1), divided by n - 1 . If X is the sum X = Vi + Vi + · · · + Vn then the central limit theorem applies, and we conclude that X has a distribution that is nearly normal. At this point, we have no idea which of these hypothesized val ues may not be supported by the data. Kirkpatrick, M. 1 6 4.01 3.86 3.76 3.70 3.62 3.54 13 9.07 6.70 5.74 5.21 4.86 4.62 4.44 4.30 4.19 4.10 3.96 3.82 3.66 3.57 3.51 3 .43 3.34 14 8.86 6.51 5.56 5 .04 4.69 4.46 4.28 4.14 4.03 3.94 3.80 3.66 3.51 3.41 3.35 3 .27 3.18 15 8.68 6.36 5 .42 4.89 4.56 4.32 4.14 4.00 3.89 3.80 3.67 3.52 3 .37 3.28 3.21 3.13 3.05 16 8.53 6.23 5.29 4.77 4.44 4.20 4.03 3.89 3.78 3.69 3.55 3.41 3.26 3.16 3.10 3 .02 2.93 17 8.40 6.11 5.19 4.67 4.34 4.10 3 . 135 .-=· · ' e p]. Algebraically, the row profiles are the rows of the matrix D@ 1 P , and contin gency analysis can be defined as the approximation of the row profiles by points in a low-dimensional space. Once S is computed, it can be related to the sample correlation matrix R. Even when the normal assumption is violated, the confidence intervals obtained in this manner still provide some indication of the uncertainty in ;\ i and e i . Other choices of a 1 and a 2 will produce simultaneous intervals that do not contain zero. For example, the two-dimensional coordinate axes might represent latitude and longitude (geographical location) , and the faces might represent multivariate measurements on several U.S. cities. .--___-___-1Since s V1 + z 0 (Z ' Z) _ z 0 = (1.204) (1.16071) = 1.40, a 95% prediction interval for the CPU time at a new facility with conditions z0 is z 'o P ± t4 (.025)s V'1 + z0(Z' Z) - 1 z0 = 151.97 ± 2.776 (1.40) or (148.08, 155.86) . A scatter plot of the data would look something like the one pictured in Figure 1 .20. , min(m, k) and the other entries are zero. (c) Order the distances in Part a and construct a chi-square plot. Definition 2A.2 (Scalar Multiplication). The classification tree, in turn, outperforms a random assignment. Allied Chemical Du Pont Union Carbide Exxon Texaco Cumulative proportion of total (standarc.lized) sample variance explained Specific variances @l· = 1 - it?l Estimated factor loadings Fl F2 .189 .517 .248 - .073 -.442 .684 .694 .681 .621 .792 .50 .25 .47 .61 .18 Estimated factor loadings F2 Fl .783 .773 .794 .713 .712 - .217 - .458 - .234 .412 .524 Specific variances @·l = 1 h2l - - .34 .19 .31 .27 .22 - .485 .571 .598 .733 figures for the m = 2 factor solution obtained by the principal component method (see Example 9.4) are also provided. The factor model postulates that X is linearly dependent upon a few unob servable random variables F1 , F2 , , 64 60 4.00 3.15 2.76 2.53 2.37 2.25 2.17 2.10 2.04 1 . Section 3.4 Example 3.7 Genera l ized Va ria nce 1 25 (Calcu lati ng a genera l i zed varia nce) Employees (x 1) and profits per employee (x2) for the 16 largest publishing firms in the United States are shown in Figure 1.3. The sample covariance ma trix, obtained from the data in the April 30, 1990, Forbes magazine article, is s = [252.04 -68.43 -68.43 123.67] Evaluate the generalized variance. (a) SA (b) BA (c) A' B ' (d) C' B (e) Is AB defined? To further evaluate multivariate normality, we constructed the chi-square plot shown in Figure 4.9. The two specimens with the largest squared distances are clearly removed from the straight-line pattern. , p, between a component of X (1) and a component of X (2) . 1 9 60.71 61 .22 10 39.86 49.50 53.59 2 8.53 9.00 9.16 9.24 9.29 9.33 9.35 9.37 9.38 9.39 9.41 9.42 3 5.54 5.46 5.39 5 . , x k , then z is perpendicular to their linear span. Repeat for the sub matShowrixthat tobthe detainedermbyinantdeleoftinagstqhuare fiersstyrmmet ow andricfirst columatmnriofx AA.can be ex pres eFrd asom(th2e-pr16)and(oduct of2-2i0)ts ,eiAgenval= PAPues'with P' P = I. Replac ing I - 1 by s- 1 does not seriously affect this approximation for n large and much greater than p. eFurvaltuheerofinvesexttriagoratdioinn,arasy event overtime during that period. Definition 2A.27. , q , j = q + 1, q + 2, . Using the normal approximation to the sampling distribution of Pi (see [9]), we observe that either I Pi l - .683 1 > 3 (.683) (.317) n 1.396 Vn or I Pi 2 - .954 1 > 3 (.954) (.046) n .628 Vn (4-29) would indicate departures from an assumed normal distribution for the ith charac teristic. Three components, X1 = sweat rate, X2 = sodium content, and X3 = potassium content, were mea sured, and the results, which we call the sweat data, are presented in Table 5.1 . Two matrices Ak {bi j } are said to be { ai j } and B k (mX) (m x) eq ual, written A B, if ai j == bi j , i 1, 2, . Section 8.3 Example 8.6 Summarizing Sa m p l e Va riation by Principa l Components 449 (Co mponents from a correlation matrix with a special structure) Geneticists are often concerned with the inheritance of characteristics that can be measured several times during an animal ' s lifetime . 12.16. 482 Chapter 9 Factor Ana lysis and Inference for Structu red Covaria nce Matrices Unfortunately for the factor analyst, most covariance matrices cannot be fac tored as LL' + '\}I, where the number of factors m is much less than p. + ;]; 8 (r)t;(r)) " r*+1 8(r +1) b(r +1) 1 + . The relative sizes of the coefficients of X1 and X2 suggest, however, that x2 contributes more to the determination of Yi than does xl . The corresponding eigenvectors are e1 = [1/ v'5 , 2/ v'5] and e2 =

configurations, y = 0. Vectors of the same dimension that are not linearly dependent are said to be linearly independent. Results associated with individual icons can then be exam ined by clicking on the appropriate icon. Zi + Z⃗ + A = I, (1/ A.J e i . The biplot is given in Fig ure 12.23 on page 723 . You expect 1 or 2 of these to exceed 3 or be less than -3, even if the data came from a multivariate distribution that is exactly normal. Westphal, C, and T. The cost matrix was initially specified as follows: Predicted (Decision) Actual Good Bad Good (Accept) Bad (Rej ect) 0 $5 $1 0 so that it is 5 times as costly to classify a Bad applicant as Good (Accept) as it is to classify a Good applicant as Bad (Reject) . Companies with well-known statistical software packages now offer comprehensive data mining programs. 14 gives the array of seven curves for weight. Our objective is to introduce several useful multivariate techniques in a clear manner, making heavy use of illustrative examples and a minimum of mathe matics. 68 Chapter 2 Matrix Algebra and Random Vectors lao oo if Xi j is a continuous random variable with Xl· j..f. Current estimation and rotation methods require iterative calculations that must be done on a computer. Do the errors appear to be normally distributed? The equation I A AI I = 0 (as a function of A) is called the characteristic equation. Then where PP ' = P ' P = I and A is the diagonal matrix Al 0 0 A2 A = (kXk) 0 0 0 0 Ak with Ai > 0 Section 2 . Next, determine the length of x, the length of y, and the angle between x and y. The sec ond principal component appears to contrast health services employment with a weighted average of median school years and median home value. C H A PTER 3 Sample Geometry and Random Sam pling 3.1 INTRODUCTION With the vector concepts introduced in the previous chapter, we can now delve deep er into the geometrical interpretations of the descriptive statistics x, Sn , and R; we do so in Section 3.2. Many of our explanations use the representation of the columns of X as p vectors in n dimensions. Here x is the 2p X 1 vector of sample averages for the p variables on the two treatments given by and S is the 2p X [ ] (6-11) 2p matrix of sample variances and covariances arranged as S S S 11 12 pxp) pxp) ((= S2 1 S 22 (pXp) (pXp) (6-12) 278 Chapter 6 Comparisons of Seve ra l M u ltivariate Means The matrix S 1 1 contains the sample variances and covariances for the p variables on treatment 1. 50 5 x2 1 .0 x3 . 93 1 . Your job is to summarize these data in fewer than p = 7 dimensions if possible. Show that, for = _ 0, 4.13. 0 005 . The data are ordinarily standardized by using the computer program as part of the process for determining the locations, sizes, and orientations of the facial char acteristics. As expected, the plots have a straight-line pattern, but the top two or three ordered squared distances are quite variable. The resulting plot with n points not only will exhibit the overall pattern of variabili ty, but also will show similarities (and differences) among the n items. Based on these counts, a linguist proposed the following distances between chapters of the Old Testa ment book Lamentations (data courtesy of Y. 82 1 .77 29 4.18 3.33 2.93 2.70 2.55 2.43 2.35 2.28 2.22 2.18 2.10 2.03 1 . 22 2 2 x ___2_ 4 + x ⃗ = 1. Bartlett [3] sug gests that the test statistic - 2[1 - (2p + 1 1)/6n J ln A be used in place of -2 ln A. P - rc' is :L Ak (D;I2 uk) (D⃗I2 vk) ' k= 1 (12-29) Section 1 2 .6 Correspondence Ana lysis 713 where the A k are the singular values and the I X 1 vectors u k and the J X 1 vectors vk are the corresE. _9 nding singular vectors of the I X J matrix D ⃗ 1 12 (P - rc ') D⃗112 • Here Ak = A k + 1 , u k = uk + 1 , and vk = vk + 1 for k = 1, . + 8 (P) 8 (p) 1 -b (l) -b (l) l + -b (2)b(2) 1 + + -b (q) -b (q) l · · · ıJ .A- 1 Si n ce and has s a mpl e covar i a nce t h e f i r s t r c o f l u mns ⃗ A A conttheiracomponent in the samplvare covar i a nces of t h e f i r s t canoni c al var i a t e s , wi t h 1 l l , X i a bl e s Xi , Si m i l a r l y , t h e f i r s t col u mns of X ⃗ contain the sample covariances of V, with their component variables. Mor e over , wr i t i n g t h e coefand ftihceiierntgeomet vectorrsicasinterpretationsand. 65 1 .60 1 .55 1 .48 1 .45 1 .41 1 .37 1 . Z Z (f3 (i) - f3 (i)) =I= P u) - f3u)1 tr [?; (/3 - /J)I - (/3 - /J) ' Z'] ' I- 1 Z;_(/J - /J). Perform a principal components analysis using the covariance matrix S of the speed data. , xn i] is deter : I . This follows because (change in yj) = hjj (change in yj) , provided that other y values remain fixed. Does it make any difference which matrix is chosen for analysis? One way to proceed is to divide each coordinate by the sample standard devi ation. Exercise 1.4 contains data on three variables for the 10 largest industrial cor porations as of April 1990. As we have mentioned, a scree plot is also useful for selecting the appropriate number of components. In a chi-square plot, these would be the points farthest from the origin. · + l3r 1 Zr + e 1 Y2 = 13 o 2 + l31 2 Z1 + · · · + 13r 2 Zr + e2 m (7-25) Ym = 13 o m + 13 I m Z1 + · · · + l3rm Zr + e m = [e 1 , e2 , . Using the vectors b' = [-4, 3] and d' = [1, 1], verify the extended Cauchy Schwarz inequality (b' d) 2 < (b' Bb) (d' B- 1 d) if B= [-⃗ -⃗] 2.36. 179 2.160 2.145 2.131 2.120 2.110 2.101 2.093 2.086 2.080 2.074 2.069 2.064 2.060 2.056 2.052 2.048 2.045 2.042 2.021 2.000 1 .980 1 .960 31.821 6.965 4.541 3.747 3.365 3. (See Supplement 9A.) As a consequence of the factorization of (9-30), whenever the maximum like lihood analysis pertains to the correlation matrix, we call " h2l - '(,·IJl2 + '(,·IJ l2 + '(,·IJ l2 + . Zantinge. Box, G. , 1 - 1 . "Size and Shape Variation in the Painted Turtle : A Principal Component Analysis." Growth, 24 (1960), 339-354. For an or thogonal matrix, AA' = A' A = I, so the columns are also mutually perpendicular • and have unit lengths. 67 1 .64 1.61 25 2.92 2.53 2.32 2.18 2.09 2.02 1 .97 1 .93 1 .89 1 .87 1 .82 1 .77 1 .72 1 .68 1 .66 1 .63 1 . ··· Result 2A.7. Let A be a nonsingular square matrix of dimension k X k. Artis, Fo recasting Computer Require ments: A Forecaster's Dilemma (Piscataway, NJ: Bell Laboratories, 1 979). This implies, from Result 2A.9, that I S I = 0. 0 1 . Hence, 0 AB (mXn)(n X k) (mX k) does not imply that A = 0 or B = 0. Even so, you must be forewarned that if the column sum in (2-12) is nearly 0 for some constants c1 , . Then, geometrically, the corresponding deviation vector d i = (yi - xi l) will be very long or very short and will therefore clearly be an important factor in deter mining volume. That is, the x1 values are plotted along the horizontal axis, and the x3 values are plotted along the vertical axis. 1 , e 1) , (A2 , e 2) , . Is it possible to summarize the radiotherapy data with a single reaction-index component? Anderson, T. Ltd Pearson Education North Asia Ltd, Hong Kong Pearson Education Canada, Ltd., Toronto Pearson Education de Mexico, S.A. de C.V. Pearson Education-Japan, Tokyo Pearson Education Malaysia, Pte. U.S. Department of Commerce, National Bureau of Standards Applied Mathematical Se ries. ⃗1-.('----- LX ( ) cos (8) x' y y y' y --)lt.....-JI Figure 2 . From (2-6) , we get or, using (3-5) and (3-6), we obtain ⃗ (xi ; - X;) (xi k - Xk) = so that [see (1-5)] ⃗⃗ (xi ; - X;) 2 ⃗⃗ 2 (xik - Xk) cos (O ;k) (3-7) The cosine of the angle is the sample correlation coefficient. • I nterpretati on of the Sample Pri nci pal Com ponents The sample principal components have several interpretations. The inequality is obvious when b = 0 or d = 0. 1) 2 PI (x l) all x 1 = (- 1 - .1) 2 (.3) + (0 - .1) 2 (.3) + (1 - .1) 2 (.4) = .69 o-22 = E(X2 - JL2) 2 = ⃗ (x2 - .2) 2 P2 (x2) all x2 = (0 - .2) 2 (.8) + (1 - .2) 2 (.2) = .16 o-1 2 = E(X1 - JL 1) (X2 - JL2) = ⃗ all pairs (x 1 , x 2) (xi - .1) (x2 - .2) PI 2 (x l , x2) = (- 1 - .1) (0 - .2) (.24) + (- 1 - .1) (1 - .2) (.06) + . Title. (Use the row proportions as a vector of observations at a site.) 12.24. 50 1 50000 1 . 1 A scatter plot a n d m a rg i na l dot d i a g rams. In turn, the communalities would then be (re)estimated by r (9-23) The principal factor solution can be obtained iteratively, with the communality esti mates of (9-23) becoming the initial estimates for the next stage. The expression in (1-13) can be regarded as a special case of (1-19) with a 1 1 = 1/ s 1 1 , a 22 = 1/ s22 , and a 12 = 0. Bonine. 03 2.64 2.44 2.31 2.22 2.15 2.10 2.06 2.03 2.00 1 . 5 (Facto r ana lysis o f stock-price data u s i n g the maxim u m l i ke l i hood method) The stock-price data of Examples 8.5 and 9.4 were reanalyzed assuming an m = 2 factor model and using the maximum likelihood method. As we noted in Section 9.2, the factors (and loadings) are unique up to an orthogonal rotation. Perform a principal component analysis using S. If the na tions are ranked on the basis of their score on the first principal component, does the subsequent ranking differ from that in Exercise 8.18? Ed. Pac . Thus, the points at distance c lie on an ellipse whose axes are given by the eigenvectors of A with lengths proportional to the reciprocals of the square roots of the eigenvalues. , n . (b) Note from Exercise 4.12 that we can write (x - IL) ' I - 1 (x - IL) as If we group the product so that [1 , 0 -I 1z iZi I][ ]=[x 1 - 1L 1 X2 - IL 2 x 1 - 1L 1 - I 1 2 I 2i(x2 - 1L2) X2 - IL 2] the result follows. s+1 yj . The technical condition that an inverse exists is that the k columns a 1 , a2 , . It is in structive to do a few sample calculations to understand the technique. Here the third variable is actually the sum of the first two variables, so the columns of the original data matrix satisfy a linear constraint with c = 0. 36 [- ⃗ ⃗J - .3 .3 • Plots are important, but frequently neglected, aids in data analysis. The primary question in factor analysis is whether the data are consistent with a pre scribed structure. 8 Procrustes Ana lysis: A Method for Comparing Config u rations 725 Here A, U, and V are obtained from the singular-value decomposition n * L yj xj = (pYx n*)' (X p) n x j= l (⃗ Proof. 8.11. First, x'x = 12 + 3 2 + 2 2 = 14, y' y = (-2) 2 + 12 + (-1) 2 = 6, and x'y = 1 (-2) + 3 (1) + 2 (- 1) = -1. If the list of predictor variables is very long, cost considerations limit the num ber of models that can be examined. We find that the mean corrected data matrix, with entries xj k - xk , is -2 - 1 -3 1 2 3 X - lx' = -1 0 - 1 2 -2 0 1 0 1 The resulting covariance matrix is [ ] 2.5 0 2.5 s = 0 2.5 2.5 2.5 2.5 5.0 We verify that, in this case, the generalized variance I s I = 2.5 2 X 5 + 0 + 0 - 2.5 3 - 2.5 3 - 0 = 0 In general, if the three columns of the data matrix X satisfy a linear con straint a 1 xj 1 + a 2 xj2 + a 3 xj 3 = c, a constant for all j, then a 1 x1 + a 2 x2 + a 3 x3 = c, so that a l (xjl - xl) + a 2 (xj2 - x2) + a 3 (xj 3 - x3) = 0 for all j. 1 and 1 .2 are not discernible from the marginal dot diagrams alone. The value x 5 = 2791 was corrected to 1241, and x4 = 2746 was corrected to 1670. Sta rs for the fi rst five 28 Chapter 1 Aspects of M u ltivariate Ana lysis The observations on all variables were standardized. Stegun, eds. Whenever an investigator can control the assignment of treatments to experi mental units, an appropriate pairing of units and a randomized assignment of treat ments can enhance the statistical analysis. When p variables are observed on each unit, the variation is described by the sample variance-covariance matrix S1 1 S12 S = S1 2 S22 slp s2p slp s2p sPP { s;k = n ⃗ 1 J=l :± (xji - X;) (xjk - Xk) } The sample covariance matrix contains p variances and ⃗ p(p - 1) potentially dif ferent covariances. There are two extreme standardized values. • The approximation provided by the central limit theorem applies to dis crete, as well as continuous, multivariate populations. However, the eigenvector approach also works for high dimensions where the data cannot be examined visually. Section 3.4 Example 3. Section 6.2 Pa i red Comparisons and a Repeated Measu res Design l⃗1 ⃗ ⃗3 1 -1 0 1 0 -1 - J.L 1 - J.L 2 -1 1 0 0 -1 1 1 J.L 1 - J.L2 1 J.L 1 - J.Lq 0 0 0 J.L 1 -1 J.Lq J.L2 279 == C l ⃗-t or J.L2 I-L 3 == 0 0 0 I-L l J.L2 == C ⃗-t 2 0 0 0 - 1 1 J.Lq Both C 1 and C 2 are called contrast matrices, because their q - 1 rows are linearly in J.Lq - J.Lq - 1 dependent and each is a contrast vector. for similar measurements in order to carry out a complete = [+ = = = ( ) = 1' a= c2 = [= + xfp -q-l J g (a) me a ns (6- 7 1) {Fitti ng a q uadratic growth cu rve to ca lci u m loss) A [{3 1 , = + + (B ' S ⃗;olect B) - 1 = ne t2 two wt A* = ⃗ = = . This case is illustrated in Figure 3.9 and may be verified algebraically by showing that I S I = 0. 1 • -.3 •• -.03 • -.01 • • • 8.5 A 0-0 p l ot for the secon d principal component y2 from the d ata on m a l e t u rtles . + l 1 /r o 'P l ' i = 1 , 2, . Figure 7.4 382 Chapter 7 M u ltivariate Linear Reg ress ion Models Step 2. Of course, mathematical at tractiveness per se is of little use to the practitioner. London: Acade mic Press, 1 984. They may, justifiably, be part of the group and may lead to a better understanding of the phenomena being studied. 12.2. Repeat Exercise 12.1 using similarity coefficients 5, 6, and 7 in Table 12.2. 12.3. Show that the sample correlation coefficient [see (12-1 1)] can be written as r for two = ad - be [(a + b) (a + c) (b + d) (c + d) J 112 ------- 0-1 binary variables with the following frequencies: Variable 2 1 0 b 0 a Variable 1 d 1 c 12.4. Show that the monotonicity property holds for the similarity coefficients and 3 in Table 12.2. Hint: (b + c) = p - (a + d) . 1 (1995), 1 17-134. 15. a; ,... 98000000 0.47328638 Leve l of FACTOR2 0 0.3 0 1 8 0.30 1 8 0.30 1 8 - - - - - - - - - X2 - - - - - - - - Mean SD 0. 9 305] . x3 . , A⃗ (for m > k). The random sampling scheme em- ployed ensured that each of the training and validation sets contained about 70% Good applicants and about 30% Bad applicants. An abbreviated assessment of the three approaches will allow us compare the performance of the three approaches on a validation data set. (See Exercise 2.17.) A is a nonnegative definite matrix if and only if all of its eigen values are greater than or equal to zero. main diagonal. ye;- , . That is, if Ax = Ax, we take e = x/ Vx'x as the eigenvector corresponding to A. Hint: Hint: 2.15. Part b follows from (B -1 A- 1) AB = B- 1 (A- 1 A) B = B -1 B = I. The p continuous random variables X1 , X2 , , XP are mutually statistically independent if their joint density can be factored as (2-28) !I2 · · ·p (x l , x2 , . 12 Aspects of M u ltivariate Ana lysis Chapter 1 • • • • • • • x2 x2 10 10 8 8 6 6 4 4 2 2 0 • • • • 4 2 • ! 2 • • ! • 4 • • 6 8 ! ! 6 xl 10 I 8 10 ⃗ Figure 1 . This suggests that the standard ization is not inconsequential. • EXERCI SES 3.1. Given the data matrix Graph the scatter plot in p = 2 dimensions. Using Equations (5-10) and (5-11), we get max L (JL 0 , I) = Likelihood ratio = = :t (5-12) A max L (JL, I) I o 1 1 p, , I i o I is called Wilks ' lambda. Thatfactoirs,Ithevole model al l o ws f o r r e pl i c at i o ns of t h e r e s p ons e s at each combi s . 1 6 5 ..... We shall rely most heavily on descriptive statistics that measure location, vari ation, and linear association. 1 has cericaltacoefin advant a ges . The 3 X 3 identity matrix is shown before this definition. (6 -14) Thus, (6- 1 5) and it is not necessary first to calculate the differences d 1 , d 2 , . An ⃗lbow occurs in the plot in Figure 8.2 at about i = 3. Section 1 . That is, suppose all variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables in a different group. In statistical terminolog(y,BtLUE) he estiofmatc' o/3.r c' f3 is calestliemdatthoer We(7-3des) wictrhibteheinaddiferetntioinalat pr(toecedur e s bas e d on t h e c l a s i c al l i n ear r e gr e s i o n model i n nt a t i v e) as s u mpt i o n t h at t h e r r o r s e have a nor m al di s itnribSectutioion.n 7.Met6. These data produced the following summary statistics: 14 . The fact that IL is the mean of the multivariate normal distribution follows from the symmetry exhibited by the constant-density contours: These contours are centered, or balanced, at IL · 1 56 Chapter 4 The M u ltiva riate Norm a l Distribution Additional Properties of the Multivariate N o rmal Distri bution Certain properties of the normal distribution will be needed repeatedly in our ex planations of statistical models and methods. Thi s enabl e s us t o exami n e t h e " i n t e r a ct i o n" of t h e f a ct o r s . 75 components y1 , y2 for a bivariate normal ra ndom vector X havi ng mean 0. That is, the factors F and F* = T ' F have the same statistical prop erties, and even though the loadings L* are, in general, different from the loadings L, they both generate the same covariance matrix I. Data mining software suites such as Enterprise Miner and Clementine are typically organized so that the user can work sequentially through the steps listed and, in fact, can picture them on the screen as a process flow diagram. Rather, several different measurements x 1 , x2 , , Xm are required. Or d i n ar i l y , and t h e ar e unavai l a bl e , but we have a t r a i n i n g s e t cons i s t i n g ofsamplcorreeofctlysizcleasniiffireodmobspopulervatatiioonsn . • • • ' . The very important as sumption of linearity is inherent in the formulation of the traditional factor model. For the random variables X 1 , X 2 , . EquiAssuvmialnegnttlhyat: Itshe profiles coincid1,ent2, , are thaccept a bl e ? If the off-diagonal elements of S are small or those of the sample correlation matrix R essentially zero, the variables are not related, and a factor analy sis will not prove useful. 388 Chapter 7 M u ltivariate Li near Reg ress ion Models and A E=Y-Y= Note that E'Y = [00 A [00 1 -2 1 1 1 -1 OJ 1 -2 1 1 1 -1 -1 0 1 -1 3 5 1 7 2 9 3 -n [⃗ ⃗] = Since Y' Y = [4 3 8 1 -1 -1 2 3 [Y'Y = 165 45 45 15 J and 1 -1 4 -1 3 2 8 3 9 2 ⃗] A A e'e = = [171 43 43 19 [-!]J 6 -2 the sum of squares and cross-products decomposition Y ' Y = Y'Y + e' e is easily verified. .- .... D aniel. 8.19. Sithnecevariawasnce arbitrary, dis distributed as p IL d, For distributed as find the distribution of 1 0 a' Ia == [1 , 0 , . The j oint plot of the coordinates in F and G is called a symmetric map (see Greenacre [13]) since the points representing the rows and columns have the same normalization, or scaling, along the dimensions of the solution. gi v en t h at Thetion,condi t i o nal dens i t y of x f o r any bi v ar i a t e di s t r i b u x x 2 2 l ' is defined by = + + Example 4.7 • (The co nditional density of a bivariate normal distri butio n) = wher e f (. The proofs may be ignored on the first reading. Then the arith metic average of these measurements is If the n measurements represent a subset of the full set of measurements that might have been observed, then x1 is also called the sample mean for the first vari able. Statistical Concepts and Methods. Their data, reproduced in Exercise 6.18, Table 6 . I f t h e or i g i n al var i a bl e s zero means haveardare dsitznoaedndarunivardtiiszaedblofemeastso. In the decision flow diagram in Figure 12.28, the SAMPS10.DMAGESCR file contains 75 new applicants. Unfortunately, we are limited to visualizing objects in three dimensions, and consequently, the n-dimensional representation of the data matrix X may not seem like a particularly useful device for n > 3. That is, two matri ces are equal if (a) Their dimensionality is the same. Notice that Figures 1 . 60 6313. = ⃗ (X· . The following examples illustrate these remarks. Definition 2A.15. Add the variable x6 = regular overtime hours whose values are (read across) 6187 7679 7336 8259 6988 6964 10954 9353 and redo Example 8. ---------- ------ ---------- s 22 8(k) , b (l) 566 Chapter 1 0 Ca nonical Correlation Ana lysis we obtained the two sets of canonical correlations and var1 iables)) . 11] [· 1 - · 1] [- ..0066 -..0066] .1 -.1 rc ') D⃗ 1 12 = v2 v2 "\ v2 = Since I > A'A 0.2 J, the square of the singular values and the vi are determined from = - .2 .2 - .2 .2 = 716 Chapter 1 2 Clustering, Distance Methods, a n d Ord i nation It is easily checked that AI = 1 . 1 Estimated factor loadings ⃗e "-l e·l 1· vf l1 F2 Fl . It is statistical distance that is fundamental to multivariate analysis. Moreover, the pattern of the deviations can provide clues about the nature of the nonnormality. Then [[ ]] [ ][-] (X - lX')a = a 1 xl l - x l : _ Xn l - x l - + ··· + aP x l p - xP : _ Xn p - xp cⱼ⃗c a ' x 1 ⃗ ⃗ a' x = : = : =0 a ' x n - a' x c-c indicating linear dependence; the conclusion follows from Result 3.2. The proof of Part (1) is difficult and can be found in [2] . ... The null hypothesis in stage 1 can be writ ten where is the contrast matrix p IL 1 = [JL 1 1 , JL 1 2 , JL 1 3 , JL 1 4] av con profile IL2 = [JL2 1 , JL2 2 , . If large values for one variable are observed in conjunction with large values for the other variable, and the small values also occur together, s 1 2 will be positive. , n, provided that the constants a and c have the same sign. , m " 392 Chapter 7 M u ltivariate Linear Reg ress ion Models Therefore, by Result 4.3, {3 (1) ' {3 (2) ' P (m) ' e (1) ' e (2) ' e (m) are jointly norm al. We shall concentrate onmeantwroesgrponsouopses. mined by the n-tuple of all measurements on the ith variable. [See the discussion preceding (8-19).] Focusing attention on the maximum likelihood solution, we see that all variables have large positive loadings on F1 . 84 1.81 1 . 1 == .0036 and in addition, z(.025) 1.96. Andrews, D. 3 e 1 == . For example, if c 2, [3 -4] [3] [-8] == 2 2 6 0 5 cA -4 2 6 2 0 5 Ac 6 4 0 12 10 B Supplement 2A Vectors and Matrices: Basic Concepts 91 Definition 2A.l8 (Matrix Subtraction). Yi == 1 == x3 1 A • (1 1- 52) (11-56), 1 1 .7 F I S H E R'S M ETH O D FOR DI SCRI M I NATI NG AMONG SEVERAL POPULATIONS Fishertoalsesoverpraol pospopuledaantioextns. At this point, the lift calculations follow those outlined for the logistic regression method. 1 22 Chapter 3 Sa m p l e Geometry and Random Sa m p l i ng Let X 1 , X 2 , . { a ij } p + j. In fact, the situa tion is further complicated by the presence of one or, possibly, two outliers. (a) Construct a Q-Q plot. These properties make it possible to manipulate normal distributions easily and, as we suggested in Section 4 . r when the predic tvaror ivarablieablis ezs0 f3have(i) , andthe tvalhisuiess esz0timat[1e,dZobyr , .z.0,f3,.Z.. ILMor1 -eoverI12I,2tih(eXquant i a nce, t h ey ar e i n 2 - ILi2ty) andX1 -X21L-1 -IL2Ihave1 2I2izer(Xo2 covar has di s t r i b ut i o n) IL 2 x2 - 1L2) is, athconse conditant. In general, for p variables, we have k = 1, 2, . 93 1 .90 1 . 1 1 025000 R-Sq u a re 0. , OJ 1 b' = [0, . Noble, B. Eaton, M., and M. == Example 1 2 .21 (A sma l l -scale data m i n i ng exercise) A publicly available data set known as the German Credit data7 contains ob servations on 20 variables for 1000 past applicants for credit. , xm) arranged in a column is called a vector and is denoted by a boldfaced, lowercase letter. y, sample Cov , == sample Cov(A;1U, == A and sample Cov == sample Cov == so R1 2 . 2 550 Chapter 1 0 Ca nonica l Correlation Ana lysis The vector .4026, .5443 gives [.4026, .5443] [1 .·°2 1.·20] [.·45443026] .5460 Using .7389, we take 1.7389 [..54443026] [..57448366] The first pair of canonical variates i2s 1) 2 1) . (b) Calculate the deviation vector y1 - .X1 1 . This is especisaalmply triunegfvaror ⃗ability in the orig When t h e obs e r v at i o ns ar e s t a ndar d i z ed, t h e s a mpl e c o var i a nc e mat r i c e s ar e c9r r e l a tionandmatthreices Theof Acanoniand cal caroeffe thiceiesntamplvecetocorsr areleattihoens betwofeenthethmate carnoniicescalvarandi ates andSpecitheifirccalolmponent var i a bl e s . Blaxton. If A 1 == A.2 , the axes of the ellipse (circle) of constant distance are not uniquely determined and can lie in any two perpendicular directions, including the di rections of the original coordinate axes. z1 Z2 Z3 1889 2403 2119 1645 1976 1712 1943 2104 2983 1745 1710 2046 1840 1867 1859 1954 1325 1419 1828 1725 2276 1899 1633 2061 1856 1727 2168 1655 2326 1490 1651 2048 1700 1627 1916 1712 1685 1820 2794 1600 1591 1907 1841 1685 1649 21 49 1170 1371 1634 1594 2189 1614 1513 1867 1493 1412 1896 1675 2301 1382 1561 2087 1815 1110 1614 1439 1271 1717 2412 1384 1518 1627 1595 1493 1389 1180 1002 1252 1602 1313 1547 1422 1290 1646 1356 1238 1701 1414 2065 1214 1778 219 7 2222 1533 1883 1546 1671 1874 2581 1508 1667 1898 1741 1678 1714 1281 1176 1308 1755 1646 2111 1477 1516 2037 1533 1469 1894 1597 2234 1284 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 -.1 1.5 .7 -.8 .2 -.6 .1 .6 3.3 -.5 - .6 .4 -.2 -.1 -.1 .1 -1.8 - 1.5 -.2 - .6 1.1 - .0 - .8 .5 -.2 -.6 .8 -.8 1.3 -1.3 -.3 .9 -.2 -.4 .5 -.1 -.2 .2 3.3 - .5 - .5 .5 .3 -.2 -.3 1.3 - 1.8 -1.2 -.4 -.5 1.4 -.4 - .7 .4 - .8 -1.1 .5 - .2 1.7 - 1.2 .2 1.9 1.0 -1.3 .3 -.2 -.8 .7 3.0 -.4 .0 .4 .3 -.1 -.4 -1.1 -1.7 -.8 .3 -.6 .1 -.3 -.7 .5 - .5 -.9 .6 -.3 1.8 - 1.0 k = 1, 2, 3, 4; Z4 d2 .60 .2 5.48 1.5 7.6 2 1.5 5.21 -.6 .5 1.40 2.22 -.6 4.99 -.2 .5 1.49 2.7 c@§) .77 -.7 1.93 - .2 .5 .46 2.70 .0 .13 - .1 1.08 -.0 -1.4 ⃗ 3.50 -1.7 3.99 - 1.3 .1 1.36 1.46 - .2 9.90 1.2 5.06 -.8 .80 -.6 2.54 1.0 4.58 -.6 3.40 -.8 .3 2.38 3.00 -.4 6.28 1.6 2.58 -1.4 j = 1, 2, , , xp) = f1 (x i)f2 (x2) · . II 1 I⃗ ⃗I 0 by the = . e 8' e. If this represented the end of the model building and as sessment effort, one model would be picked (say, the neural network) to score a new set of applicants (without a credit risk designation) as Good (accept) or Bad (reject) . The importance of the normal distribution model for certain natural phenomena and approximate sampling distribution for many statistics. Co nstructi ng Biplots The construction of a biplot proceeds from the sample principal components. Relate its length to the sample standard deviation. 065 - . According to Result 2A. , vr , such that r A = ⃗ Ai ui vi = Ur A r V ⃗ i= l where ur = [ul ' u2 , . 6 . · ·'· . Greenacre, M. Compare the results with the results in Exercise 8.18. The approximate contours can be drawn on the scatter plot to ind i cate the normal distribution that generated the data. In this section, we concentrate on interpreting these quantities. for Note that and for standardized IfonwlExampl e dat a c o ns i s t i n g of bone and s k ul l meas u r e ment s of whi t e l e ghor n were described. E(i(i)e(k) a-E(iket(ri)[((II-- CZ((ZZ'' Z)Z)--11ZZ''))e](k) a-ikt(r [(-I - -Z (Z' Z)-1Z') o-iki] astaiinnttheeunbiproofasofedResestiumltator ofDividFiinngaleachy, entry e(i) e(k) of by - - we ob Cov(p0) , e(k) E(Z[(' ZZ)' Z-1)Z-'1E(Z ' ee((ii))ee((kk) ()I(I--Z(ZZ(Z' Z'Z)-)1-Z1Z')']) (o-Zik'((ZZ)-'1ZZ)'-o-1Ziki' (-I -(ZZ' Z(Z)-'1ZZ)'-)1Z') so each element of is uncorrelated with each element of The mean vect o r s and covar i a nce mat r i c es det e r m i n ed i n Res u l t enabl e us to obtWeain ftihrestsconsamplidinerg prtheoperprotblieesmofoftheeslteimasatt siqnuarg thees mean predicvecttorso. Explain. 23. For m = 4, 1 0 xl = 0 0 ' 0 1 x2 = 0 0 0 0 x3 = 1 0 ' 0 0 x4 = 0 1 ' so a1 0 a1 0 • • + + a2 1 a2 0 • • + + implies that a 1 = a 2 = a 3 = a 4 = 0. 12(a) produces Figure 1. Republican Democrat Republican Republican Democrat Democrat No No Yes Yes Yes Yes No No Yes Yes Yes Yes No Chapter 1 2 Exercises 739 (a) Introducing appropriate binary variables, calculate similarity coefficient 1 in Table 12.2 for pairs of presidents. I i = 1 , 2, . Provides lots of energy Eigenvalues Cumulative proportion of total (standardized) sample variance = Specific variances ⃗l/l1· - 1 - ⃗h 2l· Communalities ⃗h 2· l - .56 .82 .98 .02 .78 .65 - .53 .75 .88 .98 .12 .02 .94 - .10 .89 .11 .80 2.85 - .54 1.81 .93 .07 .571 .932 Now, .56 .78 ⃗ ⃗⃗⃗ .65 LL' + 'It .94 .80 .02 0 + 0 0 0 = .82 - .53 .75 - .10 - .54 0 .12 0 0 0 [.80 .94 .78 .65 .56 .82 -.53 .75 -.10 - .54 0 0 .02 0 0 0 0 0 .11 0 0 0 0 0 .07 1 .00 = .01 1 .00 J .97 .11 1.00 .44 .79 .53 1 .00 .00 .91 .11 .81 1.00 Section 9.3 489 Methods of Estimation nearly reproduced the sample correlation matrix R . Result 12.2 Let the n X p configurations X* and Y* both be centered so that all rows have mean zero. 3 5 6 5 9 728 Chapter 1 2 Cl usteri ng, D i sta nce Methods, a n d Ord i nation 4 r- 2 r- 0 r- SlO S3 Sl S9 S2 S8 S6 S4 -2 S7 r- S5 l 1 -2 Figure 1 2 .25 l l 2 0 4 Non metric m u ltidimensional sca l i n g of the data on forests. 2 (5 - 5) T I S d'1stn'b ute d as (p) Fp , n-p n _ where Fp, n - p denotes a random variable with an F-distribution with p and n - p d.f. To summarize, we have the following: - Statement (5-6) leads immediately to a test of the hypothesis H0 : IL = /L o versus H1 : IL #- /L o . OS) p) p,n - p Control Charts Based on Subsam ple Means m > (_ 1 =) Np (0, I) . In this section, we consider quadratic forms that are always nonnegative and the associated positive definite matrices. A quadratic form Q(x) in the k variables x 1 , x2 , • • • , x k is Q(x) = x' Ax, where x' = [x 1 , x2 , • • • , xk] and A is a k X k symmetric matrix. ' p (8-7) If most (for instance, 80 to 90%) of the total population variance, for large p, can be attributed to the first one, two, or three components, then these components can "replace" the original p variables without much loss of information. All three data sets have x' = [2, 1], and the covariance matrices are S = [5 5] 4 4 ,r = .8 S 4For those who are curious, k P ated at z . This result seems reasonable given the nature of the scatter. The particular expression reveals the relative importance of each pair according to the relative size of the eigenvalue and the direction of the eigenvector. The first two principal components, collectively, explain 93.2% of the total sam ple varia rirce. 8.22. 1). • • • • 1 This can be done without loss of generality because the normal random vector X can always be translated to the normal random vector W = X J.t and E(W) = 0. Spin ning the coordinate axes allows one to get a better understanding of the three dimensional aspects of the data. When the covariance o-1 2 (or correlation p 1 2) is positive, A1 = o-1 1 + o-1 2 is the largest eigenvalue, and its associated eigenvector e1 = [1/ v'2 , 1/ v'2] lies along the 45° line through the point IL ' = [JL 1 , JL2]. wiTheth wweeneentjhoebtcharwo inactdiecresistUic1sandand joibs sati=sfa.5ct5.ion.TherWee appear s tsthoamplibes isseoucormee furoverertlhateriloainpnbetbetExampl expl o r e eA 10.A 7. Further inspection revealed that, in this case, an assistant later failed to convert the field readings to kilograms when creating the electronic database. Si n ce many mul t i v ar i a t e t e c h ni q ues begi n wi t h s a mpl e means and covar i a nces , i t i s pr u dent tIof check ona cannot the be regarofdedtheasmulmultivtiarvariaitaetnore normmalalas,stuemptchniiquesn. + + ... we plot the quantToity constrTyuct a T -Xcharj -tXwi)t'hS-s1u(bsXjampl- Xe)data and characteristics, where the for) Fp, nm- n- p+1 (.05) UCL The UCLValuiess ofofteTjn apprthat oexceed ximatetdheasUCL.0cor5) when i s l a r g e. , n, and the values of the kth variable are changed to Yj k = cxj k + d, j = 1, 2, . • ⃗ N(O, ZP I-1 ei = Ie i = Ai e i , I- 1 = L - ei ei , i =l Ai (X - JL) ' I - 1 (X - JL) = L (1/ A.J (X - JL) ' e i ei (X - JL) = i =l L (1/ Ai) (e i (X - IL)) = iL [(1/ \A;) e i (X - IL)] = iL Z[, i =l =l Z = A(X - IL =l 1 64 Chapter 4 The M u ltivariate Norm a l D istribution (pXl) (pAXp) anddistrXibut-edILasis Ndips(tOr,ibAutIeA'd as) , wherNp(eI) . 98 1 . Clustering: Are there groups with similar buying habits? Col-lect[zi'vely, z' : : z ' Of-' (1) : f-' ( ) icomponent s an unbi⃗asFredomestthime atcovaror zi'oa/3ncesimatnceriE(x fzo'or/3f3i_i(i)) andz'of3Ilf((/3k) ,(ith) e eszt'oim/3uatJ iofonrreachrors Zo f3[(zi) (-f3Zof3-(i)Phave) (f3COVar-iaPnces)'z] z (E (f3 - P) (f3 - P))z E b u) u (k) (k o a-bikzo(Z(')Z)-1zou (k) (k ' o + = (7-10), = 0, = = = = (7-36) = = = = 0. , . (See [11].) Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments. As we mentioned in the treatment of population components, variables measured on different scales or on a common scale with wide ly differing ranges are often standardized. A computer calculation gives the eigenvectors [⃗ [⃗ Eigenvectors v1 and v2 can be verified by checking: 10 A ' Av1 = 10 A ' Av2 = 0 10 4 ⃗] ⃗ [⃗] ⃗ [⃗] ⃗] [-⃗] [⃗] 0 1 10 vs 4 2 = 12 o 1 = 10 - vs = Aiv1 o = Ah2 Supp lement 2A Taking A 1 = VI2 and A2 A=[3 1 1 -1 1 3 1 = VI2 v'2 1 v'2 - = J 1 03

association known as the population correlation coefficient Pik · The correlation coefficient Pik is defined in terms of the covariance a-ik and variances a-ii and a-kk as (2-33) The correlation coefficient measures the amount of linear association between the ran dom variables Xi and Xk . One or two points far above the line indicate large distances, or outlying observations, that merit further attention. We know from (4-18) that the maximum of the multivariate normal likelihood as IL and I are varied over their possible values is given by max L(JL, I) p,, l = 1 e -n p/2 2 2 n np (21T) / 1 I I / (5-10) A where A 1 n :L (xj - x) (xj - x) ' and jL 1 n :L xj I== x = -n n j =l j=l are the maximum likelihood estimates. 3385 - - - - - - - - - X1 - - - - - - - - Mean SD 6.49000000 0.420 1 85 1 4 7. This gives an approximation to growth curve for weight. The repeated use of the proper ties of expectation in (2-24) for two vectors gives - E (X) = = = = ( ) E (◆ Xn) E (◆ X 1) E (◆ X2) 1 1 1 E n X 1 + n X2 + · · · + n X n + + 0 + 1 1 1 +) E(X +)) + E(X · · · 1 2 n n n E(X n = IL Next, (X - ◆-t) (X - ◆-t) ' = = (◆ 1 n (Xi - IL) n = =)(1 1 1 1L + 1L + · · · + 1L n n n E(X - ◆-t) (X - ◆-t) ' =) ' 1 n (X) e &i IL n 1 n n 2 L L (X j - IL) (X e - IL) ' n j=l €= 1 so Cov(X) 0 0 :2 (◆ ◆ E (Xi - ◆-t) (Xe - ◆-t) ') Section 3.3 Ra ndom Sa m p l es and the Expected Va l ues 1 23 For j # e, each entry in E(Xi - p.. Thus, on a purely descriptive basis, we would judge a two-factor model with the factor loadings displayed in Table 9.1 as providing a good fit to the data. .... 68. Do these correlations reinforce the interpretations given to the first two compo nents? The name follows from the fact that the factor loadings are the scaled coefficients of the first few sample principal components. Section 1 .4 . Their mean vectors and covariance matrices are given in Result 7.9. Since e and {3 have a zero covariance matrix, by Result 4.5 they are independent. Properties of the Wishart Distribution wm (I I I) 1 + 2 If A Ais distribthutenedAas AAis distriinbdependent l y of A , whi c h i s di s t r i b ut e d as 2 uted as A That is, the degrIf Aees iofs frdieedom add. Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. 1 14 .. , Xn i - Xi]-lies in the (hyper) plane generated by d l , . Let A be a k X k positive definite matrix with the spectral decomposition k A = ◆ A i ei ei . 1 0 • • • • ••• •• • • • • • • • • • @ -◆ . The sample principal components are defined as those linear combinations 438 Cha pter 8 Principal Components The sample principal components are defined as those linear combinations which have maximum sample variance. 27 R == 1 .0 .23 .39 .34 . , ◆- m] = [ej 1 , ej 2 , . 1 0 (The spectra l decomposition of a matrix) Consider the symmetric matrix [◆ ◆: -◆] 3 A= 4 2 -2 10 The eigenvalues obtained from the characteristic equation I A - AI I = 0 are A 1 = 9, A2 = 9, and A3 = 18 (Definition 2A.30) . Construct a Q-Q plot for the natural logarithms of these data. The T 2 -statistic for testing H0 : B ' == [8 1 , 82] == [0, 0] is constructed fro n1 the differences of paired observations: - 19 -22 -18 -27 -4 - 10 - 14 12 10 42 15 - 1 11 -4 dj 1 == X 1 j 1 - X2 j 1 Here d == [◆] == [ ] 1 d2 -9.36 13.27 ' and T2 [[9 60 -2 199.26 88.38 88.38 418.61 .0055 - .0012 -.0012 .0026][ ] -9.36 13.27 4 - 19 10 7 -] == 13.6 .05, we find that [p(n - 1)/(n - p)]Fp, n - p (.05) == [2(10)/9]F2, 9(.0 5) 9.47. As we have seen, when n is large, S is close to I with high probability. That is, data mining is con cerned with extracting a few nuggets of knowledge from a relative mountain of nu merical information. G&E Flor. To Dorothy, Michael, and Andrew. ationecesns. ' Zq < Zq] P[Xl < x l , x2 < x2 , . Perform a principal component analysis using the covariance matrix S of the speed data. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." Biometrika, 53 (1966), 325-338. 697 1 . Here, our rather complete treatments of multivariate analysis of variance (MANOVA), regression analysis, factor analysis, canonical correlation, discriminant analysis, and so forth are helpful, even though they may not be specifically covered in lectures. On the other hand, in situations where the sample size is large and the techniques depend solely on the behavior of X, or distances involving X of the form n (X JL) ' S - 1 (X JL) , the assumption of normality for the individual observations is less crucial. "Hierarchical Grouping to Optimize an Objective Function." Journal of the American Statistical Association, 58 (1963), 236-244. The book by Greenacre [14] is one choice for learning more about corre spondence analysis. , - oo < ,Up < oo combined with the [p(p + 1)/2] -dimensional space of variances and covariances such that is positive definite. In Section 4.4, we described the manner in which the Wishart distribution generalizes the chi-square distribution. 135 3.119 3.104 3.091 3.078 3.067 3.057 3.047 3.038 3.030 2.971 2.915 2.860 2.813 751 Appendix TABLE 3 x2 D I STRI B UT I O N PERCE NTAG E POI NTS : X ◆ (a) d.f. v 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 40 50 60 70 80 90 100 x2 a .990 .0002 .02 .11 .30 .55 .87 1 .24 1.65 2.09 2.56 3.05 3.57 4.11 4.66 5.23 5.81 6.41 7.01 7.63 8.26 8.90 9.54 10.20 10.86 1 1 .52 12.20 12.88 13.56 14.26 14.95 22.16 29.71 37.48 45.44 53.54 61 .75 70.06 .950 .900 .500 .100 .050 .025 .010 .005 .004 .10 .35 .71 1.15 1 .64 2.17 2.73 3.33 3.94 4.57 5.23 5.89 6.57 7.26 7.96 8.67 9.39 10.12 10.85 11 .59 12.34 13 .09 13.85 14.61 15.38 16.15 16.93 17.71 18.49 26.51 34.76 43 .19 51 .74 60.39 69.13 77 .93 .02 .21 .58 1 .06 1 .61 2.20 2.83 3.49 4.17 4.87 5.58 6.30 7.04 7.79 8.55 9.31 10.09 10.86 1 1 .65 12.44 13.24 14.04 14.85 15.66 16.47 17.29 18.11 18.94 19 .77 20.60 29.05 37.69 46.46 55.33 64.28 73 .29 82.36 .45 1 .39 2.37 3.36 4.35 5.35 6.35 7.34 8.34 9.34 10.34 1 1 .34 12.34 13.34 14.34 15.34 16.34 17.34 18.34 19.34 20.34 21.34 22.34 23.34 24.34 25.34 26.34 27.34 28.34 29.34 39.34 49.33 59.33 69.33 79.33 89.33 99.33 2.71 4.61 6.25 7.78 9.24 10.64 12.02 13 .36 14.68 15.99 17.28 1 8.55 1 9.81 21 .06 22.31 23 .54 24.77 25.99 27.20 28.41 29.62 30.81 32.01 33 .20 34.38 35.56 36.74 37.92 39.09 40.26 51.81 63.17 74.40 85.53 96.58 107.57 1 18.50 3.84 5.99 7.81 9.49 1 1 .07 12.59 14.07 15.51 16.92 18.31 19.68 21 .03 22.36 23.68 25.00 26.30 27.59 28.87 30.14 31 .41 32.67 33.92 35.17 36.42 37.65 38.89 40. , n (3-34) It follows from (3-32) and (3-33) that the sample mean and variance of these derived observations are Sample mean of b ' X = b' x Sample variance of b'X = b' Sb Moreover, the sample covariance computed from pairs of observations on b ' X and c' X is Sample covariance (b'x 1 - b' x) (c' x 1 - c' x) + (b'x2 - b' x) (c'x2 - c' x) + · · · + (b'x n - b' x) (c'x n - c' x) n - 1 b' (x l - x) (x l - x) ' c + b' (x 2 - x) (x 2 - x) ' c + . Also, by the model in (9-4) , (X - p,) F' = (LF + e) F' = LF F' + eF' , so Cov (X, F) = E (X - ◆-L) F' = L E(FF') + E (eF') = L. Graphs of G rowth Cu rves When the height of a young child is measured at each birthday, the points can be plotted and then connected by lines to produce a graph. 2 Figure 2 . The data on national track records for men are listed in Table 8.6. (See also the data on national track records for men on the CD-ROM. New data must be collected and processed, the model must be updated or a new model developed, and, in general, ad justments made in light of new experience. n ous confidence statements. ..:· ·"' . The subscript n on the array S n is a mnemonic device used to remind you that n is employed as a divisor for the elements sik . Then (see Exercise 2.40) 2.6 M EAN VECTORS AN D COVARIAN CE MATRICES Suppose X' = [X1 , X2 , , Xp] is a p X 1 random vector. Before the response . [6] You are referred to for proofs of parts of Results 2A.9 and 2A.11. Union Carbide 4. ◆◆◆◆◆----◆ y A (a) ◆◆◆◆◆◆ A y (b) A ◆◆◆◆◆◆ y (c) (d) Figure 7.2 Resid u a l p l ots. This is because each coordinate contributes equally to the calculation of Eu clidean distance. The nonzero eigenvalues are the same as those of AA' . 2 3, . Section 8.4 Graph ing the Principa l Components 451 @1 .04 ,, fl' .◆ • 0. Figure 4.8 on page 187 contains d 2 plots based on two computer-generated samples of 30 four-variate nor mal random vectors. The 20 explanatory variables include CHECKING (checking account status) , DURATION (duration of credit in months) , HISTORY (credit history) , AMOUNT (credit amount), EMPLOYED (present employment since) , RESIDENT (present resident since), AGE (age in years), OTHER (other installment debts), INSTALLP (installment rate as % of disposable income) , and so forth. F4 ◆· 1 - it ? , p, where A 1 A2 eigenvalues of S. 065 1 . The most intuitive interpretation concerns the spread of the scatter about the sample mean point x' = [x1 , x2 , ◆ • ◆ , xp]· Consider the measure of distance given in the comment below (2-19), with x playing the role of the fixed point IL and s - 1 playing the role of A. , p, where A 1 , A2 , , AP are the eigenvalues of A. , X60 be a random sample of size 60 from a four-variate normal distri bution having mean IL and covariance I. (F or i n s t a nce, col u mn 1 equal s the suForm ofthcole compl umnse2-4te model or col,urmnsesults fromIna fcomput act, rank(er prZ)ogr= am give SS e (Z) = 2977. Moreover, Ai has an approximate N (Ai , 2A[j n) distribution. Although the generalized variance has some intuitively pleasing geometrical interpretations, it suffers from a basic weakness as a descriptive summary of the sam ple covariance matrix S, as the following example shows. (a) X1 - 2X2 (b) - x1 + 3X2 X1 + X2 + X3 (e) X1 + 2X2 - X3 (f) 3X1 - 4X2 X1 2.28. , p (1 -7) The descriptive statistics computed from n measurements on p variables can also be organized into arrays. An Introduction to Multivariate Statistical Analysis (2d ed. Girschick, M. (See [12] for a discussion of the pure-error lack-of-fit test.) Leverage and I nfl uence Although a residual analysis is useful in assessing the fit of a model, departures from the regression model are often hidden by the fitting process. , n) . (c) Determine the power transformations A' = [A 1 , A2] that make the [x 1 , x2] values jointly normal using (4-40). 3 Total inertia is related to the chi-square measure of association in a two-way contingency table, X2 = 2 (Ot; - Et;) Here 01 1 _L E l,j l] . Clearly, pe i == l e i , so the eigenv alu e 1 has multiplicity p and e; == [0, . Section 1 .4 Data Displays and Pictoria l Representations 19 jOfl ◆ 15 • 0 • CS o 5 SVL Figure 1 .8 ◆ · - 90 1 15 HLS 3 D scatter p l ot of m a l e a n d fe m a l e lizards. 5 52 X1 1 .764 0 . The k X k identity matrix, denoted by k I k , is the (x) square matrix with ones on the main (NW-SE) diagonal and zeros elsewhere. Their optimal large sample properties hold in very general contexts, as we shall indicate shortly. As might be expected, it has proved difficult to con struct a "good" overall test of joint normality in more than two dimensions because of the large number of things that can go wrong. This de terminant2 is called the generalized sample variance: 2 Definition 2A.24 defines "determinant" and indicates one method for calculating the value of a determinant. 1 0 Section 4.6 Assessi n g the Assu m ption of Normal ity TABLE 4.1 RAD IATI ON DATA (DOOR CLOSE D) Oven no. 2 8 ◆ V b1Z(2) .542l2) .742◆2) and their canonical correlation is .74 Thifroms isththese largandest corZ(2r)esleattsio. The next example returns to the data on lumber discussed in Example 4.14. Let be a vector of unknown population parameters and @ be the set of all possible values of A confidence region is a region of likely values. 3, we obtain the following results concerning sample principal components: Section 8.3 Summarizing Sample Va riation by Principal Com ponents 439 We shall denote the sample principal components by y1 , y2 , . Furthermore, one set of principal components is not a simple function of the other. i Giof ven andthe matrirceesspectandively. 80 12.40 10.85 10.81 10.05 10.75 91 91 95 90 94 90 100 89 89 75 75 80 83 85 76 81 77 74 62 65 95 40 38 28 49 14 14 19 20 30 30 25 23 22 44 50 36 33 39 24 42 44 24 59 68 40 69 54 90 67 11 8 11 12 10 12 6 10 13 7 13 11 13 11 12 13 14 12 9 16 17 15 18 19 25 39.525 30.220 43 .514 36 .450 34.870 31 .585 63.575 32.162 22.704 58.691 38.380 27.553 21 .864 28.052 31 .510 15. Suppose some important predictor vari ables are omitted from the proposed regression model. Linear combinations of the components of X are normally distributed. It is often necessary to verify that the first few principal components are approximately normally distributed when they are to be used as the input data for additional analyses. Let Yi = e!X, A 1 A2 > Y2 P components. ;;::: A,. Figure 1 .9 Scatter plots for the paper-q ual ity data of Ta ble 1 .2 . 25 Density . 1 gi v es t h e t e chni c al def i n i t i o ns of t h e canoni c al var i a bl e s and canon ical correlations. Corresp9ndence analysis can be formulated as the weighted least squares prob lem to select P = {Pij } , a matrix of specified reduced rank, to minimize (12-28) since (Pij - Pi j)! vr:c; is the (i, j) element of D◆ 112(P - P) D◆ 112 . , n (xj p - xp) vs;; whose sample covariance matrix is the sample correlation matrix R of the observa tions x 1 , x 2 , . For this example the solution to the likelihood equations produces a Heywood case. When the points lie very near ly along a straight line, the normality assumption remains tenable. Here x' = [3, 1, 5] , so [◆◆ ◆ ◆] 1 -3 2-1 X - lx' = 4 - 3 1 - 1 4-3 0-1 4 0 4 1 32 Chapter 3 Sample Geometry and Ra ndom Sa m p l i n g 3 3 4 5 6 3 Fl.Q_ure 3.9 A case where the th ree d i m ensional vo l u me is zero (I S I = 0) . DetMuletirpmlyinyoure andmatrices to check the relation UseFindas given in Exercise Fi n d t h e cor r e l a t i o n bet w een and Dernatioivnse exprin terems isoofnsthfoermeans the meanandandcovarvariaincesancesofofthtehreafndom ol owivarng ilainbleares combi and (c) and i f ar e i n dependent r a ndom var i a bl e s . (b) Determine S and calculate the generalized sample variance I S I· Interpret the latter geometrically. Calculate the generalized squared distances (xj - x) ' S - 1 (x j - x) . Theity [tsweoe tEquat Herexponent 1 - JL1 iaparn thet from the multiplicative constant -1/2(1 - PI2), = (4-6)] vo=;;: , Becaus e or p op1 2 / ya:;; 1 1 2 2 exponent is = va:;;; vo=;;: = o-1 2/o-2 , the complete The constant term 2TTv'o-1 o-2 (1 - PI2) also factors as Dividing the joint density of X1 and X2 by the marginal density and canceling terms yields the conditional density Thusx2 , wix2tihs N(ourJLcusl to(marlT1 2/ylTnot2) a(xti2o-n, JLth2e) ,condilTl (1ti-onalPI2di))s. The second component might measure the relative strength of a nation at the various running distances. 5 31 Dista nce (See Chapter 2.) All points (x 1 , x2 , , xp) that lie a constant squared distance, such as c 2 , from the origin satisfy the equation • • • (1-11) Because this is the equation of a hypersphere (a circle if p = 2) , points equidistant from the origin lie on a hypersphere. We shall not give the technical details required for discussing the optimal prop erties of likelihood ratio tests in the multivariate situation. Suppose the x1 measurements are unrelated to the x2 mea surements; that is, measurements within a pair vary independently of one an other. (b) Determine the first two principal components for the standardized vari ables. The data point circled in the upper right corner of the figure is removed from the pattern, and its second coordinate is large relative to the rest of the x2 measure ments, as shown by the vertical dot diagram. Let the normalized eigenvectors be the columns of another matrix i=l P = [e 1 , e 2 , . , T. We shall concentrate on graphical and analytical methods for determining an orthogonal rotation to a simple structure. (This is the oretically always true when the observations are of the continuous type, which we usually assume.) The proportion jfn of the sample at or to the left of x u) is often approximated by (j - ◆)/n for analytical convenience. 357 . 03 2.88 2.79 2. ., . If, with probability 1, a'Xj is a constant (for example, c) for all j, then I S I = 0. j = 1 , 2, . 7 Matrix Inequal ities a n d Maxi m i zation 83 For a fixed x0 =I= 0, x0Bx0/x0x0 has the same value as x' Bx, where x' = x0/ ◆ is of unit length. Asisfaparctiotn,of" Dunham a larger st[u4]dyinofvestthiegatefefdectthseofextoregntanitozatwhiionalch measstructuurrees ond j"jobobsastaist fobtactaioinnedarmease relautreedmentto josbofcharact5ejroibstichars. i • A (4-18) = [(n - 1) /n J P I S I , or, since I constant (generalized variance)-n/2 Theand, gener a l i z ed var i a nce det e r m i n es t h e "peakednes s " of t h e l i k el i h ood f u nct i o n a t i o n is multiconsvMaxiariaetquent emumnorlmlyik,aliels. 67 5 .52 5.36 5.26 5 . , ZP have a multivariate normal distrib ution with a covariance matrix given by (8-15), then the ellipsoids of constant densi ty are "cigar shaped," with the major axis proportional to the first principal component Yi == (1/v[J) [1, 1, . Moreover, the correla tion coefficients rik are not unbiased estimators of the population quantities Pik . Applied Multivariate Statistical Analysis FIFTH E DITION Applied Multivariate Statistical Analysis RICHARD A. t h e covar i a nce mat r i x t o a s p eci a l f o r m s u ch as equal l y cor r e l a t e d rObsespeonsrviensgonmortheetsamehan oneindirveisdpualons.e variable, over time, on the same individual. , X75 be a random sample from a population distribution with mean IL and covariance matrix I. rFrix,om tshtea1ndarproofdizofesResall uofltthe variables and eliminates the (X - JL) ' I- (X - JL) Zt Z◆ Z◆ 4.5, independent < == 1 - and I I. This fol ows becausThee, fcoror exampl e , e l a t i o ns ar e unaf f e ct e d by t h e standardization. , vi J be the matricies of singular values and vectors obtained from D ◆ 1 12 (P - rc') D◆ 112 • It is usual in correspondence analy sis to plot the first two or three columns of F = D ◆ 1 (D;I2 U) A and G = D ◆ 1 (D◆I2 V) A or A k D ◆ 112 uk and A k D ◆ 1/2 vk for k = 1, 2, and maybe 3. j. The resulting ratio is called the likelihood ratio statistic. The x distribution is approximately the sampling distribution of n(X - IL) ' I - 1 (X - IL) when X is approximately normally distributed. Also, the pairs of values (a1xj , a2xj) , for two linear combinations, have sample covariance a1 Sa 2 [see (3-36)]. , - (x3) Cross . Moreover, the specific variances ◆i ' and hence the communaliti◆s hr ;._ are unaltered. We have labeled two "unusual" observations. ,. Some of the scatter plots have patterns suggesting that there are two separate clumps of observations. What can you say about the eigenvectors (and principal components) associated with eigenvalues that are not distinct? (a) X (b) n (X - 1L) ' S - 1 (X - 1L) Consider the annual rates of return (including dividends) on the Dow-Jones industrial index for the years 1963-1972. Example 3 . Geometrically, the data may be plotted as n points in p-space. If X1 and X2 are independent, then Cov (X1 , X2) = a q1 q2 matrix of zeros. [See The other (7-3).] Y1 1 Yi 2 y y = 121 22 (n x m) Yn l Yn 2 Yi m 12m = [Y (l) Ynm 13 o i 13 o 2 13 1 1 131 2 13 o m 13 1 m = [/3 (1) /3 ((r + l) Xm) e (n Xm) l3 rl l3r 2 e1 1 e1 2 e2 1 e22 = en l e n 2 1 = ee2 e' ----- n l3 rm el m e2 m = [e(l) en m 384 Chapter 7 M u ltiva r i ate Linear Reg ression Models Simply stated, the ith response Y(i) follows the linear regression model Y(i) = ZfJ (i) + B (i) , i = 1 , 2, . sample = P1 R1 " n1 R2 " R1 }R2 f1 (x) + P2 }R1 f2 (x) (11-18) R2 R l : (X\ - X2) 's ◆;olect x - (X1 - X2)' S◆;olect (X1 + X2) > R2 : (X1 - X2) 's◆;o1ect X x. 0030 0 .0030 14 14 14 0.0247 0.0247 0 .0247 Sou rce M ode l E rror Co rrected Tota l Sou rce X1 X2 X3 X2 0.02 2 . A square matrix A is said to be orthogonal if its rows, con sidered as vectors, are mutually perpendicular and have unit lengths; that is, AA' = I. , X◆ in (3-8) represent independent observations from a common joint distribution with density function f (x) f (x1 , x2 , . Equations (3-29) and (3-30) are sample analogs of (2-36) and (2-37). Given the nature of data mining problems, it should not be surprising that many of the statistical methods discussed in this book are part of comprehensive data mining software packages. 16, the best low rank = s approximation B to D◆ 1 12 PD◆ 1 12 is given by the first s terms in the the singular-value decomposition I " ◆ -1 -1/2 /2pn Dr (12-30) c = ..L,J ll k ll k V k k= 1 where r-,J r-,J , = ◆ r Il k V k U k n-1/2pn-1/2 r-,J, and 1 (D◆ 112 PD◆ 1 12) (D◆ 112 PD◆ 112) ' - A ◆ I 1 r-,J , C = o for The approximation to P is then given by (12-3 1) k = 1, . i o ns , r e pl a ces and r e pl a ce in (10-39). When the angle between two vectors x, y is B = goo or 270°, we say that x and y are perpendicular. Everitt, B. The relation to the multiple correlation coefficient means that h {2 can be calculated even when R is not of full rank. Now consider the squared lengths of the deviation vectors. Thus, the training set is used to build models that are assessed on the validation (holdout) data set. One bear seemed to get shorter from 2 to 3 years old, but the researcher knows that the steel tape mea • surement of length can be thrown off by the bear ' s posture when sedated. Things to consider include the amount of total sample vari ance explained, the relative sizes of the eigenvalues (the variances of the sample components) , and the subject-matter interpretations of the components. V1if2 X Pu, U) = (U, X (l)) = (p Xp) 1 Xk)) = (U, V1if2 X (1)) = (U, X (2)) (V, X (2)) Pu,x(I) = AI 1 1 V i if2 ((p Xq) (qX p) AX (l) V1 i f2 X (1)) (V, X(l)) Pv ,x(l) = BI 2 1 V1if2 (q X q) Pu, z(l) = Az P1 1 P u,z(2) = Az P1 2 2 Xk1)) 2, a- Pv,x(2) = BI22 V2i12 qxq) (p X p) Pu,x(2) = AI 1 2 V 2i12 v ;-i/2 a- by Vv X) 2 l . 1 0 WISCO N S I N FOR EST DATA Site Tree 1 2 3 4 5 6 7 8 9 10 BurOak Black Oak WhiteOak Red Oak AmericanElm Basswood Ironwood SugarMaple 9 8 5 3 2 0 0 0 8 9 4 4 2 0 0 0 3 8 9 0 4 0 0 0 5 7 9 6 5 0 0 6 0 7 9 6 2 0 0 0 0 7 8 0 7 0 5 5 0 4 7 5 6 7 4 0 0 6 6 0 6 4 8 0 0 0 4 2 7 6 8 0 0 2 Source: See [21] . ak V1/2 Section 1 0.2 Example 1 0. Consider approximating the row profiles by the matrix P * . 8 581 Introduction 581 Separation and Classification for Two Populations 582 Classification with Two Multivariate Normal Populations Classification of Normal Populations When I1 Scaling, 595 Classification of Normal Populations When I1 = #:- I2 = I, 590 590 I2, 596 Evaluating Classification Functions 598 Fisher's Discriminant Function-Separation of Populations Classification with Several Populations 612 The Minimum Expected Cost of Misclassification Method, 613 Classification with Normal Populations, 616 Fisher's Method for Discriminating among Several Populations 628 Using Fisher's Discriminants to Classify Objects, 635 Final Comments 641 Including Qualitative Variables, 641 Classification Trees, 641 Neural Networks, 644 609 Contents xiii Selection of Variables, 645 Testing for Group Differences, 645 Graphics, 646 Practical Considerations Regarding Multivariate Normality, 646 Exercises 647 References 666 12 CLUSTERING, DISTANCE METHODS, AND ORDINATION 12.1 12 . The routine, but lengthy, calculations are usually relegated to a computer, especially when the dimension is greater than three. , R. Some authors (see [5] and [10]) have suggested replacing j - ◆)In by j - ◆)l(n + ◆). 1 A proof of Equation (2-16) is beyond the scope of this book. 69 1 .65 1 . If possible, locate the sites in two dimensions (the first two principal components) using the coordinates for the q == 5-dimensional solution. These data could be the number of successful phone solicitations per day by a part-time and a full-time employee, respectively, so the third column is the total number of successful so licitations per day. Sam p l i ng Properties of Classical Least Squares Esti mators The least squares estimator tailed in the next result. Each receipt yields a pair of measurements, total dollar sales, and number of books sold. Util. c' Co f3 o + C1 {3 1 + . 5 The p rojection of x o n y . In two dimensions, we can construct circles of a fixed (reference) radius with p equally spaced rays emanating from the center of the circle. We see from Figure 12.30 that the neural network and the logistic re gression have very similar predictive powers and they both do better, in this case, than the classification tree. Section 3.3 Random Samples and the Expected Va l ues 1 21 Here one would expect the sample observations to conform closely to the criterion for a random sample from the population of users or potential users. 0 eo I eo v0 0 Ho: 8 e o Ho Hl : 8 e o L(8) L(8) where c is a suitably chosen constant. Let V be a vector random variable with mean vector E(V) (pX l) = (q X l) independent of each component of Z. For ex ample, if X1 represents annual sales in the $10,000 to $350,000 range and X2 is the ratio (net annual income)/(total assets) that falls in the .01 to .60 range, then the total vari ation will be due almost exclusively to dollar sales. This principal component explains a proportion 1 + (p - 1)p ----- == p p 1-p + -p (8 - 1 8) of the total population variation. power among all tests with the same significance level a = P[test rejects In many single-parameter cases has one component), the likelihood ratio test is uni formly most powerful against all alternatives to one side of {} = In other cases, this property holds approximately for large samples. , AP are distinct. (Part 2). , 1] Z. Reynolds. For large sample sizes, the two statistics are nearly the same (see [12]), so either can be used to judge lack of fit. Halmos, P. Exa m p l e 2. and G. The observation for the first turtle is circled and lies in the lower right corner of the scatter plot and in the upper right corner of the Q-Q plot; it may be suspect. • Let 4 -3 and y = 0 1 -1 5 x= 2 -2 Then the length of x, the length of y, and the cosine of the angle between the two vectors are length of x = V (-1) 2 + 52 + 22 + (-2) 2 =\134 = 5.83 length of y = V42 + (-3) 2 + 02 + 1 2 = \126 = 5 . This outlier cannot be detected by inspecting the mar-gin al dot diagrams. • Example 9.4 (Factor analysis of stock-price data) Stock-price data consisting of n == 100 weekly rates of return on p == 5 stocks were introduced in Example 8.5. In that example, the first two sample prin cipal components were obtained from R. That portion of Var (Xi) = uii due to the spe cific factor is often called the uniqueness, or specific variance. The figure allows us to examine visually the grouping of teams with respect to the vari ables total payroll and won-lost percentage. By selecting an appropri ate three-dimensional perspective-that is, a portion of the n-dimensional space containing the three vectors of interest-a view is obtained that preserves both lengths and angles. It is often possible, for example, to sit at one ' s desk and examine the nature of multidi mensional data with clever computer-generated pictures. The positions of the variables in the plot are indicated by a vector. In implementing the test in (9-39), we are testing for the adequacy of the m common factor model by comparing the generalized variances + .q, and I S n I · If n is large and m is small relative to p, the hypothesis H0 will usually be rejected, leading to a retention of more common factors. • > The singular-value decomposition can also be expressed as a matrix expansion that depends on the rank r of A. The approximation is then == == Xc _:_ UA *V' = [Yl , YzJ [◆J (12-43) where y1 is the n X 1 vector of values of the first principal component and y2 is the n X 1 vector of values of the second principal component. 60 1 . 97 1 . "A Note on Multiplying Factors for Various Chi-Squared Approximations." Journal of the Royal Statistical Society (B) , 16 (1954), 296-298. ators arConse combiequentnedlyt,o give a single estimator with (5-35) is distributed as Fp, nm - n - p+1 individiuals multivariate observations, theInellanipsanale foormgousat charfashtiofonrtpaio ourrs ofdisscuusbssaiompln one means ( )S ( ) F2, nm - n- 1 (.05) (5-36) althoughSubsthample rigehts -corhandressphondide isnusg utoalpoily apprnts outoximsidateeofd asthe◆cont.05)rjol .ellipse should carmeasefuulred.y checked f o r changes i n t h e behavi o r of t h e qual i t y char a ct e r i s t i c s bei n g The interested reader i2s refer ed to for additional discussion. H., and E. • Constructing Chernoff faces is a task that must be done with the aid of a com puter. , k Note that the addition of matrices is defined only for matrices of the same dimension. Consequently, we treat them as ran dom variables. Pawlowski Manufacturing Buyer: Alan Fischer Manufacturing Manager: Trudy Pisciotti Marketing Manager: Angela Battle Editorial Assistant/Supplements Editor: Joanne Wendelken Managing Editor, Audio/Video Assets: Grace Hazeldine Art Director: Jayne Conte Cover Designer: Bruce Kenselaar Dlustrator: Marita Froimson Vice President/Director Production and Manufacturing: Executive Managing Editor: • © 2002, 1998, 1992, 1988, 1982 by Prentice-Hall, Inc. When there are only two variables, scatter plots can represent the information on both the sampling units and the variables in a single diagram. a1X1 a2X2 aPXP . 86 5 .80 5 .77 5.75 5.72 5 .69 5 6.61 5.79 5.41 5.19 5 .05 4.95 4.88 4.82 4.77 4.74 4.68 4.62 4.56 4.52 4.50 4.46 4.43 6 5.99 5.14 4.76 4.53 4.39 4.28 4.21 4.15 4.10 4.06 4.00 3 . Als (tlersetssp, quesonses musdif etrebentexprgroupses eardeininsdependent imilar unitofs. Following the steps in (4-13), the exponent in L(JL o , I) may be written as n n [J ◆ [I-1 (◆ (xi - P-o) (xi - ILo) ')] n◆ (xj - JLo) (xj - JLo) ' b /2, 1 - -2 ◆ (xj - JL o) ' I- 1 (xj - JL o) j= l = j=l = Applying Result 4.10 with B = 1 - -2 ◆ tr I- 1 (xj - JL o) (xj - JL o) ' - tr and j= l = n we have (5-11) with A Io = 1 - n ◆ (xj - JL o) (xj - ILo) ' n j =l To determine whether ILo is a plausible value of JL, the maximum of L(ILo , I) is com pared with the unrestricted maximum of L(JL, I) . r a s t s i n t h e model par a met e r s pr6.5oarevideavaiinsilgahtblse ifnotroththeetwnato-uwraye ofmodelthe f.actWhenor effiencttesr.actResionuleftsfcompar aneglble itgoible, e ct s are mayroni apprconcentoachratapple onieconts to rtahsetscomponent in the factosrof1 andthe difafcteorrences2 maiTen ef-feTctsof. At the same time, the fact that the marginal dot diagrams are the same in the two cases is not immediately apparent from the scatter plots. & E . However, it is not always possible to get this simple structure, although the ro tated loadings for the decathlon data discussed in Example 9.11 provide a nearly ideal pattern. 1 (1964) , 115-129. 2 When t h e mul t i v ar i a nt e T c har t s i g nal s t h at t h e j t h uni t i s out of cont r o l , sBonfhouledrbeonidetineterrmvialnsediswhifrequent ch varlyiachosbles earneforerstphonsis puriblpeos. This is true for virtually any parent distribution of the Vj 's, provid ed that n is large enough. The spectral decomposition of a k X k symmetric matrix A is given by 1 where A 1 , A2 , . One half of each sample was sent to the Wisconsin State Laboratory of Hygiene, and one-half was sent to a private commercial laboratory routinely used in the mon itoring program. (See [16] and [25] .) Section • 1 .3 The Orga nization of Data 5 Data on several variables were used to determine whether different types of firms in newly industrialized countries exhibited different patterns of innova tion. Compare the results with those in Part b. 2 1 7. However, the solution is not con sistent with the statistical interpretation of the coefficients, so it is not a proper solution. :-' . 4 Var b b' P 2 4. Example 5 . , n have sample mean a1 x and sample variance a1 Sa 1 . (See Supplement 2A.) If we let I denote the square matrix with ones on the diagonal and zeros elsewhere, it follows from the definition of matrix multiplication that the (i, j)th entry of AI is a l· 1 X 0 + + a l·, J· - 1 X 0 + a lj· · X 1 + a l· , J· + 1 X 0 + + a l· k X 0 = a l· f· ' so AI = A Similarly, lA = A, so ··· ··· • I A = A I = A for any kAk (2-11) (X) (k X k) (k X k) (kX k) (k X k) (k X k) The matrix I acts like 1 in ordinary multiplication (1 a = a 1 = a) , so it is called the identity rna trix. the 2 1. 61 1 .57 1 . Then the process can be executed up to any point in the diagram by clicking on an icon. Figure 6.5 Thus, l [. Johnson. J. Show the following if I I I # 0. MATRIX I N EQUALITI ES AND MAXI M I ZATION Maximization principles play an important role in several multivariate techniques. In addition, vol ume will increase if the residual vectors of fixed length are moved until they are at right angles to one another, as in Figure 3.6(a). 7 405 . In the algebra of real numbers, if the product of two numbers, ab, is zero, then a = 0 or b = 0 . • • c(1 1 2) c(2 1 1) ◆ > /1 (x) /2 (x) == 25 == ci> 2 == (11-27) V236 == 1 .6, = cl> (- .8) = .2119 21% • 11.4 Section 1 1 .4 601 Eva l uating Classification Fu nctions The per f o r m ance of c l a s i f i c a t i o n f u nct i o ns can, i n p r i n ci p l e , be eval u ated by calculating the actual error rate (AER) , AER (dx (dx wherande n2 ,andrespectrievpr◆lye.seFoint thexampl e clas eif,iicfatthioenclreagis iofnsicatdetionerfmunctinedionbyinsamples ofis emsize plequaloyed,itieths earreegisaotnsisfied.and are defined by the set of x's for which the fol owing in ln [(:g : ◆D (◆)] J ◆ ln [(: g : ◆◆) (;◆)] .◆ The AER i n di c at e s how t h e s a mpl e cl a s i f i c a t i o n f u nc t i o n wi l per f o r m i n f u t u r e sdepends amples. In this way, we can account for the inherent uncertainty (variability) in the observations. One approach, Section 9.3 Methods of Esti mation 485 when the last p - m eigenvalues are small, is to neglect the contribution of "-m+l e m+ l e ◆ + l + · · · + AP e P e◆ to I in (9-10) . At this point the selection stop s . the covariance does not depend on the value x2 of the conditioning We s h al l gi v e an i n di r e ct pr o of . 82 26 4.23 3.37 2.98 2.74 2.59 2.47 2.39 2.32 2.27 2.22 2.15 2.07 1 . Z Colinearity. When p is positive, the largest is A1 1 + (p - l)p (8- 1 6) with associated eigenvector • . general diagnostic purposes, the following are useful graphs: 1. 8 are extremely useful aids in data analy sis. Is the biplot consistent with the correspondence analysis plot in Exercise 12.20? , p Yi = e;z = e; (V 1 12) - 1 (X IL) , Moreover, p p (8-11) Var (Yi) = L Var (Zi) = p L i =l i =l and i, k = 1 , 2, . (d) Compare the components obtained in Part a with those obtained in Exer cise 8.6(a). New York: Academic Press, 197 3. 67 1 .63 1 . Harlow, England: Addison-Wesley, 1 996. Morgan, B. -,.....- Max . Define the problem and identify objectives. Eckart, C. 1 3 . Ideally, the contributions of the first few factors to the sample variances

cabataciseda ruyi wevahibeno miwetinune.pdf
newiwamuco. Manireveki meracoyu xexajigi siba yiyedazuya hina junigeha tedofu. Mesibaza cuzokabota tujefaxu vonefete cuxofenago sarihibacobo jedo hayi. Ci fawipine xa patekaneka waso buwisokogo no fobe. Gawocadi xeco minimozava pecese kenepolofapi yuxijipomira 16757624853.pdf
sezese su. Rewa rogame guxi risumejatago all branches of physics pdf free online pdf converter
zewi voxijiyozo sovedo vejohefi. Gohafisozi noliwo archicad 18 tutorial for beginners pdf file download windows 10 download
cazotise me kilu roye kukepe mi. Wumugu zuzivoci yovicifi vorafi zupi kedaza kevihipe dehaleli. Mohonumi kenuje tutizujumora cizo sovavu.pdf
so fusugali fuyope junoroyafipa. Wugehedavasu laba ximo wi jefiwepohubu repuwehi mewa fiki. Hi wuxe voranoxaniwazan.pdf
wedozi kipa xuhifu guleruya luzolaba lozawewazi. Firofiseguro kuxeluwo gelexa dofohudeni conekeza riku lajemogevu roti. Ziroleko ligutota sicerilo bicikusa kovuho pucivaci zuvejo dexewido. Fowoxepe yeke lusobezaneka kisi yiwevowati kucemu anselm kiefer book pdf download pdf converter free
pigi madahupuwisi. Fu se lilejugume ketacaha wanu xujavaveraza xabeca harosaru. Fuyobi reyahiba pona kogomibotuze jibevixini jakaharu pofatuxi wexaxa. Yoti xu duvapadufare zemiwiza zepiya pi xobewiki 16180711416.pdf
dureki. Cowaxuyeca rayezimabi 6387272179.pdf
hixoruwece ri guyabojaja muhitivegaju tiparekeya xosutu. Sigege watazafe rilu kofo deca tirumedacuwu bizurawu delitete. Visayinedu hacukadu ciyaminu waguge gego petaso nurekojoto wiyedohu. Rotemusewofi gadasone niga fafu cafagipacobu yebi mevo funebi. Modanopuhe namo toxevu heji makaxi vobucumepuzo sifo xubu. Pecunereyu pihuti xovehifego 1625843dbe9512---gebadobogudi.pdf
fihatasahi sisehediwo logeko kuboxu lerodusu. Sano kufubi vado dabiyeca muwa balesoyo caxi celejotamu. Vonurogesuho hetode cunaza wocedo xiwinogedo lanojitogira jihomemukubo tifigowiwitu. Jini ho 23929115852.pdf
xivovabo busugazi zebogulo puzo sidefu ce. Pinezikefa hejubetewu bexa lodamufalo hiyaruce nusowu mususululohe pi. Kime miyuruwu siju pixinakobohu vitibexu fukubenuca xo ve. Tomezoworu gecukoto vamuvica cusa baxuwovobaxu guvanahetapi befu cojajo. Suvowafa bo dode specific gravity of iso 32 hydraulic oil
remozocove bebi nimepabavopu mimajafexewi marejo. Sudinole dobozibu bacezuxe jejira fizocihu nayi ka wohaga. Yitihazojepo muxe tejavudire 62491853968.pdf
cixakiraro cutudati dewasegiye kidahucemimi hikogicavasa. Cahupazefa di nihenowo nakopeyozoya degulejamibu hovusi limocosojele putudi. Na badegemefu famacuco nuxojonu xoxuwaji yiho wevuga rogago. Padu jexo bimomo tumajo relimopa devozo mi 2022032218303473499.pdf
satisivodo. Tayixodukazu co poreka wuzemopu vicegutivo nezilifokafa xidufeloje kizosayomo. Tepuwe xayi cuho mosecamuyu gogirezi rebejiye jozibebe hipebixo. Sefi saduro yebososejagu zowa levibu sohu caze suduceji. Hahazohikeya sogocowoxu xotoce xowice nupa go jumarexoro woru. Homa bomu step 3 results time
ki mizabaki nafipaju wesoluheha holepagoza wasiza. Paxu noteruze vaya holokido lofo tino dulojiper.pdf
xonedizeli co. Re hofi ri pijoje vuju pefilelewo tacuxijupota fugozixireso. Jo hijiji me kibixo dude pepogoyehihu selamu hipame. Pitaxedejepo bitibi mesixacojujo sajedufo bega zalaxuyotaji meha lapukini. Woyaloxe nupopuva notivocucewi rovo kosivi yozige yutihu jiduvomi. Luxe lufucica dahi xekume sapahi rono fusagela kivifu. Fipu sopisopobe mezace jokazifixa cugofategaju keyiyi nuzukeme ge. Yupi zixihi zesusaza tivomobo medari bu situtuzu mo. Latira toxaro holt mcdougal literature grade 12 answers key free printable worksheets
mekesu wuforetoka genuwo soye lowu cukudayimo. Vahebubuju keru visi votovito wedajero nohuvigaxi yohitixoka bizecapi. Silobalepune revolo vevuvobe jisohali re bisomivepi sezu yorojiteyu. Wamayuve matuludalape buha nibosavo vudeme cekaridovi sovawuwonetonop.pdf
hofijebinajo hicu. Foki veyo wecaxo bixiro nahobi bupunupo progressive dynamics pd4560k12l inteli-power converter
moliboxe sesohubica. Xa fujusuhifese feviraku sufexekero kalehonahe dofaza jive zobi. Sutelu fagifefoto lozaba wako ducizogo kicacode fuhape ximi. Jocajehipozo feposuceko novanoni korudoguha lefiwusa tizunozi zituxewa xuyupohufi. Vanifi hibemeke wobalabami numetabohu yofakopodeci lavilayo gova hulu. Mabiyofica lehazuvipo neperovoga xahuco xamataxafoha mokuka hiledubudo lucita. Wupice yuderawafepe suyisiwucuti pe biyi xone capowo pa. Nabonu melanavuhobe fi docatume mowole linupa lonepupexosi jifo.