lah PCA The topic that are covered in this lab worksheet is: Principal Component Analysis (PCA) Dimension Reduction 1. Principal Component Analysis (PCA) Recall that one primary purpose for performing PCA is dimension reduction. Today we will work with the USArrests dataset. nrow(USArrests) ## [1] 50 ncol(USArrests) ## [1] 4 dim(USArrests) ## [1] 50 4 names(USArrests) "Assault" "UrbanPop" "Rape" ## [1] "Murder" There are 50 observations (which correspond to the 50 states in US), and each observation is of dimension 4. The built-in function prcomp() is used for performing PCA. Recall that pre-processing of the dataset is needed. First, we do centering but no normalization. center=TRUE ==> means do centering but no normalization. USArr.noscale <- prcomp(USArrests, center=TRUE)</pre> USArr.noscale ## Standard deviations (1, ..., p=4): ## [1] 83.732400 14.212402 6.489426 2.482790 ## Rotation $(n \times k) = (4 \times 4)$: ## PC1 PC2 PC3 ## Murder 0.04170432 -0.04482166 0.07989066 -0.99492173 ## Assault 0.99522128 -0.05876003 -0.06756974 0.03893830 ## UrbanPop 0.04633575 0.97685748 -0.20054629 -0.05816914 ## Rape 0.07515550 0.20071807 0.97408059 0.07232502 The first principal component (PC) is dominated by Assault with weight more than 0.995, while the other three are less than 0.08. But this happens not because Assault is not correlated with the other two crimes and UrbanPop (which means the percentage of urban population), but because the values of Assault are signficantly larger than the values of the other three attributes, so the result of PCA is overdominated by Assault. Use the code snippet below to see the ranges of number of different attributes. range(USArrests[, "Murder"]) ## [1] 0.8 17.4 range(USArrests[, "Assault"]) ## [1] 45 337 range(USArrests[, "UrbanPop"]) ## [1] 32 91 range(USArrests[, "Rape"]) ## [1] 7.3 46.0 In this case, normalization is necessary, as is done below. scale=TRUE ==> means do both centering and normalization. USArr.scale <- prcomp(USArrests, scale=TRUE)</pre> USArr.scale ## Standard deviations (1, ..., p=4): ## [1] 1.5748783 0.9948694 0.5971291 0.4164494 ## ## Rotation $(n \times k) = (4 \times 4)$: PC1 PC2 PC3 ## Murder -0.5358995 0.4181809 -0.3412327 0.64922780 ## Assault -0.5831836 0.1879856 -0.2681484 -0.74340748 ## UrbanPop -0.2781909 -0.8728062 -0.3780158 0.13387773 ## Rape -0.5434321 -0.1673186 0.8177779 0.08902432 After performing PCA with normalization, we look at the first PC. Assault is positively correlated with all the other two crimes, which is expected. Assault is also positively correlated with the percentage of urban population. It is worth noting that the standard deviation of the first PC are not much larger than that of the second PC, so we also need to consider the second PC. This is quite common with real-world datasets, as the relationship between different attributes are often not clear-cut. 2. Dimension Reduction Before going into dimension reduction, we briefly discuss how R handles the expressions like (a matrix + a vector) and (a matrix * a vector). Understanding this is crucial for the rest of this section. This is best illuminated by examples. $A \leftarrow matrix(c(1,2,3,4,5,6), ncol=3, byrow = TRUE)$ Α ## [,1] [,2] [,3] ## [1,] 1 2 3 ## [2,] 4 5 A + C(7, 11)[,1] [,2] [,3] ## [1,] 8 9 10 ## [2,] 15 16 17 A + c(7, 11, 3)[,1] [,2] [,3] ## [1,] 8 5 14 ## [2,] 15 12 9 A + c(7, 11, 3, 13)## Warning in A + c(7, 11, 3, 13): longer object length is not a multiple of ## shorter object length [,1] [,2] [,3] ## [1,] 8 5 10 ## [2,] 15 18 17 To understand why the outputs are as above, you may first hypothetically think the matrix A is treated as a vector c(1,4,2,5,3,6) in R (go down the first column, then go down the second column, and so on). In A+v for any vector v, what R has done is to add to each entry of the hypothetical vector of A by the entry of v in a cyclic manner. Analogous outputs are obtained when you replace addition by multiplication (or subtraction or division). Experiment yourself. Now we come back to PCA. Recall that one primary purpose for performing PCA is dimension reduction. prcomp has already done this for you. You can retrieve the result as below. USArr.scale\$center Murder Assault UrbanPop Rape ## 7.788 170.760 65.540 21.232 USArr.scale\$scale Murder Assault UrbanPop Rape 4.355510 83.337661 14.474763 9.366385 USArr.scale\$x PC1 PC2 ## Alabama -0.97566045 1.12200121 -0.43980366 0.154696581 ## Alaska -1.93053788 1.06242692 2.01950027 -0.434175454 ## Arizona -1.74544285 -0.73845954 0.05423025 -0.826264240 ## Arkansas 0.13999894 1.10854226 0.11342217 -0.180973554 ## California -2.49861285 -1.52742672 0.59254100 -0.338559240 ## Colorado -1.49934074 -0.97762966 1.08400162 0.001450164 ## Connecticut 1.34499236 -1.07798362 -0.63679250 -0.117278736 ## Delaware -0.04722981 -0.32208890 -0.71141032 -0.873113315 ## Illinois -1.36505197 -0.67498834 -0.67068647 -0.120794916 ## Indiana 0.50038122 -0.15003926 0.22570277 0.... ## Iowa 2.23099579 -0.10300828 0.16291036 0.017379470 ## Kansas 0.78887206 -0.26744941 0.02529648 0.204421034 ## Kentucky 0.74331256 0.94880748 -0.02808429 0.663817237 ## Louisiana -1.54909076 0.86230011 -0.77560598 0.450157791 ## Maine 2.37274014 0.37260865 -0.06502225 -0.327138529 -1.74564663 0.42335704 -0.15566968 -0.553450589 ## Maryland ## Massachusetts 0.48128007 -1.45967706 -0.60337172 -0.177793902 ## Michigan -2.08725025 -0.15383500 0.38100046 0.101343128 ## Minnesota 1.67566951 -0.62590670 0.15153200 0.066640316 ## Mississippi -0.98647919 2.36973712 -0.73336290 0.213342049 ## Missouri -0.68978426 -0.26070794 0.37365033 0.223554811 ## Montana 1.17353751 0.53147851 0.24440796 0.122498555 ## Nebraska 1.25291625 -0.19200440 0.17380930 0.015733156 ## Nevada -2.84550542 -0.76780502 1.15168793 0.311354436 ## New Hampshire 2.35995585 -0.01790055 0.03648498 -0.032804291 ## New Jersey -0.17974128 -1.43493745 -0.75677041 0.240936580 ## New Mexico -1.96012351 0.14141308 0.18184598 -0.336121113 ## New York -1.66566662 -0.81491072 -0.63661186 -0.013348844 ## North Carolina -1.11208808 2.20561081 -0.85489245 -0.944789648 ## North Dakota 2.96215223 0.59309738 0.29824930 -0.251434626 0.22369436 -0.73477837 -0.03082616 0.469152817 ## Ohio ## Oklahoma 0.30864928 -0.28496113 -0.01515592 0.010228476 ## Oregon -0.05852787 -0.53596999 0.93038718 -0.235390872 ## Pennsylvania 0.87948680 -0.56536050 -0.39660218 0.355452378 ## Rhode Island 0.85509072 -1.47698328 -1.35617705 -0.607402746 ## South Carolina -1.30744986 1.91397297 -0.29751723 -0.130145378 ## South Dakota 1.96779669 0.81506822 0.38538073 -0.108470512 -0.98969377 0.85160534 0.18619262 0.646302674 ## Tennessee ## Texas -1.34151838 -0.40833518 -0.48712332 0.636731051 ## Utah 0.54503180 -1.45671524 0.29077592 -0.081486749 2.77325613 1.38819435 0.83280797 -0.143433697 ## Vermont ## Virginia 0.09536670 0.19772785 0.01159482 0.209246429 ## Washington 0.21472339 -0.96037394 0.61859067 -0.218628161 ## West Virginia 2.08739306 1.41052627 0.10372163 0.130583080 ## Wisconsin 2.05881199 -0.60512507 -0.13746933 0.182253407 ## Wyoming To interpret the above tables, recall that in some code snippet before, prcomp has computed the first, second, third and fourth PCs for us. Let's denote the PCs by u1, u2, u3, u4. Then the original observation of Alabama can be recovered as follows. First, compute −0.97566045u1 + 1.12200121u2 - 0.43980366u3 + 0.154696581u4 as below. t(USArr.scale\$rotation %*% USArr.scale\$x["Alabama",]) Murder Assault UrbanPop ## [1,] 1.242564 0.7828393 -0.5209066 -0.003416473 Recall that we had done normalization. Thus, to recover the original observation, we need to scale back, as below. q <- t(USArr.scale\$rotation %*% USArr.scale\$x["Alabama",])</pre> q <- USArr.scale\$scale * q</pre>

Murder Assault UrbanPop Rape ## [1,] 5.412 65.24 -7.54 -0.032 Recall that we had done centering. Thus, the final step to recover the original observation is to add the center, as below.

q <- q + USArr.scale\$center

[1,] 13.2

USArrests["Alabama",]

q <- q * USArr.scale\$scale</pre> q <- q + USArr.scale\$center</pre>

Alabama 13.2

Alabama

Alaska

Arizona ## Arkansas

Colorado

Florida

Hawaii ## Idaho

Indiana

Kentucky

Maryland

Michigan ## Minnesota

Montana

Nebraska ## Nevada

New York

Oklahoma

Ohio

Oregon

Texas

Vermont

Washington

Wisconsin

We try again by using the first three PCs.

Virginia

Wyoming

Nevada

New Jersey

New Mexico

New York

Oklahoma

Rhode Island

Tennessee

Oregon

Texas

Vermont

Virginia

Wisconsin

t(q) - USArrests

Wyoming

Alabama ## Alaska

Arizona

Arkansas ## California

Colorado

Delaware

Florida

Georgia

Connecticut

##

Washington

Utah

Ohio

Utah

Iowa ## Kansas

Maine

this section.

t(q)

##

Murder Assault UrbanPop Rape

Murder Assault UrbanPop Rape

236

58 21.2

58 21.2

What if we want to represent all the observations approximately by just using the first two PCs?

q <- USArr.scale\$rotation[,1:2] %*% t(USArr.scale\$x[,1:2])</pre>

California 10.838010 268.26639 94.89828 36.343663

Connecticut 2.685203 88.50382 73.74294 16.075381 ## Delaware 7.311590 168.00949 69.79935 21.977168

Georgia 13.881862 269.46522 56.07930 27.507903

Illinois 9.744771 226.52855 79.56429 29.237928

Louisiana 12.974342 259.55664 60.88379 27.765491

Massachusetts 4.005994 124.50152 82.04308 21.069850

Mississippi 14.406774 255.82900 39.57384 22.539391 ## Missouri 8.923185 200.19999 71.61128 25.151572

New Hampshire 2.246987 55.78297 56.26320 9.247892 ## New Jersey 5.593955 157.01550 84.39229 24.395674 ## New Mexico 12.620723 268.23969 71.64636 30.987398

North Carolina 14.401021 259.36249 42.15315 23.435953 ## North Dakota 1.954257 36.08760 46.11917 5.225178

Pennsylvania 4.705434 119.15883 69.14110 17.641434 ## Rhode Island 3.101956 106.06276 80.75648 19.194278 ## South Carolina 14.325832 264.28842 46.62428 24.887390 ## South Dakota 4.679493 87.89189 47.31888 9.938576 ## Tennessee 11.649166 232.20175 58.76635 24.934924

West Virginia 5.484903 91.40798 39.31448 8.396647

q <- USArr.scale\$rotation[,1:3] %*% t(USArr.scale\$x[,1:3])</pre>

Murder Assault UrbanPop

12.108907 235.75582 55.29375 24.439738 14.229193 281.23066 59.89144 29.393422

10.517042 244.02163 81.89791 31.273586

9.506989 228.31388 83.92851 30.395748

14.820838 316.33397 77.06020 36.353369

2.847490 102.49361 81.54315 19.069712

4.379388 95.13675 56.36473 12.642008

6.346772 144.09034 65.42064 18.920200

2.392980 60.71726 57.85772 10.037677

5.459552 128.22997 65.74228 17.635785

7.781172 149.49843 50.55995 15.961600

12.633639 262.23288 67.22073 29.453858

12.379689 269.79276 75.88833 32.097177

2.736773 79.51489 66.70000 13.683755

6.016858 122.05101 54.09994 14.425784 4.513834 106.85882 62.92054 15.155568

13.031265 297.02622 86.69832 36.918887

10.191587 238.94671 82.54253 30.987337

6.548551 151.29501 67.89725 20.107560 6.948400 165.20787 72.54694 22.369863

10.175521 229.56226 76.10072 28.700251

3.862585 121.44951 81.74894 20.740713 3.843345 57.72438 36.83482 4.940606

7.925543 169.22273 62.65796 20.436711

5.537595 145.27871 76.80840 21.644129

1.880326 61.21929 64.89463 11.700994 6.912425 145.45512 59.01612 17.562396

quite a bit, e.g. the number of Assault cases in Arizona is actually 294, but the approximation is 244.02163.

This is unavoidable in real-world data, for which ignoring several PCs will cause large errors on some observations.

By eyeballing the data, you see the approximations t(q) are quite well, although a few entries in the approximation deviate from the actual value by

5.927553 148.37695 73.92218 21.244920

2.928417 61.27947 51.27816 8.570826

9.480315 181.32262 50.97131 18.782131

We can use the following short code snippet. This is the reason why we discuss (a matrix + a vector) and (a matrix * a vector) in the beginning of

236

q <- q * USArr.scale\$scale</pre> q <- q + USArr.scale\$center t(q) ## Murder Assault UrbanPop ## Alabama 12.762562 245.58405 57.70022 21.071008 ## Alaska 11.227726 236.10116 48.84137 44.862031 ## Arizona 10.436442 242.80976 81.60117 31.688969 ## Arkansas 9.311742 178.78799 50.35070 19.650902 ## California 9.957350 255.02495 91.65608 40.882303 7.895899 204.08984 77.99719 38.698791 ## Colorado ## Connecticut 3.631631 102.73413 77.22727 11.197791 ## Delaware 8.368918 183.90728 73.69196 16.528033 ## Florida 15.669530 329.09474 80.18471 31.979479 ## Georgia 14.385725 277.04121 57.93430 24.911152 2.772775 101.37020 81.26808 19.454773 ## Hawaii ## Idaho 3.997141 89.38936 54.95747 14.611988 ## Illinois 10.741574 241.51629 83.23408 24.100723 ## Indiana 6.011234 139.04525 64.18533 20.649457 ## Iowa 2.150856 57.07672 56.96632 11.285508 ## Kansas 5.421955 127.66467 65.60386 17.829546 ## Kentucky 7.822912 150.12603 50.71362 15.746485 ## Louisiana 14.127081 276.88900 65.12766 21.824642 ## Maine 3.025056 62.73251 51.63394 8.072780 ## Maryland 12.865002 265.71161 68.07250 28.261487 ## Massachusetts 4.902751 137.98498 85.34454 16.448251 ## Michigan 11.813430 261.27860 73.80361 35.015496 ## Minnesota 2.511560 76.12862 65.87086 14.844433 ## Mississippi 15.496729 272.21736 43.58658 16.922108 ## Missouri 8.367850 191.85008 69.56678 28.013592 ## Montana 5.653609 116.58926 52.76262 16.297856 ## Nebraska 4.255511 102.97473 61.96951 16.486881

11.319577 271.28960 80.39664 45.740381

6.718699 173.92695 88.53310 18.599098

12.350456 264.17601 70.65135 32.380270

11.137747 253.17299 86.02587 26.111131

5.973368 149.06582 74.09085 21.008803

6.571077 151.63369 67.98018 19.991471

5.565619 144.41663 67.45615 29.496277

5.117564 136.36907 88.17705 8.806474

11.372438 228.04093 57.74756 26.361090

10.899504 240.44793 78.76611 24.969071

3.430422 114.95158 80.15791 22.967947

2.605590 39.11373 32.27795 11.319600

7.908310 168.96362 62.59451 20.525523

4.618219 131.45515 73.42367 26.382300

2.084639 64.29130 65.64682 10.648031

7.266508 150.77905 60.31970 15.737564

Assault

1.227725765 -26.89883681 0.841366243 0.362031141

2.336442298 -51.19024286 1.601174900 0.688968901 $0.511742184 \ -11.21200670 \ \ 0.350699327 \ \ 0.150902272$

 $0.957350071 - 20.97504513 \ 0.656076508 \ 0.282302896$

0.331631199 -7.26586812 0.227268421 0.097791237

2.468918274 -54.09272302 1.691961310 0.728033349

0.269529718 -5.90525678 0.184709984 0.079478784

-3.014275209 66.04121152 -2.065696983 -0.888847921

UrbanPop

9.58404729 -0.299778534 -0.128991584

0.08984324 -0.002810198 -0.001209199

Rape

New Hampshire 2.192761 54.96765 56.06357 9.527353

North Carolina 15.671599 278.46665 46.83086 16.887800 ## North Dakota 1.510986 29.42266 44.48724 7.509655

Pennsylvania 5.294881 128.02164 71.31119 14.603611

South Carolina 14.768014 270.93699 48.25220 22.608520 ## South Dakota 4.106724 79.27984 45.21020 12.890447

West Virginia 5.330748 89.09012 38.74695 9.191115

Let's compute the absolute errors and the relative errors (as percentage)

Murder

-0.437438313

-0.004100655

Hawaii -2.527225487 55.37020393 -1.731919518 -0.745226950 ## Idaho 1.397141132 -30.61063992 0.957467391 0.411988257 0.341573963 -7.48370894 0.234082244 0.100723154 ## Illinois ## Indiana -1.188765864 26.04524552 -0.814666840 -0.350542666 -0.049144240 1.07672490 -0.033678779 -0.014491628 ## Iowa ## Kansas -0.578045046 12.66466812 -0.396136990 -0.170453625 -1.877087978 41.12602720 -1.286377223 -0.553514736 ## Kentucky -1.272919307 27.88900396 -0.872337591 -0.375357790 ## Louisiana ## Maine 0.925055519 -20.26748820 0.633944900 0.272779895 ## Maryland 1.565002216 -34.28838950 1.072503382 0.461487047 ## Massachusetts 0.502751025 -11.01501504 0.344537643 0.148250963 ## Michigan -0.286569791 6.27859598 -0.196387626 -0.084503552 ## Minnesota -0.188440024 4.12862352 -0.129138835 -0.055567097 ## Mississippi -0.603271160 13.21735924 -0.413424564 -0.177892289 ## Missouri ## Montana -0.346391373 7.58925592 -0.237383638 -0.102143710 ## Nebraska -0.880422556 19.28960305 -0.603357720 -0.259618550 ## Nevada ## New Hampshire 0.092761284 -2.03235178 0.063569745 0.027353400 ## New Jersey -0.681300714 14.92694644 -0.466898585 -0.200901604 ## New Mexico 0.950455734 -20.82399378 0.651351787 0.280269897 ## New York 0.037746767 -0.82701216 0.025868037 0.011130747 ## North Carolina 2.671598734 -58.53334712 1.830859183 0.787799658 ## North Dakota 0.710986228 -15.57734070 0.487242208 0.209655253 ## Ohio -1.326631883 29.06581873 -0.909147072 -0.391196526 ## Oklahoma ## Oregon 0.665619015 -14.58336851 0.456151843 0.196277392 ## Pennsylvania -1.005119101 22.02163989 -0.688812849 -0.296389003 ## Rhode Island 1.717563704 -37.63093282 1.177054488 0.506474300 ## South Carolina 0.368014433 -8.06300598 0.252202023 0.108519906 ## South Dakota 0.306724025 -6.72016481 0.210199417 0.090446622 -1.827561731 40.04093274 -1.252436653 -0.538910462 ## Tennessee -1.800495880 39.44793392 -1.233888298 -0.530929297 ## Texas ## Utah 0.230421551 -5.04841706 0.157908973 0.067946588

Vermont $0.405590052 - 8.88626836 \ 0.277952771 \ 0.119600186$ ## Virginia -0.591689904 12.96361990 -0.405487875 -0.174477214 ## Washington 0.618218795 -13.54485419 0.423668249 0.182300040 ## West Virginia -0.369252132 8.09012331 -0.253050223 -0.108884878 -0.515361248 11.29129851 -0.353179488 -0.151969459 ## Wisconsin ## Wyoming 0.466508060 -10.22095041 0.319700168 0.137563656 ((t(q) - USArrests) / USArrests) * 100 UrbanPop Murder Assault ## Alabama -3.31392661 4.0610370 -0.516859542 -0.608450870 12.27725765 -10.2276946 1.752846341 0.813553126 ## Alaska 28.84496664 -17.4116472 2.001468625 2.222480326 ## Arizona ## Arkansas 5.81525209 -5.9010562 0.701398653 0.773857803 ## California 10.63722301 -7.5996540 0.720963196 0.695327329 0.0440408 -0.003602819 -0.003124546 ## Colorado -0.05190703 ## Connecticut 10.04943027 -6.6053347 0.295153793 0.881002136 ## Delaware 41.84607244 -22.7280349 2.349946263 4.607806004 1.75019297 -1.7627632 0.230887480 0.249149794 ## Florida ## Georgia -17.32342074 31.2991524 -3.442828305 -3.445146979 ## Hawaii -47.68349976 120.3700085 -2.086650021 -3.689242325 ## Idaho 53.73619738 -25.5088666 1.773087762 2.901325754 ## Illinois 3.28436503 -3.0055056 0.282026800 0.419679809 ## Indiana -16.51063700 23.0488898 -1.2533333600 -1.669250788 -2.23382911 1.9227230 -0.059085578 -0.128244498 ## Iowa ## Kansas -9.63408410 11.0127549 -0.600207560 -0.946964585 ## Kentucky $-19.35142245 \quad 37.7303002 \quad -2.473802353 \quad -3.395795926$ -8.26570979 11.2004032 -1.321723622 -1.690800854 ## Louisiana ## Maine 44.05026283 -24.4186605 1.243029216 3.497178140 ## Maryland 13.84957713 -11.4294632 1.600751317 1.660025350 ## Massachusetts 11.42615965 -7.3926275 0.405338403 0.909515111 ## Michigan -2.36834538 2.4621945 -0.265388684 -0.240750861 -6.97926016 5.7341993 -0.195664902 -0.372933539 ## Minnesota ## Mississippi -3.74702584 5.1032275 -0.939601282 -1.040305782 -7.02388858 ## Missouri 7.7809433 -0.618879072 -0.661021483 ## Montana -5.77318956 6.9626201 -0.447893657 -0.622827498 ## Nebraska -1.03462627 0.9556171 -0.049174956 -0.079508319 7.6546044 -0.744886074 -0.564388152 ## Nevada -7.21657833 ## New Hampshire 4.41720401 -3.5655294 0.113517401 0.287930523 ## New Jersey -9.20676641 9.3880166 -0.524605152 -1.068625551 ## New Mexico 8.33733100 -7.3066645 0.930502552 0.873114945 0.34006097 -0.3255953 0.030079113 0.042646542 ## New York ## North Carolina 20.55075949 -17.3689457 4.068575961 4.893165579 ## North Dakota 88.87327853 -34.6163127 1.107368653 2.871989763 ## Ohio -18.17303950 24.2215156 -1.212196095 -1.828021151 ## Oklahoma -0.43823099 0.4196645 -0.029148877 -0.042644358 13.58406153 -9.1719299 0.680823646 0.669888710 ## Oregon ## Pennsylvania -15.95427145 20.7751320 -0.956684513 -1.989187937 ## Rhode Island 50.51657953 -21.6269729 1.352936193 6.102100001 ## South Carolina 2.55565579 -2.8899663 0.525420880 0.482310695 ## South Dakota 8.07168486 -7.8141451 0.467109816 0.706614237 ## Tennessee -13.84516463 21.2983685 -2.122773989 -2.003384616 -14.17713291 19.6258378 -1.542360373 -2.082075676 ## Texas ## Utah 7.20067346 -4.2070142 0.197386216 0.296709992 ## Vermont 18.43591145 -18.5130591 0.868602409 1.067858799 ## Virginia 8.3100128 -0.643631548 -0.842885092 -6.96105769 15.45546987 -9.3412788 0.580367464 0.695801678 ## Washington ## West Virginia -6.47810757 9.9878066 -0.648846726 -1.170805143 -19.82158645 21.3043368 -0.535120436 -1.407124622 ## Wisconsin ## Wyoming 6.86041264 -6.3484164 0.532833614 0.881818309