- estimated covariance $S=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T$, diagonal is sample variance
- corr: $\rho_{ij}=\frac{\sigma_{ij}}{\sigma_i\sigma_j}$, $\sigma_i=\sqrt{\sigma_i^2}$; $R=D^{-\frac{1}{2}}SD^{-\frac{1}{2}}$, $D^{-\frac{1}{2}}=\mathrm{diag}(\frac{1}{s_1},\cdots,\frac{1}{s_p})$
- Euclidean dist: $d_{ij}=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$ → est. of R, where true $R=D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$
est. mean: $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$ (both est. are unbiased) → fix: scale vars to be dimensionless by $\times\frac{1}{s_k}$
Manhattan dist: $d_M(x_i,x_j)=\sum_{k=1}^{p}|x_{ik}-x_{jk}|$

$N(\mu,\sigma^2)$, $f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$; $\underset{\sim}{X}\sim N(\underset{\sim}{\mu},C)\Rightarrow$
$f(\underset{\sim}{x})=f(x_1,\cdots,x_p)=\frac{1}{(2\pi)^{p/2}|C|^{\frac{1}{2}}}\exp[-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})^TC^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})]$ (→ determinants; ↳ positive definite)

Two approaches to check multi-normality: ① Mahalanobis distance / ② Ad hoc.
$\underset{\sim}{X}=\begin{pmatrix}\underset{\sim}{X_1}\\ \underset{\sim}{X_2}\end{pmatrix}$ ← p-r, $\underset{\sim}{\mu}=\begin{pmatrix}\underset{\sim}{\mu_1}\\ \underset{\sim}{\mu_2}\end{pmatrix}$, $C=\begin{pmatrix}C_{11}&C_{12}\\ C_{21}&C_{22}\end{pmatrix}$, $K=C^{-1}$ (concentration matrix)
$\underset{\sim}{X_1}|\underset{\sim}{X_2}=\underset{\sim}{x_2}\sim N_r(\underset{\sim}{\mu_{1|2}},C_{1|2})$ where $\underset{\sim}{\mu_{1|2}}=\underset{\sim}{\mu_1}+C_{12}C_{22}^{-1}(\underset{\sim}{x_2}-\underset{\sim}{\mu_2})=\underset{\sim}{\mu_1}-K_{11}^{-1}K_{12}(\underset{\sim}{x_2}-\underset{\sim}{\mu_2})$ (conditional mean)
$C_{1|2}=C_{11}-C_{12}C_{22}^{-1}C_{21}=K_{11}^{-1}$ conditional covariance
Graphical Lasso: K "sparse", maximize $\ln(|K|)-\mathrm{trace}(K\hat{C})-\lambda\|K\|_1$.
scatterplots(2D) / ① find "interesting" 2D projections / ② Grand Tour Method
→ $\binom{p}{2}$ pairs! downside: miss out on higher dim structure.
"Non-interesting" projection = projections that are normal.
Kurtosis, X, μ, σ^2: $K(X)=E[(X-\mu)^4]/\sigma^4$, $K(aX+b)=K(X)$
If $X\sim N(\mu,\sigma^2)$, $K(X)=3$, long tailed dist'n, $K>3$, multi-model $K(X)>3$
$\hat{K}(\underset{\sim}{a})=\frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{a}^T\underset{\sim}{x_i}-\underset{\sim}{a}^T\underset{\sim}{\bar{x}})^4/(\underset{\sim}{a}^TS\underset{\sim}{a})^2$, find $\underset{\sim}{a}$ to optimize $\hat{K}(\underset{\sim}{a})$
Density Estimation, kernel density estimation: $\hat{f}_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K(\frac{x-x_i}{h})$
Types of kernel: ① Rectangle kernel $K(x)=\frac{1}{2}$ for $|x|\le1$
② Δ kernel $K(x)=1-|x|$ for $|x|\le1$ ③ Gaussian kernel $K(x)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$
④ Epanechnikov kernel $K(x)=\frac{3}{4}(1-x^2)$, $|x|\le1$.
kernel density est for large p is problematic ⇒ use projection pursuit
PCA approaches ① Low rank matrix approximation; ② Projection pursuit → maximize variance
assume $p<n$, $X=\begin{pmatrix}\underset{\sim}{x_1}^T\\ \vdots\\ \underset{\sim}{x_n}^T\end{pmatrix}$ rank$(X)=p$, approx X by a matrix with rank $r<p$.
- find X^* to minimize $\|X-M\|_F^2=\sum_{i=1}^{n}\sum_{j=1}^{p}|X_{ij}-M_{ij}|^2$
- Singular Value Decomposition (SVD)
Write $X=\underset{n\times p}{U}\underset{p\times p}{D}\underset{p\times p}{V^T}$, U & V have orthonormal columns. $U=(\underset{\sim}{u_1},\cdots\underset{\sim}{u_p})$, $V=(\underset{\sim}{v_1}\cdots\underset{\sim}{v_p})$
with $\underset{\sim}{u_i}^T\underset{\sim}{u_j}=\underset{\sim}{v_i}^T\underset{\sim}{v_j}=\begin{cases}1&\text{if } i=j\\0&\text{o.w.}\end{cases}$, $D=\begin{pmatrix}d_1&&0\\&\ddots&\\0&&d_p\end{pmatrix}$, $d_1\ge\cdots\ge d_p>0$
Now $D^*=\begin{pmatrix}d_1&&&0\\&\ddots&&\\&&d_r&\\0&&&0\end{pmatrix}$, $X^*=UD^*V^T$, rank$(X^*)=\min($rank$(U),$ rank$(D^*),$ rank$(V))$ $=$ rank$(D^*)=r$
$X^*V=UD^*V^TV=UD^*=(d_1\underset{\sim}{u_1}\cdots d_r\underset{\sim}{u_r}\ 0\cdots0)$
$X^*\underset{\sim}{v_1}=d_1\underset{\sim}{u_1}$, $X^*\underset{\sim}{v_2}=d_2\underset{\sim}{u_2}$, $\cdots$, $X^*\underset{\sim}{v_r}=d_r\underset{\sim}{u_r}$, $X^*\underset{\sim}{v_{r+1}}=\cdots=X^*\underset{\sim}{v_p}=\underset{\sim}{0}$
From SVD → PCA: $\tilde{X}=UDV^T$, $S=\frac{1}{n-1}\tilde{X}^T\tilde{X}=\frac{1}{n-1}VDU^TUDV^T=V(\frac{D^2}{n-1})V^T$
Sample cov. matrix, corr matrix if scaled
columns of V are eigenvectors $\underset{\sim}{v_1},\cdots,\underset{\sim}{v_p}$ of S with eigenvalues the diagonal elements of $D^2/(n-1)\Rightarrow d_1^2/(n-1),\cdots,d_r^2/(n-1)$
Principal Components: $y=\tilde{X}V=\tilde{X}(\underset{\sim}{v_1},\cdots,\underset{\sim}{v_p})=(\tilde{X}\underset{\sim}{v_1},\cdots,\tilde{X}\underset{\sim}{v_p})\Rightarrow$ PC scores

② proj. pursuit approach: $\underset{\sim}{x_1},\cdots,\underset{\sim}{x_n}$ with sample var S or correlation matrix R
find $\underset{\sim}{a}$ with $\|\underset{\sim}{a}\|^2=\underset{\sim}{a}^T\underset{\sim}{a}=1$ maxing sample variance of $\{\underset{\sim}{a}^T\underset{\sim}{x_i}\}$:
$\frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{a}^T\underset{\sim}{x_i}-\underset{\sim}{a}^T\underset{\sim}{\bar{x}})^2=\underset{\sim}{a}^TS\underset{\sim}{a}$, $S=V\Lambda V^T$, $\Lambda=D^2/(n-1)$, $\underset{\sim}{a}^TS\underset{\sim}{a}=\underset{\sim}{a}^TV\Lambda V^T\underset{\sim}{a}$
$=\sum_{k=1}^{p}\lambda_k(\underset{\sim}{v_k}^T\underset{\sim}{a})^2$, $\Lambda=\begin{pmatrix}\lambda_1&&0\\&\ddots&\\0&&\lambda_p\end{pmatrix}$, maxed out $\underset{\sim}{a}=\underset{\sim}{v_1}$
max $\underset{\sim}{a}^TS\underset{\sim}{a}$ s.t. $\|\underset{\sim}{a}\|^2=1$ & $\underset{\sim}{a}^T\underset{\sim}{v_1}=0\Rightarrow\underset{\sim}{a}=\underset{\sim}{v_2}$ etc.
Note that $\frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{v_k}^T\underset{\sim}{x_i}-\underset{\sim}{v_k}^T\underset{\sim}{\bar{x}})^2=\underset{\sim}{v_k}^TS\underset{\sim}{v_k}=\lambda_k$

- R better than S
- var with high variance dominates.
- few PCs most info.
- biplot (important)

$\lambda_1,\cdots,\lambda_g\cdots$ PCs' variances, PCA → transforms original data to obtain uncorrelated variables.
- PC scores: columns of $\tilde{X}V=UD=(\underset{\sim}{u_1},\cdots\underset{\sim}{u_p})\begin{pmatrix}d_1&&0\\&\ddots&\\0&&d_p\end{pmatrix}=(d_1\underset{\sim}{u_1}\cdots d_p\underset{\sim}{u_p})$
- PC loadings: columns of $V=(\underset{\sim}{v_1},\cdots,\underset{\sim}{v_p})$

Multidim-scaling
distance matrix $D=(d_{ij})$, $d=(\underset{\sim}{x_i}-\underset{\sim}{x_j})$, Recover $\{d_{ij}\}$ from $B=XX^T=\begin{pmatrix}b_{11}&\cdots\\ \vdots&\ddots\\ \cdots&b_{nn}\end{pmatrix}$
$d_{ij}^2=b_{ii}+b_{jj}-2b_{ij}$, SVD $\Rightarrow X=U\Lambda\overset{\frac{1}{2}}{\Lambda}V^T$, $XX^T=U\Lambda U^T$, $U=(\underset{\sim}{u_1},\cdots\underset{\sim}{u_p})$
- optimal k dim rep. of pts $(\lambda_1^{\frac{1}{2}}\underset{\sim}{u_1},\lambda_2^{\frac{1}{2}}\underset{\sim}{u_2},\cdots,\lambda_k^{\frac{1}{2}}\underset{\sim}{u_k})$
- how to get low dim rep for data? $B=U\Lambda U^T$, $\Lambda=\begin{pmatrix}\lambda_1&&0\\&\ddots&\\0&&\lambda_n\end{pmatrix}$, $\lambda_1\ge\lambda_2\ge\cdots$
if $k=2$, plot $\lambda_1^{\frac{1}{2}}\underset{\sim}{u_1}$ v.s. $\lambda_2^{\frac{1}{2}}\underset{\sim}{u_2}$

Andrew curves (idea: rep each $\underset{\sim}{x_i}$ as a function of $[0,1]$) $g_i(t)=\sum_{k=1}^{p}x_{ik}\phi_k(t)$
Need to choose $\phi_1,\cdots\phi_p$ s.t. $g_i(t)=g_j(t)$ for $t\Rightarrow x_i=x_j$
take $\phi_1(t)=\frac{1}{\sqrt{2}}$, $\phi_2(t)=\sin(2\pi t)$, $\phi_3(t)=\cos(2\pi t)$, $\sin(4\pi t)$, $\cos(4\pi t)$, $\cdots$

Practice: Q1. ① joint dist'n of multi-normal: "section". ② dist'n of $X_1+X_2+X_3$, Ψ is the sum, variance $\sigma^2=a^TCa$, a is scalar. ③ Conditional of X_1 given $X_2+X_4+X_5$, $X_3=-1$, know joint mean $\underset{\sim}{\mu}$, C_{15}, Corr$(X_1,X_5)=\rho_{15}=\sigma_{15}/\sqrt{\sigma_{11}^2\sigma_{55}^2}$
so $\mu_{1|5}=\mu_1+\rho_{15}\frac{\sigma_1}{\sigma_5}(\bar{x}-\mu_5)$, variance $\sigma_1^2(1-\rho_{15}^2)$
Q2: $\hat{R}=\begin{pmatrix}1&\rho&\cdots\\ \rho&1&\\ \vdots&&\ddots\end{pmatrix}$ p-variate, $x_1,\cdots,x_n$. Show one of PC has equal loadings, i.e. coeff of all p variables are equal.
Loadings are eigenvectors of correlation, their sum equal to 1
if one of PC has equal loadings then $\vec{v}=(1,1,\cdots1)^T$ for example is an eigenvector of $\hat{R}$, verify $\hat{R}\vec{v}=\lambda\vec{v}$ for some λ.
$\hat{R}\vec{v}=\begin{pmatrix}&&\\&&\\&&\end{pmatrix}\begin{pmatrix}1\\1\\\vdots\end{pmatrix}=(1+(p-1)\rho)\begin{pmatrix}1\\1\\\vdots\end{pmatrix}$

$\hat{R}=V\Lambda V^T$, where cols of V are PC loadings and diag Λ are eigenvalues of $\hat{R}$. Define $\underset{\sim}{v_1}$, $\underset{\sim}{v_2}$, (2 cols of V) and $A=(\lambda_1^{\frac{1}{2}}\underset{\sim}{v_1},\ \lambda_2^{\frac{1}{2}}\underset{\sim}{v_2})$
If $\lambda+\lambda_2=p$ then $\lambda_1+\cdots+\lambda_p\approx0$, so $\hat{R}\approx AA^T$.
If $p=16$, $\underset{\sim}{x}=(x_1,\cdots,x_{16})$, $\underset{\sim}{y}=(y_1,\cdots,y_{16})$. Approx to $d(\underset{\sim}{x},\underset{\sim}{y})=[\sum_{i=1}^{16}(x_i-y_i)^2]^{\frac{1}{2}}$
$\underset{\sim}{x},\underset{\sim}{y}$ can be well represented by 1st 2 PC scores:
$S_1(\underset{\sim}{x})=\frac{1}{4}(x_1+\cdots+x_{16})$ ... standardized!
$S_2(\underset{\sim}{x})=\frac{1}{4}(x_1+\cdots x_8-x_9-\cdots)$
$S_1(\underset{\sim}{y})=\cdots$, $S_2(\underset{\sim}{y})=\cdots$, $d(\underset{\sim}{x},\underset{\sim}{y})=[(S_1(\underset{\sim}{x})-S_1(\underset{\sim}{y}))^2+(S_2(\underset{\sim}{x})-S_2(\underset{\sim}{y}))^2]^{\frac{1}{2}}$

Q3. summary of loadings:
- sd are square roots of the eigenvalues & sum of eigenvalue equals to # of variables (for PCA using R):
$sd(\text{comp }1)^2+\cdots+sd(\text{comp }5)^2=5$
- in "Loadings": sum of squares of each loading is 1.
- the sign of a "loading value"
  - PC loadings are orthogonal vectors: $(\ x\ )+\cdots(\ )(\ )=1$
  - or $R\vec{v}=\lambda\vec{v}$, $=sd^2$
- sd of 5th component $\doteq0$: that particular linear combination of variables (given by PC loadings) takes the same value for all observations.

HW1 Q1. marginal; conditional dist'n: $\underset{\sim}{\mu}=\begin{pmatrix}\underset{\sim}{\mu_1}\\ \underset{\sim}{\mu_2}\end{pmatrix}$, $C=\begin{pmatrix}C_{11}&C_{12}\\ C_{21}&C_{22}\end{pmatrix}$
$\mu_{1|2}=\mu_1+C_{12}C_{22}^{-1}(x_2-\mu_2)$, $C_{1|2}=C_{11}-C_{12}C_{22}^{-1}C_{21}$
Q2: Shapiro-Wilks Test: small-p-value ⇒ problematic?
QQplot ⇒ check normality by checking diagonal straight line.

HW2 Q1: use identified observations (possible suspects) see where they fall in pairwise scatterplots for the original variables & PC scores. The PC scores are particularly useful as these pts tend to lie on the edge of point cloud, particularly for first few PCs.
Q2: Write $g_i(t)=\sum_{k=1}^{p}x_{ik}\phi_k(t)$.
Note that $\int_0^1\phi_k^2(t)\,dt=\frac{1}{2}$, $\int_0^1\phi_k(t)\int_0^1\phi_\ell(t)\,dt=0$
Thus $\int_0^1[g_i(t)-g_j(t)]^2=\int_0^1[\sum_{k=1}^{p}\sum_{\ell=1}^{p}(x_{ik}-x_{jk})(x_{i\ell}-x_{j\ell})\phi_k(t)\phi_\ell(t)]\,dt$
$=\sum\sum(\ \ x\ )\int_0^1\{\phi\ \phi\ dt\}=\frac{1}{2}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2$
mean of Andrew curve ---
"sandwich"
PCAs, (a). Loadings for 1st PC approx same - since the variables are different measures of the size of the crab. 2nd PC compares the variable RW to the other 4.
(d). Given some data, predict?
- need to first normalize variables & compute score for measurements: data $\underset{\sim}{x}$, $\frac{\underset{\sim}{x}-\underset{\sim}{\mu}}{\underset{\sim}{sds}}$, $t(\text{loadings})\%*\%(\ )$

- Projection pursuit: remove "interesting" (non-normal distribution)
- curse of dimensionality ("high dim space is mostly empty")
- can ignore irrelevant variables
- PCA transforms the original coordinate system to PCs
① origin of the new system is located in the centre of data points; ② 1st PC with highest variance ... ③ system stays perpendicular.
- Because PCA estimates how much of each old coordinates contribute to each of new ones? Values called loadings.

Example $p=2$, $R=\begin{pmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{pmatrix}$: eigenvalues solve $\det(R-\lambda I)=0$

$\lambda^2-2\lambda+1-\hat{\rho}^2=0 \Rightarrow \lambda_1=1+|\hat{\rho}|,\ \lambda_2=1-|\hat{\rho}|$

eigenvectors (loadings) $= \underset{\sim}{v}_1=\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \operatorname{sgn}(\hat{\rho})\frac{1}{\sqrt{2}} \end{pmatrix}$ $\underset{\sim}{v}_2=\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\operatorname{sgn}(\hat{\rho})\frac{1}{\sqrt{2}} \end{pmatrix}$

Joint distribution of $(X_1, X_2, X_5)^T$ is 3-variate normal with mean (μ_1, μ_2, μ_5)

covariance matrix C_{335} =

the distribution of $X_1+X_2+\cdots+X_5$ is normal with $\mu=\mu_1+\cdots+\mu_5$, variance is

$\sigma^2=(1\,1\,1\,1\,1)\,C\,(1\,1\,1\,1\,1)^T=140$.

conditional distribution: X_1 given $X_5=-1$? $X_1,\cdots,X_5$? $C_{55}=\begin{pmatrix}55 & -6\\ -6 & 60\end{pmatrix}$

Cor$(X_1,X_5)=\rho_{15}=-6/\sqrt{55\times 60}$. The conditional dist'n of X_1 given $X_5=x_5$

is normal with mean $\mu_{1|5}=\mu_1+\rho_{15}\frac{\sigma_1}{\sigma_5}(x-\mu_5)=1-\frac{6}{60}x$

variance is $\sigma_1^2(1-\rho_{15}^2)=54.4$. plug in $x=-1$, $\mu_{1|5}=1.1$

correlation matrix?

$K=C^{-1}$, linkage if not equal to 0.

$R=\begin{pmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & & \\ \vdots & & \ddots & \\ \rho & & & 1\end{pmatrix}_{(a)}$ Show PC's have equal loadings: The loadings are eigenvectors of the correlation; standardized so that their sums of squares is 1. If one of the PCs has equal loadings then (e.g.) the vector $v=(1,1,\cdots,1)^T$ is an eigenvector of $\hat{R}$. We just to need verify $Rv=\lambda v$ for some λ: $\hat{R}v=\begin{pmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & & \\ \vdots & & \ddots & \\ \rho & & & 1\end{pmatrix}\begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix}=\begin{pmatrix}1+(p-1)\rho\\ \vdots\\ 1+(p-1)\rho\end{pmatrix}$

(b). Sps p is even loadings of 1st 2 PCs are $(p^{-\frac{1}{2}},\cdots,p^{-\frac{1}{2}})^T$ & $(p^{-\frac{1}{2}},\cdots,p^{-\frac{1}{2}},-p^{-\frac{1}{2}},\cdots,-p^{-\frac{1}{2}})^T$

& $\lambda_1\geq\lambda_2$, $\lambda_1+\lambda_2=p$.

$\hat{R}=V\Lambda V^T$ where columns of V are PC loadings & diagonal elements of Λ are λ_1 & λ_2. and $A=(\lambda_1^{\frac{1}{2}}V_1\ \ \lambda_2^{\frac{1}{2}}V_2)$ (A is $p\times 2$ matrix). If $\lambda_1+\lambda_2=p$

then $\lambda_3+\cdots+\lambda_p=0$ so $\hat{R}=AA^T$

(c). If $p=16$, $\underline{x}=(x_1,\cdots,x_{16})$ & $\underline{y}=(y_1,\cdots,y_{16})$, what is $d(\underline{x},\underline{y})=\left(\sum_{i=1}^{16}(x_i-y_i)^2\right)^{\frac{1}{2}}$?

$S_1(\underline{x})=\frac{1}{4}(x_1+\cdots+x_{16})$, $S_2(\underline{x})=\frac{1}{4}(x_1+\cdots+x_8-x_9-\cdots-x_{16})$ $S_1(\underline{y})=\frac{1}{4}(y_1+\cdots+y_{16})$

$S_2(\underline{y})=\frac{1}{4}(y_1+\cdots+y_8-y_9-y_{16})$ so. $d(\underline{x},\underline{y})=\left[(S_1(\underline{x})-S_1(\underline{y}))^2+(S_2(\underline{x})-S_2(\underline{y}))^2\right]^{\frac{1}{2}}$.

PCA prob: ①The standard deviations are the square roots of the eigenvalues & the sum of the eigenvalues equals the # of variables.

for PCA using the correlation matrix)

②We know that the sum of each loading is 1. To determine the sign, we can use fact that loadings are orthogonal: squares of

$l_{11}l_{21}+l_{12}l_{22}+\cdots=0$.

③If the variance of a PC is 0 or ≈ 0, it means the particular linear combination of variables (given by PC loadings) takes the same value for all observations

If $X\sim N_p(0,C)$, $\forall a$, $a^TCa\geq 0$. $X_1\sim N_p(0,C_1)$, $X_2\sim N_p(0,C_2)$ iid $X_1-X_2\sim N_p(0,C_1+C_2)$. If $X\sim N_p(0,C)$, $a^TX\sim N(0,a^TCa)$;

If $X\sim N_p(0,C)$, then $X^TC^{-1}X$ has χ^2 dist with p dof.

If $X\sim N_p(0,I)$. then X^TX has χ^2 dist with p dof.

If $l_1,\cdots,l_4$ are the loadings then PC scores are

$l_j^Ty_i$ for $j=1,\cdots,4$, $i=1,\cdots,100$.

④The correlation matrix of PC scores is I b/c PCs are uncorrelated.

Multivariate normality: ①Assess the normality of many one-dim projections: informally, using quantile-quantile plots & more formally, using Shapiro-Wilks test. ②Compare quantiles of a χ^2 distribution with p dof to order values of $(x_i-\bar{x})^TS^{-1}(x_i-\bar{x})$ $(i=1,\cdots,n)$. If the data are multivariate normal the pts should fall close to a straight line.

Biplot: show the individual variables are correlated with the first 2 PCs.

Factor analysis: One-factor model: $\Psi=\mathrm{diag}(\psi)$ is uniqueness:

$\hat{R}=LL^T+\Psi$. p-value should >0.05. OK✓.

Two-factor model: Q rotation. $QQ^T=I$ if Q orthogonal.

new loading is LQ.

Classification: ~~$\underline{x}=\underline{x}(X_1,X_2)$ coming from~~ ① $f_1(\underline{x}),f_2(\underline{x})$ for $0<x_1^2+x_2^2<1$

$f_1(x_1,x_2)=\frac{1}{\pi}$ for $0<x_1^2+x_2^2<1$, $\lambda_1=\frac{1}{3}$ | classification rule: classifies $\underline{x}=(x_1,x_2)$ as coming from f_2 if $x_1>0$.

$f_2(x_1,x_2)=2(x_1^2+x_2^2)\frac{1}{\pi}$ for $0<x_1^2+x_2^2<1$ | $\lambda_2=\frac{2}{3}$

prob of misclassification $=\frac{1}{3}P(X_1>0\,|\,f_1)+\frac{2}{3}P(X_1<0\,|\,f_2)=\frac{1}{3}\times\frac{1}{2}+\frac{2}{3}\times\frac{1}{2}=\frac{1}{2}$

②The optimal classification rules classifies an obs $\underline{x}$ as belonging to f_1 if $f_1(\underline{x})\frac{1}{3}>2f_2(\underline{x})\frac{1}{3}$ and as belonging to f_2 if $2f_2(\underline{x})\frac{1}{3}>f_1(\underline{x})\frac{1}{3}$. In other words, the optimal rule depends on whether $x_1^2+x_2^2$ is greater than or less than $\frac{1}{2}$.

③Sps we only know the signs of X_1 & X_2, only tells us which quadrant the obs. lies in. If we define $Y_1=\mathrm{sign}(X_1)$, $Y_2=\mathrm{sign}(X_2)$ then dist of (Y_1,Y_2) is the same under f_1 & f_2.

The optimal classification rule is to always classify as belonging to f_2, rate of misclassification is $\frac{1}{3}$.

ICA vs. PCA: PCA effectively does an eigen-decomposition of the correlation or covariance matrix of the data to define new uncorrelated variables, the principal components. In many cases, the first new PCs (with large variances) will describe the structure of the data very well. The model for ICA is $X=\mu+AY$ where the components of Y are assumed to be independent, not simply uncorrelated. Like PCA, ICA can be used for dimension reduction but more generally, we are interested in all the components of the random vector Y.

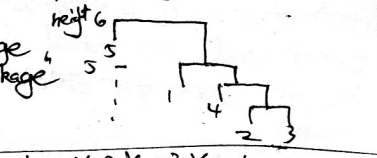
How Principal Components can be used construct est. for Factor Analysis

In the factor analysis model, the correlation matrix $R=LL^T+\Psi$. If $\hat{R}$ is the sample correlation matrix then $\hat{R}=V\Lambda V^T$ where the diagonal elements of Λ are ordered from largest to smallest. Therefore, if we want an r factor model, we can define L to be the first r columns of V multiplied by the square roots of the eigenvalues $\lambda_1^{\frac{1}{2}}\geq\cdots\geq\lambda_r^{\frac{1}{2}}$ and Ψ to be a diagonal matrix s.t. so that the diagonal elements of $LL^T+\Psi$ are 1s. (The same can be done with the sample covariance matrix making appropriate modifications).

distance matrix

	1	2	3	4
2	3			
3	9	③		
4	10	7	4	
5	16	11	8	6

single linkage
complete linkage



HW1

conditional dist'n (X1, X2) given X3=2, X4=3, X5=-1

$\mu_1=\begin{pmatrix}1\\ 2\end{pmatrix}$, $\mu_2=\begin{pmatrix}3\\ 0\\ 0\end{pmatrix}$, $C_{11}(2\times 2)$ $C_{12}=(2\times 3)$, $C_{22}=(3\times 3)$

$\mu_{1|2}=\mu_1+C_{12}C_{22}^{-1}(\underline{x}_2-\mu_2)$, $C_{1|2}=C_{11}-C_{12}C_{22}^{-1}C_{21}=(2\times 2)$

Shapiro-Wilks test: p-value <0.05 not normal.

HW2

2(a). If $\{g_j(t)\}=\frac{x_{j1}}{\sqrt{2\pi}}+x_{j2}\sin(2\pi t)+x_{j3}\cos(2\pi t)+\cdots$

Show that $2\int_0^1[g_i(t)-g_j(t)]^2dt=\sum_{k=1}^p(x_{ik}-x_{jk})^2$

Write $g_i(t)=\sum_{k=1}^p x_{ik}\phi_k(t)$ where $\phi_1(t)=\frac{1}{\sqrt{2\pi}}$, $\phi_2,\cdots,\phi_k$ are sines & cosines. Note that

$\int_0^1\phi_k^2(t)dt=\frac{1}{2}$ & $\int_0^1\phi_k(t)\int_0^1\phi_k(t)\,dt=0$ $\forall k,l$

and $\int_0^1[g_i(t)-g_j(t)]^2dt=\int_0^1\sum_{k=1}^p\sum_{l=1}^p(x_{ik}-x_{jk})(x_{il}-x_{jl})\phi_k(t)\phi_l(t)dt$

$=\sum\sum(x-x)(x-x)\int_0^1\Phi\Phi\,dt=\frac{1}{2}\sum_{k=1}^p(x_{ik}-x_{jk})^2$

HW3

Sps S symmetric positive definite $p\times p$ with $S=V\Lambda V^T$ where Λ is diagonal with $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$. V is an orthogonal matrix with columns $v_1,v_2,\cdots,v_p$. (a). Sps approximate S by $\Psi=LL^T$ where $L=(\lambda_1^{\frac{1}{2}}v_1,\cdots,\lambda_r^{\frac{1}{2}}v_r)$, Ψ is diagonal with elements of $\Psi+LL^T$ equal to those of S. If l_{ij} is $(i-j)$ th element of L, $\psi_{ii}=?$

$s_{ii}=\psi_{ii}+\sum_{j=1}^r l_{ij}^2$, $\psi_{ii}=s_{ii}-\sum_{j=1}^r l_{ij}^2$

(b). $D=\Psi+LL^T-S$. $[d_{ij}]$ show $\sum_{i=1}^p\sum_{j=1}^p d_{ij}^2\leq\lambda_{r+1}^2+\cdots+\lambda_p^2$

The diag of D are 0 by definition; if we define $D^*=LL^T-S$

D & D^* have the same off-diag elements and so $\sum_{i=1}^p\sum_{j=1}^p d_{ij}^2<\sum_{i=1}^p\sum_{j=1}^p(d_{ij}^*)^2$. Now $D^*=LL^T-S=V\Lambda^*V^T-V\Lambda V^T$

where Λ^* is $p\times p$ diagonal with $\lambda_1,\cdots,\lambda_r,0,\cdots,0$. Thus $D^*=-\sum_{j=r+1}^p\lambda_j v_j v_j^T$. The i-th row of D^* is $d_i^T=-\sum_{j=r+1}^p\lambda_j v_{ij}v_j^T$

and $d_i^Td_i=\sum_{j=r+1}^p\lambda_j v_{ij}^2$. Therefore $\sum\sum d_{ij}^2=\sum_{i=1}^p d_i^Td_i^*=\sum_{j=r+1}^p\lambda_j^2\sum_{i=1}^p v_{ij}^2$

$=\lambda_{r+1}^2+\cdots+\lambda_p^2$

(c). The result in (b) says $LL^T+\Psi$ is a good approx to S if $\lambda_{r+1}^2+\cdots+\lambda_p^2$ is small. choose r s.t. $\frac{\lambda_1^2+\cdots+\lambda_r^2}{\lambda_1^2+\cdots+\lambda_p^2}\to 1$.

one vs. two factor model

Sometimes the loadings are not easy/obvious to interpret p-value higher ↗ fit better.

LDA vs QDA.

CV (leave one out, 10-fold) → train & test, but when predicting, no CV.

QDA misclassification rate slightly higher. (model for QDA is more complicated than LDA & the variability in

estimating additional parameters may lead to LDA having a higher misclassification rate)

HW4. single-linkage: $\min\{d(A,B), d(A,C)\} = \min\{\min d(a,b), \min d(a,c)\}$
$= \min_{a\in A, b\in B\cup C} d(a,b)$
$= d(A, B\cup C)$

the other equality follows since $\min(x,y) = (x+y)\frac{1}{2} - |x-y|\frac{1}{2}$

complete-linkage is similar.

Show $d(A,C) \le d(A,B) + d(B,C)$ complete-linkage.

disjoint clusters A, B, C.

For a, b, c in A, B, C $d(a,c) \le d(a,b) + d(b,c)$

$\max_{a\in A, b\in B, c\in C} d(a,c) = d(A,C) \le \max_{a,b,c} (d(a,b) + d(b,c))$

likewise $\max_{a,b,c}\{d(a,b)+d(b,c)\} \le \max d(a,b) + \max d(b,c)$
$= d(A,B), d(B,C)$

This won't work for single-linkage.

- If EM cannot separate #k-groups after enough iterations, it is possible that using mixture model is not good. the initial clusters provided by k-means are not clear for EM (local extrema) to isolate the groups.

QDA vs LDA CLASSIFICATIONS

LDA assumes same var-cov. matrices between input variables of classes. The assumption is important for classification stage of the analysis. If the matrices substantially differ, observations will tend to be assigned to the class where variability is greater. To overcome the problem - QDA which is a modification of LDA allowing for the above heterogeneity of classes' cov matrices "more effective", say clusters have extremely different shapes. If similar shape, LDA > QDA b/c QDA has more misclassifications b/c it counts more variabilities.

PCA vs ICA see before.

Additionally. **vs FA** FA, PCA & ICA are all related, three of them seek basis vectors that the data is projected against, such that you maximize [insert-criteria-here] Think of the basis vectors are just encapsulating linear combinations.

Criteria?

① Second-order criteria:

In PCA, you are finding basis vectors that explain best the variance of your data. The first (i.e. highest ranked) basis vector is going to be the one that best fits all the variance from your data. The second one also has this criteria, but must be orthogonal to the first. and so on & so forth. (Turns out these basis vectors for PCA are nothing but the eigenvectors of your data's covariance matrix).

In FA, ~~you are finding~~ there is difference between it & PCA, because FA is generative, whereas PCA is not. I have seen FA as being described as PCA w/noise", where the noise is called "specific factors" All the same, the overall conclusion is that PCA & FA are based on second-order statistics, (covariance) and nothing more.

② Higher order Criteria's.

In ICA, you are again finding basis vectors, but this time, you want basis vectors that ~~are~~ give a result, such that this resulting vector is one of the independent components of the original data. You can do this by maximization of the absolute value of normalized kurtosis — a 4th order statistic. That is, you project your data on the same vector, and measure the kurtosis of the result. You change your basis vector a little (usually through gradient ascent), and then measure the kurtosis again, etc, etc. Eventually you will happen ~~to~~ upon a basis that gives your ~~a~~ result that has the highest possible kurtosis, and this is your indpdt components. ~~The end~~ In graph if the data is like

but ICA would be like ↙ independent

Another version of difference b/w FA & PCA

PCA involves extracting linear composites of observed variables.

Factor analysis is based on a formal model ~~of~~ predicting observed from theoretical latent ~~variables~~ factors.

1. Run FA if you assume or wish to test a theoretical model ~~for~~ of latent factors causing observed variables.
2. Run PCA if you want to simply reduce your correlated obs variables to a smaller set of important independent composite variables.

FA is a generalization of PCA which is based explicitly on ML. Like PCA, each data point is assumed to arise from sampling a pt in a subspace & then perturbing it with full-dimensional Gaussian noise. The difference is that FA allows the noise to have an arbitrary diagonal covariance matrix while PCA assumes the noise is spherical. In addition to estimating the subspace, FA ests the noise covariance matrix.

Cluster analysis vs FA

Short: Cluster analysis is about grouping subjects, FA is about grouping variables.

Long: Clustering is asking how many groups the items can be split into. Factor analysis is asking whether the features of you have are explained as combinations of some smaller set of features. You don't get any groupings out of a factor analysis, and you don't get any dimensionality reduction by doing clustering.

Cluster & Classification

Both split data into groups. With classification the groups (or classes) are specified before hand, with each training data instance belonging to a particular ~~data~~ class. It is the association b/w the instances features & the class they belong to that classification algorithms are supposed to learn. (supervised learning)

With clustering the groups (or clusters) are based on the similarities of data instances

2 is used in training & the clustering algorithm is supposed to learn the grouping itself. (unsupervised learning).

Notes

- Covariance C is non-negative definite $a^TCa \geq 0$ $\forall a$
- conditional distribution
  $X_1|X_2=x_2 \sim N(\mu_1+\rho(\frac{x_2-\mu_2}{\sigma_2})\sigma_1, \sigma_1^2(1-\rho^2))$
- $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, C)$ ① $\underset{\sim}{a}^T\underset{\sim}{X} \sim N(\underset{\sim}{a}^T\underset{\sim}{\mu}, \underset{\sim}{a}^TC\underset{\sim}{a})$
  ② $(\underset{\sim}{X}-\underset{\sim}{\mu})^TC^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu}) \sim \chi^2(p)$
- $\hat{R} \leftrightarrow \hat{C}$. $cor_{ij}=\frac{cov_{ij}}{\sqrt{\sigma_i^2}\sqrt{\sigma_j^2}}$
- Simple visualization of multivariate data (2D-scatterplot)
  pros: - easy, more informative version of a correlation matrix; non-linear pairwise dependencies revealed.
  cons: - miss out on higher dimensional structure.
- projection pursuit: "interesting" = "non-normality"
- 4th order statistic Kurtosis: $K(X)=\frac{E[(X-\mu)^4]}{\sigma^4}$ $K(aX+b)=K(X)$
  find a to maximize/or minimize kurtosis
- In PCA, λ_i's are variances.

e.g of PCA. $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$ n obs, p variables.

$X=\begin{pmatrix}\underset{\sim}{x}_1^T\\ \vdots \\ \underset{\sim}{x}_n^T\end{pmatrix}$ $\xrightarrow[\text{scaled}]{\text{centred}}$ $\tilde{X}$. PCA transforms original data to obtain p uncorrelated variables: SVD: $\tilde{X}=UDV^T \to p\times p$ $U^TU=I=VV^TV$ ($n\times p$, $p\times p$ diag)

$\tilde{X}V=UD=(\underset{\sim}{u}_1, \cdots, \underset{\sim}{u}_p)\begin{pmatrix}d_1 & & 0\\ & \ddots & \\ 0 & & d_p\end{pmatrix}$

$=(d_1\underset{\sim}{u}_1 \cdots d_p\underset{\sim}{u}_p)$, $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$

Covariance matrix

$S=\frac{1}{n-1}\sum_{i=1}^n(\underset{\sim}{x}_i-\bar{\underset{\sim}{x}})(\underset{\sim}{x}_i-\bar{\underset{\sim}{x}})^T=\frac{1}{n-1}\tilde{X}^T\tilde{X}=V\Lambda V^T$ where $\Lambda=\frac{D^2}{n-1}$

first PC maximizes $\underset{\sim}{a}^TS\underset{\sim}{a}$ s.t. $\underset{\sim}{a}^T\underset{\sim}{a}=1 \Rightarrow \underset{\sim}{a}=\underset{\sim}{v}_1$ $V=(\underset{\sim}{v}_1, \cdots, \underset{\sim}{v}_p)$

k-th PC maximizes $\underset{\sim}{a}^TS\underset{\sim}{a}$ s.t. $\underset{\sim}{a}^T\underset{\sim}{v}_j=0$ & $\underset{\sim}{a}^T\underset{\sim}{a}=1$ for $j=1, \cdots, k-1$

PC scores: columns of $\tilde{X}V$; PC loadings columns of $V=(\underset{\sim}{v}_1, \cdots, \underset{\sim}{v}_p)$

Naive PCA model

$\underset{\sim}{X}=\underset{\sim}{\mu}+L\underset{\sim}{F}$ (with noise)

loadings, factors

$\underset{\sim}{F} \sim N_r(\underset{\sim}{0}, I)$

Factor analysis model

$\underset{\sim}{X}=\underset{\sim}{\mu}+L\underset{\sim}{F}+\underset{\sim}{\varepsilon}$

where ① $\underset{\sim}{F} \sim N_r(\underset{\sim}{0}, I)$ ② $\underset{\sim}{\varepsilon} \sim N_p(\underset{\sim}{0}, \Psi)$ where $\Psi=\begin{pmatrix}\psi_1 & & 0\\ & \ddots & \\ 0 & & \psi_p\end{pmatrix}$

③ Components of $\underset{\sim}{\varepsilon}$ are independent components of $\underset{\sim}{F}$

Note: Can also specify model in terms of $Cov(\underset{\sim}{F})$, $Cov(\underset{\sim}{\varepsilon})$ and $\underset{\sim}{\varepsilon}$, $\underset{\sim}{F}$ uncorrelated.

$\Rightarrow \underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, LL^T+\Psi)$

why factor analysis?

① Unobserved factors explain relationships b/w variables

② Factor analysis is invariant (equiv) to changes in scale for each variable.

③ Relationship to graphical models:

for one-factor model: link exists b/w F & X_j if $l_j \neq 0$.

for multi-factor model: X_i, X_j are conditional independent given $F_1 \cdots$ L (loading is not uniquely determined)

Both a feature & a bug.

Generally. $\underset{\sim}{X}=\underset{\sim}{\mu}+L\underset{\sim}{F}+\underset{\sim}{\varepsilon}$. $Cov(\underset{\sim}{F})=I$, $Cov(\underset{\sim}{\varepsilon})=\Psi=\begin{pmatrix}\psi_1 & & 0\\ & \ddots & \\ 0 & & \psi_p\end{pmatrix}$

leng(p), $p\times r$, leng($r\times p$)

$Cov(\underset{\sim}{X})=LL^T+\Psi$

$Cov\begin{pmatrix}\underset{\sim}{X}\\ \underset{\sim}{F}\end{pmatrix}=\begin{pmatrix}\Psi+LL^T & L\\ L^T & I\end{pmatrix}$

$S=\frac{1}{n-1}\sum_{i=1}^n(\underset{\sim}{x}_i-\bar{\underset{\sim}{x}})(\underset{\sim}{x}_i-\bar{\underset{\sim}{x}})^T \doteq LL^T+\Psi$

$\hat{\Psi}=diag(S-LL^T)$ (but assuming $S-LL^T$ diag are positive)

$LL^T \doteq S-\Psi=V\Lambda V^T=(\underset{\sim}{v}_1 \cdots \underset{\sim}{v}_r)\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_r\end{pmatrix}\begin{pmatrix}\underset{\sim}{v}_1^T\\ \vdots \\ \underset{\sim}{v}_r^T\end{pmatrix}$

$=(\sqrt{\lambda_1}\underset{\sim}{v}_1 \cdots \sqrt{\lambda_r}\underset{\sim}{v}_r)\begin{pmatrix}\sqrt{\lambda_1}\underset{\sim}{v}_1^T\\ \vdots \\ \sqrt{\lambda_r}\underset{\sim}{v}_r^T\end{pmatrix}$

Principal factor Method

Algorithm:

① set $\hat{\Psi}=0$ ① set $S-\hat{\Psi}=V\Lambda V^T$ & $\hat{L}=(\sqrt{\lambda_1}\underset{\sim}{v}_1 \cdots \sqrt{\lambda_r}\underset{\sim}{v}_r)$

② set $\hat{\Psi}=diag(S-\hat{L}\hat{L}^T)$ ③ Iterate ① ② till convergence

Hierarchical clustering methods

Start with n clusters each with 1 observation. Group together similar obs.

① single / ② complete ③ average linkage

General algorithm:

① ↗

② Search Distance to find nearest pair of clusters (depends on approach) call these U + V

③ Merge U + V into a single cluster and redefine D (linkage)

④ Repeat ② & ③ until we have a single cluster

- Dendogram
- k-means clustering, EM:
- $f(\underset{\sim}{x})=\lambda_1 f_1(\underset{\sim}{x}, \underset{\sim}{\theta}_1)+\cdots+\lambda_k f_k(\underset{\sim}{x}, \underset{\sim}{\theta}_k)$ mixture of multivariate normals.

$f_j(\underset{\sim}{x}, \underset{\sim}{\mu}_j, C_j)=\frac{1}{(2\pi)^{\frac{p}{2}}|C_j|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_j)^TC_j^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_j))$

- Classification

Data problem: given $(g_1, \underset{\sim}{x}_1), \cdots, (g_n, \underset{\sim}{x}_n)$ as training data, determine a classification rule with small error rate.

Say, R_k are boundaries of each class.

error rate $=P(\text{error})=\sum_{j=1}^k P(G=j)P(\text{error}|G=j)$

$=\sum_{j=1}^k \lambda_j P(\underset{\sim}{X} \notin R_j | G=j)=\sum_{j=1}^k \lambda_j\{1-\int_{R_j} f_j(\underset{\sim}{x})d\underset{\sim}{x}\}$

$=1-\sum_{j=1}^k \lambda_j \int_{R_j} f_j(\underset{\sim}{x})d\underset{\sim}{x}$ → correct classification rate

Optimal classification ~~rule~~: (Baye's ~~Rule~~) classifier

~~Xing~~ Classify $\underset{\sim}{x}$ to group j if

$\forall i \neq j$ $\lambda_j f_j(\underset{\sim}{x}) > \lambda_i f_i(\underset{\sim}{x})$

$\frac{f_j(\underset{\sim}{x})}{f_i(\underset{\sim}{x})} > \frac{\lambda_i}{\lambda_j}$ $\forall i \neq j$

e.g. example from prac-final:

$\int_0^1 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{\pi} r\, d\theta\, dr = \int_0^1 \frac{\pi}{\pi} r\, dr = \frac{1}{2}$?

- naive error rate: $\frac{1}{n}\sum_{i=1}^{n} I(g_i \neq \hat{g}_i)$
- resubstitution error rate ↓ biased downside.
- fix: CV (leave one-out), estimate classifier from remaining obs. & look at error rate for obs. not used to est. the classifier,
- Two approaches (Optimal classification)

① Minimize P(error).

$\lambda_j f_j(\underset{\sim}{x}) > \lambda_i f_i(\underset{\sim}{x})$

② Look at conditional dist'n of G given $\underset{\sim}{X} = \underset{\sim}{x}$

Bayes Thm $P(G=j \mid \underset{\sim}{X}=\underset{\sim}{x}) = \frac{P(G=j, \underset{\sim}{X}=\underset{\sim}{x})}{P(\underset{\sim}{X}=\underset{\sim}{x})}$

$= \frac{P(G=j)P(\underset{\sim}{X}=\underset{\sim}{x} \mid G=j)}{P(\underset{\sim}{X}=\underset{\sim}{x})} = \frac{\lambda_j f_j(\underset{\sim}{x})}{\lambda_1 f_1(\underset{\sim}{x}) + \cdots + \lambda_K f_K(\underset{\sim}{x})}$

max this value. (same as ①?)

But more information we get)