Displays for Statistics 5401/8401

Lecture 26

November 7, 2005

Christopher Bingham, Instructor

612-625-1024, kb@umn.edu
372 Ford Hall

Class Web Page
http://www.stat.umn.edu/~kb/classes/5401

SVD: $\tilde{\mathbf{X}} = \mathbf{LTR}'$, $\mathbf{L} = [\mathbf{L}_1 \ldots \mathbf{L}_p]$, $\mathbf{R} = [\mathbf{r}_1 \ldots \mathbf{r}_p]$, $\mathbf{T} = \text{diag}[t_1,\ldots,t_p]$, $t_j = f_e\hat{\lambda}_j$, $\hat{\lambda}_j$ and $\mathbf{r}_j$ eigenvalue and eigenvector of $\mathbf{S}$.

Using the best rank m approximation to $\tilde{\mathbf{X}}$ $\tilde{\mathbf{X}}^{(m)} = \sum_{1\le j\le m}\tilde{\mathbf{Z}}_j\mathbf{r}_j'$, you can decompose $\mathbf{X}$ as

$$\begin{aligned}\mathbf{X} &= \mathbf{1}_N\overline{\mathbf{x}}' + \tilde{\mathbf{X}} = (\text{mean} + \text{residuals})\\ &= \mathbf{1}_N\overline{\mathbf{x}}' + \sum_{1\le j\le m}\tilde{\mathbf{Z}}_j\mathbf{r}_j' + \left(\sum_{m+1\le j\le p}\tilde{\mathbf{Z}}_j\mathbf{r}_j'\right)\\ &= \mathbf{1}_N\overline{\mathbf{x}}' + \tilde{\mathbf{X}}^{(m)} + \mathbf{e},\end{aligned}$$

where

$$\mathbf{e} = [e_{i\ell}] \equiv \tilde{\mathbf{X}} - \sum_{1\le j\le m}\mathbf{Z}_j\mathbf{r}_j' = \sum_{m+1\le j\le p}\tilde{\mathbf{Z}}_j\mathbf{r}_j'$$

is the "error" when $\tilde{\mathbf{X}}$ is approximated by $\tilde{\mathbf{X}}^{(m)}$. The SS of all the errors $e_{i\ell}$ is

$$\sum_{1\le i\le n}\sum_{1\le \ell\le p}e_{i\ell}^2 = \sum_{m+1\le j\le p}t_j^2 = f_e\left(\sum_{m+1\le j\le p}\hat{\lambda}_j\right)$$

The value of variable ℓ for case i is

$$x_{i\ell} = \overline{x}_\ell + \sum_{1\le j\le m}\tilde{z}_{ij}r_{\ell j} + e_{i\ell}$$

This looks sort of like a multiple regression on the $\tilde{z}_{hj}$'s.

When

$$(\sum_{m+1 \le j \le p} \hat{\lambda}_j)/(\sum_{1 \le j \le p} \hat{\lambda}_j) =$$
$$(\sum_{m+1 \le j \le p} \hat{\lambda}_j)/(\sum_{1 \le j \le p} s_{jj}) \approx 0,$$

you can "explain" a large part of the variability in $\mathbf{X}$ in terms of the principal components.

Put this way, the focus is on explaining variability rather than approximating data.

This brings us to the more usual way to define principal components.

This sees PCA (principal components analysis) as a technique to understand

- the *structure* of a variance matrix $\mathbf{\Sigma}$ as estimated by $\mathbf{S}$, or
- the *structure* of a correlation matrix $\mathbf{\rho} = [\rho_{ij}]$ as estimated by $\mathbf{R} = [r_{ij}]$.

This makes sense only in a context where the rows $\mathbf{x}_i'$ of $\mathbf{X}$ are a random sample from a population.

- This does not make sense when doing PCA of the Fisher data matrix which consists of three random samples.
- It *might* make sense when doing PCA starting with MANOVA residuals, provided the three varieties all had the same $\mathbf{\Sigma}$, which is probably not the case.

Population Principal Components

Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ be the eigenvectors of $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

The idea starts with finding a linear combination $\mathbf{v}'\mathbf{x}$, with $\|\mathbf{v}\|^2 = \sum v_i^2 = 1$, which has the largest possible variance $\sigma_v^2 = V[\mathbf{v}'\mathbf{x}] = \mathbf{v}'\mathbf{\Sigma}\mathbf{v}$.

The solution is $\mathbf{v} = \mathbf{v}_1$, the first eigen-vector of $\mathbf{\Sigma}$.

The variance of $z_1 = \mathbf{v}_1'\mathbf{x}$ is $\sigma_{v_1}^2 = \lambda_1$.

z_1 is the first *population principal component*.

- z_1 has mean $\mathbf{v}_1'\mu$
- z_1 has variance $\sigma_{z_1}^2 = \mathbf{v}_1'\mathbf{\Sigma}\mathbf{v}_1 = \lambda_1$.

If you measure "what is going on" in terms of variance, then z_1 is the linear combination that has the most "going on".

Next, you seek $\mathbf{v}$ with $\|\mathbf{v}\|^2 = 1$ so that the linear combination $z_2 = \mathbf{v}'\mathbf{x}$ has the largest variance under the restriction that z_2 is *uncorrelated* with z_1 (that is, z_2 is "new information" not included in z_1).

The solution is $z_2 = \mathbf{v}_2'\mathbf{x}$, the 2nd popul-ation principal component, with variance $\sigma_{z_2}^2 = \mathbf{v}_2'\mathbf{\Sigma}\mathbf{v}_2 = \lambda_2$.

You get the remaining *population prin-cipal components*, $z_j = \mathbf{v}_j\mathbf{x}$, $j = 3, \ldots, p$ by seeking $\mathbf{v}$ with $\|\mathbf{v}\|^2 = 1$ so that z_j is uncorrelated with previous PC's and has the largest variance. $\sigma_{z_j}^2 = \mathbf{v}_j'\mathbf{\Sigma}\mathbf{v}_j = \lambda_j$.

Except when rank(Σ) < p, *all* the z_i's will be "non trivial", that is

$$\sigma_{z_j}^2 = \lambda_j > 0, \; j = 1, \ldots, p.$$

So the population number of principal components is *always* p.

It is *meaningless* to think of the choice of how many principal components to use as an estimation problem.

It is equally meaningless to seek a test of H_0: number of PC's = q.

Q Why does the question as to how many PCs to use even arise?

A Because, properly used, ignoring all but a few PCs *may* lose very little information. When this is the case and p is large and m << p, you can achieve a very important ***dimension reduction*** while losing little information.

Summary of Properties of PCs

- $V[z_k] = \lambda_k$, $Cov[z_j, z_k] = 0$, $j \neq k$

  That is, if $\mathbf{z} = [z_1, z_2, \ldots, z_p]'$

  $$V[\mathbf{z}] = \Lambda = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_p]$$

Describing and approximating Σ

$\Sigma = \sum_{1 \le j \le p} \lambda_j \mathbf{v}_j \mathbf{v}_j'$ can be split as

$$\Sigma = \underbrace{\sum_{1 \le j \le m} \lambda_j \mathbf{v}_j \mathbf{v}_j'}_{\text{rank m}} + \underbrace{\sum_{m+1 \le j \le p} \lambda_j \mathbf{v}_j \mathbf{v}_j'}_{\text{remainder}}$$

When

$$\sum_{1 \le j \le m} \lambda_j / \sum_{1 \le j \le p} \lambda_j = \sum_{1 \le j \le m} \lambda_j / \sum_{1 \le j \le p} \sigma_{jj} \approx 1$$

or

$$\sum_{m+1 \le j \le p} \lambda_j / \sum_{1 \le j \le p} \lambda_j = \sum_{m+1 \le j \le p} \lambda_j / \sum_{1 \le j \le p} \sigma_{jj} \approx 0,$$

the remainder

$$\Sigma - \sum_{1 \le j \le m} \lambda_j \mathbf{v}_j \mathbf{v}_j' = \sum_{m+1 \le j \le p} \lambda_j \mathbf{v}_j \mathbf{v}_j'$$

will be small relative to Σ.

When

$$\sum_{m+1\le j\le p}\lambda_j/\sum_{1\le j\le p}\lambda_j = \sum_{m+1\le j\le p}\lambda_j/\sum_{1\le j\le p}\sigma_{jj} \approx 0,$$

$$\boldsymbol{\Sigma} - \sum_{1\le j\le m}\lambda_j\mathbf{v}_j\mathbf{v}_j' = \sum_{m+1\le j\le p}\lambda_j\mathbf{v}_j\mathbf{v}_j'$$

will be small relative to $\boldsymbol{\Sigma}$ and

$$\boldsymbol{\Sigma}^{(m)} \equiv \sum_{1\le j\le m}\lambda_j\mathbf{v}_j\mathbf{v}_j' \cong \boldsymbol{\Sigma}$$

$\boldsymbol{\Sigma}^{(m)}$ is a rank m approximation to $\boldsymbol{\Sigma}$ and you might say that $\boldsymbol{\Sigma}$ "almost has rank m."

It turns out that both $\boldsymbol{\Sigma}^{(m)}$ and $\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(m)}$ are variance matrices of random vectors.

Since $\lambda_j = V[\mathbf{v}_j'\mathbf{x}]$,

$$\boldsymbol{\Sigma}^{(m)} \equiv \sum_{1\le j\le m}\lambda_j\mathbf{v}_j\mathbf{v}_j' = V[\sum_{1\le j\le m}\mathbf{v}_j(\mathbf{v}_j'\mathbf{x})]$$

$$= V[\sum_{1\le j\le m}\mathbf{v}_j z_j] = V[\mathbf{x}^{(m)}],$$

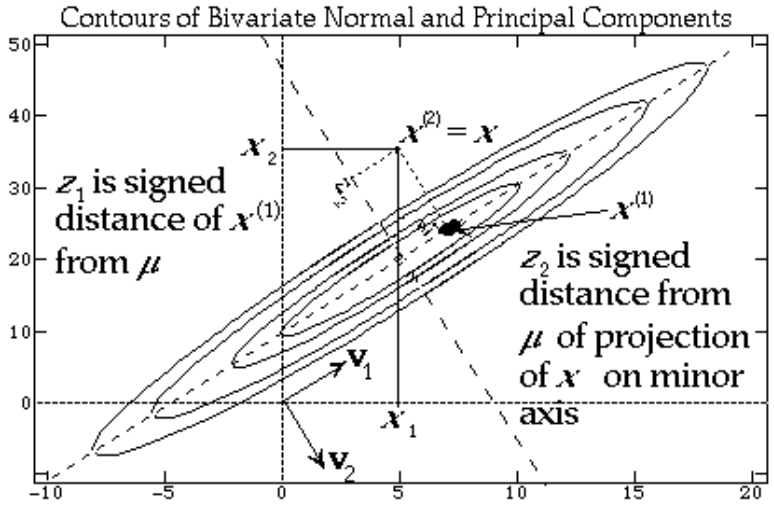
where $\mathbf{x}^{(m)} \equiv \sum_{1\le j\le m} z_j\mathbf{v}_j$ is the part of $\mathbf{x}$ where things are "going on."

In this, $z_j = \mathbf{v}_j'\mathbf{x}$ is the j^{th} population principal component, a random variable.

$\mathbf{x}^{(m)}$ is a sum of multiples of the m <u>fixed</u> eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$, with <u>random</u> multipliers $z_1, \ldots, z_m$.

This means $\mathbf{x}^{(m)}$ lies on an m-dimensional "plane" in p dimensional space. $z_1, \ldots, z_m$ specify its location in that plane.

Here is a picture to illustrate what's happening in the p = 2 case.



The ellipses are contours of the $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density with $\boldsymbol{\mu} = [5, 20]'$ when $\boldsymbol{\Sigma}$ has eigenvalues

- $\lambda_1 = 100 = 10^2$
- $\lambda_2 = 1 = 1^2$.

$\lambda_1/(\lambda_1+\lambda_2) = 0.9901$, $\lambda_2/(\lambda_1+\lambda_2) = 0.0099$.

$\mathbf{x}^{(1)}$ (the part of $\mathbf{x}$ with largest variance) is the perpendicular projection of $\mathbf{x}$ on the major axis.

How about the part of $\mathbf{x}$ in which nothing important is "going on", that is $\mathbf{x} - \mathbf{x}^{(m)}$?

$$\begin{aligned} V[\mathbf{x} - \mathbf{x}^{(m)}] &= V[\textstyle\sum_{m+1\le j\le p} z_j \mathbf{v}_j] \\ &= \textstyle\sum_{m+1\le j\le p} \lambda_j \mathbf{v}_j \mathbf{v}_j' \\ &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(m)} \end{aligned}$$

Thus the variances and covariances of the part of $\mathbf{x}$ not determined by the first m principal components come from the part of $\boldsymbol{\Sigma}$ not fit by the best rank m approximation to $\boldsymbol{\Sigma}$.

When $\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(m)}$ is small compared to $\boldsymbol{\Sigma}$, $\mathbf{x} - \mathbf{x}^{(m)}$ will be small confirming that $\mathbf{x}^{(m)}$ contains what's "going on"

The ***sample principal components*** $\hat{z}_j = \hat{\mathbf{v}}_j'\mathbf{x}$ computed using the eigenvectors $\hat{\mathbf{v}}_j = \mathbf{r}_j$ of the sample variance matrix $\mathbf{S}$ maximize the *sample* variances of linear combinations of the variables.

When $\sum_{m+1\le j\le p}\hat{\lambda}_j / \sum_{1\le j\le p} s_{jj}$ is small, the first m principal components have captured most of "what's going on" in the *sample*, and $\hat{\mathbf{\Sigma}}^{(m)} = \sum_{1\le j\le m}\hat{\lambda}_j \mathbf{r}_j\mathbf{r}_j' = \sum_{1\le j\le m}\hat{\lambda}_j \hat{\mathbf{v}}_j\hat{\mathbf{v}}_j'$ is a good *rank m approximation* to $\hat{\mathbf{\Sigma}}^{(p)} = \mathbf{S}$, in the sense that $\mathbf{S} - \hat{\mathbf{\Sigma}}^{(m)} = \sum_{m+1\le j\le p}\hat{\lambda}_j \mathbf{r}_j\mathbf{r}_j'$ is small relative to $\mathbf{S}$.

To emphasize again, the vectors $\hat{\mathbf{Z}}_j = \mathbf{X}\hat{\mathbf{v}}_j$ of PC values are closely related to the left singular vectors $\mathbf{L}_j$ of $\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}_N\overline{\mathbf{x}}' = \mathbf{LTR}'$. Specifically

$$\hat{\mathbf{Z}}_j = \mathbf{1}_N\overline{\mathbf{x}}' + t_j\mathbf{L}_j = \mathbf{1}_N\overline{\mathbf{x}}' + \sqrt{\{(N-1)\hat{\lambda}\}}\times\mathbf{L}_j$$

Correlation principal components

Because PCs are highly *dependent on scale*, in practice PC's are often computed from standardized data. This amounts to using *eigenvectors of the correlation matrix* as PC coefficients.

For example, when $s_{kk} >> s_{jj}$, $j \ne k$, $z_1 \tilde{=} C(x_k - \overline{x_k})$ and almost entirely reflect the behavior of x_k. This won't happen be the case if you the variables so their variances are the same or similar, and in particular not when you standardize all the variables.

Comment

It is misleading to say that correlation PC's maximize the "variance" of linear combinations, although this is often said.

I don't find the variance of linear combinations of standardized variables to be very interesting.

Population ***correlation*** **principal components** are based on the eigenvectors and eigenvalues of the population *correlation* matrix

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{13} & \rho_{23} & 1 & \cdots & \rho_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{1p} & \rho_{2p} & \rho_{3p} & \cdots & 1 \end{bmatrix}, \quad \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

with eigenvalues $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_p$ and eigen vectors $\mathbf{e}_1, \ldots, \mathbf{e}_p$.

$$\sum_{1 \leq i \leq p} \tau_j = \text{trace}(\boldsymbol{\rho}) = p$$

Now, when $\sum_{m+1 \leq i \leq p} \tau_j / \sum_{1 \leq i \leq p} \tau_j = \sum_{m+1 \leq i \leq p} \tau_j / p$ is small, $\boldsymbol{\rho}^{(m)} \equiv \sum_{1 \leq k \leq m} \tau_k \mathbf{e}_k \mathbf{e}_k'$ is a rank m approximation to $\boldsymbol{\rho}$.

$\boldsymbol{\rho}^{(m)}$ is *not* a correlation matrix since

$$\sum_\ell \rho_{\ell\ell}^{(m)} = \text{trace}(\boldsymbol{\rho}^{(m)}) = \sum_{1 \leq k \leq m} \tau_k < p$$

so at least one $\rho_{\ell\ell}^{(m)} < 1$.

Recall that $\boldsymbol{\rho}$ is the variance matrix of standardized vector:

$$\boldsymbol{\rho} = V[\mathbf{x}^*], \quad \mathbf{x}^* = [x_1^*, x_2^*, \ldots, x_p^*]'$$

where the x_ℓ^*'s are z-scores:

$$x_\ell^* = (x_\ell - \mu_\ell)/\sqrt{\sigma_{\ell\ell}}$$

"Unstandardizing",

$$x_\ell = \mu_\ell + \sqrt{\sigma_{\ell\ell}} x_\ell^*$$

In vector notation,

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Delta}^{-1}\mathbf{x}^*,$$

$$\boldsymbol{\Delta} = \text{diag}[1/\sqrt{\sigma_{11}}, \ldots, 1/\sqrt{\sigma_{pp}}]$$

The population *correlation* principal components are the linear combinations of $\mathbf{x}^*$

$$z_j = \mathbf{e}_j'\mathbf{x}^* = \mathbf{e}_j'\boldsymbol{\Delta}(\mathbf{x} - \boldsymbol{\mu}) = (\boldsymbol{\Delta}\mathbf{e}_j)'(\mathbf{x} - \boldsymbol{\mu}).$$

$\boldsymbol{\Delta}\mathbf{e}_j = [e_{1j}/\sqrt{\sigma_{11}}, \ldots, e_{pj}/\sqrt{\sigma_{pp}}]'$ is the vector of coefficients *in a form to multiply* x_j or $(x_j - \mu_j)$, not x_j^*.

You can approximate the standardized random vector $\mathbf{x}^*$ by

$$\mathbf{x}^{(m)*} \equiv \sum_{1 \le j \le m} z_j \mathbf{e}_j, \; z_j = \text{correlation PC}$$

You can approximate $\mathbf{x}$ itself using correlation principal components by unstandardizing $\mathbf{x}^{(m)*}$:

$$\mathbf{x}^{(m)\dagger} \equiv \boldsymbol{\mu} + \boldsymbol{\Delta}^{-1}\mathbf{x}^{(m)*} = \boldsymbol{\mu} + \boldsymbol{\Delta}^{-1}\sum_{1 \le j \le m} z_j \mathbf{e}_j$$
$$= \boldsymbol{\mu} + \sum_{1 \le j \le m} z_j (\boldsymbol{\Delta}^{-1}\mathbf{e}_j)$$

The "error" is $\mathbf{x} - \mathbf{x}^{(m)\dagger} = \sum_{m+1 \le j \le p} z_j (\boldsymbol{\Delta}^{-1}\mathbf{e}_j)$.

In a similar way as before, when

$$\sum_{1 \le k \le m} \tau_k \mathbf{e}_k \mathbf{e}_k' = \boldsymbol{\rho}^{(m)} = V[\mathbf{x}^{(m)*}] \tilde{=} \boldsymbol{\rho} = V[\mathbf{x}^*]$$

is a "good" approximation to $\boldsymbol{\rho}$, you can interpret it as "explaining" the correlations among $x_1, x_2, \ldots, x_p$ by their dependence on a smaller number m of "factors" $z_1, \ldots, z_m$ they have "in common".

Principal components are often used in place a larger number of variables.

They are usually viewed as derived variables which contain *most* of the "information" in the original variables.

1. **New response variables** $z_1, z_2, \ldots, z_m$. that might be used as the basis of a classification algorithm.
2. **Variables used as input to other procedures** You might do cluster analysis starting with $z_1, z_2, \ldots, z_m$, that is attempt to find separated groups of cases.
3. **New predictor variables** Use $z_1, z_2, \ldots, z_m$ enable you to model other response variables $\mathbf{Y}$ as $\mathbf{Y} = \hat{\mathbf{Z}}\mathbf{B}^* + \boldsymbol{\varepsilon}$ instead of $\mathbf{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{\varepsilon}$.
   Note: When $m = p$, $\hat{\mathbf{Z}}\mathbf{B}^* = \mathbf{X}\mathbf{B}$

A potential serious *problem* with any such use is that there is no guarantee that the information you *want* or need is actually in the first few PC's.

Here's a "toy" example: Suppose you want to predict y from x_1 and x_2 where the true relationship is

$$y = 50 + 3(x_1 - x_2) + \varepsilon$$

where x_1 and x_2 have high positive correlation so that $\mathrm{Var}[x_1 - x_2] << \mathrm{Var}[x_1 + x_2]$. To try to reduce the predictors from 2 to 1 ($x_1 - x_2$ is the correct choice), you might compute principal components based on x_1 and x_2 and use the "important" PC_1 as predictor, omitting PC_2.

But, when $\mathrm{Var}[x_1] \tilde{=} \mathrm{Var}[x_2]$, $z_1 = PC_1 \tilde{=} (x_1 + x_2)/\sqrt{2}$ and $z_2 = PC_2 \tilde{=} (x_1 - x_2)/\sqrt{2}$. To discard PC_2 is to lose what you need.

Inference for principal components

Not much is known except when the population is MVN. In that case, the exact distributions of eigenvalues and vectors of a sample variance matrix **S** are complicated and hard to compute.

The *large sample* normal results are fairly easy to understand.

Sampling distribution of eigenvalues

Fact: In large samples, when $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$, the sample eigenvalues $\{\hat{\lambda}_j\}$ are approximately independent $N(\lambda_j, 2\lambda_j^2/n)$ for large n.

This is correct only for "isolated" λ's. It is wrong for any $\lambda_j = \lambda_k$, $k \neq j$.

Except for large f_e, this normal approximation is not good because the distribution of $\hat{\lambda}_j$ is quite skewed.

A widely used <u>non-symmetric</u> approximate distribution of a random variable $T > 0$ is to treat it as a multiple of χ^2, specifically as $T \tilde{=} \mu_T \times \chi^2_{edf}/edf$, where

edf = **effective degrees of freedom** $\equiv 2E[T]^2/V[T]$.

When T is exactly $\mu_T \times \chi^2_f/f$, then edf = f. If not, then T has the same mean and variance as $\mu_T \times \chi^2_{edf}/edf$.

The large sample edf of $\hat{\lambda}_i$ is

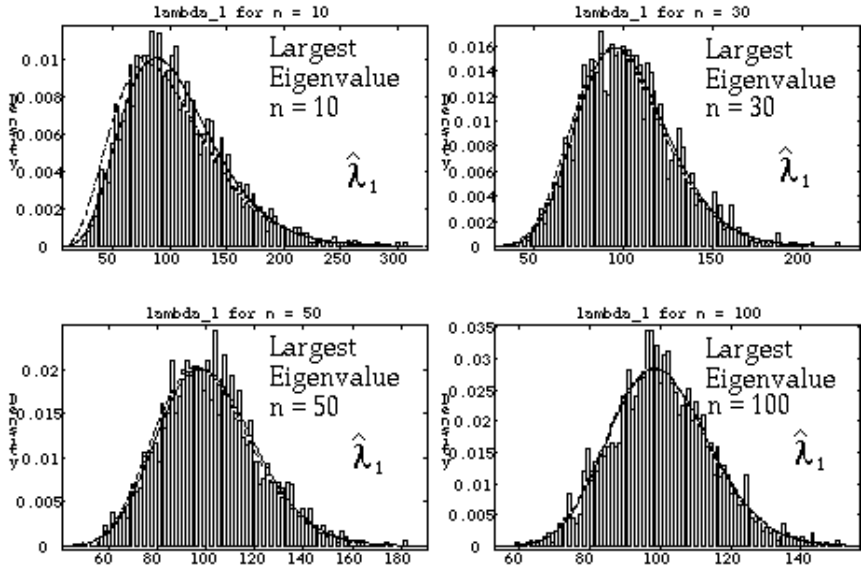
$$edf = 2\lambda_i^2/(2\lambda_i^2/n) = n$$

I made plots of simulated distributions of $\hat{\lambda}_j$ along with with the densities of $\bar{\hat{\lambda}}_j \chi^2_{edf}/edf$, $edf = 2\bar{\hat{\lambda}}_i^2/s^2_{\hat{\lambda}_1}$ (solid curves) and $\lambda_j \chi^2_n/n$ (dashed curves).

The simulations were for p = 3 and

$$\lambda_1 = 100 > \lambda_2 = 30 > \lambda_3 = 10$$

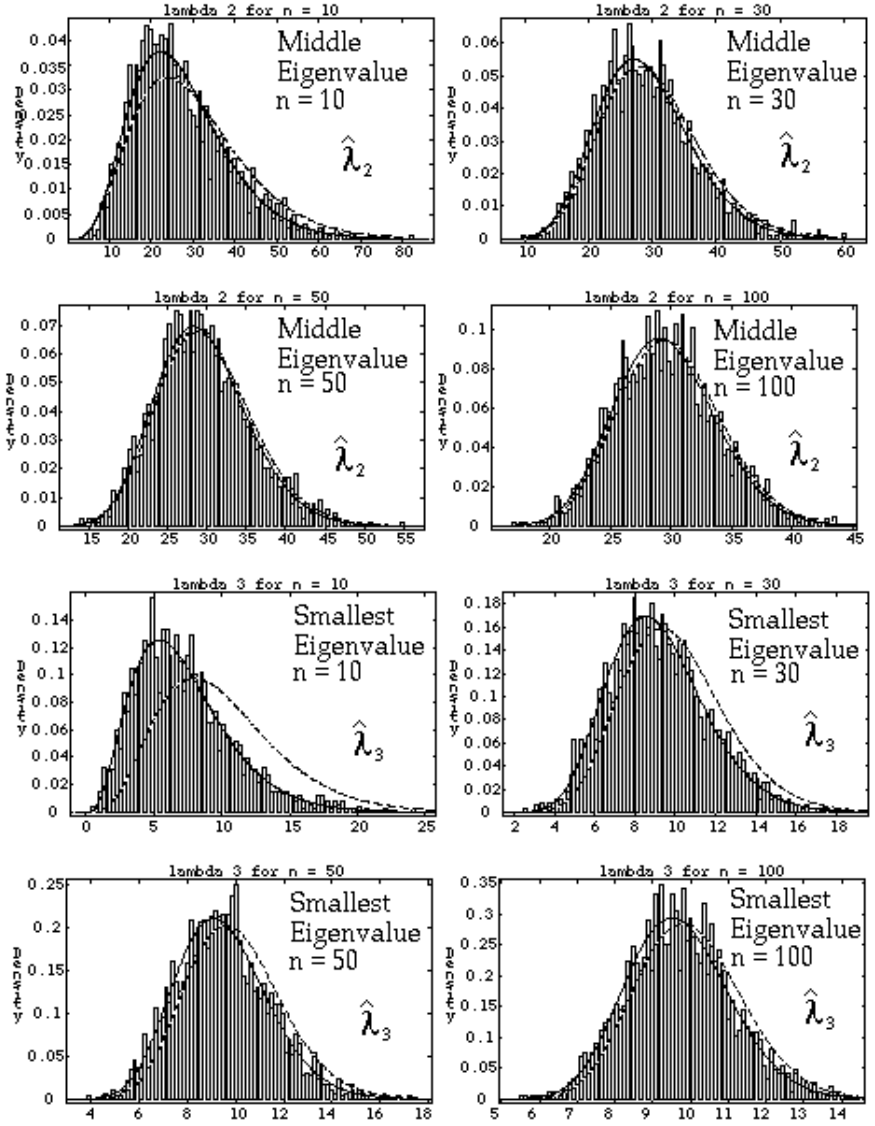
χ^2 approximations for $\hat{\lambda}_1$, the largest eigenvalue.



The solid line is the distribution using $edf = 2(\bar{\hat{\lambda}}_1)^2/s^2_{\hat{\lambda}_1}$ and $E[\hat{\lambda}_i] = \bar{\hat{\lambda}}_1$.

The dashed is the distribution using edf = n and $E[\hat{\lambda}_i] = \lambda_i$. For n = 100 they are almost indistinguishable and appear to fit the histogram well

Middle and smallest eigenvalues



Other simulations confirmed well separated eigenvalues were almost uncorrelated.

These results could in principle allow you test hypotheses about the variances of the principal components and to compute approximate confidence intervals for them.

They do *not* allow you to test hypotheses like H_0: $\lambda_2 = \lambda_3$ since these large sample results are valid only when the eigenvalues are distinct and well separated.

It is not clear to me in what circumstances I would be interested in testing hypotheses about the λ's or finding confidence intervals for them.

Sampling distribution of eigenvectors
(not presented in class)

When you use the correct coordinates, the asymptotic distribution of the eigenvectors is simpler than it appears in the text.

The "right" coordinate system has the true eigenvectors $\mathbf{v}_i$ as coordinate axes.

In this coordinate system, the coordinates of $\hat{\mathbf{v}}_j$ are

$$\tilde{v}_{ij} \equiv \mathbf{v}_i' \hat{\mathbf{v}}_j, \; i = 1,\ldots,p$$

For large n,

- $\tilde{v}_{jj} = \mathbf{v}_j' \hat{\mathbf{v}}_j = 1 + O(1/n)$ (essentially constant)
- $\{\tilde{v}_{ij}\}_{i \neq j}$ are asymptotically $N_{p-1}(\mathbf{0}, n^{-1}\mathbf{D}_j)$, where
  $\mathbf{D}_j = \mathrm{Diag}[\delta_{1j}, \delta_{2j}, \ldots, \delta_{j-1,j}, \delta_{j+1,j} \ldots, \delta_{pk}]$
  $\delta_{ij} \equiv \lambda_i\lambda_j/(\lambda_i-\lambda_j)^2$.

When $\lambda_i = \lambda_j$ this can't be valid.

Here's how you can think of this geometrically.

Since $\mathbf{v}_j$ and $\hat{\mathbf{v}}_j$ are unit vectors ($\|\mathbf{v}_j\| = \|\hat{\mathbf{v}}_j\| = 1$), they can be considered to be points on a "sphere" with radius 1 in p-dimensional space and the sampling distribution of $\hat{\mathbf{v}}_j$ is a distribution on the sphere centered at $\mathbf{v}_j$.

For large n, it is concentrated near $\mathbf{v}_j$.

Close to $\mathbf{v}_j$ (or any other point on the sphere), the surface of a sphere is almost a flat p-1 dimensional plane. The result essentially says $\hat{\mathbf{v}}_j$ is $MVN_{p-1}(\hat{\mathbf{v}}_j,(1/n)\mathbf{\Sigma}_j)$ on the surface of the sphere treated as if it were flat, where $\mathbf{\Sigma}_j$ has eigenvectors pointing in the direction of $\mathbf{v}_i$, $i \neq j$, and eigenvalues $\delta_{ij} = \lambda_i\lambda_j/(\lambda_i-\lambda_j)^2$.

This implies you can find a large sample approximate $1 - \alpha$ confidence region for $\mathbf{v}_j$ as an ellipsoid on the sphere centered at $\hat{\mathbf{v}}_j$ with principal axis pointing in the directions of $\hat{\mathbf{v}}_i$, $i \neq j$, and lengths

$$\sqrt{\{\chi_{p-1}^2(\alpha) \times \delta_{ij}\}} / \sqrt{n}$$

Since the sample eigenvectors are the coefficients of the x_i' in the principal components, you can also use these results to test that one or more of the coefficients are 0.

Again, I have had little reason ever to think this was something I wanted to do.