Displays for Statistics 5401/8401

Lecture 24

November 2, 2005

Christopher Bingham, Instructor

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

Class Web Page

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

Principal components

Principal components are specific linear combinations $z_j \equiv \mathbf{v}_j'\mathbf{x} = \sum_{1 \le \ell \le p} v_{\ell j} x_\ell$ of variables $x_1, ..., x_p$. The vectors $\mathbf{v}_j = [v_{1j}, ..., v_{pj}]'$, $1 \le j \le p$, of coefficients are chosen to have certain properties.

There are at least two ways to motivate principal components.

- Principal components are linear combinations $\mathbf{v}_j'\mathbf{x}$ of variables which have the largest variances subject to constraints on the coefficients $v_{\ell j}$:

  $$\|\mathbf{v}_j\|^2 = \mathbf{v}_j'\mathbf{v}_j = \sum_\ell v_{\ell j}^2 = 1 \text{ (normalized)}$$
  $$\mathbf{v}_j'\mathbf{v}_k = \sum_\ell v_{\ell j} v_{\ell k} = 0,\ j \ne k \text{ (orthogonal)}$$

- Principal components are linear combinations from which you can approximately reconstruct a data matrix $\mathbf{X}$ using a least squares criterion.

The two approaches agree in an important way.

Both view the first few principal components as a set, preferably small, of *new* variables z_1, z_2, ..., z_m, linearly related to x_1, x_2, ..., x_p, and which lose as little as possible "important information" in the complete data.

- The variance maximization approach equates "important information" with high variability.
- The data matrix approximation approach equates "important information" with having an approximation with small errors.

I focus on their use in *approximating* a N by p data matrix $\mathbf{X}$, because I think that is closer to the way principal components are usually used.

In many cases, the single most important description or summary of a data matrix $\mathbf{X}$ is its sample mean vector $\overline{\mathbf{x}} = \sum_i \mathbf{x}_i/N$.

I think this is because the N by p matrix

$$\mathbf{1}_N\overline{\mathbf{x}}' = \begin{bmatrix} \overline{\mathbf{x}}' \\ \overline{\mathbf{x}}' \\ \ldots \\ \overline{\mathbf{x}}' \end{bmatrix}$$

often "explains" or predicts $\mathbf{X}$ well in the following sense:

Elements of $\tilde{\mathbf{X}} \equiv \mathbf{X} - \mathbf{1}_N\overline{\mathbf{x}}'$ are often much smaller than the elements of $\mathbf{X}$.

Principal components are actually derived from a process which attempts to approximate $\tilde{\mathbf{X}}$ rather than $\mathbf{X}$. By adding $\mathbf{1}_n\overline{\mathbf{x}}'$ to an approximation $\hat{\tilde{\mathbf{X}}}$ to $\tilde{\mathbf{X}}$, you can then get an approximation $\hat{\mathbf{X}}$ for $\mathbf{X}$:

$$\hat{\mathbf{X}} = \hat{\tilde{\mathbf{X}}} + \mathbf{1}_N\overline{\mathbf{x}}'.$$

The matrix of <u>residuals</u> of $\mathbf{X}$ from $\overline{\mathbf{x}}$ is,

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}_N\overline{\mathbf{x}}' = \begin{bmatrix} (\mathbf{x}_1 - \overline{\mathbf{x}})' \\ (\mathbf{x}_2 - \overline{\mathbf{x}})' \\ (\mathbf{x}_3 - \overline{\mathbf{x}})' \\ \cdots\cdots \\ (\mathbf{x}_N - \overline{\mathbf{x}})' \end{bmatrix}, \text{ N by p}$$

When N > p, usually rank($\tilde{\mathbf{X}}$) = p. That is, you can always find p pairs of vectors,

$$\mathbf{U}_k \ (N \times 1) \text{ and } \mathbf{v}_k \ (p \times 1)$$

so that $\tilde{\mathbf{X}}$ is the sum of p *outer products* $\mathbf{U}_k\mathbf{v}_k'$ of $\mathbf{U}_k$ and $\mathbf{v}_k$:

$$\tilde{\mathbf{X}} = \sum_{1\leq k\leq p} \mathbf{U}_k\mathbf{v}_k'$$

The element in row i (case i) and column ℓ (variable ℓ) of $\tilde{\mathbf{X}}$ is

$$\tilde{x}_{i\ell} \equiv x_{i\ell} - \overline{x}_\ell = \sum_{1\leq k\leq p} u_{ik}v_{\ell k} = \sum_{1\leq k\leq p} v_{\ell k}u_{ik}$$

There are an infinite number of such sets of vectors $\{\mathbf{U}_k\}$ and $\{\mathbf{v}_k\}$.

In this representation, $\tilde{\mathbf{X}} = \sum_{1\leq k\leq p} \mathbf{U}_k\mathbf{v}_k'$, think of $\mathbf{U}_k$ and $\mathbf{v}_k$ in the following way:

- Each N by 1 $\mathbf{U}_k$ is a new *variable* with a value for each of the N cases
- Element $v_{\ell k}$ of the p by 1 vector $\mathbf{v}_k$ is a *coefficient* of $\mathbf{U}_k$, in a representation for column $\tilde{\mathbf{X}}_\ell$ of $\tilde{\mathbf{X}}$.

  Specifically, if $\tilde{\mathbf{X}}_\ell$ is column ℓ of $\tilde{\mathbf{X}}$,

$$\tilde{\mathbf{X}}_\ell = \sum_{1\leq k\leq p} v_{\ell k}\mathbf{U}_k.$$

Except for there being no constant term (intercept), this looks a little like a multiple regression of $\tilde{\mathbf{X}}_\ell$ on $\mathbf{U}_1$, ..., $\mathbf{U}_p$.

When rank($\tilde{\mathbf{X}}$) = p, when m < p you can't *exactly* <u>reproduce</u> $\tilde{\mathbf{X}}$ by $\sum_{1\leq k\leq m} \mathbf{U}_k\mathbf{v}_k'$, but it may be possible to get a good fit that is, have $\tilde{\mathbf{X}} \approx \sum_{1\leq k\leq m} \mathbf{U}_k\mathbf{v}_k'$.

That is, you *may* be able closely to *approximate* $\tilde{\mathbf{X}}$ by a sum of m < p outer products:

$\tilde{\mathbf{X}} \cong \sum_{1\le k\le m}\mathbf{U}_k\mathbf{v}_k'$ or $\tilde{\mathbf{X}}_\ell = \sum_{1\le k\le m} v_{\ell k}\mathbf{U}_k, \ell = 1,\ldots,p$

in the sense that the elements of $\tilde{\mathbf{X}} - \sum_{1\le k\le m}\mathbf{U}_k\mathbf{v}_k'$ are small.

This would be a *reduced rank approximation* (specifically a <u>rank m approximation</u>) to $\tilde{\mathbf{X}}$ because

$$\text{rank}(\textstyle\sum_{1\le i\le m}\mathbf{U}_i\mathbf{v}_i') = m < p$$

How do you find such $\mathbf{U}_i$'s and $\mathbf{v}_i$'s?

That is, for a specific m < p, how do you find

- Variables $\mathbf{U}_k$, k = 1,...,m
- Coefficient vectors $\mathbf{v}_k$, k = 1,...,m

such that $\tilde{\mathbf{X}} - \sum_{1\le k\le m}\mathbf{U}_k\mathbf{v}_k'$ is small?

Let's start with m = 1, that is, find a N×1 $\mathbf{U}$ and p×1 $\mathbf{v}$ such that

$$\tilde{\mathbf{X}}^{(1)} = \mathbf{U}\mathbf{v}' = [v_1\mathbf{U}\ v_2\mathbf{U}\ \ldots\ v_p\mathbf{U}] \cong \tilde{\mathbf{X}} \quad \text{(rank 1)}$$

that is, find numbers

- $\{u_i\}$, i = 1,...,N
- $\{v_\ell\}$, ℓ = 1,...p

so that $\tilde{x}_{i\ell}^{(1)} - u_i v_\ell$ is small.

Let's drop the ~ and just use $\mathbf{X}$, since what we are doing does not depend on working with a matrix with 0 means, although that is the usual case.

If you view $\mathbf{U}$ as a new variable, this is like finding one predictor variable $\mathbf{U}$ so that the regressions (without intercept) of each column of $\mathbf{X}$ on $\mathbf{U}$ is a good fit.

Contrast this with the usual regression situation where you are *given* a predictor variable $\mathbf{Z}$ and seek to find coefficients. Here you need to find both predictor and coefficients.

The first thought many statisticians would have would be to find **U** and **v** so that **Uv**' is close to **X** by the *least squares* criterion.

That is, find **U** and **v** so as to minimize

$$\sum_{1\le i\le N}\sum_{1\le \ell\le p}(x_{i\ell} - u_i v_\ell)^2 = \|\mathbf{X} - \mathbf{U}\mathbf{v}'\|^2$$

That's what we're going to do.

Notation: When **A** = $[a_{ij}]$ is a matrix

$$\|\mathbf{A}\|^2 = \sum_i\sum_j a_{ij}^2 = \mathrm{trace}(\mathbf{A}'\mathbf{A})$$

The ***Singular Value Decomposition*** (SVD) of **X** is the key to finding **U** and **v** to minimize $\|\mathbf{X} - \mathbf{U}\mathbf{v}'\|^2$.

The SVD of **X** is a mathematical representation of **X** which is useful in many contexts.

Statistics 5401 Lecture 24 November 2, 2005

The Singular Value Decomposition

For *any* N×p matrix **X** with $N \geq p$, there are always three matrices **L**, **R** and **T** such that **X** = **LTR**' where

- **L** = $[\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_p]$ is N × p with $\mathbf{L}'\mathbf{L} = \mathbf{I}_p$. That is, the columns $\mathbf{L}_1, \ldots, \mathbf{L}_p$ of **L** are *orthonormal*:

  $$\mathbf{L}_j'\mathbf{L}_j = 1,\ \mathbf{L}_j'\mathbf{L}_k = 0,\ j \neq k$$

  Note: When N > p, this does not mean that $\mathbf{L}^{-1} = \mathbf{L}'$, since **L** is not square.

- R = $[\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_p]$ is p × p *square* with $\mathbf{R}'\mathbf{R} = \mathbf{I}_p$, that is, the columns $\mathbf{r}_1, \ldots, \mathbf{r}_p$ of **R** are *orthonormal*:

  $$\mathbf{r}_j'\mathbf{r}_j = 1,\ \mathbf{r}_j'\mathbf{r}_k = 0,\ j \neq k$$

  Since **R** is p×p, this means that $\mathbf{R}^{-1} = \mathbf{R}'$ and $\mathbf{R}\mathbf{R}' = \mathbf{I}_p$. **R** is an orthogonal matrix.

- **T** = $\mathrm{diag}[t_1, \ldots, t_p]$, p × p, *diagonal* with $t_1 \geq t_2 \geq \ldots \geq t_p \geq 0$

Vocabulary

$\mathbf{X} = \mathbf{LTR}'$ is the *singular value decomposition* (SVD) of $\mathbf{X}$

Facts

- $\mathbf{X} = \mathbf{LTR}' = \sum_{1 \le k \le p} t_k \mathbf{L}_k \mathbf{r}_k' = \sum_{1 \le k \le p} (t_k \mathbf{L}_k) \mathbf{r}_k'$, a sum of *p outer products* of $t_k \mathbf{L}_k$ and $\mathbf{r}_k$
- The t_i's are unique
- When $t_i \ne t_j$, all $j \ne i$, $\mathbf{L}_i$ and $\mathbf{r}_i$ are unique (except for multiplication of both by -1 : $(-\mathbf{L}_i)(-\mathbf{r}_i)' = \mathbf{L}_i \mathbf{r}_i'$)

Thus the *SVD* $\mathbf{X} = \mathbf{LTR}'$ of $\mathbf{X}$ is essentially unique.

- The N by 1 vectors $\mathbf{L}_j$, j = 1,...,p are the ***left singular vectors*** of $\mathbf{X}$.
- The p by 1 vectors $\mathbf{r}_j$, j = 1,...,p are the ***right singular vectors*** of $\mathbf{X}$.
- The p scalars $t_1 \ge t_2 \ge \ldots \ge t_p \ge 0$ are the ***singular values*** of $\mathbf{X}$.

When there are only $s < p$ singular values $t_i \ne 0$, so that $t_{s+1} = \ldots = t_p = 0$,

$$\mathbf{X} = \sum_{1 \le k \le s} (t_k \mathbf{L}_k) \mathbf{r}_k'$$

a sum of only s outer products.

Fact:

- Rank($\mathbf{X}$) = s = number of non-zero singular values. When $s < p$,

$$t_s > t_{s+1} = t_{s+2} = \ldots = t_p = 0.$$

The SVD is often the best way numerically to determine the Rank($\mathbf{X}$):

- Compute $\mathbf{T}$ from $\mathbf{X}$
- Count how many diagonal elements are non-zero except for rounding error. This should be Rank($\mathbf{X}$).

Computing the SVD in MacAnova: **svd()**

Suppose `x` is a **REAL** matrix. Then `svd(x)` computes the vector $[t_1,\ldots,t_p]'$ of singular values of `x` (diag(T), *not* T)

`svd(x,left:T)` computes a structure with two components:

- `values`, vector of singular values
- `leftvectors` matrix **L** whose columns are $\mathbf{L}_1, \ldots, \mathbf{L}_p$, the *left* singular vectors

`svd(x,right:T)` computes a structure with two components:

- `values`, vector of singular values
- `rightvectors` matrix **R** whose columns are $\mathbf{r}_1, \ldots, \mathbf{r}_p$, *right* singular vectors

`svd(x,right:T,left:T)` or `svd(x,all:T)` computes a 3 component structure:

- `values`: singular values
- `leftvectors`: *left* singular vectors
- `rightvectors`: *right* singular vectors

Example

```
Cmd> x <- run(10)^run(0,3)' # powers of run(10)

Cmd> setlabels(x,\
       structure("@",vector("i^0","i^1","i^2","i^3")))

Cmd> x # 10 ny 4 matrix
              i^0         i^1         i^2         i^3
(1)             1           1           1           1
(2)             1           2           4           8
(3)             1           3           9          27
(4)             1           4          16          64
(5)             1           5          25         125
(6)             1           6          36         216
(7)             1           7          49         343
(8)             1           8          64         512
(9)             1           9          81         729
(10)            1          10         100        1000

Cmd> vals <- svd(x); vals #just Sing values
(1)      1415.4       27.14      2.2961     0.41587
```

X has rank 4, but since the two smallest singular value are so small, it is close to having rank 3, or possibly even rank 2.

```
Cmd> results <- svd(x,left:T,right:T)# or all:T

Cmd> compnames(results)
(1) "values"                Structure
(2) "leftvectors"           component
(3) "rightvectors"          names
```

- `values` p-vector, $(t_1, t_2, \ldots, t_p)$
- `leftvectors` N by p $\mathbf{L} = [\mathbf{L}_1 \ldots \mathbf{L}_p]$
- `rightvectors` p by p square $\mathbf{R} = [\mathbf{r}_1\ldots\mathbf{r}_p]$

You can construct the diagonal matrix **T** in the SVD by

```
Cmd> tmatrix <- dmat(results$values); tmatrix#Diagonal matrix T
(1,1)      1415.4          0          0          0
(2,1)           0      27.14          0          0
(3,1)           0          0     2.2961          0
(4,1)           0          0          0    0.41587
```

Here are numerical checks that the right and left singular vectors are orthonormal ($\mathbf{R}'\mathbf{R} = \mathbf{I}_p$ and $\mathbf{L}'\mathbf{L} = \mathbf{I}_p$):

```
Cmd> R <- results$rightvectors # right singular vectors

Cmd> list(R) # size is p by p
R                REAL    4     4     (labels)

Cmd> R' %*% R # = I_4
               (1)          (2)          (3)          (4)
(1)            1   4.0533e-17 -6.9218e-17  4.9043e-17
(2)   4.0533e-17            1   6.2095e-17  4.8526e-17
(3)  -6.9218e-17   6.2095e-17            1 -1.4827e-16
(4)   4.9043e-17   4.8526e-17 -1.4827e-16            1
```

This is $\mathbf{I}_4$.

```
Cmd> L <- results$leftvectors # left singular vectors

Cmd> list(L) # size is N by p
L                REAL    10    4     (labels)

Cmd> L' %*% L # = I_4
               (1)          (2)          (3)          (4)
(1)            1  -6.9389e-18 -1.1102e-16 -2.3592e-16
(2)  -6.9389e-18            1   1.6653e-16  3.8858e-16
(3)  -1.1102e-16   1.6653e-16            1 -4.4409e-16
(4)  -2.3592e-16   3.8858e-16 -4.4409e-16            1
```

This is also $\mathbf{I}_4$.

Relationship with Eigenvalues and Eigenvectors

- Each right singular vector $\mathbf{r}_j$ is an eigenvector of $\mathbf{X}'\mathbf{X}$ with eigenvalue t_j^2.

  That is, $\mathbf{r}_j$ satisfies $\mathbf{X}'\mathbf{X}\mathbf{r}_j = t_j^2\mathbf{r}_j$.

  Check: $\mathbf{X}'\mathbf{X}\mathbf{r}_j = \mathbf{RTL}'\mathbf{LTR}'\mathbf{r}_j = t_j^2\mathbf{r}_j$, because $\mathbf{L}'\mathbf{L} = \mathbf{I}_p$ and $\mathbf{R}'\mathbf{R} = \mathbf{I}_p$.

```
Cmd> sqrt(eigenvals(x' %*% x)) # numerical check
(1)      1415.4       27.14      2.2961     0.41587
```

$\mathbf{X}'\mathbf{X}$ is the $p\times p$ matrix of sums of squares (SS) and sums of products (SP) of the *columns* of $\mathbf{X}$.

And, for the case we apply this to, $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ consists of sums of squares $\sum_i(x_{i\ell}-\overline{x}_\ell)^2$ and products $\sum_i(x_{i\ell} - \overline{x}_\ell)(x_{ik} - \overline{x}_k)$. An, of course, $\mathbf{S}_x = (1/(N-1))\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$.

- Each <u>left</u> singular vector $\mathbf{L}_j$ is an eigenvector of the N by N matrix $\mathbf{XX}'$ with eigenvalue t_j^2.

  That is, $\mathbf{L}_j$ satisfies $\mathbf{XX}'\mathbf{L}_j = t_j^2\mathbf{L}_j$.

  <u>Check</u>: $\mathbf{XX}'\mathbf{L}_j = \mathbf{LTR}'\mathbf{RTL}'\mathbf{L}_j = t_j^2\mathbf{L}_j$ since $\mathbf{R}'\mathbf{R} = \mathbf{I}_p$ and $\mathbf{L}'\mathbf{L} = \mathbf{I}_p$.

```
Cmd> sqrt(round(eigenvals(x %*% x'),9)) # numerical check
(1)      1415.4       27.14      2.2961     0.41587           0
(6)           0           0           0           0           0
```

  There are N - p = 6 zero eigenvalues.

  $\mathbf{XX}'$ is the N×N matrix of SS and SP of the *rows* of $\mathbf{X}$.

Summary

If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ are eigenvectors of $\mathbf{X}'\mathbf{X}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_p \geq 0$ then

- the j[th] *singular value* is
  $t_j = \sqrt{\lambda_j}$
- the j[th] *right singular vector* is
  $\mathbf{r}_j = \mathbf{v}_j$ (could be $-\mathbf{v}_j$), j = 1, ..., p

If $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_N$ (all N by 1) are eigenvectors of $\mathbf{XX}'$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_p \geq \ldots \geq \lambda_N$ then

- the j[th] *singular value* is
  $t_j = \sqrt{\lambda_j}$, j = 1, ..., p
- the j[th] *left singular vector* is
  $\mathbf{L}_j = \boldsymbol{\ell}_j$ (could be $-\boldsymbol{\ell}_j$), j = 1, ..., p
- The remaining eigenvalues $\lambda_{p+1}, \ldots, \lambda_N$ are 0
- The remaining eigenvectors $\boldsymbol{\ell}_{p+1}, \ldots, \boldsymbol{\ell}_N$ are irrelevant

Now define p linear combinations of the columns of $\mathbf{X}$ with coefficients from the <u>right</u> singular vectors $\mathbf{r}_j$, j = 1,... p:

$$\mathbf{U}_j \equiv \mathbf{Z}_j = \mathbf{X}\mathbf{r}_j = \mathbf{LTR}'\mathbf{r}_j = t_j\mathbf{L}_j,\ j = 1,\ldots,p$$

$\mathbf{Z}_j = \mathbf{X}\mathbf{r}_j = \sum_{1\le k\le p} r_{kj}\mathbf{X}_k$ is a linear combination of the columns of $\mathbf{X}$.

The coefficients (weights) are the elements of

$\mathbf{r}_j$ = j^{th} right singular vector of $\mathbf{X}$
= j^{th} eigenvector of $\mathbf{X}'\mathbf{X}$.

Because $\mathbf{R}'\mathbf{R} = \mathbf{I}_p$, $\mathbf{R}'\mathbf{r}_j = [0\ \ 0\ \ldots\ 1\ \ldots\ 0]'$ and so

$$\mathbf{Z}_j = \mathbf{LTR}'\mathbf{r}_j = t_j\mathbf{L}_j$$

is proportional to a left singular vector of $\mathbf{X}$.

Back to low rank approximations

Q. With m = 1, what N×1 $\mathbf{U}$ and p×1 $\mathbf{v}$ minimize (make smallest) the "residual SS"

$$\|\mathbf{X} - \mathbf{U}\mathbf{v}'\|^2 = \sum_{1\le i\le N}\sum_{1\le \ell\le p}(x_{i\ell} - u_i v_\ell)^2\ ?$$

A. $\mathbf{U} = \mathbf{Z}_1 = t_1\mathbf{L}_1$ and $\mathbf{v} = \mathbf{r}_1$

That is

$$\hat{\mathbf{X}}^{(1)} \equiv \mathbf{Z}_1\mathbf{r}_1' = t_1\mathbf{L}_1\mathbf{r}_1' = \mathbf{X}\mathbf{r}_1\mathbf{r}_1'$$

is the best rank 1 approximation to $\mathbf{X}$ in the least squares sense.

This generalizes to rank m > 1:

$$\hat{\mathbf{X}}^{(m)} = \sum_{1\le j\le m}\mathbf{Z}_j\mathbf{r}_j' = \sum_{1\le j\le m} t_j\mathbf{L}_j\mathbf{r}_j' = \mathbf{X}(\sum_{1\le j\le m}\mathbf{r}_j\mathbf{r}_j')$$

is the <u>best rank m approximation</u> to $\mathbf{X}$ in the least squares sense.

How good is the approximation?

- The "residual sum of squares" is

$$\|\mathbf{X} - \hat{\mathbf{X}}^{(m)}\|^2 = \sum_{1\le i\le N}\sum_{1\le \ell\le p}(x_{i\ell} - \hat{x}_{i\ell}^{(m)})^2$$
$$= \sum_{m+1\le k\le p} t_k^2 = \sum_{m+1\le k\le p}\lambda_k$$

= the sum of the squared smallest $p - m$ singular values of $\mathbf{X}$

= sum of the smallest $p - m$ eigenvalues of $\mathbf{X}'\mathbf{X}$.

- The "total sum of squares" is

$$\|\mathbf{X}\|^2 = \sum_\ell\sum_i x_{i\ell}^2 = \sum_{1\le k\le p} t_k^2 = \sum_{1\le k\le p}\lambda_k = \text{tr } \mathbf{X}'\mathbf{X}$$

Therefore, when the ratio

$$\frac{\|\mathbf{X}-\hat{\mathbf{X}}^{(m)}\|^2}{\|\mathbf{X}\|^2} = \frac{\sum_{1\le i\le N}\sum_{1\le \ell\le p}(x_{i\ell}-\hat{x}_{i\ell}^{(m)})^2}{\sum_{1\le i\le N}\sum_{1\le \ell\le p}x_{i\ell}^2} = \frac{\sum_{m+1\le k\le p}t_k^2}{\sum_{1\le k\le p}t_k^2}$$

is small, the approximation is pretty good. This ratio is analogous to

$$SS_{residual}/SS_{total} = 1 - R^2$$

in regression.

In the rank m approximation,

$$\hat{\mathbf{X}}^{(m)} = \sum_{1\le k\le m}\mathbf{Z}_k\mathbf{r}_k' = \sum_{1\le k\le m}t_k\mathbf{L}_k\mathbf{r}_k' ,$$

column $\mathbf{X}_\ell$ of $\mathbf{X}$ is approximated by

$$\hat{\mathbf{X}}_\ell^{(m)} = \sum_{1\le k\le m}\mathbf{Z}_k r_{\ell k} = \sum_{1\le k\le m}r_{\ell k}\mathbf{Z}_k,$$

a linear combination of $\mathbf{Z}_1$, $\mathbf{Z}_2$, ..., $\mathbf{Z}_m$.

Since the $\mathbf{Z}_k$'s themselves are linear combinations of the columns of $\mathbf{X}$ ($\mathbf{Z}_k = \mathbf{X}\mathbf{r}_k$), so are the columns of $\hat{\mathbf{X}}^{(m)}$:

$$\hat{\mathbf{X}}_\ell^{(m)} = \sum_{1\le k\le m}r_{\ell k}\mathbf{X}\mathbf{r}_k$$
$$= \sum_{1\le \ell\le m}(\sum_{1\le k\le m}r_{\ell k}r_{jk})\mathbf{X}_j$$

$\sum_{1\le k\le m}r_{\ell k}r_{jk}$ is a partial sum of squares ($j = \ell$) or sum of products ($j \ne \ell$) of <u>rows</u> of $\mathbf{R}$. Since $\mathbf{RR}' = \mathbf{I}_p$,

$$\sum_{1\le k\le m}r_{\ell k}^2 = 1 - \sum_{m+1\le k\le p}r_{\ell k}^2$$
$$\sum_{1\le k\le m}r_{\ell k}r_{jk} = -\sum_{m+1\le k\le p}r_{\ell k}r_{jk}$$