Introduction to Multivariate Analysis

Liu, Xin

Fall, 2021

School of Statistics and Management
Shanghai University of Finance and Economics

Email: liu.xin@mail.shufe.edu.cn

Chapter 2 Principal Components Analysis and Factor Analysis

§2.1 Principal Components

Data (dimension) reduction:

- Suppose we have a p-dim population, denoted by X,

$$\mathbf{X} = (X_1, \cdots, X_p)',$$

whose mean vector μ and covariance matrix Σ are given by

$$\boldsymbol{\mu} = E(\mathbf{X}) = (\mu_1, \cdots, \mu_p)' \quad \text{and} \quad \Sigma = \text{Cov}(\mathbf{X}) = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'$$

- Consider a degenerate case where X follows a normal distribution with a singular covariance matrix Σ, i.e. $\det(\Sigma) = 0$. Then it is possible to represent the population by another random vector with dimension $d < p$.

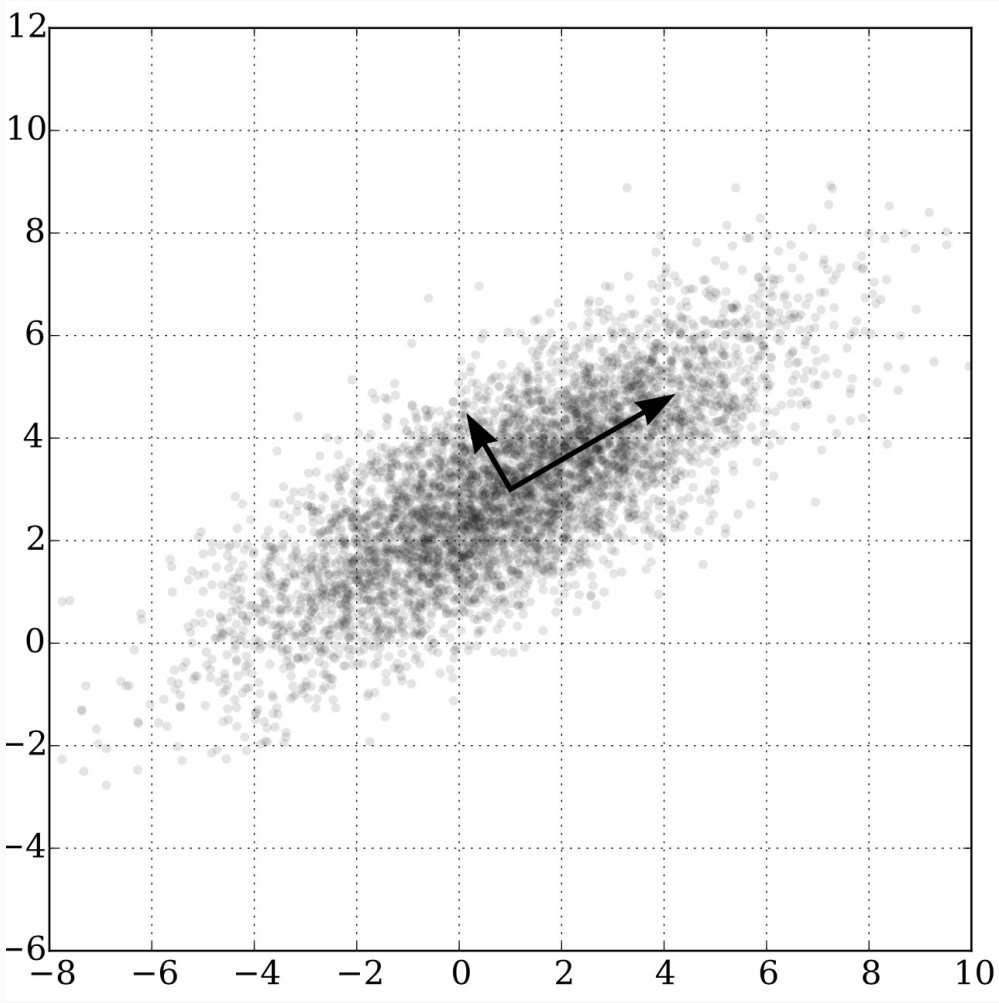


Figure 1: PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean.

- Consider the linear combinations

$$
\begin{aligned}
Z_1 &= \mathbf{v}_1'\mathbf{X} = v_{11}X_1 + v_{12}X_2 + \cdots + v_{1p}X_p \\
Z_2 &= \mathbf{v}_2'\mathbf{X} = v_{21}X_1 + v_{22}X_2 + \cdots + v_{2p}X_p \\
\cdots &= \cdots \\
Z_p &= \mathbf{v}_p'\mathbf{X} = v_{p1}X_1 + v_{p2}X_2 + \cdots + v_{pp}X_p
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathrm{Var}\,(Z_j) &= \mathbf{v}_j'\Sigma\mathbf{v}_j, \quad j = 1, \cdots, p \\
\mathrm{Cov}\,(Z_j, Z_k) &= \mathbf{v}_j'\Sigma\mathbf{v}_k, \quad \forall j \neq k
\end{aligned}
$$

- Principal component analysis (PCA, Pearson 1901) is a statistical procedure that
  - uses an orthogonal transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables (called principal components)
  - finds directions with maximum variability
- Principal components (PCs):
  - PCs are uncorrelated, orthogonal, linear combinations $Z_1, \cdots, Z_p$ whose variances are as large as possible.
  - PCs form a new coordinate system by rotating the original system constructed by $X_1, \cdots, X_p$

- The procedure seeks the direction of high variances:
  - The first PC = linear combination $Z_1 = \mathbf{v}_1'\mathbf{X}$ that maximizes $\mathrm{Var}\left(\mathbf{v}_1'\mathbf{X}\right)$ subject to $\|\mathbf{v}_1\| = 1$
  - The second PC = linear combination $Z_2 = \mathbf{v}_2'\mathbf{X}$ that maximizes $\mathrm{Var}\left(\mathbf{v}_2'\mathbf{X}\right)$ subject to $\|\mathbf{v}_2\| = 1$ and $\mathrm{Cov}\left(\mathbf{v}_1'\mathbf{X}, \mathbf{v}_2'\mathbf{X}\right) = 0$
  - The jth PC satisfies

$$\begin{aligned} &\max \mathrm{Var}\left(\mathbf{v}_j'\mathbf{X}\right) \\ \text{subject to } &\left\|\mathbf{v}_j\right\| = 1 \\ &\mathrm{Cov}\left(\mathbf{v}_l'\mathbf{X}, \mathbf{v}_j' X \mathbf{v}_j\right) = \mathbf{v}_l'\Sigma\mathbf{v}_j = 0 \\ &\text{for } 1 = 1, \ldots, j-1 \end{aligned}$$

    where $j = 2, \cdots, p$

- Assume Σ has p eigenvalue-eigenvector pairs $(\lambda, \mathbf{u})$ satisfying:

$$\Sigma \mathbf{u}_j = \lambda_j \mathbf{u}_j, \quad j = 1 \cdots, p$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and $\|\mathbf{u}_j\| = 1$ for all j. This gives the following spectral decomposition

$$\Sigma = \sum_{j=1}^{p} \lambda_j \mathbf{u}_j \mathbf{u}_j'$$

- The jth PC is given by $Z_j = \mathbf{u}_j' \mathbf{X}$ and its variance is

$$\mathrm{Var}\,(Z_j) = \mathbf{u}_j' \Sigma \mathbf{u}_j = \lambda_j$$

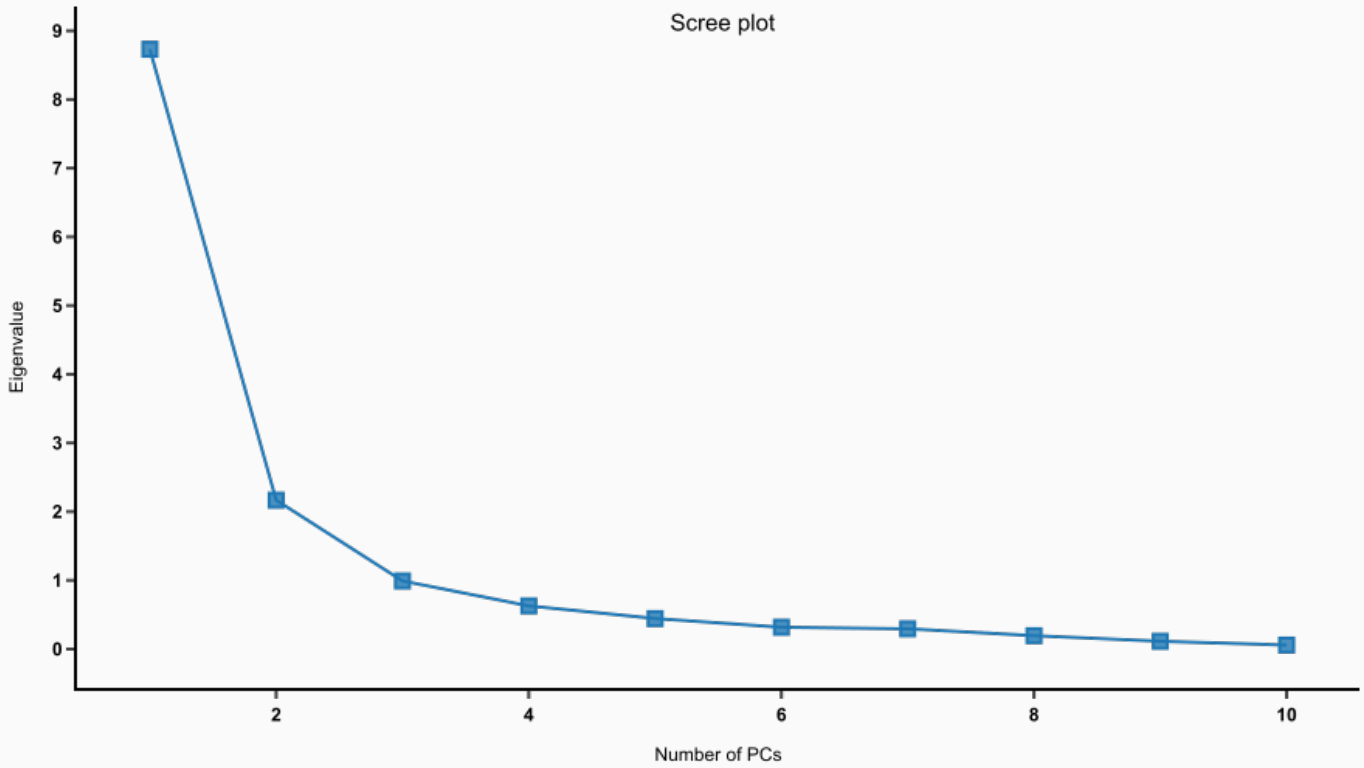
- The magnitude of u_{jk} measures the importance of the k th variable to the j th PC.

- The total (population) variance of inputs

$$\sum_{j=1}^{p} \text{Var}(X_j) = \sum_{j=1}^{p} \sigma_{jj} = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \text{Var}(Z_j)$$

- Proportion of total variance due to the j th $\text{PC} = \frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}$.
- The number of PCs are decided based on
  - the amount of total sample variance explained,
  - the variances of the sample PC, and
  - the subject-matter interpretations.

- Plot the ordered eigenvalues $\lambda_1, \cdots, \lambda_p$ and look for the elbow (bend) in the plot. The number of PCs is the point where the remaining eigenvalues are relatively small and all about the same size.



- In practice, μ and Σ are unknown and estimated from the data. Assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ of $\boldsymbol{\Sigma}$ distinct and positive.

- Sample mean:

$$\overline{\mathbf{X}} = \frac{1}{n} X' \mathbf{1}_n$$

  X is the design matrix, and 1_n is the vector of $(1, \ldots, 1)'$ of length n.

- (Unbiased) Sample variance-covariance matrix

$$\mathbf{S} = \frac{1}{n-1} X_c' X_c = \frac{1}{n-1} \sum_{i=1}^{n} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right) \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)'$$

  where X_c the centered design matrix, and $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})'$ for

$$i = 1, \cdots, n$$

  It is easy to show that

$$\mathbf{S} = \frac{1}{n-1} X' \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) X$$

§2.2 An Example

Four Measurements on Three Species of Iris (in centimeters), Fisher (1936).

Iris setosa

Iris versicolor

Iris virginica

Table 1: Iris data: S.L–Sepal length, S.W—Sepal width, P.L–Petal length, P.W—Petal width, all in cm.

	Iris setosa				Iris versicolor				Iris virginica			
	S.L	S.W	P.L	P.W	S.L	S.W	P.L	P.W	S.L	S.W	P.L	P.W
1	5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
2	4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
3	4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4	4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5	5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
46	4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
47	5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
48	4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
49	5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
50	5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

- As an example of principal component analysis we use the data set of Iris versicolor.
- There are 50 observations ($N = 50$, $n = N - 1 = 49$). Each observation consists of four measurements on a plant:
  x_1 is sepal length, x_2 is sepal width,
  x_3 is petal length, x_4 is petal width.
- The sample covariance matrix is

$$S = \frac{1}{49} \sum_{\alpha=1}^{50} (x_\alpha - \bar{x}) (x_\alpha - \bar{x})'$$

$$= \begin{pmatrix} 0.266433 & 0.085184 & 0.182899 & 0.055780 \\ 0.085184 & 0.098469 & 0.082653 & 0.041204 \\ 0.182899 & 0.082653 & 0.220816 & 0.073102 \\ 0.055780 & 0.041204 & 0.073102 & 0.039106 \end{pmatrix}$$

- We next compute $S_2 = S - l_1 b^{(1)} b^{(1)}$

$$= \begin{pmatrix} 0.0363559 & -0.0171179 & -0.0260502 & -0.0162472 \\ -0.0171179 & 0.0529813 & -0.0102546 & 0.0091777 \\ -0.0260502 & -0.0102546 & 0.0310544 & 0.0076890 \\ -0.0162472 & 0.0091777 & 0.0076890 & 0.0165574 \end{pmatrix}$$

  and iterate $z^{(j)} = S_2 z^{(j-1)}$, using $z^{(0)\prime} = (0, 1, 0, 0)$.

- In this case the iteration does not proceed as rapidly; On the last iteration, the ratios agree to within four units in the fifth significant figure. We obtain $l_2 = 0.0723828$ and

$$\boldsymbol{b}^{(2)} = \begin{pmatrix} -0.669033 \\ 0.567484 \\ 0.343309 \\ 0.335307 \end{pmatrix}$$

- The third principal component is found from $S_3 = S_2 - l_2 b^{(2;} b^{(2)\prime}$, and the fourth from $S_4 = S_3 - l_3 b^{(3)} b^{(3)\prime}$

- We use the iterative procedure to find the first principal component, by computing in turn $z^{(j)} = Sz^{(j-1)}$. As an initial approximation, we use $z^{(0)\prime} = (1, 0, 1, 0)$.
- It is not necessary to normalize the vector at each iteration; but to compare successive vectors, we compute $z_i^{(j)}/z_i^{(j-1)} = r_i^{(j)}$, each of which is an approximation to l_1, the largest root of S.
- After seven iterations,
  - $r_i^{(7)}$ agree to within two units in the fifth decimal place.
  - The ratios, $r_i^{(8)}$, agree to within two units in the sixth place.
- The value of l_1 is (nearly accurate to the sixth place) $l_1 = 0.487875$. The normalized eighth iterated vector is our estimate of $\beta^{(1)}$, namely,

$$b^{(1)} = \begin{pmatrix} 0.6867244 \\ 0.3053463 \\ 0.6236628 \\ 0.2149837 \end{pmatrix}$$

- The results may be summarized as follows:

$$(l_1, l_2, l_3, l_4) = (0.4879, 0.0724, 0.0548, 0.0098)$$

$$B = \begin{pmatrix} 0.6867 & -0.6690 & -0.2651 & 0.1023 \\ 0.3053 & 0.5675 & -0.7296 & -0.2289 \\ 0.6237 & 0.3433 & 0.6272 & -0.3160 \\ 0.2150 & 0.3353 & 0.0637 & 0.9150 \end{pmatrix}$$

- The sum of the four roots is $\sum_{i=1}^{4} l_i = 0.6249$, compared with the trace of the sample covariance matrix, tr $S = 0.624824$.
  - The first accounts for 78% of the total variance in the four measurements;
  - the last accounts for a little more than 1%.
- In fact, the variance of $0.7x_1 + 0.3x_2 + 0.6x_3 + 0.2x_4$ (an approximation to the first principal component) is 0.478, which is almost 77% of the total variance.

§1.3 Statistical Inference

- Let $X_1, X_2, \ldots, X_N \overset{iid}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $n = N - 1$, where the covariance matrix Σ has the spectral decompostion

$$\boldsymbol{\Sigma} = \sum_{j=1}^{p} \lambda_j \mathbf{u}_j \mathbf{u}_j'.$$

Then

$$B(n) \triangleq \sqrt{n}\left[\mathbf{S} - \Sigma\right] \xrightarrow{D} N(0, C)$$

where the covariance function is $\mathrm{E}[b_{ij}(n) b_{kl}(n)] \to \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$.

- Let $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p\right)'$ be the vector of eigenvalues for $\mathbf{S}$. Then

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{D} \mathrm{N}_p\left(\mathbf{0}, 2\Lambda^2\right),$$

  where Λ is the diagonal matrix constructed from the eigenvalues of Σ.

- Let $\hat{\mathbf{u}}_i$ be the eigenvectors of $\mathbf{S}$. Then

$$\sqrt{n}\left(\hat{\mathbf{u}}_i - \mathbf{u}_i\right) \xrightarrow{D} \mathrm{N}_p\left(\mathbf{0}, \mathbf{E}_i\right)$$

  where

$$\mathbf{E}_i = \lambda_i \sum_{k \neq i} \frac{\lambda_k}{\left(\lambda_k - \lambda_i\right)^2} \mathbf{u}_k \mathbf{u}_k'$$

  For large n, $\hat{\lambda}_i$ is approximately independent of the distribution for $\hat{\mathbf{u}}_i$

Confidence Region for a Characteristic Vector

- We use the asymptotic distribution of the sample characteristic vectors to obtain a large-sample confidence region for the $\mathbf{u}_i$ [Anderson (1963a).]
- Recall that

$$\sqrt{n}\left(\hat{\mathbf{u}}_i - \mathbf{u}_i\right) \xrightarrow{D} \mathrm{N}_p\left(\mathbf{0}, \mathbf{E}_i\right), \quad \mathbf{E}_i = \lambda_i \sum_{k \neq i} \frac{\lambda_k}{\left(\lambda_k - \lambda_i\right)^2} \mathbf{u}_k \mathbf{u}_k'.$$

- The covariance matrix $\mathbf{E}_i$ can be written

$$\mathbf{E}_i = \mathbf{U}\Delta_i^2\mathbf{U}' = \mathbf{U}_i^*\mathbf{\Delta}_i^{*2}\mathbf{U}_i^{*\prime}$$

where

  - Δ_i is the $p \times p$ diagonal matrix with 0 as the i th diagonal element and $\sqrt{\lambda_i\lambda_j}/\left(\lambda_i - \lambda_j\right)$ as the j th diagonal element, $j \neq i$;
  - Δ_i^* is the $(p-1) \times (p-1)$ diagonal matrix obtained from Δ_i by deleting the i th row and column; and
  - $\mathbf{U}_i^*$ is the $p \times (p-1)$ matrix formed by deleting the i th column from $\mathbf{u}$.

- Then we have

$$\Delta_i^{*-1}\mathbf{U}_i^{*\prime}\sqrt{n}\,(\hat{\mathbf{u}}_i-\mathbf{u}_i)\xrightarrow{D}N(0,I_{p-1})$$

and thus

$$n\,(\hat{\mathbf{u}}_i-\mathbf{u}_i)'\,\mathbf{U}_i^{*}\Delta_i^{*-2}\mathbf{U}_i^{*\prime}\,(\hat{\mathbf{u}}_i-\mathbf{u}_i)$$

has a limiting χ^2 -distribution with $p-1$ degrees of freedom.

- The matrix of the quadracic form is

$$\begin{aligned}\mathbf{U}_i^{*}\Delta_i^{*-2}\mathbf{U}_i^{*\prime}&=\textstyle\sum_{j=1}^{p}\mathbf{u}_j\left(\frac{\lambda_i}{\lambda_j}-2+\frac{\lambda_j}{\lambda_i}\right)\mathbf{u}_j'-\mathbf{u}_i\left(\frac{\lambda_i}{\lambda_i}-2+\frac{\lambda_i}{\lambda_i}\right)\mathbf{u}_i'\\&=\lambda_i\Sigma^{-1}-2I+(1/\lambda_i)\,\Sigma\end{aligned}$$

- We can replace Σ and λ_i by the consistent estimators S and l_i to obtain

$$\begin{aligned}T_n=&n\,(\hat{\mathbf{u}}_i-\mathbf{u}_i)'\left[l_iS^{-1}-2I+(1/l_i)\,S\right](\hat{\mathbf{u}}_i-\mathbf{u}_i)\\=&n\left[l_i\mathbf{u}_i'\boldsymbol{S}^{-1}\mathbf{u}_i+(1/l_i)\,\mathbf{u}_i'S\mathbf{u}_i-2\right]\end{aligned}$$

which has a limiting χ^2 -distribution with $p-1$ degrees of freedom.

- A confidence region for the i th characteristic vector of Σ with confidence $1-\varepsilon$ consists of the intersection of
  - $\mathbf{u}_i'\mathbf{u}_i = 1$ and the set of
  - $\mathbf{u}_i$ such that T_n is less than $\chi^2_{p-1}(\varepsilon)$,

  where $\Pr\left(\chi^2_{p-1} > \chi^2_{p-1}(\varepsilon)\right\} = \varepsilon$
- This approach also provides a test of the null hypothesis

$$H_0 : \ \mathbf{u}_i = \mathbf{u}_{0i}$$

The hypothesis is rejected if T_n with $\mathbf{u}_i$ replaced by $\mathbf{u}_{0i}$ exceeds $\chi^2_{p-1}(\varepsilon)$.

Confidence limits on the Characteristic Roots

- We now consider a confidence interval for the entire set of characteristic roots of Σ, namely, $\lambda_1 \geq \cdots \geq \lambda_p$ [Anderson (1965a)].
- We use the facts that
  - $\mathbf{u}_i'\mathbf{\Sigma}\mathbf{u}_i = \lambda_i, \quad i = 1, p$, and
  - $\mathbf{u}_1\mathbf{\Sigma}\mathbf{u}_p = 0 = \mathbf{u}_1'\mathbf{u}_p$.

  Then $\mathbf{u}_1'\boldsymbol{X}$ and $\mathbf{u}_p'\boldsymbol{X}$ are uncorrelated and have variances λ_1 and λ_p, respectively.
- Hence $n\mathbf{u}_1'S\mathbf{u}_1/\lambda_1$ and $n\mathbf{u}_p'S\mathbf{u}_p/\lambda_p$ are independently distributed as χ^2 with n degrees of freedom.

- Let l and u be two numbers such that

$$1-\varepsilon = \Pr\left\{nl \le \chi_n^2\right\} \Pr\left\{\chi_n^2 \le nu\right\}$$

Then

$$\begin{aligned}
1-\varepsilon &= \Pr\left\{l \le \frac{\boldsymbol{u}_1'\boldsymbol{S}\boldsymbol{u}_1}{\lambda_1}, \frac{\boldsymbol{u}_p'\boldsymbol{S}\boldsymbol{u}_p}{\lambda_p} \le u\right\} \\
&= \Pr\left\{\frac{\boldsymbol{u}_p'\boldsymbol{S}\boldsymbol{u}_p}{u} \le \lambda_p, \lambda_1 \le \frac{\boldsymbol{u}_1'\boldsymbol{S}\boldsymbol{u}_1}{l}\right\} \\
&\le \Pr\left\{\min_{\boldsymbol{b}'\boldsymbol{b}=1} \frac{\boldsymbol{b}'\boldsymbol{S}\boldsymbol{b}}{u} \le \lambda_p, \lambda_1 \le \max_{\boldsymbol{b}'\boldsymbol{b}=1} \frac{\boldsymbol{b}'\boldsymbol{S}\boldsymbol{b}}{l}\right\} \\
&= \Pr\left\{\frac{l_p}{u} \le \lambda_p \le \lambda_1 \le \frac{l_1}{l}\right\}
\end{aligned}$$

- A confidence interval for the characteristic roots of Σ with confidence at least $1-\varepsilon$ is

$$l_p/u \le \lambda_p \le \lambda_1 \le l_1/l$$

Testing a Hypothesis about the Sum of the Smallest Characteristic Roots

- An investigator may raise the question
  - whether the last $p - m$ principal components may be ignored, that is,
  - whether the first m principal components furnish a good approximation to X.
- He may want to do this if the sum of the variances of the last principal components is less than some specified amount, say γ.

- Consider the null hypothesis

$$H_0 : \lambda_{m+1} + \cdots + \lambda_p \geq \gamma$$

where γ is specified, against the alternative that the sum is less than γ.

- If the characteristic roots of Σ are different, it follows that

$$\sqrt{n}\left(\sum_{i=m+1}^{p} l_i - \sum_{i=m+1}^{p} \lambda_i\right) \xrightarrow{D} N\left(0, 2\sum_{i=m+1}^{p} \lambda_i^2\right)$$

- The variance can be consistently estimated by $2\Sigma_{i=m+1}^{p} l_i^2$. Then a rejection region with (large-sample) significance level α is

$$\sum_{i=m+1}^{p} l_i < \gamma - \frac{\sqrt{2\sum_{i=m+1}^{p} l_i^2}}{\sqrt{n}} Z_\alpha$$

where Z_α is the upper significance point of the standard normal distribution for significance level α.

Testing a Hypothesis about the Sum of the Smallest Characteristic Roots Relative to the Sum of All the Roots

- The investigator may want to ignore the last $p - m$ principal components if their sum is small relative to the sum of all the roots (which is the trace of the covariance matrix).

- Consider the null hypothesis

$$H_0 : f(\lambda) = \frac{\lambda_{m+1} + \cdots + \lambda_p}{\lambda_1 + \cdots + \lambda_p} \geq \delta$$

  where δ is specified, against the alternative that $f(\lambda) < \delta$.

- We use the fact that

$$\frac{\partial f(\lambda)}{\partial \lambda_i} = -\frac{\lambda_{m+1} + \cdots + \lambda_p}{(\lambda_1 + \cdots + \lambda_p)^2}, \quad i = 1, \ldots, m$$

$$\frac{\partial f(\lambda)}{\partial \lambda_i} = \frac{\lambda_1 + \cdots + \lambda_m}{(\lambda_1 + \cdots + \lambda_p)^2}, \quad i = m+1, \ldots, p$$

- Then the asymptotic variance of $f(\lambda)$ is

$$\sigma^2 = 2\left(\frac{\delta}{\operatorname{tr}\Sigma}\right)^2\left(\lambda_1^2 + \cdots + \lambda_m^2\right) + 2\left(\frac{1-\delta}{\operatorname{tr}\Sigma}\right)^2\left(\lambda_{m+1}^2 + \cdots + \lambda_p^2\right)$$

  when equality holds in H_0.

- The null hypothesis H_0 is rejected if $\sqrt{n}[f(\lambda) - \delta]$ is less than the appropriate significance point of the standard normal distribution times σ with λ 's replaced by l 's and tr Σ by tr S.

- Alternatively one can construct a large-sample confidence region for $f(\lambda)$.

$$\frac{\sum_{i=1n+1}^{p}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \leq \frac{\sum_{i=m+1}^{p} l_i}{\sum_{i=1}^{p} l_i} + Z_\alpha \frac{\left[2\left(\sum_{i=m+1}^{p} l_i\right)^2 \sum_{i=1}^{m} l_i^2 + 2\left(\sum_{i=1}^{m} l_i\right)^2 \sum_{i=m+1}^{p} l_i^2\right]^{\frac{1}{2}}}{\sqrt{n}\left(\sum_{i=1}^{p} l_i\right)^2}$$

- If the right-hand side is sufficiently small, the investigator may be willing to let the first principal components represent the entire vector of measure- ments.

Testing Equality of the Smallest Roots

- We consider testing the null hypothesis that $\lambda_{m+1} = \cdots = \lambda_p$. That is equivalent to the null hypothesis that

$$H_0 : \Sigma = \Phi + \sigma^2 I,$$

  where Φ is positive semidefinite of rank m.

- The criterion here is the $\frac{1}{2}N$ th power of

$$\frac{\prod_{i=m+1}^{p} l_i}{\left(\sum_{i-m+1}^{p} l_i\right)^{p-m}} (p-m)^{p-m},$$

  which is also the likelihood ratio criterion, but we shall not derive it. [See Anderson (1963a).]

- By taking the logarithm and multipling $-n$, under the null hypothesis, we have

$$T_n = -n \log \prod_{i=m+1}^{p} l_i + n(p-m) \log \frac{\sum_{i=m+1}^{p} l_i}{p-m}$$

  has a limiting χ^2 -distribution with $\frac{1}{2}(p-m+2)(p-m-1)$ degrees of freedom.

- The hypothesis is rejected if T_n is greater than the upper-tailed significance point of the x^2 -distribution.

- If the hypothesis is not rejected, the investigator may consider the last $p-m$ principal components to be composed entirely of error.

§2.2 Factor Analysis

- C. Spearman (1904), General Intelligence, Objectively Determined and Measured, The American Journal of Psychology.
- Children's performance in mathematics (X_1), French (X_2) and English (X_3) was measured. Correlation matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & 0.67 & 0.64 \\ & 1 & 0.67 \\ & & 1 \end{bmatrix}$$

Assume the following model:

$$X_1 = \lambda_1 f + \epsilon_1, \quad X_2 = \lambda_2 f + \epsilon_2, \quad X_3 = \lambda_3 f + \epsilon_3$$

where f is an underlying "common factor" ("general ability"), $\lambda_1, \lambda_2, \lambda_3$ are "factor loadings" and $\epsilon_1, \epsilon_2, \epsilon_3$ are random disturbance terms.

- Model:

$$X_i = \mu_i + \lambda_i f + \epsilon_i, \quad i = 1, 2, 3$$

  with the unobservable factor $f =$ "General ability"

- The variation of ϵ_i consists of two parts:
  - a part that represents the extent to which an individual's matehematics ability, say, differs from her general ability
  - a "measurement error" due to the experimental setup, since examination is only an approximate measure of her ability in the subject

- The relative sizes of $\lambda_i f$ and ϵ_i tell us to which extent variation between individuals can be described by the factor.

- Factor analysis is based on a model in which the observed vector X is partitioned into

$$X = \Lambda f + U + \mu$$

  - an unobserved systematic part $\Lambda f + \mu$ and an unobserved error part U.
  - The components of the error vector U are considered as uncorrelated or independent,
  - while the systematic part Λf is taken as a linear combination of a relatively small number of unobserved factor variables f.

- The analysis separates the effects of the factors, which are of basic interest, from the errors.

- Let the observable vector X be written as

$$X = \Lambda f + U + \mu$$

where

  - $X, U,$ and μ are column vectors of p components,
  - f is a column vector of $m(\leq p)$ components, and
  - Λ is a $p \times m$ matrix.

- We assume that

  - U is distributed independently of f and with mean $\mathbb{E}U = 0$ and
  - covariance matrix $\mathbb{E}UU' = \Psi$, which is diagonal.
  - The vector f will be treated alternatively as a random vector and as a vector of parameters that varies from observation to observation.

- We can distinguish between two kinds of models.
  - In one we consider the vector f to be a random vector, and
  - in the other we consider f to be a vector of nonrandom quantities that varies from one individual to another.
  - In the second case, it is more accurate to write $X_\alpha = \Lambda f_\alpha + U + \mu$.
- In principle,
  - the model with random factors is appropriate when different samples consist of different individuals;
  - the nonrandom factor model is suitable when the specific individuals involved and not just the structure are of interest.

- We will focus on the model of random factors in the following. Recall that the model is

$$X = \Lambda f + U + \mu.$$

There is a fundamental indeterminacy in this model.

  - Let $f^* = C^{-1}f$ and $\Lambda^* = \Lambda C$, then the model can be written as

$$X = \Lambda^* f^* + U + \boldsymbol{\mu}$$

where C is a nonsingular $m \times m$ matrix.

  - The model with Λ and f is equivalent to the model with Λ^* and f^*; that is, by observing X we cannot distinguish between these two models.

- Additional restrictions are needed for the model identification.

- For the model of random factors, the covariance matrix of the observed X is

$$\mathbf{\Sigma} = \mathbf{\Lambda}\Phi\mathbf{\Lambda}' + \mathbf{\Psi} \ .$$

- We impose conditions on $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ to make them just identified. For convenience we suppose that
  - $\Phi = I$ (i.e., the factors are orthogonal or uncorrelated) and that
  - $\mathbf{\Gamma} = \mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ is diagonal.
  - If the diagonal elements of Γ are ordered and different $(\gamma_{11} > \gamma_{22} > \cdots > \gamma_{mm})$, Λ is uniquely determined.

- Note that alternative conditions are that the first m rows of Λ form a lower triangular matrix. (This condition is implied by the so-called centroid method.)

Maximum Likelihood Estimators under Normal Populations [Lawley (1940)].

- For $x_1, \ldots, x_N$ i.i.d. observations on X, the likelihood function is

$$L = (2\pi)^{-\frac{1}{2}pN} |\Sigma|^{-\frac{1}{2}N} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha - \mu)'\,\Sigma^{-1}\,(x_\alpha - \mu)\right]$$

- The maximum likelihood estimator of the mean $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{x} = (1/N)\Sigma_{\alpha=1}^{N}\boldsymbol{x}_\alpha$ Let

$$A = \sum_{\alpha=1}^{N}(x_\alpha - \bar{x})\,(x_\alpha - \bar{x})'$$

- Next we shall maximize

$$-\frac{1}{2}pN\log 2\pi - \frac{1}{2}N\log|\Sigma| - \frac{1}{2}\operatorname{tr} A\Sigma^{-1}$$

- From $\Sigma\Sigma^{-1} = I$, we obtain for any parameter θ

$$\frac{\partial \Sigma^{-1}}{\partial \theta} = -\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta}\Sigma^{-1}$$

- Then the partial derivative of the likelihood function with regard to ψ_{ii}, a diagonal element of Ψ, is $-N/2$ times

$$\sigma^{ii} - \sum_{k,j=1}^{p} c_{kj}\sigma^{ji}\sigma^{ik}$$

where $\boldsymbol{\Sigma}^{-1} = (\sigma^{ij})$ and $(c_{ij}) = C = (1/N)\boldsymbol{A}$.

- Therefore we have

$$\operatorname{diag}\boldsymbol{\Sigma}^{-1} = \operatorname{diag}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1} \quad \text{or} \quad \operatorname{diag}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{C})\boldsymbol{\Sigma}^{-1} = \operatorname{diag}\mathbf{0}.$$

- The derivative with respect to $\lambda_{k\tau}$ is $-N$ times

$$\sum_{j=1}^{p} \sigma^{kj}\lambda_{j\tau} - \sum_{h,g,j=1}^{p} \sigma^{kh}c_{hg}\sigma^{gi}\lambda_{j\tau}, \quad k = 1,\ldots,p, \quad \tau = 1,\ldots,m,$$

which yields

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$$

from which and the identity $\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})\,\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{\Gamma} + \boldsymbol{I})$, we have further

$$\boldsymbol{\Lambda}(\boldsymbol{\Gamma} + \boldsymbol{I}) = C\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} \quad \text{or} \quad (C - \Psi)\Psi^{-1}\Lambda = \Lambda\Gamma$$

- Next we want to show that

$$\Sigma^{-1} - \Sigma^{-1}C\Sigma^{-1} = \Sigma^{-1}(\Sigma - C)\Sigma^{-1} = \Psi^{-1}(\Sigma - C)\Psi^{-1}$$

  when $\boldsymbol{\Sigma^{-1}\Lambda = \Sigma^{-1}C\Sigma^{-1}\Lambda}$.

- Multiply the latter by Σ on the left and on the right to obtain

$$\begin{aligned}\boldsymbol{\Sigma\Psi^{-1}(\Sigma - C)\Psi^{-1}\Sigma} &= \left(\boldsymbol{\Lambda\Lambda' + \Psi}\right)\boldsymbol{\Psi}^{-1}\left(\boldsymbol{\Psi + \Lambda\Lambda' - C}\right)\boldsymbol{\Psi}^{-1}\left(\boldsymbol{\Lambda\Lambda' + \Psi}\right)\\ &= \boldsymbol{\Psi + \Lambda\Lambda' - C}\end{aligned}$$

  becuase

$$\begin{aligned}\boldsymbol{\Lambda\Lambda'\Psi^{-1}}\left(\boldsymbol{\Psi + \Lambda\Lambda' - C}\right) &= \boldsymbol{\Lambda\Lambda' + \Lambda\Gamma\Lambda' - \Lambda\Lambda'\Psi^{-1}C}\\ &= \Lambda\left\{(I + \Gamma)\Lambda' - \Lambda'\Psi^{-1}C\right]\\ &= 0\end{aligned}$$

  by virtue of $\boldsymbol{\Lambda(\Gamma + I) = C\Psi^{-1}\Lambda}$. Thus

$$\boldsymbol{\Sigma^{-1}(\Sigma - C)\Sigma^{-1} = \Psi^{-1}(\Sigma - C)\Psi^{-1}}$$

- Then we get diag $\Psi^{-1}(\Sigma - C)\Psi^{-1} =$ diag 0. since Ψ is diagonal this equation is equivalent to

$$\text{diag}\left(\Lambda\Lambda' + \Psi\right) = \text{diag}\, C$$

- The estimators $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$ are determined by

$$\mathbf{\Lambda}(\mathbf{\Gamma} + \mathbf{I}) = C\mathbf{\Psi}^{-1}\mathbf{\Lambda}$$
$$\text{diag}\left(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}\right) = \text{diag}\, C$$

and the requiremert that $\mathbf{\Gamma} = \mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ is diagonal.

Asymptotic Distributions of the Estimators

- If
  - Λ and Ψ are identified by $\Lambda'\Psi^{-1}\Lambda$ being diagonal,
  - the diagonal elements are different and ordered, and
  - $C \xrightarrow{p} \Psi + \Lambda\Lambda'$,

  then

$$\hat{\Psi} \xrightarrow{p} \Psi \quad \text{and} \quad \hat{\Lambda} \xrightarrow{p} \Lambda$$

- A sufficient condition for $C \to \Sigma$ is that $(f' \; U')'$ has a distribution with finite second-order moments.

- Let
$$(\theta_{ij}) = \Theta = \Psi - \Lambda\left(\Lambda'\Psi^{-1}\Lambda\right)^{-1}\Lambda'$$
If
  - $\left(\theta_{ij}^2\right)$ is nonsingular,
  - $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are identified by the condition that $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is diagonal and the diagonal elements are different and ordered,
  - $C \xrightarrow{p} \Psi + \boldsymbol{\Lambda}\Lambda'$, and
  - $\sqrt{N}(\boldsymbol{C} - \boldsymbol{\Sigma})$ has a limiting normal distribution,

- then
$$\sqrt{N}(\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}) \quad \text{and} \quad \sqrt{N}(\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi})$$
have a limiting normal distribution.

- For example, $\sqrt{N}(C - \Sigma)$ will have a limiting distribution if $(f' \;\; U')'$ has a distribution with finite fourth moments.

Test of the Hypothesis That the Model Fits

- We shall derive the likelihood ratio test that the model fits; that is, that for a specified m the covariance matrix can be written as $\Sigma = \Psi + \Lambda\Lambda'$ for some diagonal positive definite Ψ and some $p \times m$ matrix Λ.

- To evaluate the maximized likelihood function we calculate

$$\begin{aligned}
\operatorname{tr} C\hat{\boldsymbol{\Sigma}}^{-1} &= \operatorname{tr} C\hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\Sigma}} - \hat{\Lambda}\hat{\Lambda}'\right)\hat{\psi}^{-1} \\
&= \operatorname{tr}\left[C\hat{\psi}^{-1} - \left(C\hat{\Sigma}^{-1}\hat{\Lambda}\right)\hat{\Lambda}'\hat{\psi}^{-1}\right] \\
&= \operatorname{tr}\left[C\hat{\psi}^{-1} - \hat{\Lambda}\hat{\Lambda}'\hat{\psi}^{-1}\right] \\
&= \operatorname{tr}\left[\left(\hat{\Lambda}\hat{\Lambda}' + \hat{\psi}\right)\hat{\psi}^{-1} - \hat{\Lambda}\hat{\Lambda}'\hat{\psi}^{-1}\right] \\
&= p
\end{aligned}$$

  where we have used $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$.

- Next we find

$$
\begin{aligned}
|\hat{\mathbf{\Sigma}}| &= |\hat{\Lambda}\hat{\Lambda} + \hat{\Psi}| \\
&= |\hat{\Psi}^{\frac{1}{2}}| \cdot |\hat{\Psi}^{-\frac{1}{2}}\hat{\Lambda}\hat{\Lambda}\hat{\Psi}^{-\frac{1}{2}} + I_p| \cdot |\hat{\Psi}^{\frac{1}{2}}| \\
&= |\hat{\Psi}| \cdot \left|\hat{\Lambda}'\hat{\Psi}^{-\frac{1}{2}}\hat{\mathbf{\Psi}}^{-\frac{1}{2}}\hat{\Lambda} + I_m\right| \\
&= |\hat{\Psi}| \cdot \left|\hat{\Gamma} + I_m\right| \\
&= \prod_{i=1}^{p} \hat{\psi}_{ii} \prod_{j=1}^{m} (\hat{\gamma}_j + 1)
\end{aligned}
$$

- Then using the identity

$$
\mathbf{\Psi}^{-\frac{1}{2}}(C - \mathbf{\Psi})\Psi^{-\frac{1}{2}} \left(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Lambda}\right) = \left(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Lambda}\right)\mathbf{\Gamma}
$$

we have

$$
\frac{|C|}{|\hat{\Psi}|} = \prod_{i=1}^{p}(1 + \hat{\gamma}_i) \quad \text{and thus} \quad |\hat{\Sigma}| = \frac{|C|\Pi_{j\in S}(1 + \hat{\gamma}_j)}{\Pi_{i=1}^{p}(1 + \hat{\gamma}_i)} = \frac{|C|}{\Pi_{j\notin S}(1 + \hat{\gamma}_j)}
$$

where S is the set of indices corresponding to the roots in $\hat{\Gamma}$.

- The logarithm of the maximized likelihood function is

$$-\frac{1}{2}pN\log 2\pi - \frac{1}{2}N\log|C| - \frac{1}{2}N\sum_{j\notin S}\log\left(1+\hat{\gamma}_j\right) - \frac{1}{2}Np$$

- The likelihood ratio criterion is

$$\frac{\max_{\mu.\Lambda,\Psi} L\left(\mu,\Psi+\Lambda\Lambda'\right)}{\max_{\mu,\Sigma} L(\boldsymbol{\mu},\boldsymbol{\Sigma})} = \frac{|C|^{\frac{1}{2}N}}{|\hat{\psi}+\hat{\Lambda}\hat{\Lambda}|^{\frac{1}{2}\cdot N}} = \prod_{j=m+1}^{p}\left(1+\hat{\gamma}_j\right)^{\frac{1}{2}N}$$

  We can use -2 times the logarithm of the likelihood ratio criterion:

$$-N\sum_{j=m+1}^{p}\log\left(1+\hat{\gamma}_j\right)$$

  and reject the null hypothesis if the statistic is too large.

- If the regularity conditions, the LRT statistic under the null hypothesis is χ^2 with degrees of freedom $\frac{1}{2}\left[(p-m)^2-p-m\right]$, which is the number of elements of Σ plus the number of identifying restrictions minus the number of parameters in $\boldsymbol{\Psi}$ and Λ.

§1.3 Spiked Population Covariance Model in High dimensions

- Independence components Model
  - Assumption (a). The sample and population sizes n, p both tend to infinity, and in such a way that $c_n := p/n \to c \in (0, \infty)$.
  - Assumption (b). The population is

    $$\mathbf{x} = \Sigma_p^{1/2} \mathbf{z}$$

    where $\mathbf{z} = (z_1, \ldots, z_p)'$ is a set of i.i.d. random varibles with finite fourth moments.
  - Assumption (c). The PSD H_p of Σ_p weakly converges to a probability distribution H, as $p \to \infty$.

- High-dimensional results for $\mathrm{E}(\mathbf{x}) = 0, \mathrm{Cov}(\mathbf{x}) = \mathbf{I}_p$:
  - $p/n \to c > 0$.
  - We have that
    $$F_n \xrightarrow[w]{a.s.} \nu \text{ (MP law).}$$
  - $\nu(dx) = f(x)dx + (1 - 1/c)\delta_0(dx)1_{\{c>1\}}$ where
    $$f(x) = \frac{\sqrt{(b-x)(x-a)}}{2\pi cx} 1_{[a,b]}(x),$$
    where $a = (1 - \sqrt{c})^2$ and $b = (1 + \sqrt{c})^2$.

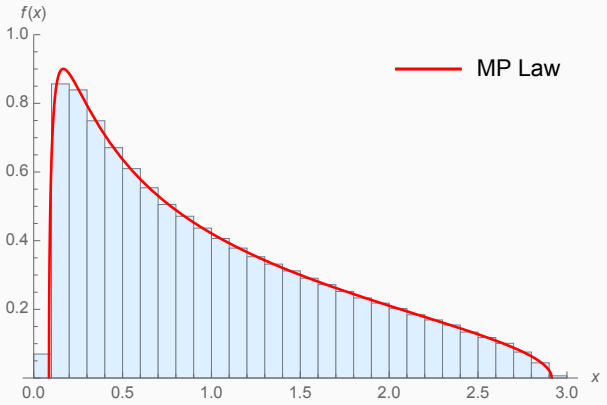


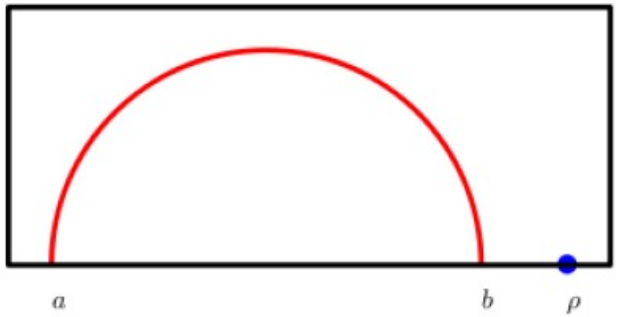
Figure 2: The Marčenko-Pastur law (red line). The dimensions are $(p, n, c) = (500, 1000, 0.5)$ and $rep = 100$.

- Considering Johnstone's spiked population model (Johnstone, 2001) with r spikes $\tau_1 > \tau_2 > \cdots > \tau_r > 1$, the covariance matrix can be represented as

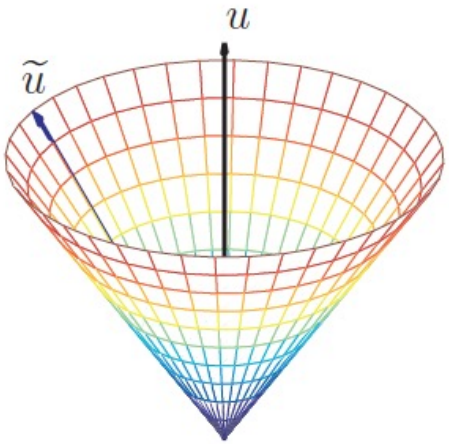
$$\Sigma_p = \sum_{j=1}^{r} (\tau_j - 1)\, \mathbf{u}_j \mathbf{u}_j' + I_p$$

  where $\mathbf{u}_i$'s are unit vectors satisfying $\mathbf{u}_i' \mathbf{u}_j = 0$ for $1 \leq i \neq j \leq m$.

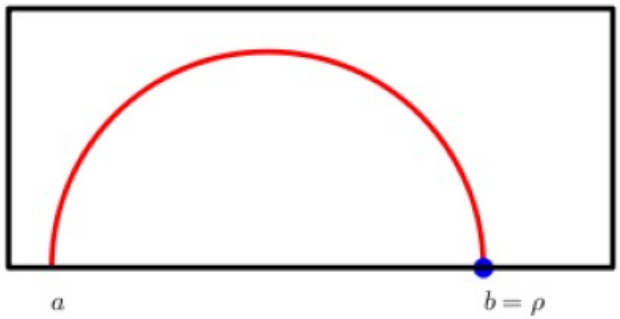
- For the simpliest case where $r = 1$, our aim is to finds the limits of the largest sample eigenvalue and its associated eigenvector.



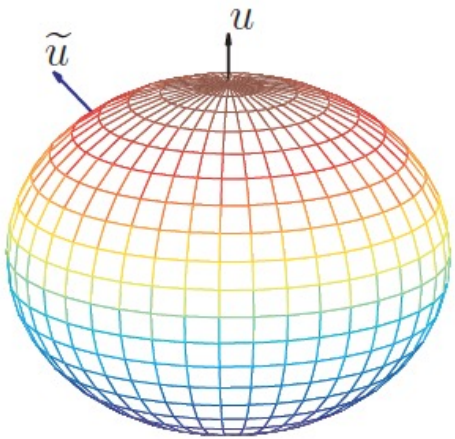
(a) Largest eigenvalue $\rho > b$ in blue when $\theta > \theta_c$



(b) Associated eigenvector when $\theta > \theta_c$



(c) Largest eigenvalue $\rho = b$ in blue when $\theta \leq \theta_c$



(d) Associated eigenvector when $\theta \leq \theta_c$

Eigenvalue phase transition

- Let
$$\psi(x) = x + c\int \frac{tx}{x-t} dH(t)$$
For a spike eigenvalue $\tau_j \in G_1$, the corresponding sample eigenvalue λ_j of S_n converges
$$\lambda_j \xrightarrow{a.s.} \begin{cases} \psi(\tau_j) & \psi'(\tau_j) > 0 \\ \gamma_j & \psi'(\tau_j) \leq 0 \end{cases}$$
where γ_j satisfies $F_{c,H}(\gamma_j) = H(-\infty, \tau_j)$

§1.4 D. Passemier, Z. Li and J. Yao. (2017). On estimation of the noise variance in high dimensional probabilistic principal component analysis, J. R. Statist. Soc. B 79, Part 1, pp. 51–67.

Probabilistic principal component analysis

- The observation vectors $\{\mathbf{x}_i\}_{1\leqslant i\leqslant n}$ are p dimensional and satisfy the equation

$$\mathbf{x}_i = \mathbf{\Lambda}\mathbf{f}_i + \mathbf{e}_i + \boldsymbol{\mu}, \quad i = 1, \ldots, n$$

  Here,

  - $\mathbf{f}_i$ are m -dimensional *PCs* with $m \ll p$,
  - Λ is a $p \times m$ matrix of loadings and
  - μ represents the general mean and
  - $\{\mathbf{e}_i\}_{1\leqslant i\leqslant n}$ are a sequence of independent errors with covariance matrix $\mathbf{\Psi} = \sigma^2\mathbf{I}_p$. The parameter σ^2 is the unknown noise variance.

- The parameter σ^2 is the unknown noise variance.

- To ensure the identification of the model, constraints must be introduced on the parameters.
- A traditional choice is (Anderson (2003), chapter 14)
  - (a) $\mathbb{E}(\mathbf{f}_i) = \mathbf{0}$ and $\mathbb{E}(\mathbf{f}_i \mathbf{f}_i') = \mathbf{I}$, and
  - (b) the matrix $\mathbf{\Gamma} := \mathbf{\Lambda}'\mathbf{\Lambda}$ is diagonal with distinct diagonal elements.
- Therefore, the population covariance matrix of $\{\mathbf{x}_i\}_{1 \leqslant i \leqslant n}$ is

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \sigma^2 \mathbf{I}.$$

Finding a reliable estimator of the noise variance σ^2 is a non-trivial issue for high dimensional data which we now pursue.

- Let $\bar{\mathbf{x}}$ be the sample mean and define the sample covariance matrix

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})'$$

  Let $\lambda_{n,1} \geqslant \lambda_{n,2} \geqslant \ldots \geqslant \lambda_{n,p}$ be the eigenvalues of $\mathbf{S}_n$.

- Under the normality assumption on both $\{\mathbf{f}_i\}$ and $\{\mathbf{e}_i\}$, the maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_{n,i}$$

- In the classic setting where the dimension p is fixed, as $n \to \infty$

$$\sqrt{n} \left(\hat{\sigma}^2 - \sigma^2 \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, s^2 \right), \quad s^2 = \frac{2\sigma^4}{p-m}$$

- The PPCA model is a spiked population model (Johnstone, 2001) since the eigenvalues of the population covariance matrix Σ are

$$\begin{aligned}\text{spec}(\mathbf{\Sigma}) &= (\alpha_1, \ldots, \alpha_m, \underbrace{0, \ldots, 0}_{p-m}) + \sigma^2(\underbrace{1, \ldots, 1}_{p}) \\ &= \sigma^2(\alpha_1^*, \ldots, \alpha_m^*, \underbrace{1, \ldots, 1}_{p-m})\end{aligned}$$

  where $\{\alpha_i\}$ are m non-null eigenvalues of $\mathbf{\Lambda\Lambda}'$ and the notation $\alpha_i^* = \alpha_i/\sigma^2 + 1$ is used.

- Let
$$\phi(\alpha) = \alpha + \frac{c\alpha}{\alpha - 1}, \quad \alpha \neq 1$$
Following Baik and Silverstein (2006), assume that $\alpha_1^* \geqslant \ldots \geqslant \alpha_m^* > 1 + \sqrt{c}$, i.e. all the eigenvalues α_i are greater than $\sigma^2\sqrt{c}$.

- It is then known that, for the spiked sample eigenvalues $\{\lambda_{n,i}\}_{1 \leqslant i \leqslant m}$ of $\mathbf{S}_n$, almost surely,
$$\lambda_{n,i} \to \sigma^2 \phi\left(\alpha_i^*\right) = \psi\left(\sigma^2, c, \alpha_i\right) = \alpha_i + \sigma^2 + \sigma^2 c\left(1 + \frac{\sigma^2}{\alpha_i}\right)$$

- In addition, the CLT for the spiked eigenvalues was established in Bai and Yao (2008):
$$\sqrt{n}\left\{\lambda_{n,i} - \sigma^2 \phi\left(\alpha_i^*\right)\right\}$$
is asymptotically Gaussian.

- Consider the PPCA model with population covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \sigma^2 \mathbf{I}_p$$

where both the PCs and the noise are Gaussian.

- Assume that $p \to \infty, n \to \infty$ and $c_n = p/(n-1) \to c > 0$, and the non-null eigenvalues of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ $\{\alpha_i\}$ satisfy $\alpha_i \geqslant \sigma^2\sqrt{c}(1 \leqslant i \leqslant m)$

- Then, we have

$$\frac{p-m}{\sigma^2\sqrt{(2c}}\left(\hat{\sigma}^2 - \sigma^2\right) + b\left(\sigma^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

where $b\left(\sigma^2\right) = \sqrt{(c/2}\left\{m + \sigma^2 \sum_{i=1}^{m}(1/\alpha_i)\right\}$

- As the bias depends on σ^2 which we want to estimate, a natural correction is to use the plug-in estimator

$$\hat{\sigma}_*^2 = \hat{\sigma}^2 + \frac{b\left(\hat{\sigma}^2\right)}{p-m}\hat{\sigma}^2\sqrt{2c_n}$$

  This estimator will be hereafter referred as the bias-corrected estimator.

- We have

$$\frac{p-m}{\sigma^2\sqrt{2c_n}}\left(\hat{\sigma}_*^2 - \sigma^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

Monte Carlo experiments

- We first check by simulation the effect of bias correction obtained in $\hat{\sigma}_*^2$ and its asymptotic normality.
- Independent Gaussian samples of size n are considered in three different settings:
  - model 1 $\text{spec}(\Sigma) = (25, 16, 9, 0, \ldots, 0) + \sigma^2(1, \ldots, 1), \sigma^2 = 4, c = 1$
  - model 2 $\text{spec}(\Sigma) = (4, 3, 0, \ldots, 0) + \sigma^2(1, \ldots, 1), \sigma^2 = 2, c = 0.2$
  - model 3 $\text{spec}(\Sigma) = (12, 10, 8, 8, 0, \ldots, 0) + \sigma^2(1, \ldots, 1), \sigma^2 = 3, c = 1.5$

Model	*p*	*n*	*Empirical bias*	*Theoretical bias*	*|Difference|*
1	100	100	−0.1556	−0.1589	0.0024
	400	400	−0.0391	−0.0388	0.0003
	800	800	−0.0197	−0.0193	0.0003
2	20	100	−0.0625	−0.0704	0.0052
	80	400	−0.0166	−0.0162	0.0027
	200	1000	−0.0064	−0.0064	0.0011
3	150	100	−0.1609	−0.1634	0.0025
	600	400	−0.0401	−0.0400	0.0001
	1500	1000	−0.0161	−0.0159	0.0002

Figure 3: Comparison between the empirical and the theoretical bias

- Next, we compare our bias-corrected estimator $\hat{\sigma}^2_*$ with the $\mathrm{MLE}\hat{\sigma}^2$ and other three existing estimators in the literature.

- The estimator $\hat{\sigma}^2_{\mathrm{KN}}$ of Kritchman and Nadler (2008) is defined as the solution of the following non-linear system of $m+1$ equations involving the $m+1$ unknowns $\hat{\rho}_1, \ldots, \hat{\rho}_m$ and $\hat{\sigma}^2_{\mathrm{KN}}$

$$\hat{\sigma}^2_{\mathrm{KN}} - \frac{1}{p-m}\left\{\sum_{j=m+1}^{p}\lambda_{n,j} + \sum_{j=1}^{m}(\lambda_{n,j} - \hat{\rho}_j)\right\} = 0$$

and

$$\hat{\rho}_j^2 - \hat{\rho}_j\left(\lambda_{n,j} + \hat{\sigma}^2_{\mathrm{KN}} - \hat{\sigma}^2_{\mathrm{KN}}\frac{p-m}{n}\right) + \lambda_{n,j}\hat{\sigma}^2_{\mathrm{KN}} = 0, \quad j = 1, \ldots, m$$

- The estimator $\hat{\sigma}^2_{\mathrm{US}}$ of Ulfarsson and Solo (2008) is defined as the ratio

$$\hat{\sigma}^2_{\mathrm{US}} = \frac{\text{median}\left(\lambda_{n,m+1}, \ldots, \lambda_{n,p}\right)}{m_{p/n,1}}$$

  where $m_{\alpha,1}$ is the median of the Marčenko-Pastur distribution $F_{\alpha,1}$.

- The estimator $\hat{\sigma}^2_{\text{median}}$ of Johnstone and Lu(2009) is defined as the median of the p sample variances (the data $\{x_{ij}\}$ are assumed centred):

$$\hat{\sigma}^2_{\text{median}} = \text{median}\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2, \quad 1 \leqslant j \leqslant p\right)$$

Model	p	n	σ^2	$\hat{\sigma}^2$	$\hat{\sigma}^2_{\text{KN}}$	$\hat{\sigma}^2_{\text{US}}$	$\hat{\sigma}^2_{\text{median}}$
1	100	100	4	7.8232	1.0130	14.6394	1.5085
	400	400		8.5905	0.9980	25.5941	1.6429
	800	800		8.1162	1.0019	39.9444	1.6639
2	20	100	2	1.7045	1.0220	2.4980	1.5926
	80	400		2.0406	1.0045	3.8686	1.5433
	200	1000		1.9729	1.0011	3.8731	1.5427
3	150	100	3	19.2114	1.2292	41.7319	1.4274
	600	400		20.8471	0.9958	48.3130	1.6096
	1500	1000		21.6207	1.0001	51.9302	1.8071

Figure 4: Comparison between four existing estimators and the proposed estimators in terms of ratios of MSEs

- In Bai and Ng (2002), three criteria are applicable to PPCA models

$$\mathrm{PC}_j(m) = V\left(m, \hat{F}^m\right) + m \cdot \hat{\sigma}^2_{\mathrm{BN}} \cdot g_j(N, T), \quad j \in \{1, 2, 3\}$$

where

$$V\left(m, \hat{F}^m\right) = (NT)^{-1} \Sigma_{i=1}^{N} \hat{\mathbf{e}}_i' \hat{\mathbf{e}}_i$$

and $g_j(N, T)$ denote the penalty functions

$$g_1(N, T) = \frac{N+T}{NT} \ln\left(\frac{NT}{N+T}\right)$$
$$g_2(N, T) = \frac{N+T}{NT} \ln(\tilde{N})$$
$$g_3(N, T) = \frac{\ln(\tilde{N})}{\tilde{N}}$$

with $\tilde{N} = \min\{N, T\}$. (Here we use T denotes the dimension p.)

- The corresponding estimators of the number of PCs are

$$\hat{m}_j = \arg \min_{0 \leqslant m \leqslant m_0} \mathrm{PC}_j(m), \quad j \in \{1, 2, 3\},$$

where m_0 is a predetermined maximum value of m.

- We substitute $\hat{\sigma}_*^2$ for empirical $V\left(m, \hat{F}^m\right)$ and $\hat{\sigma}_{\mathrm{BN}}^2$ in the criteria PC_j s. The modified criteria and estimators using $\hat{\sigma}_*^2(m)$ are thus

$$\mathrm{PC}_j^*(m) = \hat{\sigma}_*^2(m) + m\hat{\sigma}_*^2(m_0)\, g_j(N, T)$$

and

$$\hat{m}_j^* = \arg \min_{0 \leqslant m \leqslant m_0} \mathrm{PC}_j^*(m), \quad j \in \{1, 2, 3\}$$

- We introduce a new penalty function

$$g(N, T) = \frac{(c + 2\sqrt{c})\left(1 + T/N^{1+\delta}\right)}{N}$$

and define the new criterion

$$\mathrm{PC}^*(m) = \hat{\sigma}_*^2(m) + m\hat{\sigma}_*^2(m_0)\, g(N, T)$$

Here $\delta > 0$ is a small prefixed constant.

N	T	PC^*	PC_1^*	PC_2^*	PC_3^*	PC_1	PC_2	PC_3
100	40	1.00	1.00	1.00	1.00	1.17 (0.37)	1.01 (0.10)	3.78 (0.75)
100	60	1.00	1.00	1.00	1.00	1.00	1.00	3.63 (0.76)
200	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1000	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2000	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	100	1.00	1.00	1.00	1.00	1.00	1.00	5.36 (0.80)
40	100	1.00	1.00	1.00	1.00	1.79 (0.72)	1.19 (0.40)	4.91 (0.90)
60	100	1.00	1.00	1.00	1.00	1.01 (0.08)	1.00	4.30 (0.85)
60	200	1.00	1.00	1.00	1.00	1.00	1.00	1.02 (0.16)
10	50	7.19 (1.92)	8.00	8.00	8.00	8.00	8.00	8.00
10	100	4.95 (3.13)	8.00	8.00	8.00	8.00	8.00	8.00
20	100	1.00 (0.03)	1.01 (0.15)	1.01 (0.12)	1.08 (0.53)	6.96 (0.88)	6.35 (0.98)	7.84 (0.40)
100	10	2.10 (2.52)	1.08 (0.73)	1.03 (0.50)	1.15 (1.01)	8.00	8.00	8.00
100	20	1.00	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	5.88 (0.76)	5.12 (0.77)	7.35 (0.63)

Figure 5: Comparison between $\mathrm{PC}^*, \mathrm{PC}_j^*$ and PC_j for $m = 1$ and $\theta = 1$

N	T	PC^*	PC_1^*	PC_2^*	PC_3^*	PC_1	PC_2	PC_3
100	40	5.00	4.91 (0.30)	4.81 (0.41)	4.99 (0.11)	5.00 (0.03)	5.00	5.59 (0.57)
100	60	5.00	5.00 (0.04)	4.99 (0.11)	5.00	5.00	5.00	5.58 (0.57)
200	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00
500	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00
1000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00
2000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00
100	100	5.00	5.00	5.00	5.00	5.00	5.00	6.84 (0.65)
40	100	4.98 (0.14)	4.97 (0.17)	4.92 (0.27)	5.00 (0.04)	5.02 (0.12)	5.00	6.22 (0.66)
60	100	5.00	5.00 (0.04)	4.99 (0.08)	5.00	5.00	5.00	6.03 (0.64)
60	200	5.00	5.00	5.00	5.00	5.00	5.00	6.03 (0.03)
10	50	7.47 (1.00)	8.00	8.00	8.00	8.00	8.00	8.00
10	100	5.77 (1.53)	8.00	8.00	8.00	8.00	8.00	8.00
20	100	3.74 (0.84)	4.74 (0.51)	4.62 (0.57)	4.92 (0.45)	7.11 (0.63)	6.65 (0.64)	7.85 (0.37)
100	10	6.66 (1.97)	4.59 (1.99)	4.35 (1.91)	4.88 (2.09)	8.00	8.00	8.00
100	20	4.68 (0.51)	3.86 (0.79)	3.69 (0.81)	4.13 (0.73)	6.74 (0.63)	6.19 (0.62)	7.77 (0.43)

Figure 6: Comparison between $\mathrm{PC}^*, \mathrm{PC}_j^*$ and PC_j for $m = 5$ and $\theta = 3$

Real data analysis

- The first data set contains stock returns. Following Bai and Ng (2002),
  - we extract data from the Center for Research in Security Prices US stock database by using the monthly returns for all common stocks listed in the New York Stock Exchange, Amex and Nasdaq over 20 years (from January 1991 to December 2010).
  - Stocks that do not trade for a cumulative 2 years during the period have been deleted.
- The final data set includes 1913 stocks with 240 monthly returns for each of them ($T = 240; N = 1913$).

- The second data set is the functional magnetic resonance imaging data set. This data set is from http://afni.nimh.nih.gov/afni/ .
  - A human brain was scanned when the person performed finger-thumb opposition.
  - There are $T = 124$ observations on 21 brain slices. We pick out one brain slice and keep only the variables (pixels) that significantly corresponded to brain tissue, so that $N = 1126$ variables are selected.
  - We transform both data sets so that each series has mean 0.

Data set	PC^*	PC_1^*	PC_2^*	PC_3^*	PC_1	PC_2	PC_3
1 ($m_0 = 15$)	2	2	2	2	4	4	6
2 ($m_0 = 20$)	13	18	18	19	20	20	20

Figure 7: Comparison between the modified and the original criteria

Application to the goodness-of-fit test of a probabilistic principal component analysis model

- We consider the following goodness-of-fit test for the PPCA model. The null hypothesis is

$$\mathcal{H}_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \sigma^2 \mathbf{I}_p$$

  where the number of PCs m is specified.

- Following Anderson and Rubin (1956), the LRT statistic is

$$T_n = -nL^*, \quad L^* = \sum_{j=m+1}^{p} \log\left(\frac{\lambda_{n,j}}{\hat{\sigma}^2}\right)$$

  and $\hat{\sigma}^2$ is the MLE of the variance.

- Keeping p fixed while letting $n \to \infty$, classical low dimensional theory states that

$$T_n \to \chi^2_q,$$

  where $q = p(p+1)/2 + m(m-1)/2 - pm - 1$

- Assume that $p/c \to c \in (0,1)$, then we have

$$v(c)^{-1/2}\left\{L^* - m(c) - ph\left(c_n\right) + \eta + (p-m)\log(\beta)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

where

$$\begin{aligned}
m(c) =& \frac{\log(1-c)}{2} \\
h\left(c_n\right) =& \frac{c_n - 1}{c_n}\log\left(1-c_n\right) - 1 \\
\eta =& \sum_{i=1}^{m}\log\left(1 + c\sigma^2\alpha_i^{-1}\right) \\
\beta =& 1 - \frac{c}{p-m}\left(m + \sigma^2\sum_{i=1}^{m}\alpha_i^{-1}\right) \\
v(c) =& -2\log(1-c) + \frac{2c}{\beta}\left(\frac{1}{\beta} - 2\right)
\end{aligned}$$

- We present some simulation experiments to compare the classical LRT and the CLRT.
- We consider again models 1 and 2 that were described before and a new model (model 4):
  - model 1, $\text{spec}(\Sigma) = (25, 16, 9, 0, \ldots, 0) + \sigma^2(1, \ldots, 1), \sigma^2 = 4, c = 0.9$
  - model 2 $\text{spec}(\Sigma) = (4, 3, 0, \ldots, 0) + \sigma^2(1, \ldots, 1), \sigma^2 = 2, c = 0.2$
  - model 4 $\text{spec}(\Sigma) = (8, 7, 0, \ldots, 0) + \sigma^2(1, \ldots, 1), \sigma^2 = 1,$ varying c

Model	*p*	*n*	*Empirical size of CLRT*	*Empirical size of LRT*
1	90	100	0.0522	0.9997
	180	200	0.0515	1.0000
	720	800	0.0483	1.0000
2	20	100	0.0375	0.0321
	80	400	0.0440	0.0368
	200	1000	0.0481	0.0514
4	5	500	0.0122	0.0475
	10	500	0.0217	0.0482
	50	500	0.0421	0.0419
	100	500	0.0438	0.0424
	200	500	0.0498	0.2216
	250	500	0.0501	0.7416
	300	500	0.0461	0.9991

Figure 8: Comparison of the empirical size of the LRT and CLRT in various settings