Factor Analysis and Singularities

Robin Evans, October 2009

www.stat.washington.edu/˜rje42

Motivation

- Dimensionality reduction.
- Measuring genuine hidden variables.

Motivation

- Dimensionality reduction.
- Measuring genuine hidden variables.

Example. (Test scores)

Suppose 250 students in a school take tests in 12 different subjects.

Could assume normality: model as 12-variate normal distribution, measure means, variances, correlations, etc.

What does this tell us? Perhaps scores can be modelled in a simpler way?

Motivation

- Dimensionality reduction.
- Measuring genuine hidden variables.

Example. (Anxiety Data)

$n = 335$ male subjects in BC asked $p = 20$ questions about exam stress.

$x_1 =$ 'lack of confidence during tests'.
$x_2 =$ 'uneasy, upset feeling'.
$\vdots \qquad \vdots$
$x_{20} =$ 'nervous during tests, forget facts'.

Much similarity in questions, so unsurprisingly much correlation.

Example: Anxiety Data

$$\boldsymbol{S} = \begin{pmatrix} 1.000 & 0.510 & 0.240 & 0.423 & 0.307 & 0.286 & & 0.380 \\ 0.510 & 1.000 & 0.296 & 0.407 & 0.232 & 0.336 & & 0.326 \\ 0.240 & 0.296 & 1.000 & 0.368 & 0.404 & 0.271 & & 0.294 \\ 0.423 & 0.407 & 0.368 & 1.000 & 0.347 & 0.342 & \cdots & 0.530 \\ 0.307 & 0.232 & 0.404 & 0.347 & 1.000 & 0.338 & & 0.300 \\ 0.286 & 0.336 & 0.271 & 0.342 & 0.338 & 1.000 & & 0.405 \\ & & & \vdots & & & \ddots & \\ 0.380 & 0.326 & 0.294 & 0.530 & 0.300 & 0.405 & & 1.000 \end{pmatrix}$$

Example: Anxiety Data

$$\boldsymbol{S} = \begin{pmatrix} 1.000 & 0.510 & 0.240 & 0.423 & 0.307 & 0.286 & & 0.380 \\ 0.510 & 1.000 & 0.296 & 0.407 & 0.232 & 0.336 & & 0.326 \\ 0.240 & 0.296 & 1.000 & 0.368 & 0.404 & 0.271 & & 0.294 \\ 0.423 & 0.407 & 0.368 & 1.000 & 0.347 & 0.342 & \cdots & 0.530 \\ 0.307 & 0.232 & 0.404 & 0.347 & 1.000 & 0.338 & & 0.300 \\ 0.286 & 0.336 & 0.271 & 0.342 & 0.338 & 1.000 & & 0.405 \\ & & & \vdots & & & \ddots & \\ 0.380 & 0.326 & 0.294 & 0.530 & 0.300 & 0.405 & & 1.000 \end{pmatrix}$$

Suppose $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Use $\boldsymbol{S}$ to estimate $\boldsymbol{\Sigma}$. And then...?

The Model

Latent variables (*factors*) $y_1, \ldots, y_q$.

Factor loadings $\boldsymbol{\Lambda} = (\lambda_{ij})$, $i = 1, \ldots, p$, $j = 1, \ldots, q$.

$$
\begin{aligned}
x_1 &= \mu_1 + \lambda_{11} y_1 + \cdots + \lambda_{1q} y_q + \epsilon_1 \\
x_2 &= \mu_2 + \lambda_{21} y_1 + \cdots + \lambda_{2q} y_q + \epsilon_2 \\
&\vdots \\
x_p &= \mu_p + \lambda_{p1} y_1 + \cdots + \lambda_{pq} y_q + \epsilon_p
\end{aligned}
$$

Where $y_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1)$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \boldsymbol{\Psi})$ ($\boldsymbol{\Psi}$ diagonal).

The Model

In matrix notation:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda} \boldsymbol{y} + \boldsymbol{\epsilon}$$

Where

$$\boldsymbol{y} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \boldsymbol{I})$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(0, \boldsymbol{\Psi})$$

The Model

In matrix notation:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y} + \boldsymbol{\epsilon}$$

Where

$$\boldsymbol{y} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \boldsymbol{I})$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(0, \boldsymbol{\Psi})$$

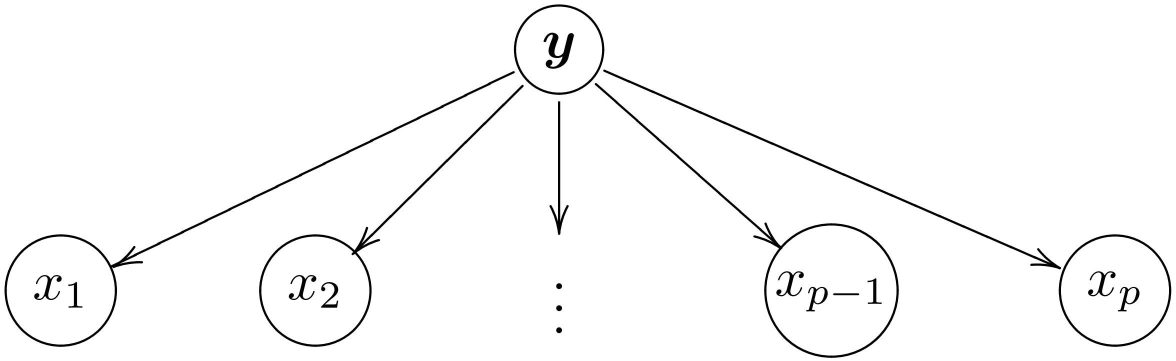
So

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y}, \boldsymbol{\Psi})$$

and unconditionally:

$$\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T)$$

The Model



Conditional independence representation; $\boldsymbol{y}$ here is still a vector.

Rotations

So, instead of taking $\hat{\mathbf{\Sigma}} = \mathbf{S}$ we can try to find $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$ instead. Then $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Psi}} + \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T$.

This restricts the space and so lowers the dimension of the model.

Rotations

So, instead of taking $\hat{\mathbf{\Sigma}} = \mathbf{S}$ we can try to find $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$ instead. Then $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Psi}} + \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T$.

This restricts the space and so lowers the dimension of the model.

One problem: let U be $q \times q$ orthogonal ($U^T U = I$) and notice that

$$\begin{aligned}(\mathbf{\Lambda} U)(\mathbf{\Lambda} U)^T &= \mathbf{\Lambda} U U^T \mathbf{\Lambda}^T \\ &= \mathbf{\Lambda}\mathbf{\Lambda}^T,\end{aligned}$$

so we cannot distinguish between rotations of $\mathbf{\Lambda}$! This amounts to changing the basis of our latent variables.

In practice people choose a rotation to aid interpretation (Abdi, 2003).

Rotations

One choice is to require that $\mathbf{\Gamma} = \mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda}$ is diagonal. Assuming that the entries of $\mathbf{\Psi}$ are distinct, this will prevent arbitrary rotations.

This interpretation means that the components of $\boldsymbol{y}$ are independent conditional on $\boldsymbol{x}$.

Rotations

One choice is to require that $\mathbf{\Gamma} = \mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda}$ is diagonal. Assuming that the entries of $\mathbf{\Psi}$ are distinct, this will prevent arbitrary rotations.

This interpretation means that the components of $\boldsymbol{y}$ are independent conditional on $\boldsymbol{x}$.

Another method is to try and create zeroes (or small values) in $\mathbf{\Lambda}$. This can aid interpretation.

Similarly, we can minimise some criterion; e.g. varimax, oblimin. See example later.

Fitting

Generally use maximum likelihood estimation to find $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$.

EM-algorithm is conceptually simpler.

Pick a starting value for $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Lambda}}$, then iterate:

1. **E-step** — calculate $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{\Sigma}}]$;
2. **M-step** — estimate $\hat{\boldsymbol{\Sigma}}$ using 'complete' data;

Likelihood is guaranteed to increase at each iteration.

Command `factanal()` in `R` finds MLE.

Example: Anxiety Data

Loadings for two factors as fitted by `factanal()`.

Unrotated

var	Factor 1	Factor 2
1	0.62	−0.07
2	0.62	−0.16
3	0.54	0.25
4	0.65	0.09
5	0.51	0.50
6	0.49	0.20
7	0.68	0.29

OBLIMIN

var	Factor 1	Factor 2
1	0.57	0.09
2	0.66	−0.04
3	0.16	0.48
4	0.42	0.31
5	−0.13	0.80
6	0.17	0.40
7	0.23	0.56

Testing

It will be crucial to know whether q independent normal random variables are sufficient to describe the model.

Testing

It will be crucial to know whether q independent normal random variables are sufficient to describe the model.

Let $\Theta_q = \{\mathbf{\Sigma} = \mathbf{\Psi} + \mathbf{\Lambda}\mathbf{\Lambda}^T \; : \; \mathbf{\Lambda} \in M_{p\times q}, \mathbf{\Psi} \in M_{p\times p}^{\text{diag}+}\}$, meaning the set of covariance matrices formed by at most q factors.

Suppose we wish to test

$$H_0 : \mathbf{\Sigma} \in \Theta_q \quad \text{vs} \quad H_1 : \mathbf{\Sigma} \in \Theta \setminus \Theta_q,$$

where $\Theta = M_{p\times p}^{\text{posdef}}$.

Testing

It will be crucial to know whether q independent normal random variables are sufficient to describe the model.

Let $\Theta_q = \{\boldsymbol{\Sigma} = \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T \ : \ \boldsymbol{\Lambda} \in M_{p\times q}, \boldsymbol{\Psi} \in M_{p\times p}^{\text{diag}+}\}$, meaning the set of covariance matrices formed by at most q factors.

Suppose we wish to test

$$H_0 : \boldsymbol{\Sigma} \in \Theta_q \quad \text{vs} \quad H_1 : \boldsymbol{\Sigma} \in \Theta \setminus \Theta_q,$$

where $\Theta = M_{p\times p}^{\text{posdef}}$.

We can use the likelihood ratio statistic

$$LR = 2(l(\boldsymbol{S}) - l(\hat{\boldsymbol{\Sigma}}))$$

where $\hat{\boldsymbol{\Sigma}}$ is the MLE under H_0.

Testing

Standard asymptotics suggest that $LR \xrightarrow{\mathcal{D}} \chi^2_{r-r_0}$ as sample size $n \longrightarrow \infty$, where

$$r = \frac{1}{2}p(p+1)$$
$$r_0 = pq + p - \frac{1}{2}q(q-1)$$

are the number of free parameters in the saturated model and null model respectively.

Testing

Standard asymptotics suggest that $LR \xrightarrow{\mathcal{D}} \chi^2_{r-r_0}$ as sample size $n \longrightarrow \infty$, where

$$r = \frac{1}{2}p(p+1)$$
$$r_0 = pq + p - \frac{1}{2}q(q-1)$$

are the number of free parameters in the saturated model and null model respectively.

If correct, repeatedly calculate LR by simulating from Θ_q: values should be consistent with $\chi^2_{r-r_0}$;

i.e. p-values should be uniform.

Example: Anxiety Data

q	χ^2	d.f.	AIC	p-value
1	450.7	170	110.7	$< 10^{-6}$
2	287.5	151	-14.50	$< 10^{-6}$
3	223.0	133	-43.00	$< 10^{-5}$
4	171.5	116	-60.47	$< 10^{-3}$
5	132.9	100	-67.09	0.015
6	99.45	85	-70.55	0.136
7	67.23	71	-74.77	0.605
8	47.85	58	-68.15	0.826

A Simulated Example (1)

$p = 4$, $q = 1$, $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Psi} = \frac{1}{3}\boldsymbol{I}$, so:

$$
\begin{aligned}
x_1 &= \lambda_{11} y + \epsilon_1 \\
&\vdots \\
x_4 &= \lambda_{41} y + \epsilon_4
\end{aligned}
$$

Where $y \sim \mathrm{N}(0, 1)$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \frac{1}{3}\boldsymbol{I})$.

A Simulated Example (1)

$p = 4$, $q = 1$, $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Psi} = \frac{1}{3}\boldsymbol{I}$, so:

$$x_1 = \lambda_{11} y + \epsilon_1$$

$$\vdots$$

$$x_4 = \lambda_{41} y + \epsilon_4$$

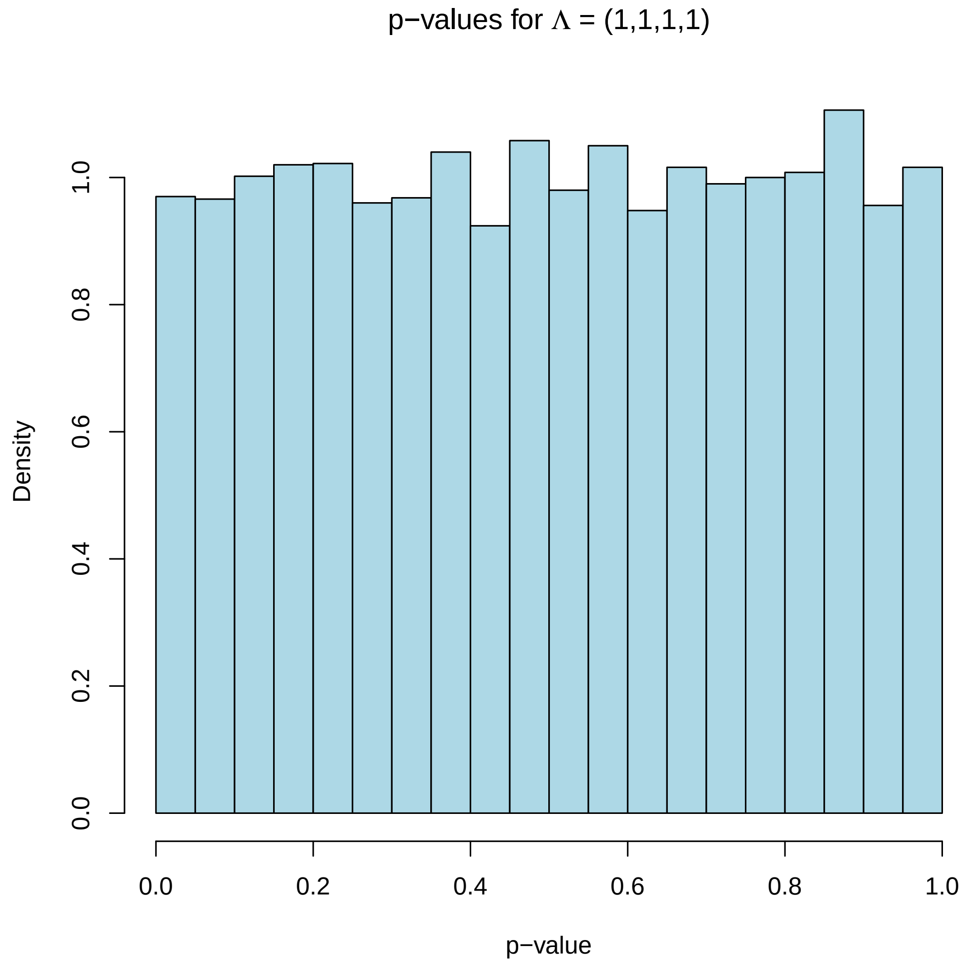
Where $y \sim \mathrm{N}(0, 1)$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \frac{1}{3}\boldsymbol{I})$.

First take $\boldsymbol{\Lambda} = (1, 1, 1, 1)^T$, so

$$\boldsymbol{x} = y\mathbf{1} + \boldsymbol{\epsilon}.$$

Use $n = 1\ 000$ and $N = 10\ 000$ repetitions.

A Simulated Example (1)



A Simulated Example (2)

Now proceed as above but with $\mathbf{\Lambda} = (1, 1, 0, 0)^T$, so

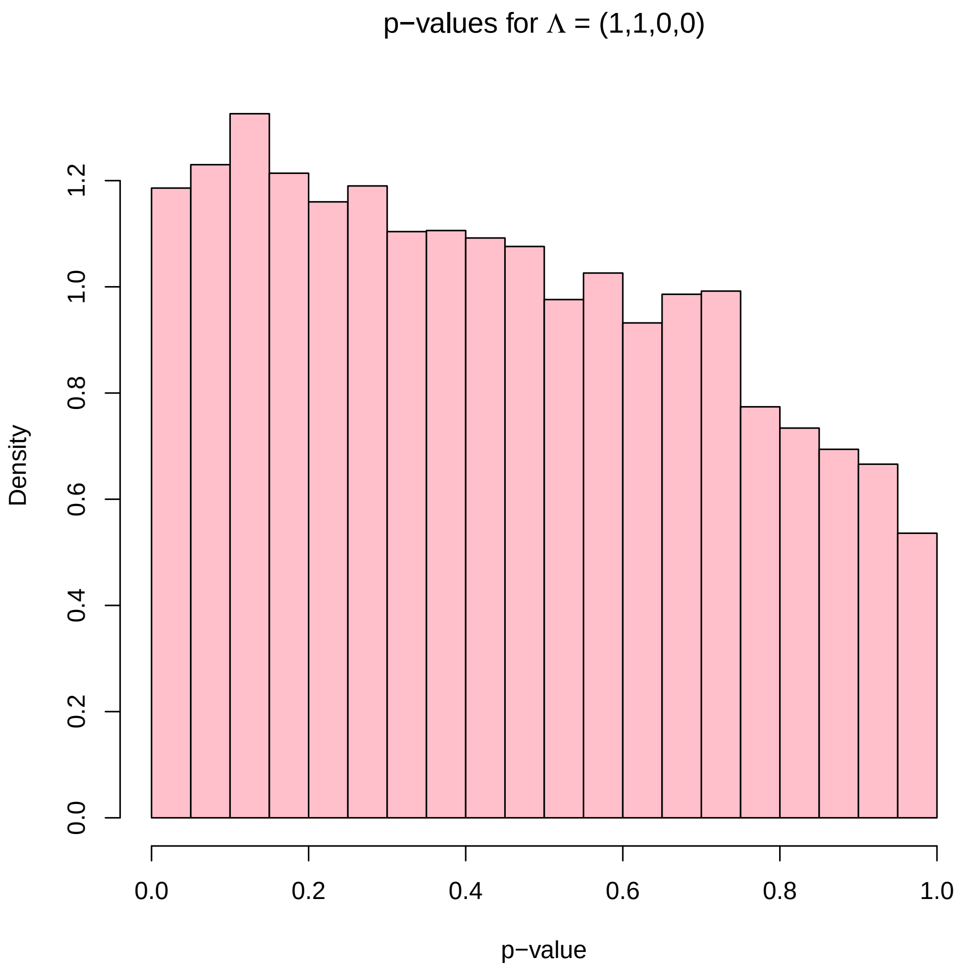
$$x_1 = y + \epsilon_1$$

$$x_2 = y + \epsilon_2$$

$$x_3 = \epsilon_3$$

$$x_4 = \epsilon_4$$

A Simulated Example (2)



A Simulated Example (2)

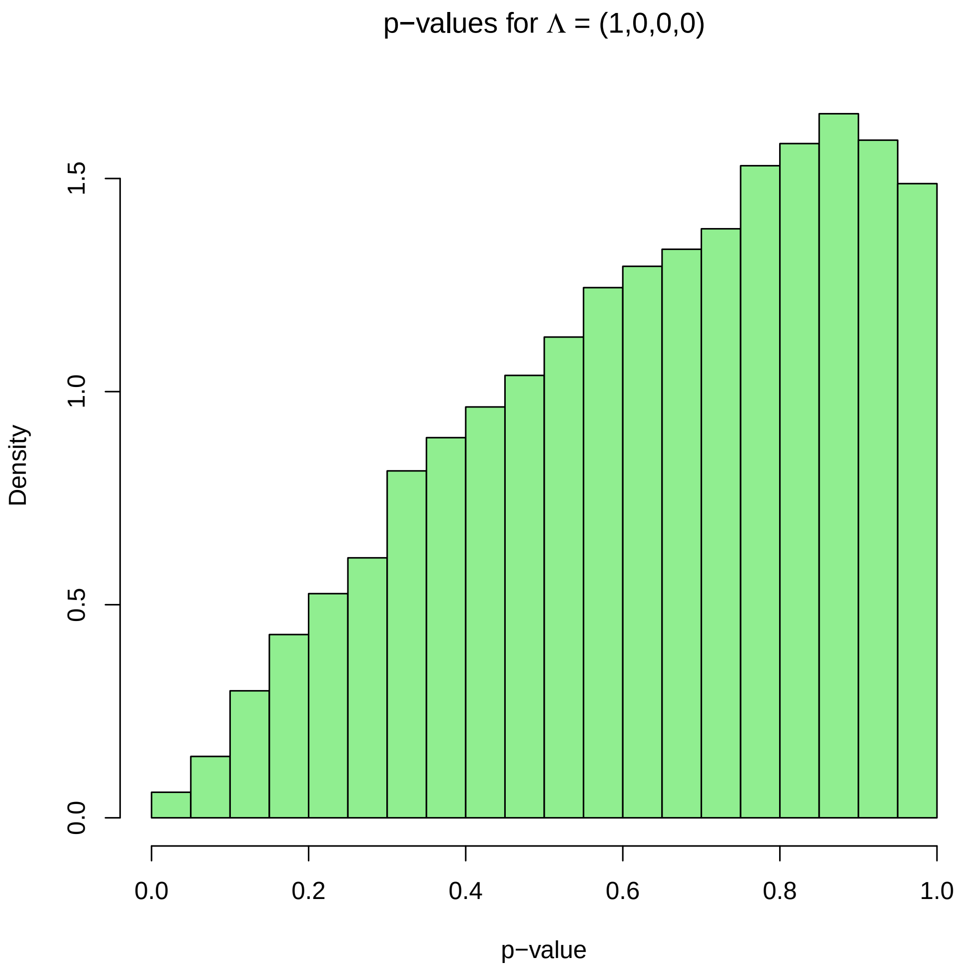
The true p-values are incorrect, even for large sample ($n = 1\ 000$):

Nominal	**Actual**
0.5	0.584
0.1	0.123
0.05	0.063
0.01	0.0116
0.005	0.006

Likelihood ratio statistic does not have expected distribution. Our test would reject the null hypothesis too often.

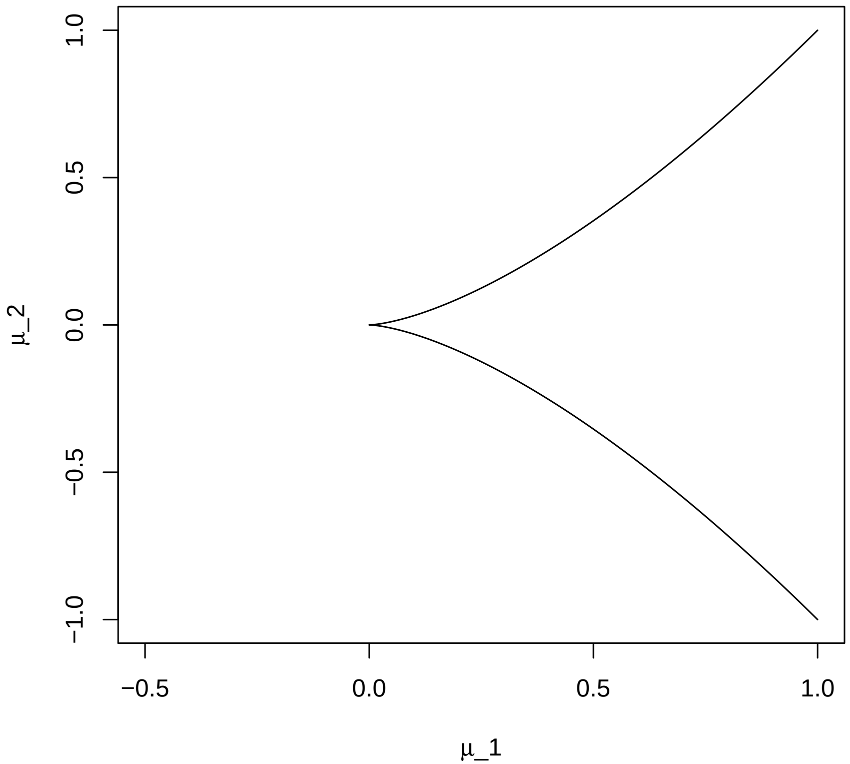
Why does this happen?

A Simulated Example (3)



Singularities — Non-Factor Analysis Example

Suppose $\Theta = \mathbb{R}^2$, and $\Theta_0 = \{(t^2, t^3) : t \in \mathbb{R}\}$.



Singularities — Non-Factor Analysis Example

Suppose $\Theta = \mathbb{R}^2$, and $\Theta_0 = \{(t^2, t^3) : t \in \mathbb{R}\}$.

The point $(0, 0)$ is not locally smooth.

As we 'zoom in', the space looks more and more like a half line.

What happens to the likelihood ratio test of $\theta \in \Theta_0$ vs $\theta \in \Theta$ in this case?

Another Simulation

We generate $n = 100$ points from a $\mathrm{N}(\boldsymbol{\mu}, \left(\begin{smallmatrix} 0.1 & 0 \\ 0 & 0.1 \end{smallmatrix}\right))$ distribution.

It is easy to see that the MLE for $\boldsymbol{\mu}$ is the closest point in Θ_0 to the sample mean $\overline{\boldsymbol{X}}_n$.

Another Simulation

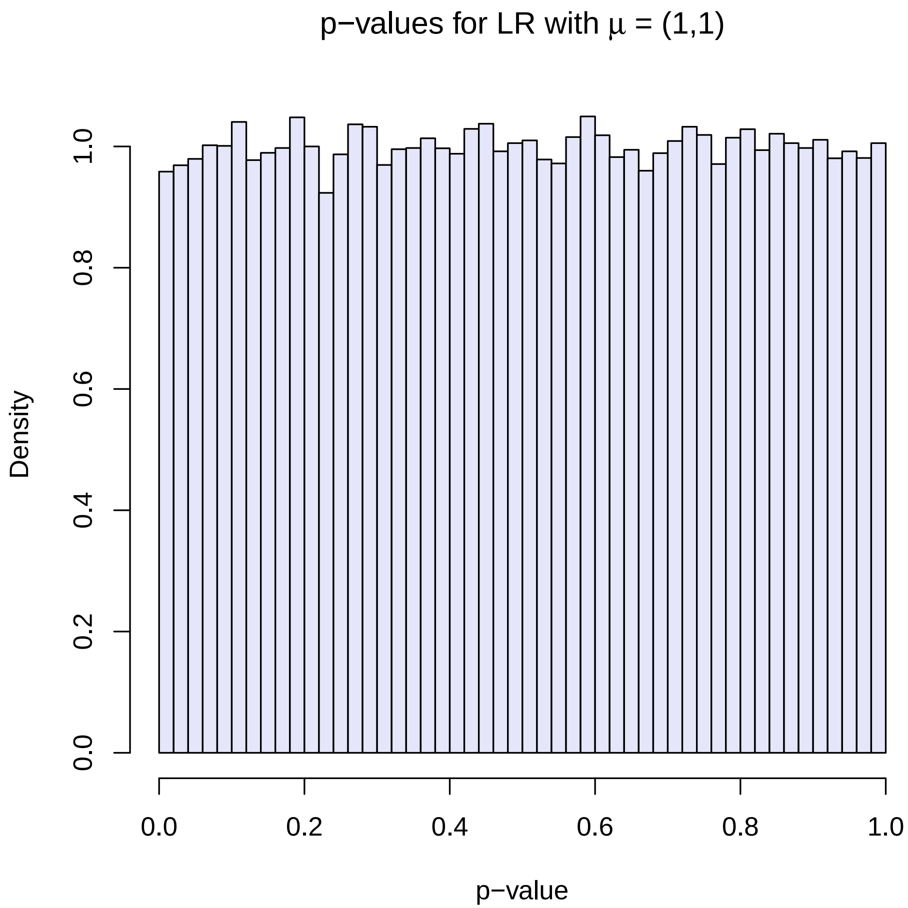
We generate $n = 100$ points from a $\mathrm{N}(\boldsymbol{\mu}, \left(\begin{smallmatrix} 0.1 & 0 \\ 0 & 0.1 \end{smallmatrix}\right))$ distribution.

It is easy to see that the MLE for $\boldsymbol{\mu}$ is the closest point in Θ_0 to the sample mean $\overline{\boldsymbol{X}}_n$.

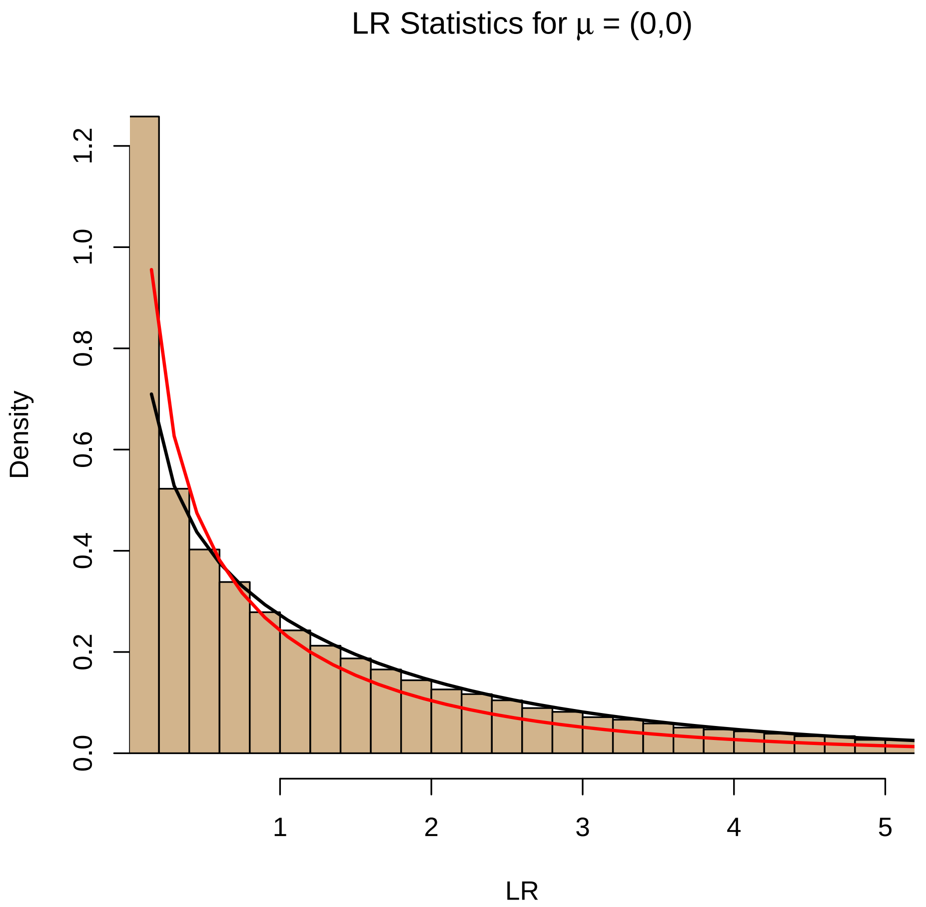
Treat covariance as known. We try with:

- $\boldsymbol{\mu} = (1,1)^T$ (smooth point),
- $\boldsymbol{\mu} = (0,0)^T$ (not a smooth point).

Then over $N = 100,000$ repetitions, record likelihood ratio test statistic of $\theta \in \Theta_0$ vs $\theta \in \Theta$.



χ^2_1 p-values for $\boldsymbol{\mu} = (1, 1)$ are clearly uniform.



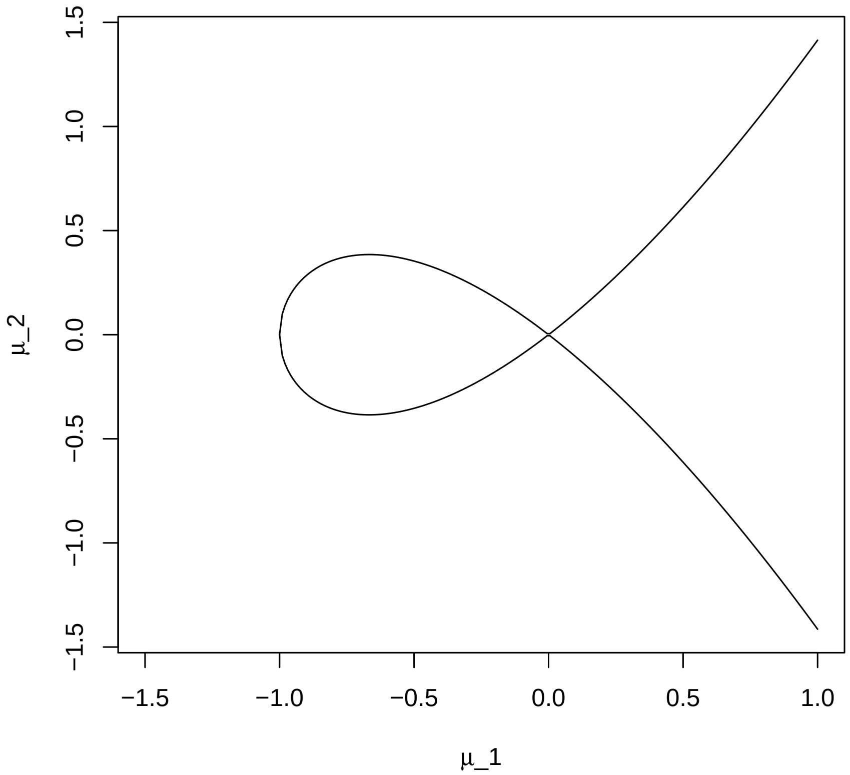
Red line is χ^2_1 density, black $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$ mixture density.

p-values:

Nominal	**Actual**
0.05	0.096
0.01	0.022
0.005	0.0114
0.001	0.0025

Singularities

Suppose $\boldsymbol{\mu} = \{(t^2 - 1, t(t^2 - 1)) : t \in \mathbb{R}\}$.



Singularities

Suppose $\boldsymbol{\mu} = \{(t^2 - 1, t(t^2 - 1)) : t \in \mathbb{R}\}$.

The point $(0, 0)$ is not locally smooth.

As we 'zoom in', the space looks more and more like two straight lines intersecting.

References

Bartholemew and Knott. (1999). *Latent variable models and factor analysis* (2nd ed.), London: Arnold, pp 41–76

Bartholomew, Steele, Moustaki and Galbraith. (2008). *Analysis of Multivariate Social Science Data* (2nd ed.), Chapman & Hall

Crader & Butler (1996). The validity of students' teaching evaluation scores: the Wimberly-Faulkner-Moxley questionnaire, *Ed. and Psych. Measurement.* **56** 304–14.

Drton, M. (2009). Likelihood ratio tests and singularities, *Ann. Stat.* **37** (2) 979–1012

R Code (data analysis)

```
library(GPArotation)       # CONTAINS OBLIMIN FUNC

# data: www.cmm.bristol.ac.uk/team/amssd-downloads.shtml
x = c(1.0000, .5101, 1.0000, .2399,...
dat = matrix(0, 20, 20)
dat[upper.tri(dat,diag=T)] = x
dat = dat + t(dat) - diag(rep(1,20))

out1 = factanal(factors = 2, covmat = dat, n.obs = 335,
    rotation="none")
out2 = factanal(factors = 2, covmat = dat, n.obs = 335,
    rotation="oblimin")
```

R Code (simulations)

```
library(MASS)

N = 1e4;  n = 1000
mu = rep(0, 4)

Ga = c(1,1,1,1) # CHANGE THIS FOR DIFFERENT CASES
De = diag(rep(1/3,4))
Si = De + outer(Ga, Ga)

out = numeric(N)

for (i in 1:N) {
  x = mvrnorm(n, mu, Si)
  out[i] = factanal(x, 1)$PVAL
}
```