MATH 250: Mathematical Data Visualization

Dimension reducation and statistical inference

Peter A. Gao

2024-03-04

San José State University

Outline

- What is inference?
- Review of multivariate statistics and probability
- The multivariate normal distribution
    - Conditional distributions
- Factor analysis and latent variables?
- Canonical correlation analysis
- Examples and comparison with PCA

Statistical Inference

Statistical inference: studying data to draw conclusions about some population or probabilistic data-generating process.

Example of inferential methods: hypothesis testing, confidence intervals

In particular, statisticians emphasize the importance of quantifying the **uncertainty** of estimates (often under many assumptions)

SVD and PCA: not inferential methods?

SVD is a matrix decomposition–in general no assumptions about how a given m by n matrix A has been generated.

Even when applying SVD to conduct PCA for a data matrix X, we do not need to place any distributional assumptions on X, making SVD and PCA just descriptive, not inferential.

We do not make any statements about whether PCs observed in the data are connected to “real” PCs in some hypothetical population or data-generating process.

Covariance/correlation of two random variables

Definition

The **covariance** between two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Definition

The **correlation** between two random variables X and Y is defined as

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Covariance matrix

Remark

If $\mathbf{x}$ is a p-dimensional random vector, its **covariance matrix** is defined to be

$$\mathrm{Cov}(\mathbf{x}) = E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^\top]$$

Thus, the covariance matrix is positive semidefinite.

$$\begin{aligned}\mathrm{Cov}(\mathbf{x}) &= E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^\top] \\ &= \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1, X_p) & \mathrm{Cov}(X_2, X_p) & \cdots & \mathrm{Var}(X_p) \end{bmatrix}\end{aligned}$$

Covariance matrix

Remark

As a result,

$$\mathrm{Cov}(A\mathbf{x}+\mathbf{b}) = A[\mathrm{Cov}(\mathbf{x})]A^\top$$

Remark

$$\mathrm{Cor}(\mathbf{x}) = \begin{bmatrix} 1 & \mathrm{Cor}(X_1,X_2) & \cdots & \mathrm{Cor}(X_1,X_p) \\ \mathrm{Cor}(X_1,X_2) & 1 & \cdots & \mathrm{Cor}(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cor}(X_1,X_p) & \mathrm{Cor}(X_2,X_p) & \cdots & 1 \end{bmatrix}$$

Sample covariance

If we have a sample of iid random vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathbf{x}$, we can combine them into an $n \times p$ data matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Definition

The **sample covariance** of $\mathbf{x}$ is defined as

$$\widehat{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top = \frac{1}{n} X_c^\top X_c$$

where $\overline{\mathbf{x}}$ is the sample mean and X_c is a centered version of X (Exercise).

Exercises

1. Give an example of two uncorrelated variables that are dependent.
2. Show that $\mathrm{Cov}(A\mathbf{x} + \mathbf{b})$ is still positive semidefinite.

Exercises

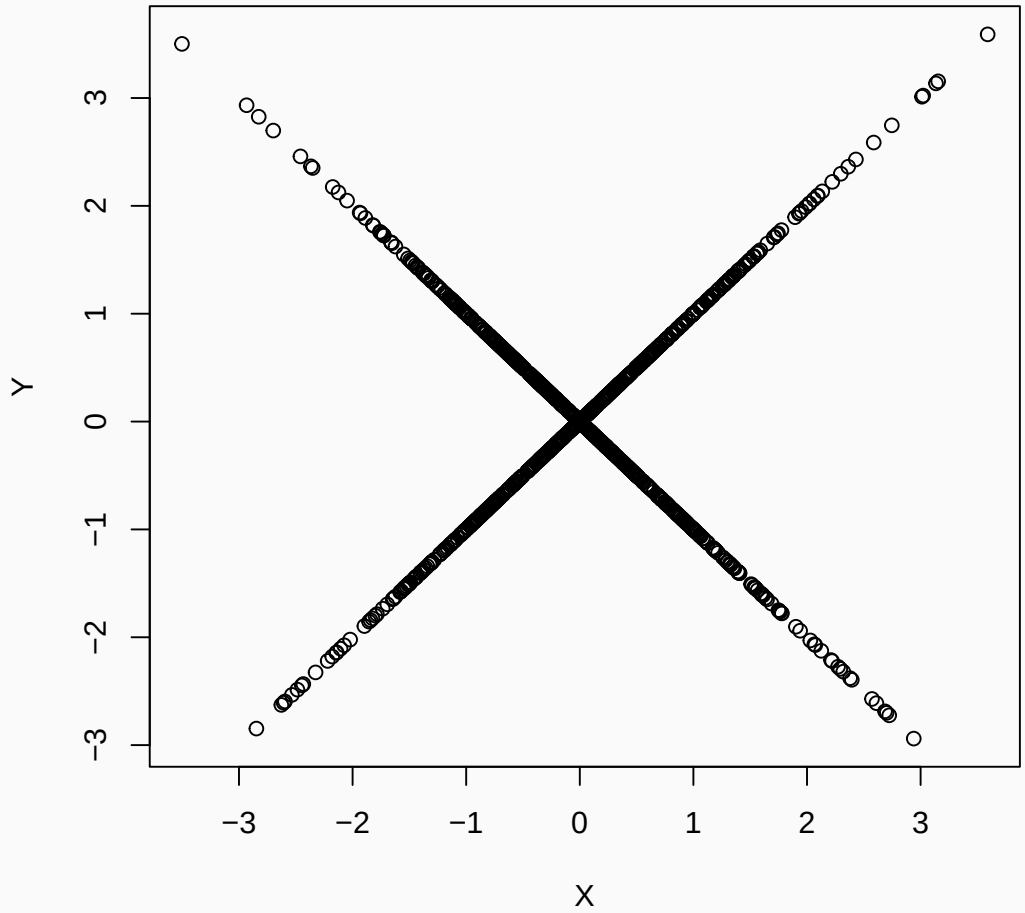
Let W be -1 with probability 1/2 and 1 with probability 1/2. Let $X \sim N(0, 1)$ and $Y = WX$.

```
n <- 1000
W <- sample(c(-1, 1), n, replace = T)
X <- rnorm(n)
Y <- W * X
cor(X, Y)
```

```
[1] 0.05753462
```

Exercises

```
plot(X, Y)
```



Standard Gaussian random vectors

The standard Gaussian random vector generalizes the univariate standard Gaussian:

Definition

Let $\mathbf{z} = (Z_1, Z_2, \dots, Z_k)^\top$ be a k-dimensional random vector. We say $\mathbf{z}$ is a **standard Gaussian random vector** if $Z_1, \dots, Z_k$ are independent $N(0,1)$ random variables.

Multivariate Gaussian distribution

The multivariate Gaussian distribution is a generalization of the univariate Gaussian distribution.

Definition

Let $\mathbf{y} = (Y_1, Y_2, \dots, Y_k)^\top$ be a k-dimensional random vector. We say $\mathbf{y}$ follows a **multivariate Gaussian** distribution if there exists a mean vector μ and matrix $A \in \mathbb{R}^{k \times l}$ such that

$$\mathbf{y} = \mu + A\mathbf{z}$$

for a standard normal random vector $\mathbf{z} \in \mathbb{R}^l$.

In particular, we say $\mathbf{y}$ has mean μ and covariance $\Sigma = \mathrm{Cov}(\mu + A\mathbf{z}) = AA^\top$.

Multivariate Gaussian distribution

If $\mathbf{y}$ is a multivariate Gaussian random vector with mean μ and covariance matrix Σ, its probability density function is:

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu)\right\}$$

What happens if Σ is not positive definite?

Other characteristics

An equivalent definition: $\mathbf{y}$ is a multivariate Gaussian random variable if every every linear combination of its components is a univariate Gaussian.

Bivariate Gaussian distribution

Marginal distributions

Suppose $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)^\top \in \mathbb{R}^k$ is jointly Gaussian:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Then $\mathbf{y}_1 \sim N(\mu_1, \Sigma_{11})$ and $\mathbf{y}_2 \sim N(\mu_2, \Sigma_{22})$

What does Σ_{12} represent?

Conditional Gaussians

Suppose we know $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)^\top \in \mathbb{R}^k$ is jointly Gaussian. If we observe $\mathbf{y}_2$, what is the conditional distribution of $\mathbf{y}_1$?

$$\mathbf{y}_1 \mid \mathbf{y}_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2)$$

and

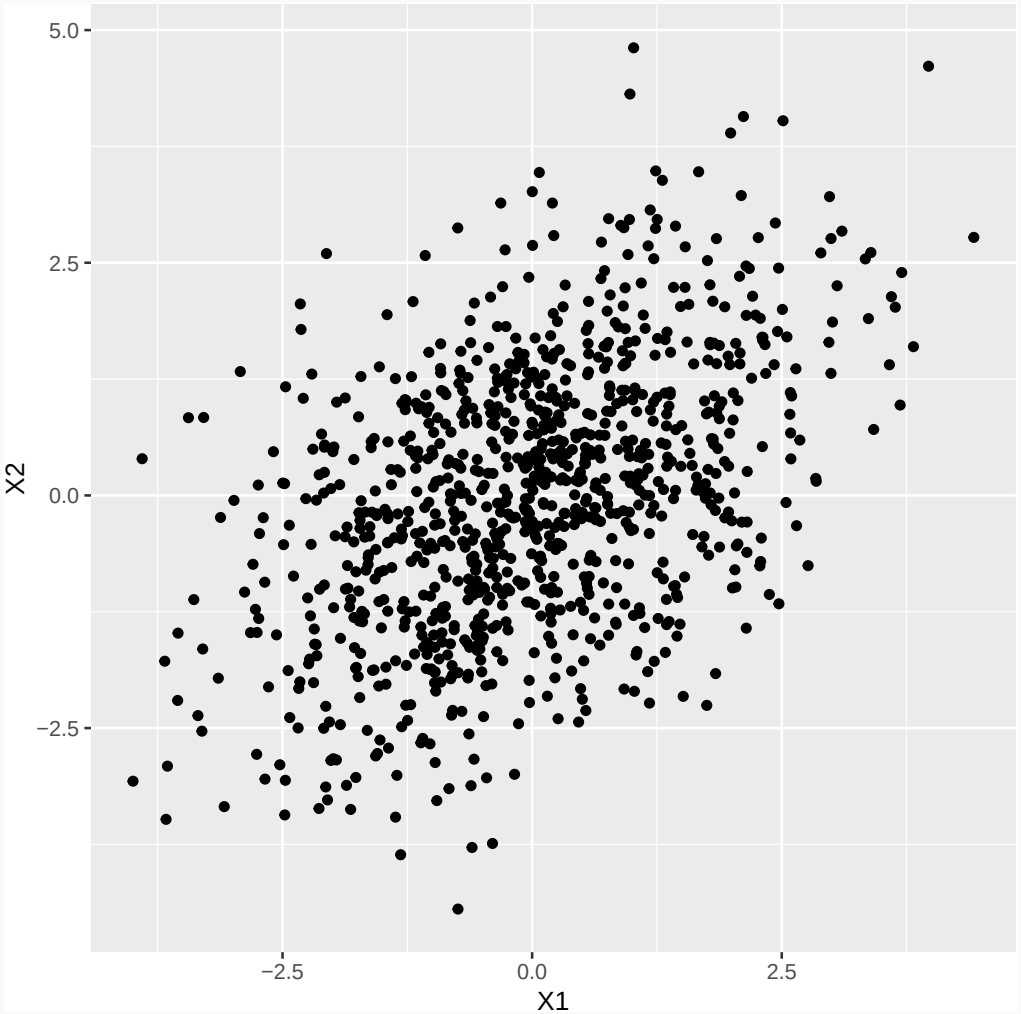
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Conditional Gaussians

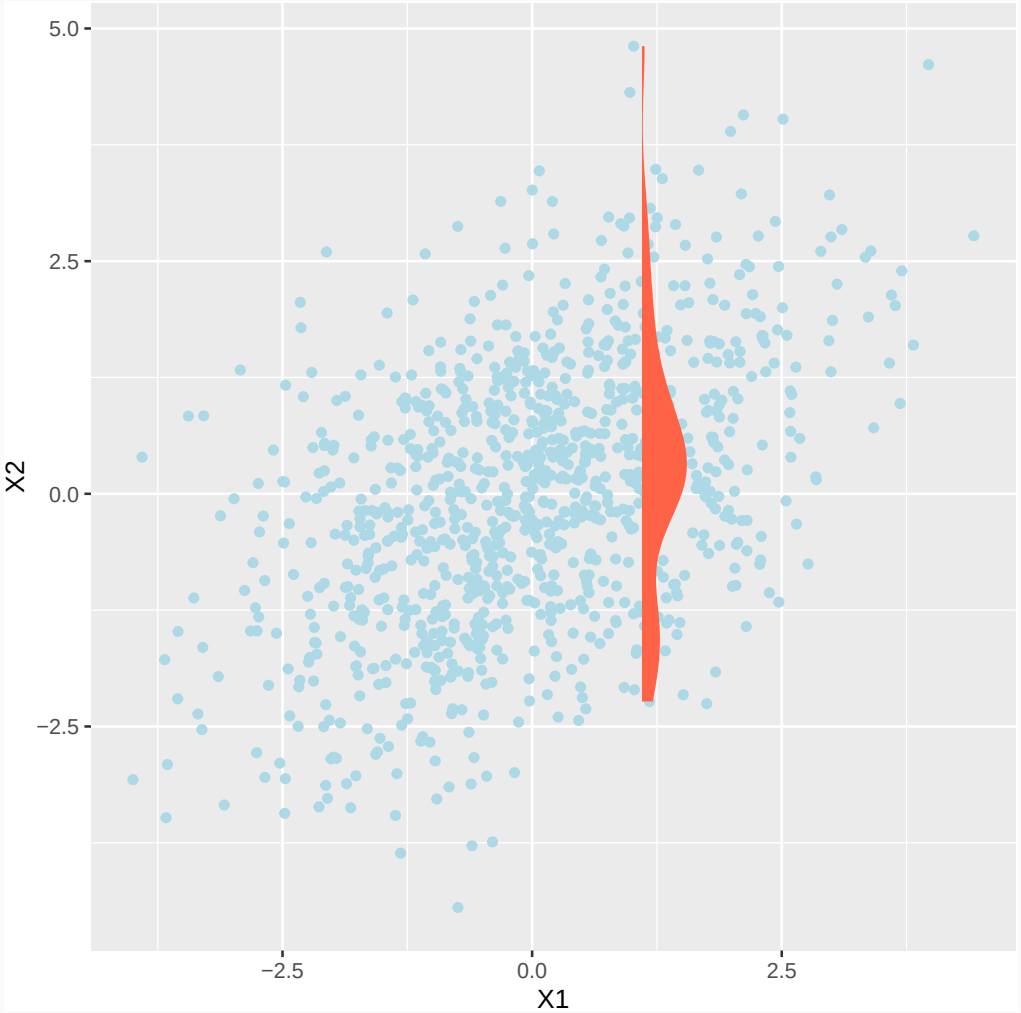
```
Sigma <- matrix(c(2, 1, 1, 2), nrow = 2)
Z <- matrix(rnorm(2000), ncol = 2) # independent Gaussians
X <- Z %*% chol(Sigma) # rotate to get correlated Gaussians
X_plot <- data.frame(X)
```

Conditional Gaussians

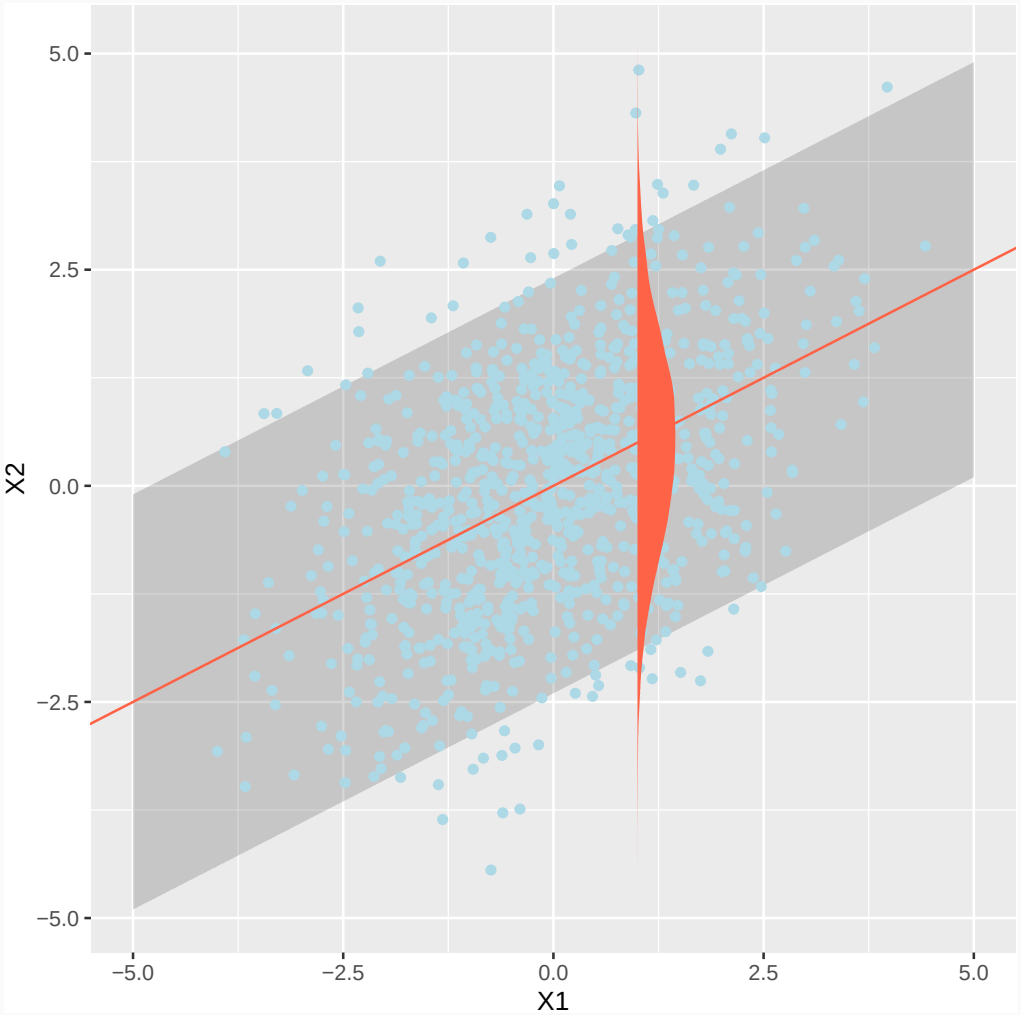
```
library(ggplot2)
ggplot(X_plot, aes(x = X1, y = X2)) + geom_point()
```



Empirical conditional distribution



Theoretical conditional distribution



Precision matrix

Let the precision matrix be defined $\Lambda = \Sigma^{-1}$. Keeping the same block dimensions as in the last slide:

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

Note that $\Lambda_{11} \neq \Sigma_{11}^{-1}$ and $\Lambda_{22} \neq \Sigma_{22}^{-1}$ in general.

In what cases will $\Lambda_{11} = \Sigma_{11}^{-1}$ and $\Lambda_{22} = \Sigma_{22}^{-1}$?

Rewriting the conditional normal

We can rewrite the conditional distribution $\mathbf{y}_1 \mid \mathbf{y}_2$ in terms of this precision matrix:

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2) = \mu_1 + \Lambda_{11}^{-1}\Lambda_{12}(\mathbf{y}_2 - \mu_2)$$

and

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1}$$

In some cases, working with the precision matrix will be easier because even dense covariance matrices may have sparse inverses.

Schur Complements

Theorem

Suppose M is a block matrix:

$$M = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$$

Then

$$M^{-1} = \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix}$$

where $M/H = E - FH^{-1}G$, assuming H is invertible.

We say that M/H is the **Schur complement** of M with respect to H.

Schur Complements

Theorem

Equivalently,

$$M^{-1} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix}$$

where $M/E = H - GE^{-1}F$, assuming E is invertible.

We say that M/E is the **Schur complement** of M with respect to E.

Schur Complement: Proof

First, we block diagonalize M, assuming H is invertible.

Observe that

$$\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} E - FH^{-1}G & 0 \\ G & H \end{bmatrix}$$

and

$$\underbrace{\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} E & F \\ G & H \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix}}_{Z} = \underbrace{\begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix}}_{W}$$

Schur Complement: Proof

Invert both sides:

$$
\begin{aligned}
Z^{-1}M^{-1}X^{-1} &= W^{-1} \\
M^{-1} &= ZW^{-1}X \\
&= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix}
\end{aligned}
$$

The proof for (M/E) is similar.

Deriving the conditional multivariate Gaussian distribution

Suppose $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)^\top \in \mathbb{R}^k$ is jointly Gaussian:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Then

$$p(\mathbf{y}_1, \mathbf{y}_2) \propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} \mathbf{y}_1 - \mu_1 \\ \mathbf{y}_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_1 - \mu_1 \\ \mathbf{y}_2 - \mu_2 \end{bmatrix} \right\}$$

Deriving the conditional multivariate Gaussian distribution

$$p(\mathbf{y}_1, \mathbf{y}_2) \propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} \mathbf{y}_1 - \mu_1 \\ \mathbf{y}_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \begin{bmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \right.$$
$$\left. \times \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 - \mu_1 \\ \mathbf{y}_2 - \mu_2 \end{bmatrix} \right\}$$
$$= \exp\left\{ -\frac{1}{2}(\mathbf{y}_1 - \mu_{1|2})^\top (\Sigma/\Sigma_{22})^{-1}(\mathbf{y}_1 - \mu_{1|2}) \right\}$$
$$\times \exp\left\{ -\frac{1}{2}(\mathbf{y}_2 - \mu_2)^\top \Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2) \right\}$$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2)$$

Factor analysis

Factor analysis is a method for modeling correlations between observed variables in terms of a smaller number of unobserved latent variables, called factors.

In particular, given a vector of p observed variables $\mathbf{x}$, we assume that:

$$\mathbf{x} \mid \mathbf{z} \sim N(W\mathbf{z} + \mu, \Psi)$$
$$\mathbf{z} \sim N(\mu_0, \Sigma_0)$$

where W is a $p \times q$ **factor loading** matrix and Ψ is a diagonal $p \times p$ covariance matrix.

Factor analysis

Note that the marginal distribution of $\mathbf{x}$ is still Gaussian:

$$\mathbf{x} \sim N(W\mu_0 + \mu, \Psi + W\Sigma_0 W^\top)$$

Without loss of generality, we can assume $\mu_0 = \mathbf{0}$ and $\Sigma_0 = I$.

Factor analysis

This simplification yields

$$\mathbf{x} \mid \mathbf{z} \sim N(W\mathbf{z} + \mu, \Psi)$$
$$\mathbf{z} \sim N(\mathbf{0}, I)$$

and

$$\mathbf{x} \sim N(\mu, \Psi + WW^\top)$$

where W is a $p \times q$ **factor loading** matrix and Ψ is a diagonal $p \times p$ covariance matrix.

Factor analysis

In essence, instead of modeling $\mathbf{x} \sim N(\mu, \Sigma)$ for some general positive definite Σ, we assume a low-rank structure for the covariance matrix:

$$\Sigma = \mathrm{Cov}(\mathbf{x}) = WW^\top + \Psi$$

For any individual component of $\mathbf{x}$, this implies that

$$\mathrm{Var}(x_i) = \sum_{k=1}^{q} w_{ik}^2 + \psi_d$$

where the sum $\sum_{k=1}^{q} w_{ik}^2$ is called the **communality** (variance due to the common factors) and ψ_d is called the **uniqueness.**

Motivation and comparison with PCA

PCA aims to find a lower-dimensional representation of the data that minimizes the mean-squared distance from the original data to the projected data. However, the correlations between the principal components may not be the same as the correlations between the original covariate vectors.

Factor analysis aims to choose a lower dimensional representation that preserves correlations when proejcting to the new feature space.

Motivation and comparison with PCA

Observe also that the factor model is a **probabilistic** data-generating model. Whereas PCA is only a descriptive technique, the factor model assumes a distribution for the existing observed data and gives a way to generate new observations.

The requirement, of course, is that the model is correct.

Definitions

Observed variables: the original covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$

Latent factors: the lower dimensional vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$

Loading matrix: The matrix W containing the contributions of each variable to the latent factors.

Typical modeling assumptions

1. The observed variables are scaled and centered to have mean zero and variance 1.
2. The latent factors also have mean zero and variance 1.
3. The latent factors are uncorrelated across individuals and and across columns.

Rotation of factors

Let Q be a q by q orthogonal matrix. Then, note that we can rewrite

$$W\mathbf{z} = (WQ)(Q^\top \mathbf{z}) = H\mathbf{y}$$

where H is a new loading matrix and $\mathbf{y} = Q^\top \mathbf{z}$ is a transformed vector of latent factors.

Observe that this rotation by Q does not change the covariance model:

$$\mathrm{Cov}(\mathbf{x}) = WW^\top + \Psi$$

Rotation of factors

This means that the factor analysis model is **unidentifiable** in the sense that we can apply any orthogonal transformation to our loadings and factors without actually changing the fit of our model.

As such, it is necessary to place some restrictions on our choice of W and $\mathbf{z}$ in order to get consistent results.

Resolving unidentifiability

Examples of restrictions include:

- Restricting W to have orthonormal columns
- Placing a sparsity-inducing prior on the entries of W
- Choosing an informative rotation Q:
  - **Varimax**: maximize the sum of the variances of the squared loadings

Example: Pastry texture

To illustrate these ideas, Hartmann, Krois, and Rudolph (2023) give a simple example using simulated data about pastry texture:

Figure 1: Croissants by Herry Wibisono

Example: Pastry texture

```
food <- read.csv("https://userpage.fu-berlin.de/soga/data/raw-data/food
                 row.names = "X")
str(food)
```

```
'data.frame':   50 obs. of  5 variables:
 $ Oil      : num  16.5 17.7 16.2 16.7 16.3 19.1 18.4 17.5 15.7 16.4 ...
 $ Density : int  2955 2660 2870 2920 2975 2790 2750 2770 2955 2945 ...
 $ Crispy   : int  10 14 12 10 11 13 13 10 11 11 ...
 $ Fracture: int  23 9 17 31 26 16 17 26 23 24 ...
 $ Hardness: int  97 139 143 95 143 189 114 63 123 132 ...
```

Example: Pastry texture

```
food_fa <- factanal(food, factors = 2)
food_fa
```

```

Call:
factanal(x = food, factors = 2)

Uniquenesses:
     Oil  Density   Crispy Fracture Hardness
   0.334    0.156    0.042    0.256    0.407

Loadings:
          Factor1 Factor2
Oil       -0.816
Density    0.919
Crispy    -0.745   0.635
Fracture   0.645  -0.573
Hardness          0.764
```

Example: Pastry texture

```
food_fa$uniquenesses
```

```
      Oil   Density    Crispy  Fracture  Hardness
0.3338599 0.1555255 0.0422238 0.2560235 0.4069459
```

```
# communalities
1 - food_fa$uniquenesses
```

```
      Oil   Density    Crispy  Fracture  Hardness
0.6661401 0.8444745 0.9577762 0.7439765 0.5930541
```

Example: Pastry texture

```
apply(food_fa$loadings ^ 2, 1, sum)
```

```
      Oil   Density    Crispy  Fracture  Hardness
0.6661398 0.8444745 0.9577762 0.7439766 0.5930539
```

Example: Pastry texture

The `factanal()` function also gives the results of a log-likelihood ratio hypothesis test based on the chi-square distribution where the null hypothesis is the number of factors is sufficient to capture the variation in the dataset. In this case, we fail to reject the null, meaning we do not have reason to increase the number of factors. (Why is there 1 degree of freedom?)

```
# chi-square statistic
food_fa$STATISTIC
```

```
objective
0.2708945
```

```
# degrees of freedom
food_fa$dof
```

```
[1] 1
```

```
# p-value
food_fa$PVAL
```

```
objective
0.6027324
```

Example: Pastry texture

```
W <- food_fa$loadings
Psi <- diag(food_fa$uniquenesses)
R <- food_fa$correlation
Sigma <- W %*% t(W) + Psi
round(R - Sigma, 6)
```

```
                Oil   Density    Crispy  Fracture  Hardness
Oil       0.000000  0.000001 -0.002613 -0.018220 -0.000776
Density   0.000001  0.000000 -0.001081 -0.007539 -0.000320
Crispy   -0.002613 -0.001081  0.000000  0.000000  0.000005
Fracture -0.018220 -0.007539  0.000000  0.000000  0.000033
Hardness -0.000776 -0.000320  0.000005  0.000033  0.000000
```

How many factors?

Some ways to choose the number of factors n:

1. Conduct PCA on the data X and choose n to be the first number for which the cumulative proportion of variance explained exceeds 80-90%. Scree plots can also be used.

```
food_pca <- prcomp(food, scale = T, center = T)
# cumulative proportion of variance
cumsum(food_pca$sdev ^ 2) / sum(food_pca$sdev ^ 2)
```

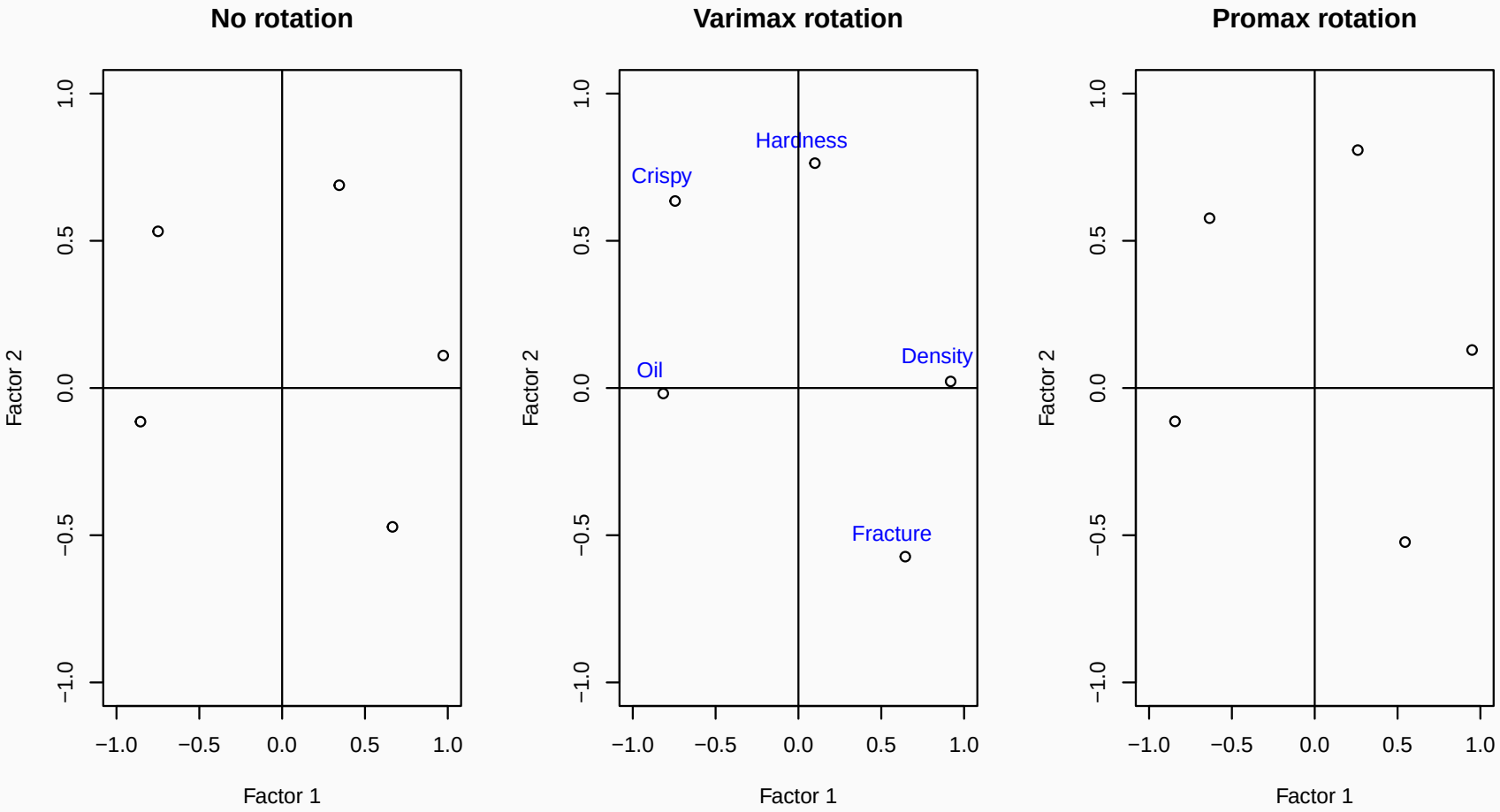
```
[1] 0.6062426 0.8653838 0.9273937 0.9757777 1.0000000
```

2. **Kaiser's rule**: Keep a factor if the sum of squared factor loadings is greater than one.

```
apply(food_fa$loadings ^ 2, 2, sum)
```

```
 Factor1  Factor2
2.489879 1.315542
```

Example: Pastry texture



Exploratory vs. Confirmatory Factor Analysis

Exploratory Factor Analysis: A descriptive method that identifies a lower-dimensional latent factor space that captures correlations between the original variables. Can be viewed as a generalization of PCA.

Confirmatory Factor Analysis: Checking whether the model with some proposed latent factors actually fits the data in question.

Factor analysis for causal inference

It is tempting to apply factor analysis for causal inference (ex. your decisions or answers on a personality test stem from a handful of fundamental personality traits).

However, it is rare that the factor analysis model proposed is exactly correct. If your model is incorrect, then you can quickly obtain false and misleading conclusions.

Even if your model is correct, the associations may not be causal. Other factors may be at play as well.

Estimation

The introduction of a probabilistic model for $\mathbf{x}$ enables us to use maximum likelihood estimation. (Why can't we use MLE for PCA?)

We assume $\mathbf{x} \sim N(\mu, \Psi + WW^\top)$ and then use direct numerical maximization or a two-stage procedure (like expectation-maximization (EM)).

If we have a guess of the noise covariance $\widehat{\Psi}$, we can also apply PCA to the reduced sample correlation matrix

$$\widehat{\mathrm{Cor}}(\mathbf{x}) - \widehat{\Psi}$$

Note that in either case, we must estimate a large number of parameters that comprise W.

Comparison with PCA

1. Factor analysis assumes a data generating model for the observed data $\mathbf{x}$. As such, factor models can be tested by making predictions for new observations.
2. PCA can be applied to any sample covariance matrix directly; factor analysis assumes that $\mathrm{Cov}(\mathbf{x}) = WW^\top + \Psi$.
3. Under the factor model, the principal directions of PCA converge to the eigenvectors of $\mathrm{Cov}(\mathbf{x}) = WW^\top + \Psi$, while the latent factors represent the eigenvectors of $WW^\top$.
4. Differences in factor scores represent differences in the expected values of the data vectors, while differences in principal components represent realized differences in the data vectors.

Pitfalls

As with PCA, there is a significant risk of overinterpretation and making causal inferences where there are none to be made.

A famous example of how this can lead to scientific disagreements is provided by Shalizi (n.d.)

Spearman's g

In the early 20th century, Charles Spearman noticed that students' grades in different subjects (ex. math, English, etc.) were correlated.

He hypothesized that the **reason** these grades were correlated was that they were all correlated with some unobserved factor, which he called "general intelligence," or g.

In other words, he assumed that $\mathbf{x} = \mu + wg + \varepsilon$, where $\mathbf{x}$ is the vector of grades.

He observed that this model fit his data on student grades well and concluded that g must exist and **cause** student grades.

Spearman's one-factor model

Take a closer look at Spearman's model:

$$\mathbf{x} = \mu + wg + \varepsilon$$

where $\mathbf{x}$ is the vector of grades, $\mathbf{w}$ is the vector of weights, and g is a single number. If this model is true, then $\mathrm{Cov}(x_i, x_j) = E(X_i, X_j) = w_i w_j$.

Spearman was interested in "tetrad equations." Under this model, for any distinct i, j, k, l,

$$\frac{\mathrm{Cov}(x_i, x_j)/\mathrm{Cov}(x_k, x_j)}{\mathrm{Cov}(x_i, x_l)/\mathrm{Cov}(x_k, x_l)} = \frac{w_i w_j / w_k w_j}{w_i w_l / w_k w_l} = 1$$

Spearman found that these tetrad equations were approximately satisfied in his grades dataset, concluding that g must exist.

Spearman's one-factor model

However, Spearman's's model is just one of many models that could have explained the correlations in his dataset.

Remember, if we rotate the weight matrix W by some orthogonal rotation Q, we can get another equally well-fitting factor model.

It can be shown that any linear Gaussian factor model with q latent variables is equivalent to a mixture (of Gaussians) model with $q + 1$ clusters – but these have very different interpretations.

The Thomson Sampling Model

Cosma Shalizi gives a walkthrough of another potential explanation, based on a sampling model proposed by Thomson (2014). Suppose that instead of $q < p$, we have more factors than features in reality and each feature is a random sum of latent factors.

Suppose that there are many different characteristics (factors) representing different mental abilities and that each time a student takes a class or an exam, a random subset of these factors is used to determine the final score.

For a fixed number of tests, some of these abilities will be shared across the tests and some will be unique to only one test.

Via simulation, Shalizi shows that applying factor analysis to this data yields a well-fitting factor model.

The Thomson Sampling Model

$$X_{ij} = \sum_{k=1}^{q} A_{ik} T_{kj} + \varepsilon_{ij}$$

where A_{ik} represent independent latent random variables with mean zero and variance 1, ε_{ij} represent independent noise, and T_{kj} represent independent Bernoulli(z_j) variables.

The Thomson Sampling Model

Then (exercise)

$$\mathrm{Cov}(X_{ia}, X_{ib}) = E(X_{ia}, X_{ib}) = \sum_{k=1}^{q} E(T_{ka} T_{kb})$$

and $E(X_{ia}, X_{ib}) = q z_a z_b$.

As such, this is another model for which the tetrad equations hold.

The Thomson Sampling Model

Did Spearman's data arise because the one-factor model is true? Or because the Thompson sampling model is true? Based on Spearman's analysis, we can't tell.

As such, even when a factor model fits well, we may learn very little about the true nature of the data-generating process.

References

Hartmann, Kai, Joachim Krois, and Annette Rudolph. 2023. "Statistics and Geodata Analysis Using R." *Department of Earth Sciences, Freie Universitaet Berlin*. https://www.geo.fu-berlin.de/en/v/soga-r/Advances-statistics/Multivariate-approaches/Factor-Analysis/A-Simple-Example-of-Factor-Analysis-in-R/index.html.

Shalizi, Cosma Rohilla. n.d. *Advanced Data Analysis from an Elementary Point of View (Draft)*.