Applied Statistical Methods II

Some additional topics I

Method of moments (MoM)

In general:

- Let $\boldsymbol{Y} \in \mathbb{R}^n$ be observed data and $\boldsymbol{\theta} \in \mathbb{R}^p$ be a population parameter of interest
- Define a function $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^q$ s.t.

$$E\left\{\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right)\right\} = \mathbf{0}_q$$

- Idea: like the score function, the function $\boldsymbol{f}$ identifies $\boldsymbol{\theta}$.
  What is the small value of q that will identify $\boldsymbol{\theta}$?
- Examples:
  - Estimating the mean: if $EY_i = \mu$, $f\left(\boldsymbol{Y}, \mu\right) = \sum_{i=1}^{n} Y_i - n\mu$.
  - Variance: if $Y_i \sim (\mu, \sigma^2)$,
    $f\left(\boldsymbol{Y}, \sigma^2\right) = (n-1) - \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$
  - Any score function: $\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right) = \nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{\theta}; \boldsymbol{Y}\right)$.
- Benefit of MoM: we don't have to know the distribution of $\boldsymbol{Y}$! Only need its moments.

Method of moments (MoM)

In general:

- Let $\boldsymbol{Y} \in \mathbb{R}^n$ be observed data and $\boldsymbol{\theta} \in \mathbb{R}^p$ be a population parameter of interest
- Define a function $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^q$ s.t.

$$E\left\{\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right)\right\} = \mathbf{0}_q$$

- Idea: like the score function, the function $\boldsymbol{f}$ identifies $\boldsymbol{\theta}$. What is the small value of q that will identify $\boldsymbol{\theta}$?
- Examples:
  - Estimating the mean: if $EY_i = \mu$, $f\left(\boldsymbol{Y}, \mu\right) = \sum_{i=1}^{n} Y_i - n\mu$.
  - Variance: if $Y_i \sim (\mu, \sigma^2)$,
    $f\left(\boldsymbol{Y}, \sigma^2\right) = (n-1) - \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$
  - Any score function: $\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right) = \nabla_{\theta} \ell\left(\boldsymbol{\theta}; \boldsymbol{Y}\right)$.
- Benefit of MoM: we don't have to know the distribution of $\boldsymbol{Y}$! Only need its moments.

Method of moments (MoM)

In general:

- Let $\boldsymbol{Y} \in \mathbb{R}^n$ be observed data and $\boldsymbol{\theta} \in \mathbb{R}^p$ be a population parameter of interest
- Define a function $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^q$ s.t.

$$E\left\{\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right)\right\} = \mathbf{0}_q$$

- Idea: like the score function, the function $\boldsymbol{f}$ identifies $\boldsymbol{\theta}$. What is the small value of q that will identify $\boldsymbol{\theta}$?
- Examples:
  - Estimating the mean: if $EY_i = \mu$, $f\left(\boldsymbol{Y}, \mu\right) = \sum_{i=1}^{n} Y_i - n\mu$.
  - Variance: if $Y_i \sim (\mu, \sigma^2)$, $f\left(\boldsymbol{Y}, \sigma^2\right) = (n-1) - \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$
  - Any score function: $\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right) = \nabla_{\theta} \ell\left(\boldsymbol{\theta}; \boldsymbol{Y}\right)$.
- Benefit of MoM: we don't have to know the distribution of $\boldsymbol{Y}$! Only need its moments.

Method of moments (MoM)

In general:

- Let $\boldsymbol{Y} \in \mathbb{R}^n$ be observed data and $\boldsymbol{\theta} \in \mathbb{R}^p$ be a population parameter of interest
- Define a function $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^q$ s.t.

$$E\left\{\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right)\right\} = \mathbf{0}_q$$

- Idea: like the score function, the function $\boldsymbol{f}$ identifies $\boldsymbol{\theta}$. What is the small value of q that will identify $\boldsymbol{\theta}$?
- Examples:
  - Estimating the mean: if $EY_i = \mu$, $f\left(\boldsymbol{Y}, \mu\right) = \sum_{i=1}^{n} Y_i - n\mu$.
  - Variance: if $Y_i \sim (\mu, \sigma^2)$, $f\left(\boldsymbol{Y}, \sigma^2\right) = (n-1) - \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$
  - Any score function: $\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right) = \nabla_{\theta} \ell\left(\boldsymbol{\theta}; \boldsymbol{Y}\right)$.
- Benefit of MoM: we don't have to know the distribution of $\boldsymbol{Y}$! Only need its moments.

Method of moments (MoM)

In general:

- Let $\boldsymbol{Y} \in \mathbb{R}^n$ be observed data and $\boldsymbol{\theta} \in \mathbb{R}^p$ be a population parameter of interest
- Define a function $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^q$ s.t.

$$E\left\{\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right)\right\} = \mathbf{0}_q$$

- Idea: like the score function, the function $\boldsymbol{f}$ identifies $\boldsymbol{\theta}$. What is the small value of q that will identify $\boldsymbol{\theta}$?
- Examples:
  - Estimating the mean: if $EY_i = \mu$, $f\left(\boldsymbol{Y}, \mu\right) = \sum_{i=1}^{n} Y_i - n\mu$.
  - Variance: if $Y_i \sim (\mu, \sigma^2)$,
    $f\left(\boldsymbol{Y}, \sigma^2\right) = (n-1) - \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$
  - Any score function: $\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right) = \nabla_{\theta} \ell\left(\boldsymbol{\theta}; \boldsymbol{Y}\right)$.
- Benefit of MoM: we don't have to know the distribution of $\boldsymbol{Y}$! Only need its moments.

MINQUE (C.R. Rao, 1973)

Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n$, $\boldsymbol{\epsilon} \sim (0, \sum_{r=1}^{b} \theta_r \boldsymbol{B}_r)$.

- Estimate θ_r with method of moments.
- We'll use MINQUE: Minimum Norm Quadratic Unbiased Estimation
- Idea: For $\boldsymbol{A}_s \in \mathbb{R}^{n \times n}$,

$$E(\boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}) = \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} \boldsymbol{\beta} + \sum_{r=1}^{b} \theta_r \operatorname{Tr}(\boldsymbol{B}_r \boldsymbol{A}_s)$$

- We will set $\hat{\theta}_s = \boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}$. What conditions do we need on $\boldsymbol{A}_s$ such that $E\left(\hat{\theta}_s\right) = \theta_s$?
- $\boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} = \boldsymbol{0}_p$.
- $\operatorname{Tr}(\boldsymbol{B}_r \boldsymbol{A}_s) = 1\{r = s\}$.
- Is $\boldsymbol{A}_s$ symmetric? How about positive semi-definite?
- Question: how do we choose $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_b$?

MINQUE (C.R. Rao, 1973)

Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n$, $\boldsymbol{\epsilon} \sim (0, \sum_{r=1}^{b} \theta_r \boldsymbol{B}_r)$.

- Estimate θ_r with method of moments.
- We'll use MINQUE: Minimum Norm Quadratic Unbiased Estimation
- Idea: For $\boldsymbol{A}_s \in \mathbb{R}^{n \times n}$,

$$E(\boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}) = \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} \boldsymbol{\beta} + \sum_{r=1}^{b} \theta_r \operatorname{Tr}(\boldsymbol{B}_r \boldsymbol{A}_s)$$

- We will set $\hat{\theta}_s = \boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}$. What conditions do we need on $\boldsymbol{A}_s$ such that $E\left(\hat{\theta}_s\right) = \theta_s$?
- $\boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} = \boldsymbol{0}_p$.
- $\operatorname{Tr}(\boldsymbol{B}_r \boldsymbol{A}_s) = 1\{r = s\}$.
- Is $\boldsymbol{A}_s$ symmetric? How about positive semi-definite?
- Question: how do we choose $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_b$?

MINQUE (C.R. Rao, 1973)

Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n$, $\boldsymbol{\epsilon} \sim (0, \sum_{r=1}^{b} \theta_r \boldsymbol{B}_r)$.

- Estimate θ_r with method of moments.
- We'll use MINQUE: Minimum Norm Quadratic Unbiased Estimation
- Idea: For $\boldsymbol{A}_s \in \mathbb{R}^{n \times n}$,

$$E(\boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}) = \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} \boldsymbol{\beta} + \sum_{r=1}^{b} \theta_r \operatorname{Tr}(\boldsymbol{B}_r \boldsymbol{A}_s)$$

- We will set $\hat{\theta}_s = \boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}$. What conditions do we need on $\boldsymbol{A}_s$ such that $E\left(\hat{\theta}_s\right) = \theta_s$?
- $\boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} = \boldsymbol{0}_p$.
- $\operatorname{Tr}(\boldsymbol{B}_r \boldsymbol{A}_s) = 1\{r = s\}$.
- Is $\boldsymbol{A}_s$ symmetric? How about positive semi-definite?
- Question: how do we choose $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_b$?

MINQUE (cont.)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n, \boldsymbol{\epsilon} \sim (0, \sum_{r=1}^{b} \theta_r \boldsymbol{B}_r)$$

- $\hat{\theta}_s = \boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}$, $\boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} = \boldsymbol{0}_p$, $\mathrm{Tr}\left(\boldsymbol{B}_r \boldsymbol{A}_s\right) = 1\left\{r = s\right\}$
- We'll choose $\boldsymbol{A}_s$ such that $\hat{\theta}_s$ has minimal variance.
- Under assumptions of normality:
  $\mathrm{Var}\left(\hat{\theta}_s\right) = 2\sum_{r,t} \theta_r \theta_t \mathrm{Tr}\left(\boldsymbol{A}_s \boldsymbol{B}_r \boldsymbol{A}_s \boldsymbol{B}_t\right)$
- Problem: $\mathrm{Var}\left(\hat{\theta}_s\right)$ depends on $\boldsymbol{\theta}$, the parameter we're trying to estimate!
- Any ideas how to circumvent this?
  1. Get a consistent, but inefficient, estimator $\hat{\theta}^{(0)}$ using $\boldsymbol{A}_1^{(0)}, \ldots, \boldsymbol{A}_b^{(0)}$.
  2. Re-compute $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ by minimizing $\mathrm{Var}\left(\hat{\theta}_1^{(0)}\right), \ldots, \mathrm{Var}\left(\hat{\theta}_b^{(0)}\right)$.

MINQUE (cont.)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n, \boldsymbol{\epsilon} \sim (0, \sum_{r=1}^{b} \theta_r \boldsymbol{B}_r)$$

- $\hat{\theta}_s = \boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}$, $\boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} = \boldsymbol{0}_p$, $\mathrm{Tr}\left(\boldsymbol{B}_r \boldsymbol{A}_s\right) = 1\left\{r = s\right\}$
- We'll choose $\boldsymbol{A}_s$ such that $\hat{\theta}_s$ has minimal variance.
- Under assumptions of normality:
  $\mathrm{Var}\left(\hat{\theta}_s\right) = 2\sum_{r,t} \theta_r \theta_t \mathrm{Tr}\left(\boldsymbol{A}_s \boldsymbol{B}_r \boldsymbol{A}_s \boldsymbol{B}_t\right)$
- Problem: $\mathrm{Var}\left(\hat{\theta}_s\right)$ depends on $\boldsymbol{\theta}$, the parameter we're trying to estimate!
- Any ideas how to circumvent this?
  1. Get a consistent, but inefficient, estimator $\hat{\theta}^{(0)}$ using $\boldsymbol{A}_1^{(0)}, \ldots, \boldsymbol{A}_b^{(0)}$.
  2. Re-compute $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ by minimizing $\mathrm{Var}\left(\hat{\theta}_1^{(0)}\right), \ldots, \mathrm{Var}\left(\hat{\theta}_b^{(0)}\right)$.

MINQUE (cont.)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n,\ \boldsymbol{\epsilon} \sim (0, \sum_{r=1}^{b} \theta_r \boldsymbol{B}_r)$$

- $\hat{\theta}_s = \boldsymbol{Y}^T \boldsymbol{A}_s \boldsymbol{Y}$, $\boldsymbol{X}^T \boldsymbol{A}_s \boldsymbol{X} = \boldsymbol{0}_p$, $\mathrm{Tr}\left(\boldsymbol{B}_r \boldsymbol{A}_s\right) = 1\left\{r = s\right\}$
- We'll choose $\boldsymbol{A}_s$ such that $\hat{\theta}_s$ has minimal variance.
- Under assumptions of normality:
  $\mathrm{Var}\left(\hat{\theta}_s\right) = 2\sum_{r,t} \theta_r \theta_t \mathrm{Tr}\left(\boldsymbol{A}_s \boldsymbol{B}_r \boldsymbol{A}_s \boldsymbol{B}_t\right)$
- Problem: $\mathrm{Var}\left(\hat{\theta}_s\right)$ depends on $\boldsymbol{\theta}$, the parameter we're trying to estimate!
- Any ideas how to circumvent this?
  1. Get a consistent, but inefficient, estimator $\hat{\boldsymbol{\theta}}^{(0)}$ using $\boldsymbol{A}_1^{(0)}, \ldots, \boldsymbol{A}_b^{(0)}$.
  2. Re-compute $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ by minimizing $\mathrm{Var}\left(\hat{\theta}_1^{(0)}\right), \ldots, \mathrm{Var}\left(\hat{\theta}_b^{(0)}\right)$.

Generalized method of moments

- Let $\boldsymbol{Y} \in \mathbb{R}^n$ be observed data and $\boldsymbol{\theta} \in \mathbb{R}^p$ be a population parameter of interest.
- Specify the function $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^q$ such that $E\left\{\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right)\right\} = \mathbf{0}_q$.
- Question: how do we estimate and perform inference on $\boldsymbol{\theta}$ when the distribution of $\boldsymbol{Y}$ is unknown?
- The problem was solved for fixed dimensions p and q in Hansen (1982), which won him the Nobel Prize in economics in 2013.
- Let $\boldsymbol{W} \in \mathbb{R}^{q \times q}$ be p.d. By moment condition, we consider the class of estimators

$$\hat{\theta}_W = \underset{\theta}{\operatorname{argmin}}\, \boldsymbol{f}\left(\boldsymbol{Y}, \theta\right)^T \boldsymbol{W} \boldsymbol{f}\left(\boldsymbol{Y}, \theta\right).$$

- We require $q \geq p$ (Why?) If $q = p$, $\boldsymbol{f}\left(\boldsymbol{Y}, \hat{\theta}_W\right) = \mathbf{0}_p$, and , $\hat{\theta}_W$ is invariant to $\boldsymbol{W}$.

Generalized method of moments

- Let $\boldsymbol{Y} \in \mathbb{R}^n$ be observed data and $\boldsymbol{\theta} \in \mathbb{R}^p$ be a population parameter of interest.
- Specify the function $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^q$ such that $E\left\{\boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right)\right\} = \mathbf{0}_q$.
- Question: how do we estimate and perform inference on $\boldsymbol{\theta}$ when the distribution of $\boldsymbol{Y}$ is unknown?
- The problem was solved for fixed dimensions p and q in Hansen (1982), which won him the Nobel Prize in economics in 2013.
- Let $\boldsymbol{W} \in \mathbb{R}^{q \times q}$ be p.d. By moment condition, we consider the class of estimators

$$\hat{\boldsymbol{\theta}}_W = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right)^T \boldsymbol{W} \boldsymbol{f}\left(\boldsymbol{Y}, \boldsymbol{\theta}\right).$$

- We require $q \geq p$ (Why?) If $q = p$, $\boldsymbol{f}\left(\boldsymbol{Y}, \hat{\boldsymbol{\theta}}_W\right) = \mathbf{0}_p$, and , $\hat{\boldsymbol{\theta}}_W$ is invariant to $\boldsymbol{W}$.

Generalized method of moments

$E\left\{\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)\right\} = \mathbf{0}_q,$
$\hat{\theta}_W = \text{argmin}_{\boldsymbol{\theta}}\, \boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)^T \boldsymbol{W}\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)$

- If $q > p$, then $\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right) \neq 0$ for any $\boldsymbol{\theta}$. How do we select $\boldsymbol{W}$?
- By same technique as MINQUE: Minimize the asymptotic variance of $\hat{\theta}_W$ (which depends on $\boldsymbol{\theta}$).
  1. Start with $\boldsymbol{W} = I_q$. $\hat{\theta}_{I_q}$ is a $\sqrt{n}$-consistent estimator for $\boldsymbol{\theta}$.
  2. Set $\boldsymbol{W} = \boldsymbol{W}\left(\hat{\theta}_{I_q}\right)$ to be the optimal $\boldsymbol{W}$.
  3. Using this $\boldsymbol{W}$, $\hat{\theta}_W$ has the smallest possible asymptotic variance!
- Called **Hansen's two-step estimator**. Used all the time in economics, and recently to study non-random missing data in mass spectrometry data (McKennan et al., 2020).

Generalized method of moments

$E\left\{\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)\right\} = \mathbf{0}_q$,
$\hat{\boldsymbol{\theta}}_W = \text{argmin}_{\boldsymbol{\theta}}\, \boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)^T \boldsymbol{W}\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)$

- If $q > p$, then $\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right) \neq 0$ for any $\boldsymbol{\theta}$. How do we select $\boldsymbol{W}$?
- By same technique as MINQUE: Minimize the asymptotic variance of $\hat{\boldsymbol{\theta}}_W$ (which depends on $\boldsymbol{\theta}$).
  1. Start with $\boldsymbol{W} = I_q$. $\hat{\boldsymbol{\theta}}_{I_q}$ is a $\sqrt{n}$-consistent estimator for $\boldsymbol{\theta}$.
  2. Set $\boldsymbol{W} = \boldsymbol{W}\left(\hat{\boldsymbol{\theta}}_{I_q}\right)$ to be the optimal $\boldsymbol{W}$.
  3. Using this $\boldsymbol{W}$, $\hat{\boldsymbol{\theta}}_W$ has the smallest possible asymptotic variance!
- Called **Hansen's two-step estimator**. Used all the time in economics, and recently to study non-random missing data in mass spectrometry data (McKennan et al., 2020).

Generalized method of moments

$E\left\{\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)\right\} = \mathbf{0}_q,$
$\hat{\theta}_W = \text{argmin}_{\boldsymbol{\theta}}\, \boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)^T \boldsymbol{W}\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right)$

- If $q > p$, then $\boldsymbol{f}\left(\boldsymbol{Y},\boldsymbol{\theta}\right) \neq 0$ for any $\boldsymbol{\theta}$. How do we select $\boldsymbol{W}$?
- By same technique as MINQUE: Minimize the asymptotic variance of $\hat{\theta}_W$ (which depends on $\boldsymbol{\theta}$).
  1. Start with $\boldsymbol{W} = I_q$. $\hat{\theta}_{I_q}$ is a $\sqrt{n}$-consistent estimator for $\boldsymbol{\theta}$.
  2. Set $\boldsymbol{W} = \boldsymbol{W}\left(\hat{\theta}_{I_q}\right)$ to be the optimal $\boldsymbol{W}$.
  3. Using this $\boldsymbol{W}$, $\hat{\theta}_W$ has the smallest possible asymptotic variance!
- Called **Hansen's two-step estimator**. Used all the time in economics, and recently to study non-random missing data in mass spectrometry data (McKennan et al., 2020).

Factor Analysis and dimension reduction in noisy data

Overview

- Likely the most important tool in all of high dimensional statistics.
- Used in every discipline: machine learning, genetics, economics, sociology, psychology, education, political science, and many others...
- Comes in many flavors: linear, non-linear, finite dimensional, infinite dimensional (i.e. factor analysis in a RKHS)
- Core idea: find a small number of **factors** that explain most of the variation in the data.
- In the machine learning/computing science community: often treat this as an optimization problem.
  - Problem: we have no way of understanding the statistical uncertainty in our estimates.
- We will study this in the context of a statistical model.
  - This will allow us to justify our choice of estimators.
  - Study the statistical uncertainty.

Overview

- Likely the most important tool in all of high dimensional statistics.
- Used in every discipline: machine learning, genetics, economics, sociology, psychology, education, political science, and many others...
- Comes in many flavors: linear, non-linear, finite dimensional, infinite dimensional (i.e. factor analysis in a RKHS)
- Core idea: find a small number of **factors** that explain most of the variation in the data.
- In the machine learning/computing science community: often treat this as an optimization problem.
  - Problem: we have no way of understanding the statistical uncertainty in our estimates.
- We will study this in the context of a statistical model.
  - This will allow us to justify our choice of estimators.
  - Study the statistical uncertainty.

Linear factor analysis in noisy data I

- $\boldsymbol{y}_g = \boldsymbol{C}\ell_g + \boldsymbol{e}_g \in \mathbb{R}^n$ is the expression of gene $g = 1, \ldots, p$ (or brain region g, survey question g, etc) across n samples.
  - Will assume that $\boldsymbol{e}_g \sim (\boldsymbol{0}, \sigma_g^2 I_n)$ and $\boldsymbol{e}_g$ is independent of $\boldsymbol{e}_h$
  - p and n are large. Think $p \gtrsim n$, and maybe $p >> n$.
- The **factors** $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ are **shared** across all genes $g = 1, \ldots, p$.
- The **loadings** $\ell_g \in \mathbb{R}^K$ are the effects of $\boldsymbol{C}$ on the expression of gene g. Examples:
  - Some of the columns of $\boldsymbol{C}$ might correspond to biological factors like cell type. If a person has more T cells, their expression will look different from someone with more B cells.
  - Maybe some are related to disease (sick vs. healthy), technical factors (processing batch), etc.

Linear factor analysis in noisy data II

- $K << \min(n, p)$, i.e. want to explain the variation in $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ with a small number of factors.

- $\boldsymbol{C}$ can induce dependencies across genes. Assuming rows $i = 1, \ldots, n$ of $\boldsymbol{C}$ are i.i.d with $\text{Var}(\boldsymbol{C}_{i\cdot}) = \boldsymbol{\Psi}$:

$$\text{Cov}\left(y_{gi}, y_{hi}\right) = \boldsymbol{\ell}_g^T \boldsymbol{\Psi} \boldsymbol{\ell}_h, \quad g \neq h = 1, \ldots, p$$

- We only observe $\boldsymbol{y}_g$. Our goal is to recover $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_p$ and $\boldsymbol{C}$.

Fundamental problem in factor analysis

$\boldsymbol{y}_g = \boldsymbol{C}\boldsymbol{\ell}_g + \boldsymbol{e}_g \in \mathbb{R}^n$ for $g = 1, \ldots, p$

$$\boldsymbol{Y}_{p\times n} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_p^T \end{bmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{\ell}_1^T \\ \vdots \\ \boldsymbol{\ell}_p^T \end{pmatrix}}_{\boldsymbol{L}_{p\times K}} \underbrace{\begin{pmatrix} \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_n \end{pmatrix}}_{\boldsymbol{C}^T} + \underbrace{\begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_p^T \end{pmatrix}}_{\boldsymbol{E}_{p\times n}}$$

- Recall: $p \gtrsim n$, maybe $p >> n$.
- Recall: we only observe $\boldsymbol{Y}$. Goal is to recover $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{L} \in \mathbb{R}^{p\times K}$, $K << \min(n, p)$
- Problem: without further assumptions, $\boldsymbol{L}$ and $\boldsymbol{C}$ are not identifiable! Why? We can only recover $\boldsymbol{L}\boldsymbol{C}^T$:

$$E\left(\boldsymbol{Y} \mid \boldsymbol{C}\right) = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{R}\left(\boldsymbol{C}\boldsymbol{R}^{-T}\right)^T, \quad \forall \text{ invertible } \boldsymbol{R} \in \mathbb{R}^{K\times K}$$

- Are $\text{im}(\boldsymbol{C})$ or $\text{im}(\boldsymbol{L})$ identifiable?
- Different types of factor analyses place different assumptions on $\boldsymbol{L}$ and $\boldsymbol{C}$ to make them identifiable/interpretable.

Fundamental problem in factor analysis

$\boldsymbol{y}_g = \boldsymbol{C}\ell_g + \boldsymbol{e}_g \in \mathbb{R}^n$ for $g = 1, \ldots, p$

$$\boldsymbol{Y}_{p\times n} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_p^T \end{bmatrix} = \underbrace{\begin{pmatrix} \ell_1^T \\ \vdots \\ \ell_p^T \end{pmatrix}}_{\boldsymbol{L}_{p\times K}} \underbrace{\begin{pmatrix} \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_n \end{pmatrix}}_{\boldsymbol{C}^T} + \underbrace{\begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_p^T \end{pmatrix}}_{\boldsymbol{E}_{p\times n}}$$

- Recall: $p \gtrsim n$, maybe $p >> n$.
- Recall: we only observe $\boldsymbol{Y}$. Goal is to recover $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{L} \in \mathbb{R}^{p\times K}$, $K << \min(n, p)$
- Problem: without further assumptions, $\boldsymbol{L}$ and $\boldsymbol{C}$ are not identifiable! Why? We can only recover $\boldsymbol{L}\boldsymbol{C}^T$:

$$E\left(\boldsymbol{Y} \mid \boldsymbol{C}\right) = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{R}\left(\boldsymbol{C}\boldsymbol{R}^{-T}\right)^T, \quad \forall \text{ invertible } \boldsymbol{R} \in \mathbb{R}^{K\times K}$$

- Are im($\boldsymbol{C}$) or im($\boldsymbol{L}$) identifiable?
- Different types of factor analyses place different assumptions on $\boldsymbol{L}$ and $\boldsymbol{C}$ to make them identifiable/interpretable.

Fundamental problem in factor analysis

$\boldsymbol{y}_g = \boldsymbol{C}\ell_g + \boldsymbol{e}_g \in \mathbb{R}^n$ for $g = 1, \ldots, p$

$$\boldsymbol{Y}_{p\times n} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_p^T \end{bmatrix} = \underbrace{\begin{pmatrix} \ell_1^T \\ \vdots \\ \ell_p^T \end{pmatrix}}_{\boldsymbol{L}_{p\times K}} \underbrace{\begin{pmatrix} \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_n \end{pmatrix}}_{\boldsymbol{C}^T} + \underbrace{\begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_p^T \end{pmatrix}}_{\boldsymbol{E}_{p\times n}}$$

- Recall: $p \gtrsim n$, maybe $p >> n$.
- Recall: we only observe $\boldsymbol{Y}$. Goal is to recover $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{L} \in \mathbb{R}^{p\times K}$, $K << \min(n, p)$
- Problem: without further assumptions, $\boldsymbol{L}$ and $\boldsymbol{C}$ are not identifiable! Why? We can only recover $\boldsymbol{L}\boldsymbol{C}^T$:

$$E\left(\boldsymbol{Y} \mid \boldsymbol{C}\right) = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{R}\left(\boldsymbol{C}\boldsymbol{R}^{-T}\right)^T, \quad \forall \text{ invertible } \boldsymbol{R} \in \mathbb{R}^{K\times K}$$

- Are im($\boldsymbol{C}$) or im($\boldsymbol{L}$) identifiable?
- Different types of factor analyses place different assumptions on $\boldsymbol{L}$ and $\boldsymbol{C}$ to make them identifiable/interpretable.

Fundamental problem in factor analysis

$\boldsymbol{y}_g = \boldsymbol{C}\ell_g + \boldsymbol{e}_g \in \mathbb{R}^n$ for $g = 1, \ldots, p$

$$\boldsymbol{Y}_{p\times n} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_p^T \end{bmatrix} = \underbrace{\begin{pmatrix} \ell_1^T \\ \vdots \\ \ell_p^T \end{pmatrix}}_{\boldsymbol{L}_{p\times K}} \underbrace{\begin{pmatrix} \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_n \end{pmatrix}}_{\boldsymbol{C}^T} + \underbrace{\begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_p^T \end{pmatrix}}_{\boldsymbol{E}_{p\times n}}$$

- Recall: $p \gtrsim n$, maybe $p >> n$.
- Recall: we only observe $\boldsymbol{Y}$. Goal is to recover $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{L} \in \mathbb{R}^{p\times K}$, $K << \min(n, p)$
- Problem: without further assumptions, $\boldsymbol{L}$ and $\boldsymbol{C}$ are not identifiable! Why? We can only recover $\boldsymbol{L}\boldsymbol{C}^T$:

$$E\left(\boldsymbol{Y} \mid \boldsymbol{C}\right) = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{R}\left(\boldsymbol{C}\boldsymbol{R}^{-T}\right)^T, \quad \forall \text{ invertible } \boldsymbol{R} \in \mathbb{R}^{K\times K}$$

- Are im($\boldsymbol{C}$) or im($\boldsymbol{L}$) identifiable?
- Different types of factor analyses place different assumptions on $\boldsymbol{L}$ and $\boldsymbol{C}$ to make them identifiable/interpretable.

Fundamental problem in factor analysis

$\boldsymbol{y}_g = \boldsymbol{C}\ell_g + \boldsymbol{e}_g \in \mathbb{R}^n$ for $g = 1, \ldots, p$

$$\boldsymbol{Y}_{p\times n} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_p^T \end{bmatrix} = \underbrace{\begin{pmatrix} \ell_1^T \\ \vdots \\ \ell_p^T \end{pmatrix}}_{\boldsymbol{L}_{p\times K}} \underbrace{\begin{pmatrix} \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_n \end{pmatrix}}_{\boldsymbol{C}^T} + \underbrace{\begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_p^T \end{pmatrix}}_{\boldsymbol{E}_{p\times n}}$$

- Recall: $p \gtrsim n$, maybe $p >> n$.
- Recall: we only observe $\boldsymbol{Y}$. Goal is to recover $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{L} \in \mathbb{R}^{p\times K}$, $K << \min(n, p)$
- Problem: without further assumptions, $\boldsymbol{L}$ and $\boldsymbol{C}$ are not identifiable! Why? We can only recover $\boldsymbol{L}\boldsymbol{C}^T$:

$$E\left(\boldsymbol{Y} \mid \boldsymbol{C}\right) = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{R}\left(\boldsymbol{C}\boldsymbol{R}^{-T}\right)^T, \quad \forall \text{ invertible } \boldsymbol{R} \in \mathbb{R}^{K\times K}$$

- Are im($\boldsymbol{C}$) or im($\boldsymbol{L}$) identifiable?
- Different types of factor analyses place different assumptions on $\boldsymbol{L}$ and $\boldsymbol{C}$ to make them identifiable/interpretable.

Fundamental problem in factor analysis

$\boldsymbol{y}_g = \boldsymbol{C}\ell_g + \boldsymbol{e}_g \in \mathbb{R}^n$ for $g = 1, \ldots, p$

$$\boldsymbol{Y}_{p\times n} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_p^T \end{bmatrix} = \underbrace{\begin{pmatrix} \ell_1^T \\ \vdots \\ \ell_p^T \end{pmatrix}}_{\boldsymbol{L}_{p\times K}} \underbrace{\begin{pmatrix} \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_n \end{pmatrix}}_{\boldsymbol{C}^T} + \underbrace{\begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_p^T \end{pmatrix}}_{\boldsymbol{E}_{p\times n}}$$

- Recall: $p \gtrsim n$, maybe $p >> n$.
- Recall: we only observe $\boldsymbol{Y}$. Goal is to recover $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{L} \in \mathbb{R}^{p\times K}$, $K << \min(n, p)$
- Problem: without further assumptions, $\boldsymbol{L}$ and $\boldsymbol{C}$ are not identifiable! Why? We can only recover $\boldsymbol{L}\boldsymbol{C}^T$:

$$E\left(\boldsymbol{Y} \mid \boldsymbol{C}\right) = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{R}\left(\boldsymbol{C}\boldsymbol{R}^{-T}\right)^T, \quad \forall \text{ invertible } \boldsymbol{R} \in \mathbb{R}^{K\times K}$$

- Are im($\boldsymbol{C}$) or im($\boldsymbol{L}$) identifiable?
- Different types of factor analyses place different assumptions on $\boldsymbol{L}$ and $\boldsymbol{C}$ to make them identifiable/interpretable.

Fundamental problem in factor analysis

$\boldsymbol{y}_g = \boldsymbol{C}\ell_g + \boldsymbol{e}_g \in \mathbb{R}^n$ for $g = 1, \ldots, p$

$$\boldsymbol{Y}_{p\times n} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_p^T \end{bmatrix} = \underbrace{\begin{pmatrix} \ell_1^T \\ \vdots \\ \ell_p^T \end{pmatrix}}_{\boldsymbol{L}_{p\times K}} \underbrace{\begin{pmatrix} \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_n \end{pmatrix}}_{\boldsymbol{C}^T} + \underbrace{\begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_p^T \end{pmatrix}}_{\boldsymbol{E}_{p\times n}}$$

- Recall: $p \gtrsim n$, maybe $p >> n$.
- Recall: we only observe $\boldsymbol{Y}$. Goal is to recover $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{L} \in \mathbb{R}^{p\times K}$, $K << \min(n, p)$
- Problem: without further assumptions, $\boldsymbol{L}$ and $\boldsymbol{C}$ are not identifiable! Why? We can only recover $\boldsymbol{L}\boldsymbol{C}^T$:

$$E\left(\boldsymbol{Y} \mid \boldsymbol{C}\right) = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{C}^T = \boldsymbol{L}\boldsymbol{R}\left(\boldsymbol{C}\boldsymbol{R}^{-T}\right)^T, \quad \forall \text{ invertible } \boldsymbol{R} \in \mathbb{R}^{K\times K}$$

- Are im($\boldsymbol{C}$) or im($\boldsymbol{L}$) identifiable?
- Different types of factor analyses place different assumptions on $\boldsymbol{L}$ and $\boldsymbol{C}$ to make them identifiable/interpretable.

PCA in noisy data I

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, entries of $\boldsymbol{E}$ are independent, $\boldsymbol{E}_{gi} \sim (0, \sigma_g^2)$.

- PCA can be interpreted as a particular parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$
- Recall goal of PCA: identify most important sources of variation of $\boldsymbol{Y}$
  - PCs can be ordered by corresponding eigenvalue.

PCA in noisy data I

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, entries of $\boldsymbol{E}$ are independent, $\boldsymbol{E}_{gi} \sim (0, \sigma_g^2)$.

- PCA can be interpreted as a particular parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$
- Recall goal of PCA: identify most important sources of variation of $\boldsymbol{Y}$
  - PCs can be ordered by corresponding eigenvalue.

PCA in noisy data II

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, entries of $\boldsymbol{E}$ are independent, $\boldsymbol{E}_{gi} \sim (0, \sigma_g^2)$

- Claim: there exists a parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$ s.t.
  1. $E(\boldsymbol{Y} \mid \boldsymbol{C}) = \boldsymbol{L}\boldsymbol{C}^T$.
  2. $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, $\lambda_1 \geq \cdots \geq \lambda_K > 0$.
     - To identify $\boldsymbol{C}_{\cdot 1}, \ldots, \boldsymbol{C}_{\cdot K}$ up to sign, need to assume $\lambda_1 > \cdots > \lambda_K > 0$. Why?
  3. λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- To show this: compute the singular value decomposition of $E(\boldsymbol{Y} \mid \boldsymbol{C}) = \boldsymbol{L}\boldsymbol{C}^T$!
- kth factor is $\boldsymbol{C}_{\cdot k}$; has loading $\boldsymbol{L}_{\cdot k}$ with $p^{-1}\boldsymbol{L}_{\cdot k}^T\boldsymbol{L}_{\cdot k} = \lambda_k$. Here λ_k is the average effect size for the kth factor.
- Interpretation: $\boldsymbol{C}_{\cdot k}$ is the kth (out of K) most important factor.
- Goal: estimate $\boldsymbol{C}_{\cdot k}$ and $\boldsymbol{L}_{\cdot k}$.

PCA in noisy data II

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, entries of $\boldsymbol{E}$ are independent, $\boldsymbol{E}_{gi} \sim (0, \sigma_g^2)$

- Claim: there exists a parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$ s.t.
  1. $E(\boldsymbol{Y} \mid \boldsymbol{C}) = \boldsymbol{L}\boldsymbol{C}^T$.
  2. $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, $\lambda_1 \geq \cdots \geq \lambda_K > 0$.
     - To identify $\boldsymbol{C}_{\cdot 1}, \ldots, \boldsymbol{C}_{\cdot K}$ up to sign, need to assume $\lambda_1 > \cdots > \lambda_K > 0$. Why?
  3. λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- To show this: compute the singular value decomposition of $E(\boldsymbol{Y} \mid \boldsymbol{C}) = \boldsymbol{L}\boldsymbol{C}^T$!
- kth factor is $\boldsymbol{C}_{\cdot k}$; has loading $\boldsymbol{L}_{\cdot k}$ with $p^{-1}\boldsymbol{L}_{\cdot k}^T\boldsymbol{L}_{\cdot k} = \lambda_k$. Here λ_k is the average effect size for the kth factor.
- Interpretation: $\boldsymbol{C}_{\cdot k}$ is the kth (out of K) most important factor.
- Goal: estimate $\boldsymbol{C}_{\cdot k}$ and $\boldsymbol{L}_{\cdot k}$.

PCA in noisy data II

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, entries of $\boldsymbol{E}$ are independent, $\boldsymbol{E}_{gi} \sim (0, \sigma_g^2)$

- Claim: there exists a parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$ s.t.
  1. $E(\boldsymbol{Y} \mid \boldsymbol{C}) = \boldsymbol{L}\boldsymbol{C}^T$.
  2. $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, $\lambda_1 \geq \cdots \geq \lambda_K > 0$.
     - To identify $\boldsymbol{C}_{\cdot 1}, \ldots, \boldsymbol{C}_{\cdot K}$ up to sign, need to assume $\lambda_1 > \cdots > \lambda_K > 0$. Why?
  3. λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- To show this: compute the singular value decomposition of $E(\boldsymbol{Y} \mid \boldsymbol{C}) = \boldsymbol{L}\boldsymbol{C}^T$!
- kth factor is $\boldsymbol{C}_{\cdot k}$; has loading $\boldsymbol{L}_{\cdot k}$ with $p^{-1}\boldsymbol{L}_{\cdot k}^T\boldsymbol{L}_{\cdot k} = \lambda_k$. Here λ_k is the average effect size for the kth factor.
- Interpretation: $\boldsymbol{C}_{\cdot k}$ is the kth (out of K) most important factor.
- Goal: estimate $\boldsymbol{C}_{\cdot k}$ and $\boldsymbol{L}_{\cdot k}$.

PCA in noisy data: estimation

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$

- $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \text{diag}\,(\lambda_1, \ldots, \lambda_K)$, $\lambda_1 > \cdots > \lambda_K > 0$
- λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- Claim: the first K eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}$ accurately estimate $\boldsymbol{C}$. Idea:
  1. $E\left(p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + E\left(p^{-1}\boldsymbol{E}^T\boldsymbol{E}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + (p^{-1}\sum_{g=1}^p \sigma_g^2) I_n = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + \bar{\sigma}^2 I_n$
  2. This implies eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y} \approx$ eigenvectors of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$!
- Note that argument relied on $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$. What if samples are related, i.e. $\boldsymbol{E}_{g\cdot} \sim (0, \boldsymbol{V}_g)$, $\boldsymbol{V}_g \neq \sigma_g^2 I_n$?
- How should we estimate $\boldsymbol{L}$? OLS using estimated design matrix $\hat{\boldsymbol{C}}$!

PCA in noisy data: estimation

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$

- $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \text{diag}\left(\lambda_1, \ldots, \lambda_K\right)$, $\lambda_1 > \cdots > \lambda_K > 0$
- λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- Claim: the first K eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}$ accurately estimate $\boldsymbol{C}$. Idea:
  1. $E\left(p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + E\left(p^{-1}\boldsymbol{E}^T\boldsymbol{E}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + (p^{-1}\sum_{g=1}^p \sigma_g^2) I_n = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + \bar{\sigma}^2 I_n$
  2. This implies eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y} \approx$ eigenvectors of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$!
- Note that argument relied on $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$. What if samples are related, i.e. $\boldsymbol{E}_{g\cdot} \sim (0, \boldsymbol{V}_g)$, $\boldsymbol{V}_g \neq \sigma_g^2 I_n$?
- How should we estimate $\boldsymbol{L}$? OLS using estimated design matrix $\hat{\boldsymbol{C}}$!

PCA in noisy data: estimation

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$

- $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \text{diag}\left(\lambda_1, \ldots, \lambda_K\right)$, $\lambda_1 > \cdots > \lambda_K > 0$
- λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- Claim: the first K eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}$ accurately estimate $\boldsymbol{C}$. Idea:
  1. $E\left(p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + E\left(p^{-1}\boldsymbol{E}^T\boldsymbol{E}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + (p^{-1}\sum_{g=1}^p \sigma_g^2) I_n = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + \bar{\sigma}^2 I_n$
  2. This implies eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y} \approx$ eigenvectors of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$!
- Note that argument relied on $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$. What if samples are related, i.e. $\boldsymbol{E}_{g\cdot} \sim (0, \boldsymbol{V}_g)$, $\boldsymbol{V}_g \neq \sigma_g^2 I_n$?
- How should we estimate $\boldsymbol{L}$? OLS using estimated design matrix $\hat{\boldsymbol{C}}$!

PCA in noisy data: estimation

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$

- $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \text{diag}\left(\lambda_1, \ldots, \lambda_K\right)$, $\lambda_1 > \cdots > \lambda_K > 0$
- λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- Claim: the first K eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}$ accurately estimate $\boldsymbol{C}$. Idea:
  1. $E\left(p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + E\left(p^{-1}\boldsymbol{E}^T\boldsymbol{E}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + (p^{-1}\sum_{g=1}^p \sigma_g^2) I_n = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + \bar{\sigma}^2 I_n$
  2. This implies eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y} \approx$ eigenvectors of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$!
- Note that argument relied on $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$. What if samples are related, i.e. $\boldsymbol{E}_{g\cdot} \sim (0, \boldsymbol{V}_g)$, $\boldsymbol{V}_g \neq \sigma_g^2 I_n$?
- How should we estimate $\boldsymbol{L}$? OLS using estimated design matrix $\hat{\boldsymbol{C}}$!

PCA in noisy data: estimation

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$

- $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \text{diag}\left(\lambda_1, \ldots, \lambda_K\right)$, $\lambda_1 > \cdots > \lambda_K > 0$
- λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- Claim: the first K eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}$ accurately estimate $\boldsymbol{C}$. Idea:
  1. $E\left(p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + E\left(p^{-1}\boldsymbol{E}^T\boldsymbol{E}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + (p^{-1}\sum_{g=1}^p \sigma_g^2) I_n = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + \bar{\sigma}^2 I_n$
  2. This implies eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y} \approx$ eigenvectors of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$!
- Note that argument relied on $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$. What if samples are related, i.e. $\boldsymbol{E}_{g\cdot} \sim (0, \boldsymbol{V}_g)$, $\boldsymbol{V}_g \neq \sigma_g^2 I_n$?
- How should we estimate $\boldsymbol{L}$? OLS using estimated design matrix $\hat{\boldsymbol{C}}$!

PCA in noisy data: estimation

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$

- $n^{-1}\boldsymbol{C}^T\boldsymbol{C} = I_K$ and $p^{-1}\boldsymbol{L}^T\boldsymbol{L} = \text{diag}\left(\lambda_1, \ldots, \lambda_K\right)$, $\lambda_1 > \cdots > \lambda_K > 0$
- λ_k and $n^{-1/2}\boldsymbol{C}_{\cdot k}$ are the kth eigenvalue and eigenvector of $n^{-1}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$
- Claim: the first K eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}$ accurately estimate $\boldsymbol{C}$. Idea:
  1. $E\left(p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + E\left(p^{-1}\boldsymbol{E}^T\boldsymbol{E}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + (p^{-1}\sum_{g=1}^p \sigma_g^2) I_n = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + \bar{\sigma}^2 I_n$
  2. This implies eigenvectors of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y} \approx$ eigenvectors of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$!
- Note that argument relied on $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$. What if samples are related, i.e. $\boldsymbol{E}_{g\cdot} \sim (0, \boldsymbol{V}_g)$, $\boldsymbol{V}_g \neq \sigma_g^2 I_n$?
- How should we estimate $\boldsymbol{L}$? OLS using estimated design matrix $\hat{\boldsymbol{C}}$!

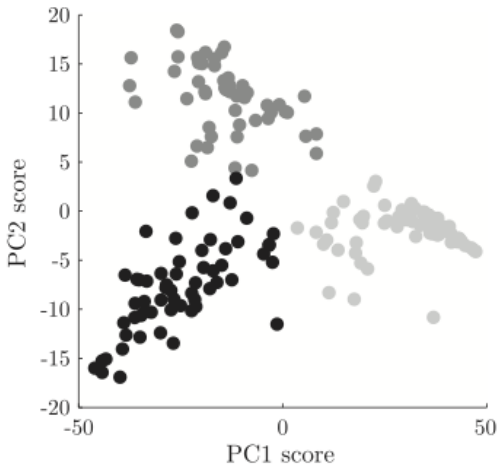
Asymptotic properties (McKennan & Nicolae, 2018)

$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}^T_{n\times K} + \boldsymbol{E}_{p\times n}$, $\boldsymbol{E}_{g\cdot} \sim (0, \sigma_g^2 I_n)$

- Consider the case $p \gtrsim n$ (we could have $p >> n$).
- $\text{Corr}\left(\hat{\boldsymbol{C}}_{k\cdot}, \boldsymbol{C}_{k\cdot}\right) = 1 - O_P\left\{(\lambda_k np)^{-1/2} + (\lambda_k p)^{-1}\right\}$
  - Blessing of dimensionality: as p gets larger, $\hat{\boldsymbol{C}}_{k\cdot}$ is more accurate!
  - This is a common theme in factor analysis problems.
- Estimate for ℓ_g is just as accurate as when $\boldsymbol{C}$ is known (under some assumptions)!
- PCA is incredibly powerful, and can accurately recover factors and loadings.

Example 1: clustering patients based on gene expression

- Expression of $p \approx 15,000$ genes measured on $n = 180$ lung cancer patients.
- Lung cancer has many sub-types (i.e. not just 1 type of lung cancer).
- Goal: can we classify patients based on expression? If we can, this can lead to personalized treatments.

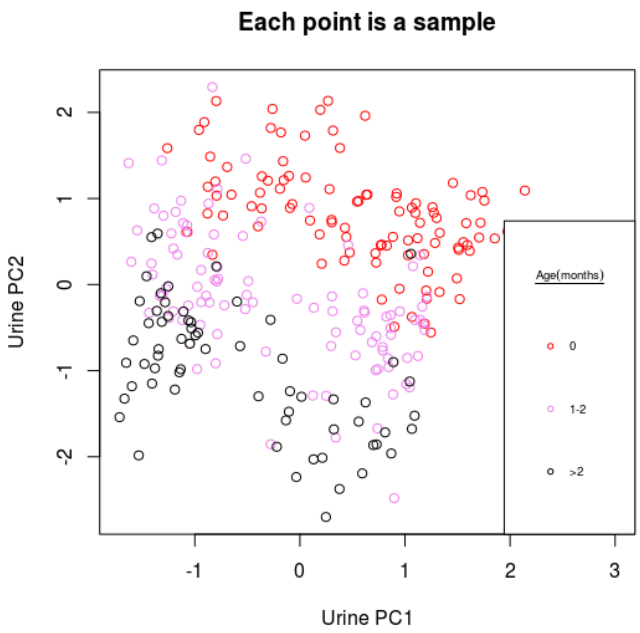


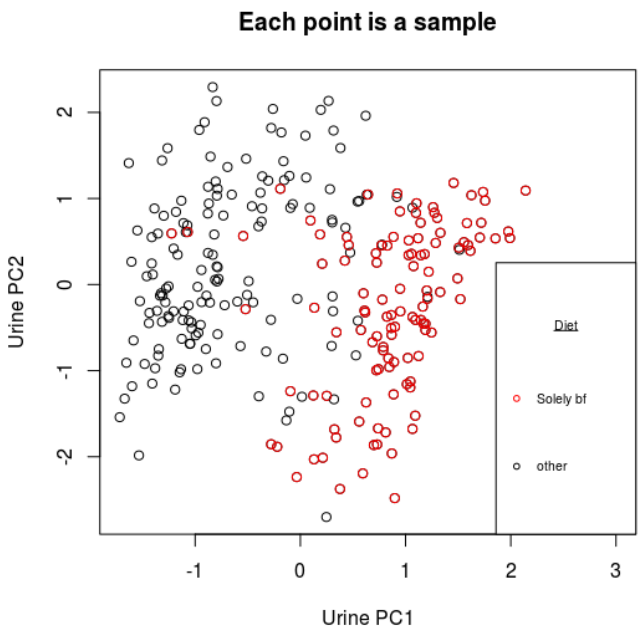
[1]

[1] Shen et al., 2016

Example 2: identifying confounders in metabolomics

- Metabolomics is the study of small molecule metabolites in tissues/bodily fluids.
  - Represent the end products of all cellular processes.
- What are the major sources of variation?
- Measured concentration of $p = 1,138$ metabolites in the urine of $n = 228$ infants.





[2]

[2] McKennan et al., 2020

Example 3: a bit of background

- Recall from high school biology: DNA is made up of the four letter alphabet A, C, T, G.
  - At each locus, inherit 1 copy from mother, and 1 from father
- Single nucleotide polymorphism (SNP): A single base pair (out of $\approx 3 \times 10^9$ base pairs) in the genome that shows variation across populations.
- There are hundreds of millions of SNPs in the human genome.
  - These are random mutations that have been inherited over thousands of generations.
  - At (nearly) every SNP, there is a major (i.e. most frequent) allele and a minor (i.e. less frequent allele).
- Genotype at SNP g in individual i can be written as

$$\begin{aligned} \boldsymbol{Y}_{gi} =& 1\,\{i \text{ inherited minor allele at } g \text{ from mother}\} \\ &+ 1\,\{i \text{ inherited minor allele at } g \text{ from father}\} \end{aligned}$$

- $\boldsymbol{Y}_{gi} \in \{0, 1, 2\}$

Example 3: a bit of background

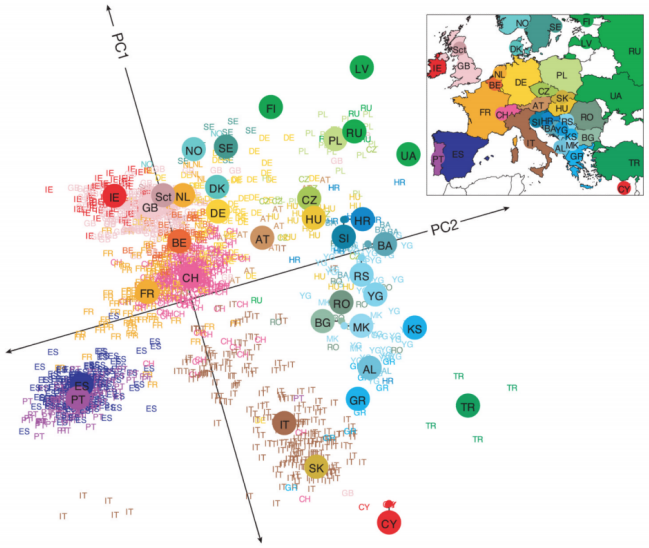
- Recall from high school biology: DNA is made up of the four letter alphabet A, C, T, G.
  - At each locus, inherit 1 copy from mother, and 1 from father
- Single nucleotide polymorphism (SNP): A single base pair (out of $\approx 3 \times 10^9$ base pairs) in the genome that shows variation across populations.
- There are hundreds of millions of SNPs in the human genome.
  - These are random mutations that have been inherited over thousands of generations.
  - At (nearly) every SNP, there is a major (i.e. most frequent) allele and a minor (i.e. less frequent allele).
- Genotype at SNP g in individual i can be written as

$$\begin{aligned} \boldsymbol{Y}_{gi} =& 1\,\{i \text{ inherited minor allele at } g \text{ from mother}\} \\ &+ 1\,\{i \text{ inherited minor allele at } g \text{ from father}\} \end{aligned}$$

- $\boldsymbol{Y}_{gi} \in \{0, 1, 2\}$

Example 3: genes mirror geography!!

- Define $\boldsymbol{Y} \in \mathbb{R}^{p \times n}$ to be the genotype matrix at $p = 197,146$ SNPs measured in $n = 1,387$ Europeans.
- $Y_{gi} \in \{0, 1, 2\}$ is the genotype.
- Goal: can we cluster individuals based on genotype?
- Intuition: individuals with similar genotypes are more related, and should cluster together.



[3]

[3] Novembre et al., 2008

Assessing the angle between subspaces

- Given $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$, how should we assess the angle θ between $\text{im}(\boldsymbol{C})$ and $\text{im}\left(\hat{\boldsymbol{C}}\right)$?
- To define a geometric angle between spaces, they must intersect. Do vector subspaces always intersect?
- When $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$ are vectors, this is easy.
  - $\cos(\theta) = \frac{|\boldsymbol{C}^T \hat{\boldsymbol{C}}|}{\|\boldsymbol{C}\|_2 \|\hat{\boldsymbol{C}}\|_2}$
- The more general case when $\boldsymbol{C}, \hat{\boldsymbol{C}} \in \mathbb{R}^{n \times K}$
  - $\theta = 0 \Leftrightarrow \text{im}(\boldsymbol{C}) = \text{im}\left(\hat{\boldsymbol{C}}\right)$
  - $\theta = \pi/2$ if there exists $v \in \text{im}(\boldsymbol{C})$ s.t. v is orthogonal to $\text{im}\left(\hat{\boldsymbol{C}}\right)$.
- $\cos(\theta) = \min_{\hat{v} \in \text{im}(\hat{\boldsymbol{C}})} \left\{ \max_{v \in \text{im}(\boldsymbol{C})} \left(\frac{\hat{v}^T v}{\|\hat{v}\|_2 \|v\|_2} \right) \right\}$
  - Also called the first **principal angle** between $\text{im}\left(\hat{\boldsymbol{C}}\right)$ and $\text{im}(\boldsymbol{C})$.

Assessing the angle between subspaces

- Given $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$, how should we assess the angle θ between $\text{im}\left(\boldsymbol{C}\right)$ and $\text{im}\left(\hat{\boldsymbol{C}}\right)$?
- To define a geometric angle between spaces, they must intersect. Do vector subspaces always intersect?
- When $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$ are vectors, this is easy.
  - $\cos\left(\theta\right) = \frac{|\boldsymbol{C}^T \hat{\boldsymbol{C}}|}{\|\boldsymbol{C}\|_2 \|\hat{\boldsymbol{C}}\|_2}$
- The more general case when $\boldsymbol{C}, \hat{\boldsymbol{C}} \in \mathbb{R}^{n \times K}$
  - $\theta = 0 \Leftrightarrow \text{im}\left(\boldsymbol{C}\right) = \text{im}\left(\hat{\boldsymbol{C}}\right)$
  - $\theta = \pi/2$ if there exists $v \in \text{im}\left(\boldsymbol{C}\right)$ s.t. v is orthogonal to $\text{im}\left(\hat{\boldsymbol{C}}\right)$.
- $\cos\left(\theta\right) = \min\limits_{\hat{v} \in \text{im}(\hat{\boldsymbol{C}})} \left\{ \max\limits_{v \in \text{im}(\boldsymbol{C})} \left(\frac{\hat{v}^T v}{\|\hat{v}\|_2 \|v\|_2} \right) \right\}$
  - Also called the first **principal angle** between $\text{im}\left(\hat{\boldsymbol{C}}\right)$ and $\text{im}\left(\boldsymbol{C}\right)$.

Assessing the angle between subspaces

- Given $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$, how should we assess the angle θ between $\text{im}\left(\boldsymbol{C}\right)$ and $\text{im}\left(\hat{\boldsymbol{C}}\right)$?
- To define a geometric angle between spaces, they must intersect. Do vector subspaces always intersect?
- When $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$ are vectors, this is easy.
  - $\cos\left(\theta\right) = \frac{|\boldsymbol{C}^T \hat{\boldsymbol{C}}|}{\|\boldsymbol{C}\|_2 \|\hat{\boldsymbol{C}}\|_2}$
- The more general case when $\boldsymbol{C}, \hat{\boldsymbol{C}} \in \mathbb{R}^{n \times K}$
  - $\theta = 0 \Leftrightarrow \text{im}\left(\boldsymbol{C}\right) = \text{im}\left(\hat{\boldsymbol{C}}\right)$
  - $\theta = \pi/2$ if there exists $v \in \text{im}\left(\boldsymbol{C}\right)$ s.t. v is orthogonal to $\text{im}\left(\hat{\boldsymbol{C}}\right)$.
- $\cos\left(\theta\right) = \min\limits_{\hat{v} \in \text{im}\left(\hat{\boldsymbol{C}}\right)} \left\{ \max\limits_{v \in \text{im}\left(\boldsymbol{C}\right)} \left(\frac{\hat{v}^T v}{\|\hat{v}\|_2 \|v\|_2} \right) \right\}$
  - Also called the first **principal angle** between $\text{im}\left(\hat{\boldsymbol{C}}\right)$ and $\text{im}\left(\boldsymbol{C}\right)$.

Assessing the angle between subspaces

- Given $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$, how should we assess the angle θ between $\text{im}\left(\boldsymbol{C}\right)$ and $\text{im}\left(\hat{\boldsymbol{C}}\right)$?
- To define a geometric angle between spaces, they must intersect. Do vector subspaces always intersect?
- When $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$ are vectors, this is easy.
  - $\cos\left(\theta\right) = \frac{|\boldsymbol{C}^T \hat{\boldsymbol{C}}|}{\|\boldsymbol{C}\|_2 \|\hat{\boldsymbol{C}}\|_2}$
- The more general case when $\boldsymbol{C}, \hat{\boldsymbol{C}} \in \mathbb{R}^{n \times K}$
  - $\theta = 0 \Leftrightarrow \text{im}\left(\boldsymbol{C}\right) = \text{im}\left(\hat{\boldsymbol{C}}\right)$
  - $\theta = \pi/2$ if there exists $v \in \text{im}\left(\boldsymbol{C}\right)$ s.t. v is orthogonal to $\text{im}\left(\hat{\boldsymbol{C}}\right)$.
- $\cos\left(\theta\right) = \min\limits_{\hat{v} \in \text{im}\left(\hat{\boldsymbol{C}}\right)} \left\{ \max\limits_{v \in \text{im}\left(\boldsymbol{C}\right)} \left(\frac{\hat{v}^T v}{\|\hat{v}\|_2 \|v\|_2} \right) \right\}$
  - Also called the first **principal angle** between $\text{im}\left(\hat{\boldsymbol{C}}\right)$ and $\text{im}\left(\boldsymbol{C}\right)$.

Other applications: adjusting for latent confounders in multivariate regression models

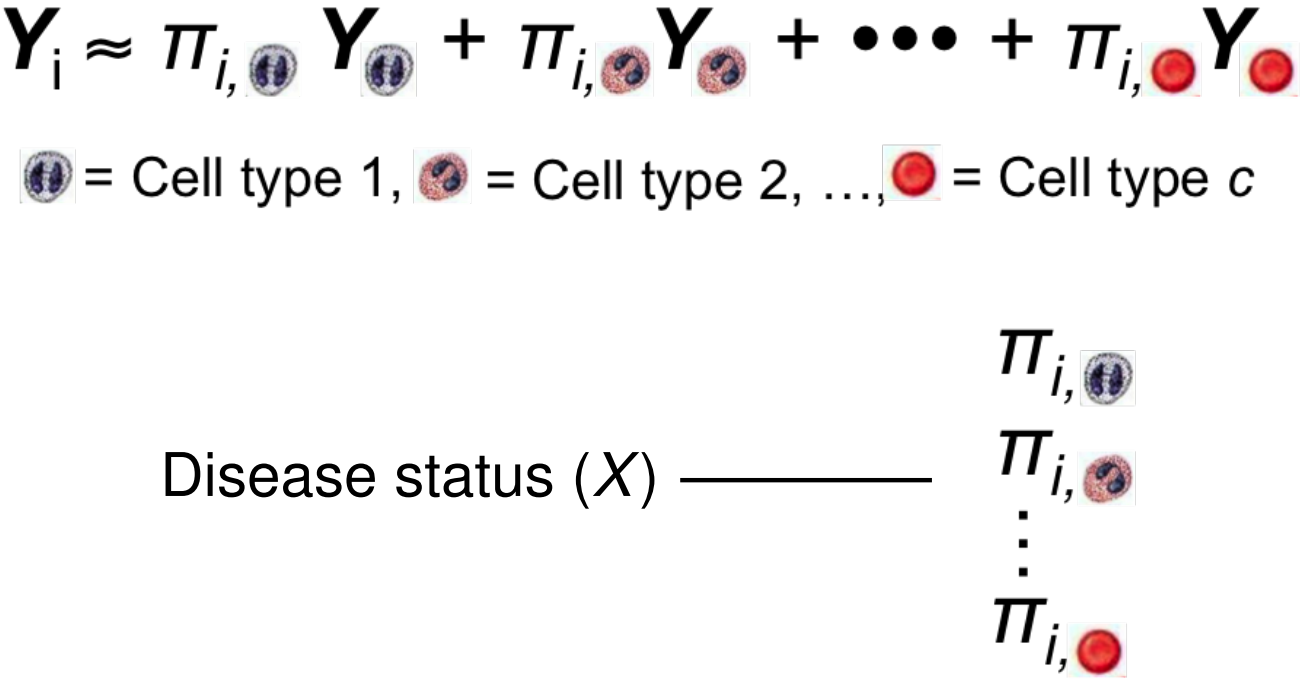
$$Y_i \approx \pi_{i,\bigcirc} Y_{\bigcirc} + \pi_{i,\bigcirc} Y_{\bigcirc} + \bullet\bullet\bullet + \pi_{i,\bigcirc} Y_{\bigcirc}$$

$\bigcirc$ = Cell type 1, $\bigcirc$ = Cell type 2, …, $\bigcirc$ = Cell type c

Disease status (X) ——— $\pi_{i,\bigcirc}$, $\pi_{i,\bigcirc}$, ⋮, $\pi_{i,\bigcirc}$

- If we want to understand relationship between disease status and expression, must account for cell type.
- Problem: we only observed express (Y) and disease status (X). Cell type is unobserved!

Other applications: adjusting for latent confounders in multivariate regression models



- If we want to understand relationship between disease status and expression, must account for cell type.
- Problem: we only observed express (Y) and disease status (X). Cell type is unobserved!

A model for data with latent confounders

- You now have the tools to model these data!
  - $\boldsymbol{Y}_{g\cdot} \in \mathbb{R}^n$: expression at gene g. We measure (i.e. observed) expression.
  - $\boldsymbol{X} \in \mathbb{R}^n$: covariate of interest, i.e. disease status. This is observed and non-random.
  - $\boldsymbol{C} \in \mathbb{R}^{n \times K}$: cell type and other latent confounders (e.g. diet, batch).
  - We will ignore other observed nuisance covariates not of interest (e.g. the intercept).
- $\boldsymbol{Y}_{g\cdot} = \boldsymbol{X}\beta_g + \boldsymbol{C}\ell_g + \boldsymbol{e}_g$, $\boldsymbol{e}_g \sim (\boldsymbol{0}, \sigma_g^2 I_n)$.
  - Goal: Estimate β_g.
  - Problem: $\boldsymbol{C}$ may be correlated with $\boldsymbol{X}$!
  - Other problem: $\boldsymbol{C}$ can induce correlations across $\hat{\beta}_1, \ldots, \hat{\beta}_p$ (recall Benjamini-Hochberg & other FDR controlling procedures fail when test statistics are correlated).
- We will look at this in depth next time...

A model for data with latent confounders

- You now have the tools to model these data!
  - $\boldsymbol{Y}_{g\cdot} \in \mathbb{R}^n$: expression at gene g. We measure (i.e. observed) expression.
  - $\boldsymbol{X} \in \mathbb{R}^n$: covariate of interest, i.e. disease status. This is observed and non-random.
  - $\boldsymbol{C} \in \mathbb{R}^{n \times K}$: cell type and other latent confounders (e.g. diet, batch).
  - We will ignore other observed nuisance covariates not of interest (e.g. the intercept).
- $\boldsymbol{Y}_{g\cdot} = \boldsymbol{X}\beta_g + \boldsymbol{C}\ell_g + \boldsymbol{e}_g$, $\boldsymbol{e}_g \sim (\boldsymbol{0}, \sigma_g^2 I_n)$.
  - Goal: Estimate β_g.
  - Problem: $\boldsymbol{C}$ may be correlated with $\boldsymbol{X}$!
  - Other problem: $\boldsymbol{C}$ can induce correlations across $\hat{\beta}_1, \ldots, \hat{\beta}_p$ (recall Benjamini-Hochberg & other FDR controlling procedures fail when test statistics are correlated).
- We will look at this in depth next time...