Applied Statistical Methods II

Some additional topics II

How do we choose the number of latent factors

- $\boldsymbol{Y}_{p \times n} = \boldsymbol{L}_{p \times K} \boldsymbol{C}_{n \times K}^T + \boldsymbol{E}_{p \times n}$
- The rows $g = 1, \ldots, p$ of $\boldsymbol{Y}$ are genes (or survey questions, companies, etc.).
  - We assume each row behaves as a standard linear model, $\boldsymbol{E}_{g\cdot} \sim (\boldsymbol{0}, \sigma^2 I_n)$
  - Entries of $\boldsymbol{E}$ are independent.
- The columns $i = 1, \ldots, n$ of $\boldsymbol{Y}$ are samples (or individuals, time, etc.)
- Ultimate goal is to understand $\boldsymbol{L} \in \mathbb{R}^{p \times K}$ and $\boldsymbol{C} \in \mathbb{R}^{n \times K}$.
- Problem: we don't know K! How should we estimate it?
- Choosing K is likely the most important part of exploratory factor analysis (Brown, 2014).

How do we choose the number of latent factors

- $\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}_{n\times K}^{T} + \boldsymbol{E}_{p\times n}$
- The rows $g = 1, \ldots, p$ of $\boldsymbol{Y}$ are genes (or survey questions, companies, etc.).
  - We assume each row behaves as a standard linear model, $\boldsymbol{E}_{g\cdot} \sim (\boldsymbol{0}, \sigma^2 I_n)$
  - Entries of $\boldsymbol{E}$ are independent.
- The columns $i = 1, \ldots, n$ of $\boldsymbol{Y}$ are samples (or individuals, time, etc.)
- Ultimate goal is to understand $\boldsymbol{L} \in \mathbb{R}^{p\times K}$ and $\boldsymbol{C} \in \mathbb{R}^{n\times K}$.
- Problem: we don't know K! How should we estimate it?
- Choosing K is likely the most important part of exploratory factor analysis (Brown, 2014).

Using a scree plot

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

- Plot the eigenvalues of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}$.
- Recall

$$E\left(p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + E(p^{-1}\boldsymbol{E}^T\boldsymbol{E}) \underbrace{=}_{\bar{\sigma}^2 = p^{-1}\sum_{g=1}^{p}\sigma_g^2} \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + \bar{\sigma}^2 I_n$$

- Idea: if the eigenvalues of the signal $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$ are large, they should be much bigger than those from the noise $p^{-1}\boldsymbol{E}^T\boldsymbol{E}$.
- Look for an **elbow** in the scree plot: a point below which the eigenvalues are “small” and can be ignored.
- Problems: qualitative, usual does not work in real data.

Using a scree plot

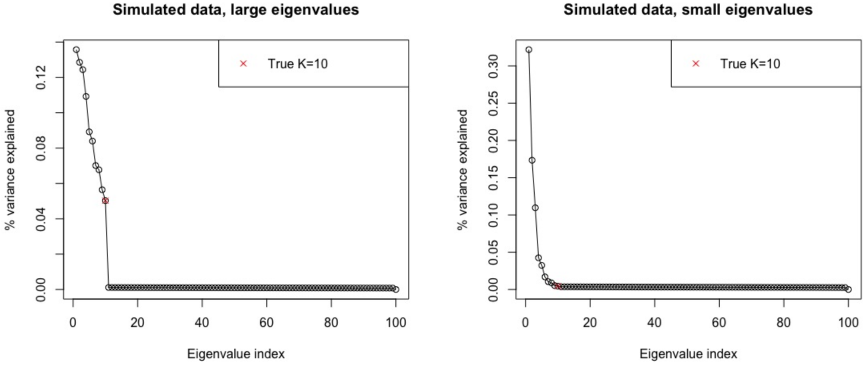
$$\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}_{n\times K}^T + \boldsymbol{E}_{p\times n},\ \boldsymbol{E}_{g\cdot} \sim (\boldsymbol{0}, \sigma^2 I_n)$$

- Plot the eigenvalues of $p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}$.
- Recall

$$E\left(p^{-1}\boldsymbol{Y}^T\boldsymbol{Y}\right) = \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + E(p^{-1}\boldsymbol{E}^T\boldsymbol{E})$$
$$\underbrace{=}_{\bar{\sigma}^2 = p^{-1}\sum_{g=1}^{p}\sigma_g^2} \boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T + \bar{\sigma}^2 I_n$$

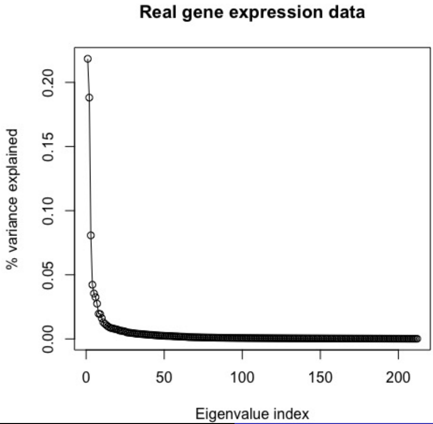
- Idea: if the eigenvalues of the signal $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$ are large, they should be much bigger than those from the noise $p^{-1}\boldsymbol{E}^T\boldsymbol{E}$.
- Look for an **elbow** in the scree plot: a point below which the eigenvalues are "small" and can be ignored.
- Problems: qualitative, usual does not work in real data.

Using a scree plot: simulated data



Using a scree plot: real gene expression data

- Real gene expression data from Knowles et al., 2018.
- Measured the expression of $p = 12,317$ genes in $n = 217$ samples.
- What value of K would you use?



More quantitative approaches to choose K

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

- Bai & Ng (2002): A BIC-like criteria.
  - $f(k) = R(k) + kP(p,n)$. Choose k to **minimize** $f(k)$.
  - $R(k) = (np)^{-1}\| \boldsymbol{Y} - \hat{\boldsymbol{L}}^{(k)} \left\{ \hat{\boldsymbol{C}}^{(k)} \right\}^T \|_F^2$ are the sum of squared residuals for PCA's estimates when it is assumed $K = k$.
  - $P(p,n)$ is a penalty function that satisfies $P(p,n) \to 0$ and $\min(n,p)P(p,n) \to \infty$ as $n,p \to \infty$.
  - Example: $P(p,n) = 1/\log\{\min(n,p)\}$
  - This is a groundbreaking paper.
  - **Serious problem**: This only works in theory and in practice if the eigenvalues of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$ are **very large** (i.e. $\asymp n$).
  - In these cases, we might as well use a scree plot!
- There are lots of similar methods (Ahn & Horenstein, 2013; Lu & Su, 2016; Li et al., 2018).
- All suffer from the same problems.

More quantitative approaches to choose K

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

- Bai & Ng (2002): A BIC-like criteria.
  - $f(k) = R(k) + kP(p,n)$. Choose k to **minimize** $f(k)$.
  - $R(k) = (np)^{-1}\| \boldsymbol{Y} - \hat{\boldsymbol{L}}^{(k)} \left\{ \hat{\boldsymbol{C}}^{(k)} \right\}^T \|_F^2$ are the sum of squared residuals for PCA's estimates when it is assumed $K = k$.
  - $P(p,n)$ is a penalty function that satisfies $P(p,n) \to 0$ and $\min(n,p)P(p,n) \to \infty$ as $n, p \to \infty$.
  - Example: $P(p,n) = 1/\log\{\min(n,p)\}$
  - This is a groundbreaking paper.
  - **Serious problem**: This only works in theory and in practice if the eigenvalues of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$ are **very large** (i.e. $\asymp n$).
  - In these cases, we might as well use a scree plot!
- There are lots of similar methods (Ahn & Horenstein, 2013; Lu & Su, 2016; Li et al., 2018).
- All suffer from the same problems.

More quantitative approaches to choose K

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

- Bai & Ng (2002): A BIC-like criteria.
  - $f(k) = R(k) + kP(p,n)$. Choose k to **minimize** $f(k)$.
  - $R(k) = (np)^{-1}\| \boldsymbol{Y} - \hat{\boldsymbol{L}}^{(k)} \left\{ \hat{\boldsymbol{C}}^{(k)} \right\}^T \|_F^2$ are the sum of squared residuals for PCA's estimates when it is assumed $K = k$.
  - $P(p,n)$ is a penalty function that satisfies $P(p,n) \to 0$ and $\min(n,p)P(p,n) \to \infty$ as $n, p \to \infty$.
  - Example: $P(p,n) = 1/\log\{\min(n,p)\}$
  - This is a groundbreaking paper.
  - **Serious problem**: This only works in theory and in practice if the eigenvalues of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$ are **very large** (i.e. $\asymp n$).
  - In these cases, we might as well use a scree plot!
- There are lots of similar methods (Ahn & Horenstein, 2013; Lu & Su, 2016; Li et al., 2018).
- All suffer from the same problems.

More quantitative approaches to choose K

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

- Bai & Ng (2002): A BIC-like criteria.
  - $f(k) = R(k) + kP(p,n)$. Choose k to **minimize** $f(k)$.
  - $R(k) = (np)^{-1}\| \boldsymbol{Y} - \hat{\boldsymbol{L}}^{(k)}\left\{\hat{\boldsymbol{C}}^{(k)}\right\}^T \|_F^2$ are the sum of squared residuals for PCA's estimates when it is assumed $K = k$.
  - $P(p,n)$ is a penalty function that satisfies $P(p,n) \to 0$ and $\min(n,p)P(p,n) \to \infty$ as $n,p \to \infty$.
  - Example: $P(p,n) = 1/\log\{\min(n,p)\}$
  - This is a groundbreaking paper.
  - **Serious problem**: This only works in theory and in practice if the eigenvalues of $\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^T\boldsymbol{L}\right)\boldsymbol{C}^T$ are **very large** (i.e. $\asymp n$).
  - In these cases, we might as well use a scree plot!
- There are lots of similar methods (Ahn & Horenstein, 2013; Lu & Su, 2016; Li et al., 2018).
- All suffer from the same problems.

Parallel analysis (PA) to choose K

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

- A more data-driven approach, based on permutation.
- $\boldsymbol{Y} = \textbf{Signal} + \textbf{Noise}$.
- Idea: we only include a factor k if it's eigenvalue is suitably greater than that of the noise.
- Goal: need to understand singular values of the noise.
- **Option 1**: use random matrix theory (RMT) to understand the singular values of $\boldsymbol{E}$.
  - There are some really beautiful results here. See Chapter 5 of Eldar & Kutyniok (2012) for some of them.
  - Problem: results are sensitive to distributional assumptions on $\boldsymbol{E}$.
- **Option 2**: Use the observed data to estimate singular values of $\boldsymbol{E}$.
  - Called parallel analysis (PA).
  - A very old idea (dates back to at least 1992). The first theoretical results were recently published (Dobriban, 2020).

Parallel analysis (PA) to choose K

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

- A more data-driven approach, based on permutation.
- $\boldsymbol{Y}$ = **Signal** + **Noise**.
- Idea: we only include a factor k if it's eigenvalue is suitably greater than that of the noise.
- Goal: need to understand singular values of the noise.
- **Option 1**: use random matrix theory (RMT) to understand the singular values of $\boldsymbol{E}$.
  - There are some really beautiful results here. See Chapter 5 of Eldar & Kutyniok (2012) for some of them.
  - Problem: results are sensitive to distributional assumptions on $\boldsymbol{E}$.
- **Option 2**: Use the observed data to estimate singular values of $\boldsymbol{E}$.
  - Called parallel analysis (PA).
  - A very old idea (dates back to at least 1992). The first theoretical results were recently published (Dobriban, 2020).

Parallel analysis (PA) to choose K

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

- A more data-driven approach, based on permutation.
- $\boldsymbol{Y}$ = **Signal** + **Noise**.
- Idea: we only include a factor k if it's eigenvalue is suitably greater than that of the noise.
- Goal: need to understand singular values of the noise.
- **Option 1**: use random matrix theory (RMT) to understand the singular values of $\boldsymbol{E}$.
  - There are some really beautiful results here. See Chapter 5 of Eldar & Kutyniok (2012) for some of them.
  - Problem: results are sensitive to distributional assumptions on $\boldsymbol{E}$.
- **Option 2**: Use the observed data to estimate singular values of $\boldsymbol{E}$.
  - Called parallel analysis (PA).
  - A very old idea (dates back to at least 1992). The first theoretical results were recently published (Dobriban, 2020).

PA: the algorithm

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

- Mean center the rows of $\boldsymbol{Y}$, i.e. $\boldsymbol{Y} \leftarrow \boldsymbol{Y}\left(I_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^T\right)$
- Let $\delta_1, \delta_2, \ldots, \delta_{\min(n,p)}$ be the singular values of $\boldsymbol{Y}$.
- Idea: choose $\hat{K}$ s.t. $\delta_{\hat{K}}$ is suitably larger than singular values of the noise $\boldsymbol{E}$.
- To approximate the singular values of the noise $\boldsymbol{E}$:
  1. "Break" the signal $\boldsymbol{L}\boldsymbol{C}^T$ by independently permuting entries in the rows of $\boldsymbol{Y}$: $\tilde{\boldsymbol{E}}_{g\cdot}^{(b)} = \Pi_g \boldsymbol{Y}_{g\cdot}$, Π_g a random permutation matrix.
  2. Let $\delta_1^{(b)}, \delta_2^{(b)}, \ldots, \delta_{\min(n,p)}^{(b)}$ be the s.v. of $\tilde{\boldsymbol{E}}^{(b)}$.
  3. Repeat for $b = 1, \ldots, B$. This gives you an approximate dist'n of the $\sigma_k\left(\boldsymbol{E}\right)$ (the kth singular value of $\boldsymbol{E}$).
- Keep factor k if $\delta_k > 95\%$ of the of the $\delta_k^{(b)}$'s.
- **Problem**: $\tilde{\boldsymbol{E}}^{(b)}$ is often very different from $\boldsymbol{E}$.
- Subject to **eigenvalue shadowing**: $\tilde{\boldsymbol{E}}^{(b)}$ is contaminated by factors with large eigenvalues.

PA: the algorithm

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

- Mean center the rows of $\boldsymbol{Y}$, i.e. $\boldsymbol{Y} \leftarrow \boldsymbol{Y}\left(I_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^T\right)$
- Let $\delta_1, \delta_2, \ldots, \delta_{\min(n,p)}$ be the singular values of $\boldsymbol{Y}$.
- Idea: choose $\hat{K}$ s.t. $\delta_{\hat{K}}$ is suitably larger than singular values of the noise $\boldsymbol{E}$.
- To approximate the singular values of the noise $\boldsymbol{E}$:
  1. "Break" the signal $\boldsymbol{LC}^T$ by independently permuting entries in the rows of $\boldsymbol{Y}$: $\tilde{\boldsymbol{E}}_{g\cdot}^{(b)} = \boldsymbol{\Pi}_g \boldsymbol{Y}_{g\cdot}$, $\boldsymbol{\Pi}_g$ a random permutation matrix.
  2. Let $\delta_1^{(b)}, \delta_2^{(b)}, \ldots, \delta_{\min(n,p)}^{(b)}$ be the s.v. of $\tilde{\boldsymbol{E}}^{(b)}$.
  3. Repeat for $b = 1, \ldots, B$. This gives you an approximate dist'n of the $\sigma_k(\boldsymbol{E})$ (the kth singular value of $\boldsymbol{E}$).
- Keep factor k if $\delta_k > 95\%$ of the of the $\delta_k^{(b)}$'s.
- **Problem**: $\tilde{E}^{(b)}$ is often very different from $\boldsymbol{E}$.
- Subject to **eigenvalue shadowing**: $\tilde{\boldsymbol{E}}^{(b)}$ is contaminated by factors with large eigenvalues.

PA: the algorithm

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

- Mean center the rows of $\boldsymbol{Y}$, i.e. $\boldsymbol{Y} \leftarrow \boldsymbol{Y}\left(I_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^T\right)$
- Let $\delta_1, \delta_2, \ldots, \delta_{\min(n,p)}$ be the singular values of $\boldsymbol{Y}$.
- Idea: choose $\hat{K}$ s.t. $\delta_{\hat{K}}$ is suitably larger than singular values of the noise $\boldsymbol{E}$.
- To approximate the singular values of the noise $\boldsymbol{E}$:
  1. "Break" the signal $\boldsymbol{LC}^T$ by independently permuting entries in the rows of $\boldsymbol{Y}$: $\tilde{\boldsymbol{E}}_{g\cdot}^{(b)} = \boldsymbol{\Pi}_g \boldsymbol{Y}_{g\cdot}$, $\boldsymbol{\Pi}_g$ a random permutation matrix.
  2. Let $\delta_1^{(b)}, \delta_2^{(b)}, \ldots, \delta_{\min(n,p)}^{(b)}$ be the s.v. of $\tilde{\boldsymbol{E}}^{(b)}$.
  3. Repeat for $b = 1, \ldots, B$. This gives you an approximate dist'n of the $\sigma_k(\boldsymbol{E})$ (the kth singular value of $\boldsymbol{E}$).
- Keep factor k if $\delta_k >$ 95% of the of the $\delta_k^{(b)}$'s.
- **Problem**: $\tilde{E}^{(b)}$ is often very different from $\boldsymbol{E}$.
- Subject to **eigenvalue shadowing**: $\tilde{\boldsymbol{E}}^{(b)}$ is contaminated by factors with large eigenvalues.

PA: the algorithm

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

- Mean center the rows of $\boldsymbol{Y}$, i.e. $\boldsymbol{Y} \leftarrow \boldsymbol{Y}\left(I_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^T\right)$
- Let $\delta_1, \delta_2, \ldots, \delta_{\min(n,p)}$ be the singular values of $\boldsymbol{Y}$.
- Idea: choose $\hat{K}$ s.t. $\delta_{\hat{K}}$ is suitably larger than singular values of the noise $\boldsymbol{E}$.
- To approximate the singular values of the noise $\boldsymbol{E}$:
  1. "Break" the signal $\boldsymbol{LC}^T$ by independently permuting entries in the rows of $\boldsymbol{Y}$: $\tilde{\boldsymbol{E}}_{g\cdot}^{(b)} = \boldsymbol{\Pi}_g \boldsymbol{Y}_{g\cdot}$, $\boldsymbol{\Pi}_g$ a random permutation matrix.
  2. Let $\delta_1^{(b)}, \delta_2^{(b)}, \ldots, \delta_{\min(n,p)}^{(b)}$ be the s.v. of $\tilde{\boldsymbol{E}}^{(b)}$.
  3. Repeat for $b = 1, \ldots, B$. This gives you an approximate dist'n of the $\sigma_k(\boldsymbol{E})$ (the kth singular value of $\boldsymbol{E}$).
- Keep factor k if $\delta_k >$ 95% of the of the $\delta_k^{(b)}$'s.
- **Problem**: $\tilde{\boldsymbol{E}}^{(b)}$ is often very different from $\boldsymbol{E}$.
- Subject to **eigenvalue shadowing**: $\tilde{\boldsymbol{E}}^{(b)}$ is contaminated by factors with large eigenvalues.

Other algorithms

$\boldsymbol{Y}_{p \times n} = \boldsymbol{L}_{p \times K} \boldsymbol{C}_{n \times K}^T + \boldsymbol{E}_{p \times n}$, $\boldsymbol{E}_{g\cdot} \sim (\mathbf{0}, \sigma^2 I_n)$. Assume rows of $\boldsymbol{E}$ are independent.

- Bi-cross validation (Wang & Owen, 2016; McKennan & Nicolae, 2019).
- Idea: partition columns of $\boldsymbol{Y}$ into training (f) and test ($-f$) sets

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_{(-f)} \\ \boldsymbol{Y}_f \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{(-f)} \boldsymbol{C}^T \\ \boldsymbol{L}_f \boldsymbol{C}^T \end{bmatrix} + \begin{bmatrix} \boldsymbol{E}_{(-f)} \\ \boldsymbol{E}_f \end{bmatrix}$$

- Key observation: $\boldsymbol{E}_{(-f)}$ is independent of $\boldsymbol{E}_f$!
- Can estimate $\boldsymbol{C}$ from $\boldsymbol{Y}_f$.
- Test estimate using $\boldsymbol{Y}_{(-f)}$.
- In my (biased) opinion, this is the most reliable way to estimate K.
- Has really beautiful theory (McKennan & Nicolae, 2019).