Sparse Principal Component Analysis

Motivation

- **Example 1**: Consider a study that tries to examine the correlations between bacteria in soil, where there could be 1000's of species of bacteria, but only a few samples are available.
- **Example 2**: Consider a psychology study that wishes to measure satisfaction and anxiety.

Both of these studies present hard problems for statistical analysis.

Factor Analysis

$$\mathbf{d} = \mathbf{\Lambda} \mathbf{F} + \mathbf{e}$$

Example with 2 factors and 8 observed variables

$$
\begin{array}{cccccc}
\mathbf{d} & = & \mathbf{\Lambda} & \mathbf{F} & + & \mathbf{e}
\end{array}
$$

$$
\begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{pmatrix}
=
\begin{pmatrix}
\lambda_{11} & \lambda_{12} \\
\lambda_{21} & \lambda_{22} \\
\lambda_{31} & \lambda_{32} \\
\lambda_{41} & \lambda_{42} \\
\lambda_{51} & \lambda_{52} \\
\lambda_{61} & \lambda_{62} \\
\lambda_{71} & \lambda_{27} \\
\lambda_{81} & \lambda_{82}
\end{pmatrix}
\begin{pmatrix} F_1 \\ F_2 \end{pmatrix}
+
\begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_{i,8} \end{pmatrix}.
$$

The Model: $\mathbf{d} = \mathbf{\Lambda F} + \mathbf{e}$

Assuming that the factors, and observable data are standardized:

$$\begin{aligned} cov(\mathbf{F}) &= \boldsymbol{I} \\ cov(\mathbf{e}) &= \mathbf{\Omega} \\ cov(\mathbf{d}) &= \mathbf{\Sigma} \\ &= \mathbf{\Lambda}\boldsymbol{I}\mathbf{\Lambda}^\top + \mathbf{\Omega} \end{aligned}$$

$\mathbf{z} = \mathbf{\Lambda F} + \mathbf{e}$

$$\begin{array}{rcl} d_1 & = & \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1p}F_p + e_1 \\ d_2 & = & \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2p}F_p + e_2 \\ \vdots & & \vdots \\ d_k & = & \lambda_{k1}F_1 + \lambda_{k2}F_2 + \cdots + \lambda_{kp}F_p + e_k \end{array}$$

$$\begin{array}{rcl} Var(d_1) & = & \lambda_{11}^2 + \lambda_{12}^2 + \cdots + \lambda_{1p}^2 + \omega_1 \\ Var(d_2) & = & \lambda_{21}^2 + \lambda_{22}^2 + \cdots + \lambda_{2p}^2 + \omega_2 \\ \vdots & & \vdots \\ Var(d_k) & = & \lambda_{k1}^2 + \lambda_{k2}^2 + \cdots + \lambda_{kp}^2 + \omega_k \end{array}$$

$Var(d_j) = 1$, so $\omega_j = 1 - \lambda_{j1}^2 - \lambda_{j2}^2 - \cdots - \lambda_{jp}^2$

Communality and Uniqueness

$Var(d_j) = \lambda_{j1}^2 + \lambda_{j2}^2 + \cdots + \lambda_{jp}^2 + \omega_j = 1$

- The explained variance in d_j is $\lambda_{j1}^2 + \lambda_{j2}^2 + \cdots + \lambda_{jp}^2$. It is called the *communality*.
- To get the communality, add the squared factor loadings in row j of $\mathbf{\Lambda}$.

If we could estimate the factor loadings

- We could estimate the correlation of each observable variable with each factor.
- We could assess how much of the variance in each observable variable comes from each factor.
- This could reveal what the underlying factors are, and what they mean.

So how do we estimate the factot loadings.

Link Between Factor analysis and Principal Component Analysis

Let $\mathbf{d} \sim \mathbf{MVN_p}(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ is a covariance matrix.
$\Sigma = CDC^T$
$\mathrm{z} = \mathrm{C}^T d$ are the principal components of $\mathbf{d}$

$$
\begin{aligned}
\mathbf{d} &= \mathbf{Cz} \\
&= \mathbf{C}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}\mathbf{z} \\
&= \underbrace{\mathbf{C}\mathbf{D}^{\frac{1}{2}}}_{k\times k}\underbrace{\mathbf{z}_2}_{k\times 1} \\
&= (\underbrace{\boldsymbol{\Lambda}}_{k\times t} \mid \underbrace{\mathbf{M}}_{k\times(k-t)}) \left(\frac{\mathbf{f}}{\mathbf{g}}\right) \begin{array}{l} \leftarrow t \times 1 \\ \leftarrow (k-t) \times 1 \end{array} \\
&= \boldsymbol{\Lambda}\mathbf{f} + \mathbf{Mg} \\
&= \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e}
\end{aligned}
$$

Okay .... How am I supposed to interpret this?

- Principal components won't tell you what they represent
- Initial Strategy: Try to figure that out from the factor loadings.

Question	PC1	PC2
How much do you sleep?	0.87	0.00
How stressed out are you about work?	0.95	0.00
Do you enjoy what you do?	0.00	0.88
How much social interaction do you get?	0.00	0.75

Okay .... How am I supposed to interpret this?

This one is less clear ...

Question	PC1	PC2
How much do you sleep?	0.32	0.53
How stressed out are you about work?	0.35	0.43
Do you enjoy what you do?	0.41	0.38
How much social interaction do you get?	0.80	0.15

The lesson: Sparsity makes things clearer, so now the question is: how do we make the principal components sparse

Method 1: Sparse Principal Component Analysis

Let Z_i be the $i'th$ principal component.

$$\hat{\beta} = \underset{\beta}{\text{argmin}} ||Z_i - \mathbf{D}\beta||^2 + \lambda||\beta||^2 + \lambda_1||\beta||_1. \tag{1}$$

Where $||\beta|| = \sum_{j=1}^{p} |\beta_j|$ and $\hat{V}_i = \frac{\hat{\beta}}{||\hat{\beta}||}$ gives the resulting approximated loadings for the ith principal component.

Adjusted total variance

To compute the amount of variance that is explained by the simplified components, we need to remove the co-linearity between the possibly correlated components:

$$\tilde{Z}_j = \hat{Z}_j - \mathbf{H}_{1,\ldots,j-1}\hat{Z}_j.$$

$\mathbf{H}_{1,\ldots,j-1}$ is the projection matrix onto the previous j-1 principal components. Then:

$$\sum_{j=1}^{k} ||\tilde{Z}_j||^2$$

gives the total variance explained by the first k components

Method 2: Rotating the Factors

$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Omega} = \mathbf{\Lambda}\mathbf{R}^\top\mathbf{R}\mathbf{\Lambda}^\top + \mathbf{\Omega}$

Post-multiplication of $\mathbf{\Lambda}$ by $\mathbf{R}^\top$ is often called "rotation of the factors."

$$\begin{aligned} \mathbf{d} &= \mathbf{\Lambda}\mathbf{F} + \mathbf{e} \\ &= (\mathbf{\Lambda}\mathbf{R}^\top)(\mathbf{R}\mathbf{F}) + \mathbf{e} \\ &= \mathbf{\Lambda}_2\mathbf{F}' + \mathbf{e}. \end{aligned}$$

- $\mathbf{F}' = \mathbf{R}\mathbf{F}$ is a set of *rotated* factors.
- All rotations of the factors produce the same covariance matrix of the observable data.

Strategy: Find a nice Rotation

Varimax Rotation

- The original idea was to maximize the variability of the *squared* loadings in each column.

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.87 & 0.00 \\ -0.95 & 0.00 \\ 0.79 & 0.00 \\ 0.00 & 0.88 \\ 0.00 & 0.75 \\ 0.00 & -0.94 \\ 0.00 & -0.82 \end{pmatrix}$$

- The results weren't great, so they fixed it up, expressing each squared factor loading as a proportion of the communality.
- Note that the criterion depends on the factor loadings only through the λ_{ij}^2.
- In practice, varimax rotation tends to maximize the squared loading of each observable variable with *just one underlying factor*.

The Varimax Criterion

The Varimax method simplifies the loadings of the principal components through an orthogonal rotation:

$$R_{VARIMAX} = \underset{R}{\operatorname{argmax}} \frac{1}{k} \sum_{j=1}^{k} \left(\frac{1}{p} \sum_{i=1}^{p} (VR)_{ij}^{4} - \left(\frac{1}{p} \sum_{i=1}^{p} (VR)_{ij}^{2} \right)^{2} \right)$$

Where $V = \Lambda H^{-1}$ where H is a diagonal matrix for the containing the communalities of each vactor as it's entries.

Model Comparison

- Generate data under a 2 factor model, with 8 observable variables.
- Each time, change the factor loadings
- Change the number of non zero loadings too.
- Simulate the data many times for each model
- Average out the results to account for error

Table of Numerical Values

Non Zero	MSE (SPCA)	MSE (Varimax)	Non Zero (SPCA)	Varimax (Non Zero)
0	3.48	1.75	10	11.6
1	4.17	2.02	10	11.0
2	3.51	1.83	9.33	11.9
3	4.30	2.01	9.33	11.6
4	4.11	1.83	8.67	11.5
5	5.07	2.09	8.67	11.1
6	3.63	2.17	8	11.1
7	4.43	2.42	8	10.9

Table: Simulation Results

References

- Kaiser, Henry F. "The varimax criterion for analytic rotation in factor analysis." Psychometrika 23.3 (1958): 187-200.
- Zou, Hui, Trevor Hastie, and Robert Tibshirani. "Sparse principal component analysis." Journal of computational and graphical statistics 15.2 (2006): 265-286.

Copyright Information