More Principal Components Ideas

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

"Sparse" PCs objective

In standard principal components analysis, the $\mathbf{v}_j$ are sometimes called "loadings" because (in light of the fact that $\mathbf{z}_j = \mathbf{X}\mathbf{v}_j$) they specify what linear combinations of variables x_j are used in making the various principal component vectors. If the $\mathbf{v}_j$ were "sparse" (had lots of 0's in them) interpretation of these loadings would be easier. So people have made proposals of alternative methods of defining "principal components" that will tend to produce sparse results. One due to Zou is as follows.

A 1st "sparse PC direction"

One might call a $\mathbf{v} \in \Re^p$ a first sparse principal component "direction" (it won't typically be a unit vector) if it is part of a minimizer (over choices of $\mathbf{v} \in \Re^p$ and $\boldsymbol{\theta} \in \Re^p$ with $\|\boldsymbol{\theta}\| = 1$) of the criterion

$$\sum_{i=1}^{N} \left\|\mathbf{x}_i - \boldsymbol{\theta}\mathbf{v}'\mathbf{x}_i\right\|^2 + \lambda \|\mathbf{v}\|^2 + \lambda_1 \|\mathbf{v}\|_1 \qquad (1)$$

for $\|\cdot\|_1$ the l_1 norm on $\Re^p$ and constants $\lambda \geq 0$ and $\lambda_1 \geq 0$. The last term in this expression is analogous to the lasso penalty on a vector of regression coefficients and produces the same kind of tendency to "0 out" entries that we saw in that context. If $\lambda_1 = 0$, the optimizing $\mathbf{v}$ is proportional to the ordinary first principal component direction. In fact, if $\lambda = \lambda_1 = 0$ and $N > p$, $\mathbf{v} = \boldsymbol{\theta}$ and the ordinary first principal component direction *is* the optimizer.

Multiple "sparse PC directions"

For multiple components, an analogue of the first case is a set of K vectors $\mathbf{v}_k \in \Re^p$ organized into a $p \times K$ matrix $\mathbf{V}$ that is part of a minimizer (over choices of $p \times K$ matrices $\mathbf{V}$ and $p \times K$ matrices $\mathbf{\Theta}$ with $\mathbf{\Theta}'\mathbf{\Theta} = \mathbf{I}$) of the criterion

$$\sum_{i=1}^{N} \left\| \mathbf{x}_i - \mathbf{\Theta}\mathbf{V}'\mathbf{x}_i \right\|^2 + \lambda \sum_{k=1}^{K} \left\| \mathbf{v}_k \right\|^2 + \sum_{k=1}^{K} \lambda_{1k} \left\| \mathbf{v}_k \right\|_1 \qquad (2)$$

for constants $\lambda \geq 0$ and $\lambda_{1k} \geq 0$. Zou has apparently provided effective algorithms for optimizing criteria (1) or (2).

Non-negative matrix factorization objective

There are contexts (for example, when data are counts) where it may not make intuitive sense to center inherently non-negative variables, so that $\mathbf{X}$ is naturally non-negative, and one might want to find non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ such that

$$\underset{N\times p}{\mathbf{X}} \approx \underset{N\times r}{\mathbf{W}}\ \underset{r\times p}{\mathbf{H}}$$

Here the emphasis might be on the columns of $\mathbf{W}$ as representing "positive components" of the (positive) $\mathbf{X}$, just as the columns of the matrix $\mathbf{UD}$ in SVD's provide the principal components of $\mathbf{X}$. Various optimization criteria could be set to guide the choice of $\mathbf{W}$ and $\mathbf{H}$.

Possible precise formulations

One might try to minimize

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\left(x_{ij}-(\mathbf{WH})_{ij}\right)^2$$

or maximize

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\left(x_{ij}\ln(\mathbf{WH})_{ij}-(\mathbf{WH})_{ij}\right)$$

over non-negative choices of **W** and **H**, and various algorithms for doing these have been proposed. (Notice that the second of these criteria is an extension of a loglikelihood for independent Poisson variables with means entries in **WH** to cases where the x_{ij} need only be non-negative, not necessarily integer.)

Fundamental limitations

While at first blush this enterprise seems sensible, there is a lack of uniqueness in a factorization producing a product $\mathbf{WH}$, and therefore how to interpret the columns of one of the many possible $\mathbf{W}$'s is not clear. (An easy way to see the lack of uniqueness is this. Suppose that all entries of the product $\mathbf{WH}$ are positive. Then for $\mathbf{E}$ a small enough (but not $\mathbf{0}$) matrix, all entries of $\mathbf{W}^* \equiv \mathbf{W}(\mathbf{I}+\mathbf{E}) \neq \mathbf{W}$ and $\mathbf{H}^* \equiv (\mathbf{I}+\mathbf{E})^{-1}\mathbf{H} \neq \mathbf{H}$ are positive, and $\mathbf{W}^*\mathbf{H}^* = \mathbf{WH}$.) Lacking some natural further restriction on the factors $\mathbf{W}$ and $\mathbf{H}$ (beyond non-negativity) it seems the practical usefulness of this basic idea is also lacking.

Archetypal analysis

Another approach to finding an interpretable factorization of $\mathbf{X}$ was provided by Cutler and Breiman in their "archetypal analysis." Again one means to write

$$\underset{N\times p}{\mathbf{X}} \approx \underset{N\times r}{\mathbf{W}}\ \underset{r\times p}{\mathbf{H}}$$

for appropriate $\mathbf{W}$ and $\mathbf{H}$. But here two restrictions are imposed, namely

1. the rows of $\mathbf{W}$ are probability vectors (so that the approximation to $\mathbf{X}$ is in terms of convex combinations/weighted averages of the rows of $\mathbf{H}$), and
2. $\underset{r\times p}{\mathbf{H}} = \underset{r\times N}{\mathbf{B}}\ \underset{N\times p}{\mathbf{X}}$ where the rows of $\mathbf{B}$ are probability vectors (so that the rows of $\mathbf{H}$ are in turn convex combinations/weighted averages of the rows of $\mathbf{X}$).

The r rows of $\mathbf{H} = \mathbf{BX}$ are the "prototypes" (?archetypes?) used to represent the data matrix $\mathbf{X}$.

Optimization problem

With this notation and restrictions, (stochastic matrices) $\mathbf{W}$ and $\mathbf{B}$ are chosen to minimize

$$\|\mathbf{X} - \mathbf{WBX}\|^2$$

It's clearly possible to rearrange the rows of a minimizing $\mathbf{B}$ and make corresponding changes in $\mathbf{W}$ without changing $\|\mathbf{X} - \mathbf{WBX}\|^2$. So strictly speaking, the optimization problem has multiple solutions. But in terms of the *set of rows* of $\mathbf{H}$ (a set of prototypes of size r) it's possible that this optimization problem often has a unique solution. (Symmetries induced in the set of N rows of $\mathbf{X}$ can be used to produce examples where it's clear that genuinely different sets of prototypes produce the same minimal value of $\|\mathbf{X} - \mathbf{WBX}\|^2$. But it seems likely that real data sets will usually lack such symmetries and lead to a single optimizing set of prototypes.)

Limitations

Emphasis in this version of the "approximate **X**" problem is on the set of prototypes as "representative data cases." This has to be taken with a grain of salt, since they are nearly always near the "edges" of the data set. This should be no surprise, as line segments between extreme cases in $\Re^p$ can be made to run close to cases in the "middle" of the data set, while line segments between interior cases in the data set will never be made to run close to extreme cases.

Independent Component Analysis set-up

Suppose that $\mathbf{X}$ is of rank p and has been centered. Based on the SVD

$$\underset{N\times p}{\mathbf{X}} = \underset{N\times p}{\mathbf{U}}\ \underset{p\times p}{\mathbf{D}}\ \underset{p\times p}{\mathbf{V}'}$$

consider the "sphered" version of the data matrix

$$\mathbf{X}^* = \sqrt{N}\mathbf{X}\mathbf{V}\mathbf{D}^{-1}$$

so that the sample correlation matrix is

$$\frac{1}{N}\left(\mathbf{X}^{*\prime}\mathbf{X}^*\right) = \mathbf{I}$$

The columns of $\mathbf{X}^*$ are then scaled principal components of the (centered) data matrix and we operate with and on $\mathbf{X}^*$. (For simplicity of notation, we'll henceforth drop the "$*$" on $\mathbf{X}$.)

ICA objective

ICA is an attempt to find latent probabilistic structure in terms of independent variables to account for the principal components. In particular (in its linear form) ICA attempts to model the N (transposed) rows of $\mathbf{X}$ as iid of the form

$$\underset{p\times 1}{\mathbf{x}_i} = \underset{p\times p}{\mathbf{A}}\,\underset{p\times 1}{\mathbf{s}_i} \tag{1}$$

for iid vectors $\mathbf{s}_i$, where the (marginal) distribution of the vectors $\mathbf{s}_i$ is one of independence of the p coordinates/components and the matrix $\mathbf{A}$ is an unknown parameter. We'll assume that $\mathrm{Cov}\mathbf{x} = \mathbf{I}$ and without any loss of generality assume that the covariance matrix for $\mathbf{s}$ is not only diagonal, but that $\mathrm{Cov}\mathbf{s} = \mathbf{I}$. Since then $\mathbf{I} = \mathrm{Cov}\mathbf{x} = \mathbf{A}\left(\mathrm{Cov}\mathbf{s}\right)\mathbf{A}' = \mathbf{A}\mathbf{A}'$, $\mathbf{A}$ must be orthogonal, and so

$$\mathbf{A}'\mathbf{x} = \mathbf{s}$$

ICA objective cont.

If one can estimate $\mathbf{A}$ with an orthogonal $\widehat{\mathbf{A}}$ then $\widehat{\mathbf{s}_i} \equiv \widehat{\mathbf{A}}' \mathbf{x}_i$ serves as an estimate of what vector of independent components led to the ith row of $\mathbf{X}$ and indeed

$$\widehat{\mathbf{S}} \equiv \mathbf{X}\widehat{\mathbf{A}}$$

has columns that provide predictions of the N (row) p-vectors $\mathbf{s}_i'$, and we might thus call those the "independent components" of $\mathbf{X}$ (just as we term the columns of $\mathbf{XV}$ the principal components of $\mathbf{X}$). There is a bit of arbitrariness in the representation (1) because the ordering of the coordinates of $\mathbf{s}$ and the corresponding rows of $\mathbf{A}$ is arbitrary. But this is no serious concern.

Estimating $\boldsymbol{A}$ and K-L divergence

The question is what one might use as a method to estimate **A** in (1). There are several possibilities. One discussed in HTF is related to entropy and Kullback-Leibler divergence. If one assumes that a (p-dimensional) random vector **Y** has a density f, with marginal densities $f_1, f_2, \ldots, f_p$ then an "independence version" of the distribution of **Y** has density $\prod_{j=1}^{p} f_j$ and the K-L divergence of the distribution of **Y** from its independence version is

$$K\left(f, \prod_{j=1}^{p} f_j\right) = \int f(\mathbf{y}) \ln\left(\frac{f(\mathbf{y})}{\prod_{j=1}^{p} f_j(y_j)}\right) d\mathbf{y}$$

Estimating $\boldsymbol{A}$ and K-L divergence cont.

Then

$$
\begin{aligned}
K\left(f, \prod_{j=1}^{p} f_j\right) &= \int f(\mathbf{y}) \ln f(\mathbf{y})\, d\mathbf{y} - \sum_{j=1}^{p} \int f(\mathbf{y}) \left(\ln f_j(y_j)\right) d\mathbf{y} \\
&= \int f(\mathbf{y}) \ln f(\mathbf{y})\, d\mathbf{y} - \sum_{j=1}^{p} \int f_j(y_j) \left(\ln f_j(y_j)\right) dy_j \\
&= \sum_{j=1}^{p} \mathcal{H}(Y_j) - \mathcal{H}(\mathbf{Y})
\end{aligned}
$$

for $\mathcal{H}$ the entropy function for a random argument, and this K-L divergence is a kind of (non-negative) difference between the information carried by $\mathbf{Y}$ (jointly) and the sum across the components of their individual information contents.

ICA matrix optimization problem

If one then thinks of $\mathbf{s}$ as random and of the form $\mathbf{A}'\mathbf{x}$ for random $\mathbf{x}$, it is perhaps sensible to seek an orthogonal $\mathbf{A}$ to minimize (for for $\mathbf{a}_j$ the jth column of $\mathbf{A}$)

$$\begin{aligned}\sum_{j=1}^{p} \mathcal{H}(s_j) - \mathcal{H}(\mathbf{s}) &= \sum_{j=1}^{p} \mathcal{H}(\mathbf{a}_j'\mathbf{x}) - \mathcal{H}(\mathbf{A}'\mathbf{x}) \\ &= \sum_{j=1}^{p} \mathcal{H}(\mathbf{a}_j'\mathbf{x}) - \mathcal{H}(\mathbf{x}) - \ln|\det \mathbf{A}| \\ &= \sum_{j=1}^{p} \mathcal{H}(\mathbf{a}_j'\mathbf{x}) - \mathcal{H}(\mathbf{x})\end{aligned}$$

Approximate ICA objective criterion

This is equivalent (for orthogonal $\mathbf{A}$) to maximization of

$$C(\mathbf{A}) = \sum_{j=1}^{p} \left(\mathcal{H}(z) - \mathcal{H}\left(\mathbf{a}_j' \mathbf{x}\right) \right) \tag{2}$$

for z standard normal and a common approximation is

$$\left(\mathcal{H}(z) - \mathcal{H}\left(\mathbf{a}_j' \mathbf{x}\right) \right) \approx \left(\mathrm{E}G(z) - \mathrm{E}G\left(\mathbf{a}_j' \mathbf{x}\right) \right)^2$$

for $G(u) \equiv \frac{1}{c} \ln \cosh(cu)$ for a $c \in [1, 2]$. Then, criterion (2) has the empirical approximation

$$\widehat{C}(\mathbf{A}) = \sum_{j=1}^{p} \left(\mathrm{E}G(z) - \frac{1}{N} \sum_{i=1}^{N} G\left(\mathbf{a}_j' \mathbf{x}_i'\right) \right)^2$$

where, $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. $\widehat{\mathbf{A}}$ can be taken to be an optimizer of $\widehat{C}(\mathbf{A})$.

ICA interpretation

Ultimately, this development produces a rotation matrix that makes the p entries of rotated and scaled principal component score vectors "look as independent as possible." This is thought of as resolution of a data matrix into its "independent sources" and as a technique for "blind source separation."