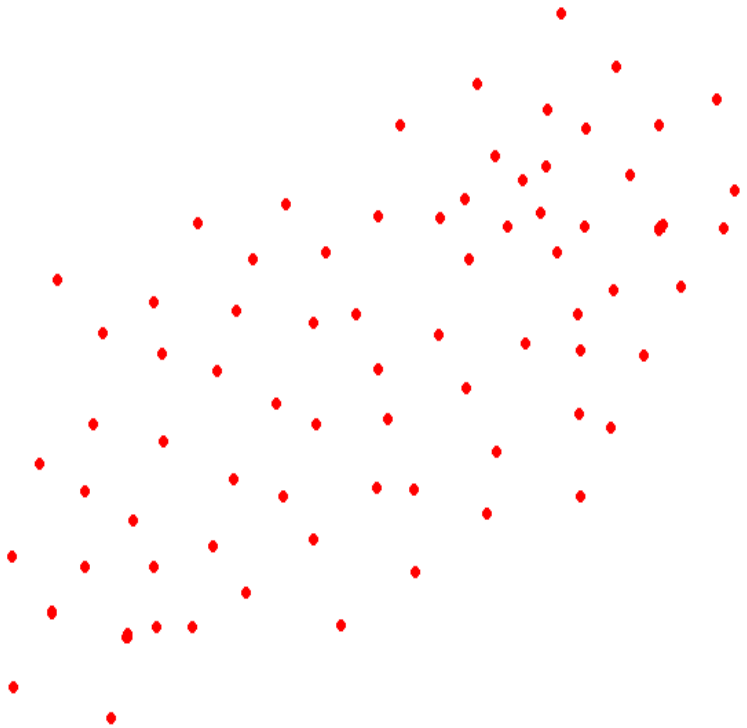
Methods for Sparse PCA

May 4, 2012

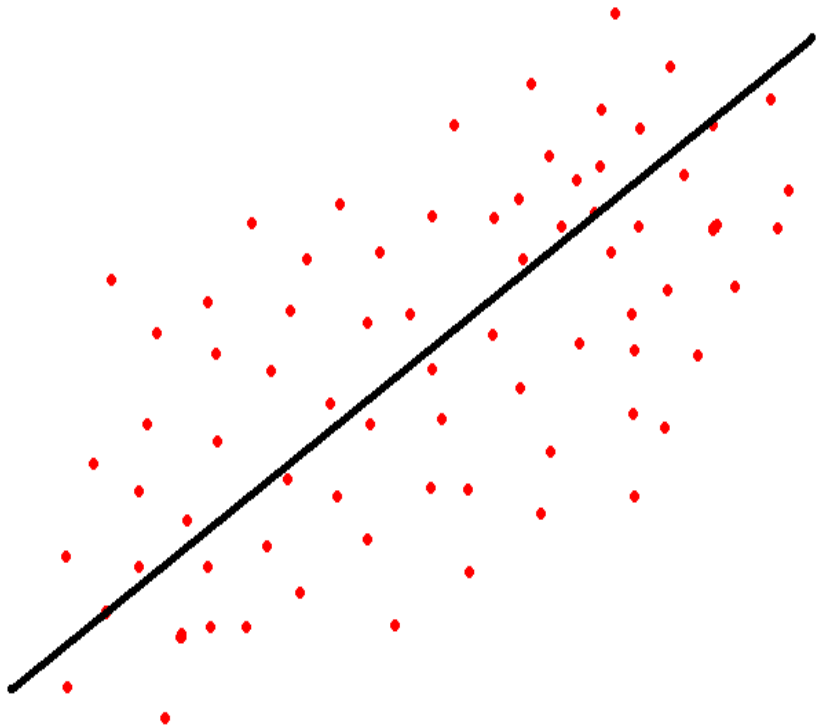
Principal Components Analysis

Principal Components Analysis is a popular tool for exploratory data analysis and dimension reduction in applied statistics.

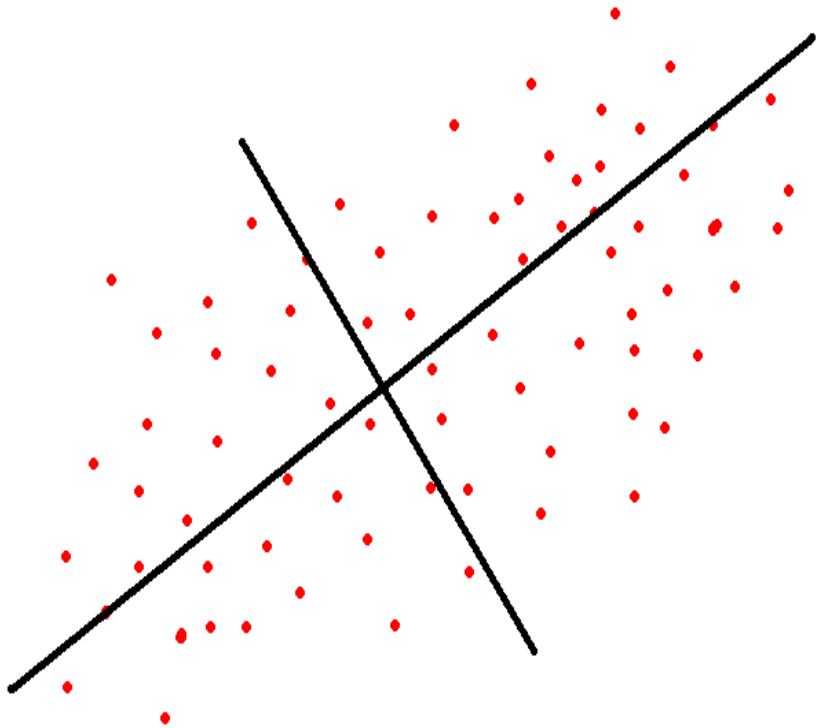
Principal Components Analysis: Example



Principal Components Analysis: Example



Principal Components Analysis: Example



Notation

Let $\mathbf{X}$ be a $n \times p$ matrix with standardized columns, that is: $\sum_{j=1}^{p} X_{ij} = 0$, $\sum_{i=1}^{n} X_{ij}^2 = 1$.

Three Ways to Arrive at First Principal Component

1. Maximal variance
2. Minimal reconstruction error
3. Best rank-1 approximation

Maximal Variance Approach

The first PC, $\mathbf{v}$, is the direction of **maximal variance**:

$$\mathbf{v} = \text{argmax}_{\mathbf{v}} \mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v} \text{ subject to } ||\mathbf{v}||_2 = 1$$

Minimal Reconstruction Error Approach

The first PC, **v**, **minimizes the reconstruction error**:

$$(\mathbf{u}, \mathbf{v}) = \text{argmin}_{\mathbf{u},\mathbf{v}} ||\mathbf{X} - \mathbf{X}\mathbf{v}\mathbf{u}^T||_F^2 \text{ subject to } ||\mathbf{u}||_2 = ||\mathbf{v}||_2 = 1$$

Best Rank-1 Approximation Approach

The first PC, **v**, **follows from the best rank-1 approximation**:

$$(\mathbf{u}, \mathbf{v}, d) = \text{argmin}_{\mathbf{u},\mathbf{v},d} ||\mathbf{X} - d\mathbf{u}\mathbf{v}^T||_F^2 \text{ subject to } ||\mathbf{u}||_2 = ||\mathbf{v}||_2 = 1$$

Principal Components: Three Approaches

$$\text{minimize}_{\mathbf{u},\mathbf{v}}\{||\mathbf{X} - \mathbf{X}\mathbf{v}\mathbf{u}^T||_F^2 + \lambda_1||\mathbf{v}||^2\}$$
$$\text{subject to } ||\mathbf{u}||^2 = 1$$

$$\text{maximize}_{\mathbf{v}}\{\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}\}$$
$$\text{subject to } ||\mathbf{v}||^2 = 1$$

$$\text{minimize}_{d,\mathbf{u},\mathbf{v}}\{||\mathbf{X} - d\mathbf{u}\mathbf{v}^T||_F^2\}$$
$$\text{subject to } ||\mathbf{u}||^2 = ||\mathbf{v}||^2 = 1$$

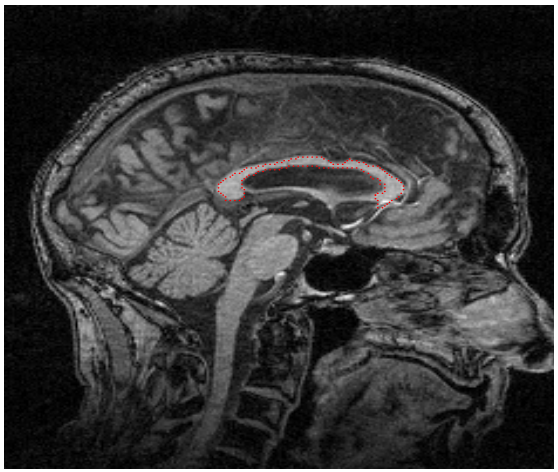
Sparse Principal Components Analysis

Suppose we want **sparse principal components**.

e.g. - Gene expression data - want to identify a **sparse** set of genes along which most of the variation in the data is really taking place.

Example

From a study involving 569 elderly persons



An example of a mid-aggital brain slice, with the corpus collosum annotated with landmarks.

Example- continued

Walking Speed

Verbal Fluency

Principal Components **Sparse Principal Components**

Standard and sparse principal components from a study of the corpus callosum variation. The shape variations corresponding to significant principal components (red curves) are overlaid on the mean CC shape (black curves).

Three Ways to Arrive at First Sparse Principal Component

1. Maximal variance ... subject to L_1 penalty
2. Minimal reconstruction error ... subject to L_1 penalty
3. Best rank-1 approximation ... subject to L_1 penalty

Maximal Variance Approach

$$\mathbf{v} = \text{argmax}_{\mathbf{v}} \mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v} \text{ subject to } ||\mathbf{v}||_2 = 1, ||\mathbf{v}||_1 \leq c$$

Citation: "SCoTLASS" method of Jolliffe et al. (2003)

Maximal Variance Approach

1. Criterion follows naturally from maximal variance description of principal components.
2. But, we are **maximizing** a **convex** function subject to a penalty... Not convex

Citation: Trendafilov and Jolliffe (2006)

Minimal Reconstruction Error Approach

$$(\mathbf{u}, \mathbf{v}) = \text{argmin}_{\mathbf{u},\mathbf{v}} ||\mathbf{X} - \mathbf{X}\mathbf{v}\mathbf{u}^T||_F^2 + \lambda_1 ||\mathbf{v}||_1 + \lambda_2 ||\mathbf{v}||^2 \text{ subject to } ||\mathbf{u}||_2 = 1$$

Citation: "SPCA" method of Zou, Hastie, and Tibshirani (2006)

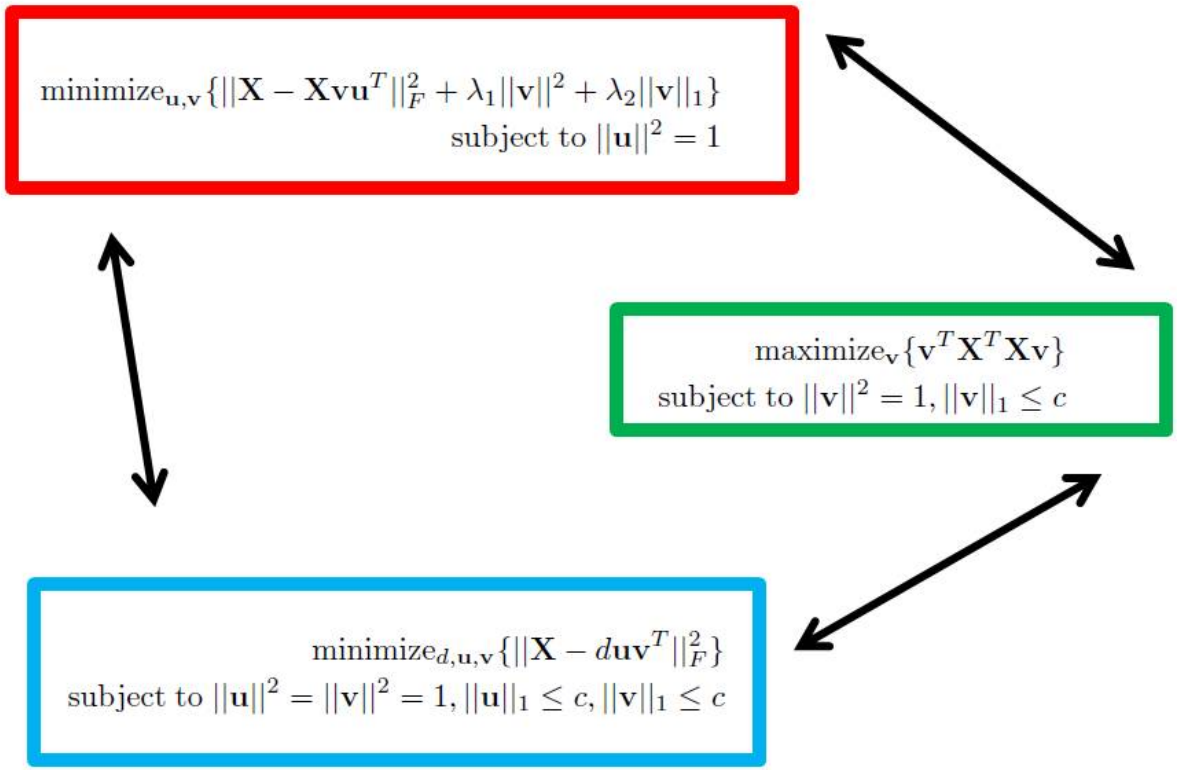
Iterative algorithm to solve for $\mathbf{u}$ and $\mathbf{v}$.

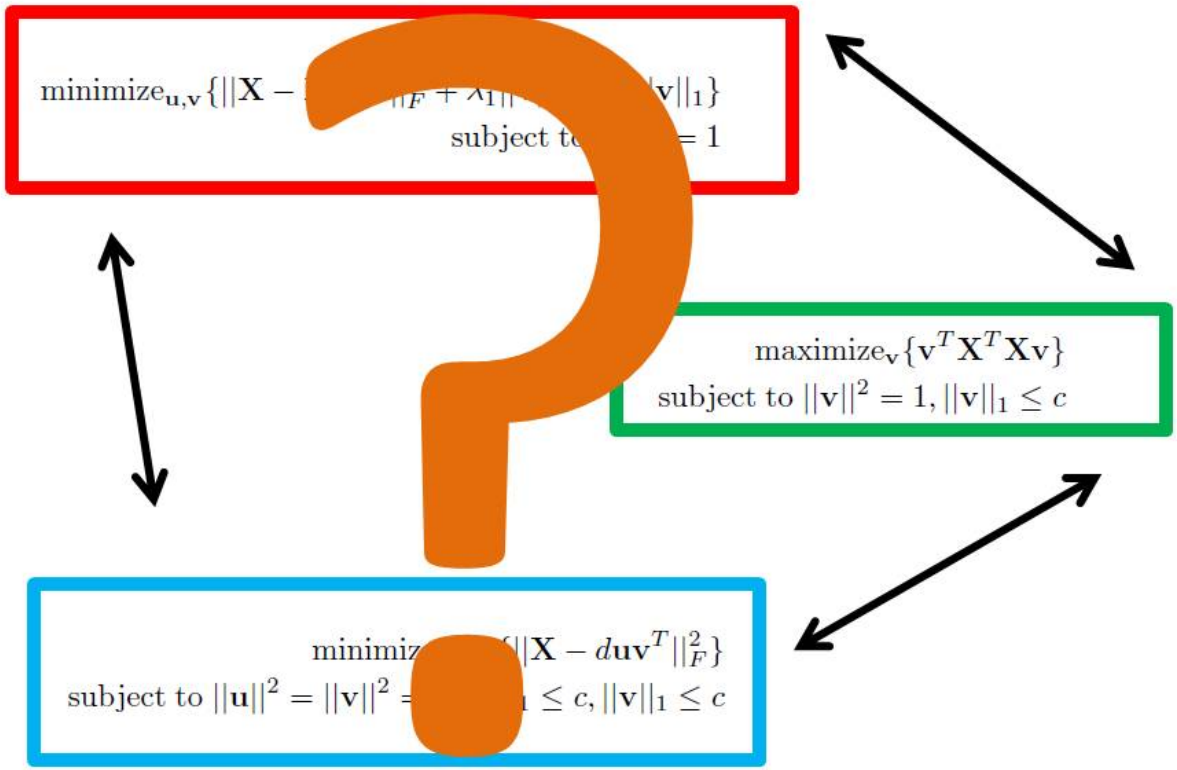
Rank-1 Matrix Approximation

$$(\mathbf{u}, \mathbf{v}, d) = \text{argmin}||\mathbf{X} - d\mathbf{u}\mathbf{v}^T||_F^2 \text{ subject to } ||\mathbf{u}||_2 = ||\mathbf{v}||_2 = 1, ||\mathbf{u}||_1 \leq c_1, ||\mathbf{v}||_1 \leq c_2$$

Citations: "Low rank matrix decomposition" of Shen and Huang (2008); "Penalized matrix decomposition" of Witten, Hastie, and Tibshirani (2008)

Fast iterative algorithm to solve for **u** and **v** using soft thresholding


$\text{minimize}_{\mathbf{u},\mathbf{v}}\{||\mathbf{X} - \mathbf{X}\mathbf{v}\mathbf{u}^T||_F^2 + \lambda_1||\mathbf{v}||^2 + \lambda_2||\mathbf{v}||_1\}$
subject to $||\mathbf{u}||^2 = 1$
$\text{maximize}_{\mathbf{v}}\{\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}\}$
subject to $||\mathbf{v}||^2 = 1, ||\mathbf{v}||_1 \leq c$
$\text{minimize}_{d,\mathbf{u},\mathbf{v}}\{||\mathbf{X} - d\mathbf{u}\mathbf{v}^T||_F^2\}$
subject to $||\mathbf{u}||^2 = ||\mathbf{v}||^2 = 1, ||\mathbf{u}||_1 \leq c, ||\mathbf{v}||_1 \leq c$


maximize$_{\mathbf{v}}\{\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}\}$
subject to $||\mathbf{v}||^2 = 1, ||\mathbf{v}||_1 \leq c$

Rank-1 Approximation leads to Maximal Variance Approach

It is not hard to show that we can re-write the criterion for the rank-1 approximation in a way that looks more like a variance criterion:

$$\begin{aligned}(\mathbf{u},\mathbf{v}) &= \text{argmin}||\mathbf{X} - d\mathbf{u}\mathbf{v}^T||_F^2 \text{ subject to } ||\mathbf{u}||_2 = ||\mathbf{v}||_2 = 1, ||\mathbf{u}||_1 \leq c_1, ||\mathbf{v}||_1 \leq c_2 \\ &= \text{argmax } \mathbf{u}^T\mathbf{X}\mathbf{v} \text{ subject to } ||\mathbf{u}||_2 = ||\mathbf{v}||_2 = 1, ||\mathbf{u}||_1 \leq c_1, ||\mathbf{v}||_1 \leq c_2\end{aligned}$$

Rank-1 Approximation leads to Maximal Variance Approach

Suppose we apply the Rank-1 approximation to $\mathbf{X}$.

$$(\mathbf{u}, \mathbf{v}, d) = \text{argmin}||\mathbf{X} - d\mathbf{u}\mathbf{v}^T||_F^2 \text{ subject to } ||\mathbf{u}||_2 = ||\mathbf{v}||_2 = 1, ||\mathbf{v}||_1 \leq c$$

Then, the solution $\mathbf{v}$ solves maximal variance criterion.
So, rather than solving maximal variance criterion by maximizing a convex function, we can use the quick iterative algorithm for the sparse rank-1 approximation.

Minimal Reconstruction Error as a Variance criterion

In a similar way, one can also show equivalence between minimal reconstruction error and maximal variance criterion, if we add an L_1 constraint on $\mathbf{u}$ to the former.

Conclusions

1. There is no unique definition of sparse PCA: 3+ methods have been proposed.
2. There exist previously unknown connections between these (seemingly different) methods; in fact, they are almost identical!!
3. These connections have not only improved our understanding of each of the different methods, but have resulted in a new fast algorithm for a previously very difficult problem (Maximal Variance Criterion).

References

1. Jolliffe, Trendafilov, and Uddin (2003) 'A modified principal component technique based on the lasso', *Journal of Computational and Graphical Statistics* **12** 531-547.
2. Trendafilov and Jolliffe (2006) 'Projected gradient approach to the numerical solution of the SCoTLASS', *Computational Statistics and Data Analysis* **50** 242-253.
3. Zou, Hastie, and Tibshirani (2006) 'Sparse principal component analysis' *Journal of Computational and Graphical Statistics* **15** 262-286.
4. Shen and Huang (2008) 'Sparse principal component analysis via regularized low rank matrix approximation' *Journal of Multivariate Analysis*.
5. Witten, Hastie, and Tibshirani (2008) 'A penalized matrix decomposition, with applications to canonical correlation analysis and principal components', Submitted.