# Sparse Principal Component Analysis Incorporating Stability Selection

Martin Sill*

German Cancer Research Center, Heidelberg, Germany
m.sill@dkfz.de

Principal component analysis (PCA) is a popular dimension reduction method that approximates a numerical data matrix by seeking principal components (PC), i.e. linear combinations of variables that captures maximal variance. Since each PC is a linear combination of all variables of a data set, interpretation of the PCs can be difficult, especially in high-dimensional data. In order to find 'sparse' PCs that are linear combinations of only a subset of possibly relevant variables and therefore easier to interpret, several sparse PCA approaches have been proposed in the recent years. Typically, these methods use the singular value decomposition (SVD) to calculate PCs. Sparsity is attained by relating the SVD to linear regression and perform a variable selection using penalty terms similar to those in penalized linear regression models. Our approach combines such a regularized SVD with stability selection. Stability selection is a general approach that combines variable selection methods, e.g. penalized regression models, with resampling techniques to control the error of falsely selecting irrelevant variables. Thus our new approach is able to find sparse PCs that are linear combinations of subsets of variables selected with respect to Type I error control. The performance of the proposed method will be compared with other sparse PCA approaches by a simulation study. Application of the method will be demonstrated using high-dimensional molecular data.

Keywords: SVD, high-dimensional, penalization, resampling, variable selection

## 1 Introduction

Suppose $\mathbf{X}$ is an $p \times n$ data matrix with entries $x_{ij}$ and indices $i = 1, \cdots, p$ and $j = 1, \cdots, n$ and rank $r$. PCA seeks for a number of $K \leq r$ linear combinations of the $p$ variables that capture maximal variance:

$$\tilde{\mathbf{u}}_{\mathbf{k}} = \mathbf{X}^T \mathbf{v}_k = \sum_{i=1}^{p} v_{k,i} \mathbf{x}_i, \tag{1}$$

where $\tilde{\mathbf{u}}_k$ is the $k$th principal component (PC) with index $k = 1, \cdots, K$ and $\mathbf{v}_k$ is the so called loadings vector. This vector has unit length and maximizes the variance of the $k$th PC. The coefficients of the loadings vector are interpreted as the contribution of each variable to the $k$th PC. Typically, the PCs are uncorrelated, e.g. the first PC points in the direction of maximal variance and the second PC shows in the direction of maximal variance orthogonal to the first PC and so on. A PCA can be performed by either an eigenvalue decomposition of the covariance matrix or by singular value decomposition (SVD).
The SVD of $\mathbf{X}$ is:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{2}$$

where $\mathbf{U}$ is a $n \times r$ orthogonal matrix and the column vectors $\mathbf{u}_k$ are the PCs scaled to unit length. $\mathbf{V}$ is $p \times r$ orthogonal matrix which columns $\mathbf{v}_k$ are the loadings vectors. $\mathbf{D}$ is a diagonal matrix and the diagonal entries $d_1, \cdots, d_r$ are the singular values, where $d_k\mathbf{u}_k = \tilde{\mathbf{u}}_k$ is the $k$th PC with variance $d_k^2$. Typically, we are interested in a low rank approximation of $\mathbf{X}$, e.g. the first few PCs that explain most of the variance. It is known that the SVD gives the closest rank one approximation of $\mathbf{X}$ with respect to the Frobenius norm (Eckart and Young, 1936):

$$(d, \mathbf{u}, \mathbf{v}) = \arg\min_{d,\mathbf{u},\mathbf{v}} \left\|\mathbf{X} - d\mathbf{u}\mathbf{v}^T\right\|_F^2, \tag{3}$$

where $\|\cdot\|_F^2$ indicates the squared Frobenius norm, which is the sum of squared elements of the matrix. Shen and Huang (2006) and later Lee et al. (2010) showed that with $\mathbf{u}$ fixed the minimization in Equation (3) can be formulated as a least square regression. For a fixed $\mathbf{u}$, the least squares coefficient vector of regressing the columns of $\mathbf{X}$ on $\mathbf{u}$ is $\tilde{\mathbf{v}} = d\mathbf{v}$. The ordinary least squares estimator (OLS) for $\tilde{\mathbf{v}}$ is $\hat{\tilde{\mathbf{v}}} = \mathbf{X}\mathbf{u}$. Without loss of generality, holding $\mathbf{v}$ fixed the OLS for $\tilde{\mathbf{u}}$ is $\hat{\tilde{\mathbf{u}}} = \mathbf{X}^T\mathbf{v}$ With this connection to least squares regression it is straightforward to use penalization terms to impose sparsity on $\tilde{\mathbf{v}}$.

$$(\mathbf{u}, \hat{\tilde{\mathbf{v}}}) = \arg\min_{\mathbf{u},\tilde{\mathbf{v}}} \left\|\mathbf{X} - \mathbf{u}\tilde{\mathbf{v}}^T\right\|_F^2 + \lambda P(\tilde{\mathbf{v}}), \tag{4}$$

where $P(\tilde{\mathbf{v}})$ is a penalization term that induces sparsity on $\tilde{\mathbf{v}}$ and $\lambda$ is a tuning parameter that determines the strength of the penalization. To demonstrate our approach we use the lasso penalty $P(\tilde{\mathbf{v}}) = |\tilde{\mathbf{v}}|$ (Tibshirani,1994), however other sparsity inducing penalization terms are conceivable. With the lasso penalization term in Equation (4) a soft-thresholding estimator (Tibshirani, 1994) can be derived that estimates the elements of $\hat{\tilde{\mathbf{v}}}$:

$$\hat{\tilde{v}}_i = \text{sign}\left\{(\mathbf{X}\mathbf{u})_i\right\}\left(|(\mathbf{X}\mathbf{u})_i| - \lambda\right)_+. \tag{5}$$

Lee et al. (2010) proposed an algorithm that finds solutions to the minimization in Equation (4) by alternating between the following two steps until convergence:

1. $\hat{\tilde{v}}_i = \text{sign}\left\{(\mathbf{X}\mathbf{u})_i\right\}\left(|(\mathbf{X}\mathbf{u})_i| - \lambda\right)_+ \qquad \mathbf{v} = \hat{\tilde{\mathbf{v}}}/\|\hat{\tilde{\mathbf{v}}}\|$
2. $\hat{\tilde{\mathbf{u}}} = \mathbf{X}^T\mathbf{v} \qquad \mathbf{u} = \hat{\tilde{\mathbf{u}}}/\|\hat{\tilde{\mathbf{u}}}\|$

To choose an optimal penalization parameter $\lambda$, Lee et al. (2010) proposed to use the Bayesian information criterion (BIC). The BIC is a model selection criterion that judges the quality of model by the goodness of fit of the model but also penalizes for it's complexity, e.g. the number of parameters in the model. Subsequent PCs are fitted by subtracting the rank one approximation from the data matrix and applying the algorithm to the residual matrix. A poor approximation might induce noise structure when subtracted from the data matrix and this will influence the fitting process of the following PCs. Choosing sparse PC solutions by the BIC will somehow guarantee that the corresponding penalized linear regression model has good prediction accuracy and hence gives a good rank one approximation.

In contrast to the approach described so far, we propose to estimate the coefficients that are truly active in the loadings vectors by applying stability selection (Meinshausen and Bühlmann, 2010). The stability selection is a general approach that combines variable selection methods such as the penalized regression models with resampling. By applying the corresponding variable selection method to subsamples that were drawn without replacement, selection probabilities for each variable can be estimated as the proportion of subsamples where the variable is included in the fitted model. These selection probabilities are used to define a set of stable variables. Meinshausen and Bühlmann (2010) provide a theoretical framework for controlling Type I error rates of falsely assigning variables to the estimated set of stable variables. The selection probability of each feature along the regularization path, e.g. along the range of possible penalization parameters $\Lambda = \{\lambda_1, \lambda_2..., \lambda_L\}$, is called stability path. Given an arbitrary threshold $\pi_{thr} \in (0.5, 1)$ and the set of penalization parameters $\Lambda$, the set of stable features estimated with stability selection is:

$$\hat{S}^{stable}_{\tilde{v}_i} = \left\{ i : \max_{\lambda_l \in \Lambda} \hat{\Pi}^{\lambda_l}_i \geq \pi_{thr} \right\}, \tag{6}$$

where $\hat{\Pi}^{\lambda_l}_i$ denotes the estimated selection probability of the $i$th coefficient at $\lambda_l$. Then according to Theorem 1 in Meinshausen and Bühlmann (2010), the expected number of falsely selected features $E(V)$ will be bounded by:

$$E(V) \leq \frac{1}{(2\pi_{thr} - 1)} \frac{q^2_\Lambda}{p}, \tag{7}$$

where $q_\Lambda$ is the average of the number of non-zero coefficients w.r.t. to the drawn subsamples. Interpreting Equation (7) the expected number of falsely selected variables decreases by either reducing the average number of selected variables $q_\Lambda$ or by increasing the threshold $\pi_{thr}$. Suppose that $\pi_{thr}$ is fixed, then $E(V)$ can be controlled by limiting $q_\Lambda$ by the length of the regularization path $\Lambda$. In multiple testing the expected number of falsely selected variables is also known as the per-family error rate (PFER) and if divided by the total number of variables $p$ will become the per-comparison error rate (PCER). The stability selection allows to control these Type I error rates. For instance, suppose the threshold $\pi_{thr} = 0.8$ is fixed, then choosing $\Lambda$ such that $q_\Lambda \leq \sqrt{0.6p}$ will control $E(V) \leq 1$. Moreover, by choosing $\Lambda$ so that $q_\Lambda \leq \sqrt{0.6p\alpha}$ will control the family wise error rate (FWER) at level $\alpha$, $P(|V| > 0) \leq \alpha$.

Here we propose to estimate the stable set of coefficients in the loadings vector $\hat{S}^{stable}_{\tilde{v}_i}$ by applying stability selection to the penalized regression used to estimate $\tilde{\mathbf{v}}$. Moreover, we replace the soft-thresholding step in the algorithm of Lee et al. (2010) by a hard-thresholding rule that sets all coefficients to zero which are not in the estimated stable set. Then the two steps of the algorithm are as follows:

1. $\hat{\tilde{v}}_i = \mathbf{1}(i \in \hat{S}^{stable}_{\tilde{v}_i})\tilde{v}_i \qquad \mathbf{v} = \hat{\tilde{\mathbf{v}}}/||\hat{\tilde{\mathbf{v}}}||$
2. $\hat{\tilde{\mathbf{u}}} = \mathbf{X}^T\mathbf{v} \qquad \mathbf{u} = \hat{\tilde{\mathbf{u}}}/||\hat{\tilde{\mathbf{u}}}||$

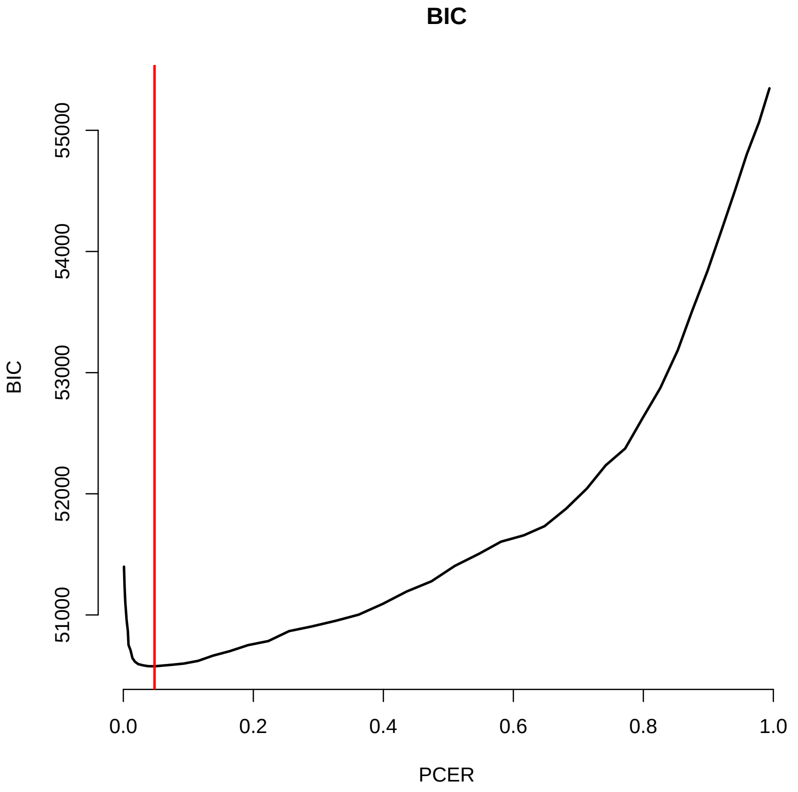

Figure 1: An simulation example that shows how the BIC relates to the PCER of falsely selecting variables into stable sets. For a PCER of $0.001$ to $1$ the non-zero coefficients of the loadings vector of sparse PCs have been estimated by stability selection. The sparse PC solution that minimizes the BIC is based on a stable set of coefficients for which the expected number of falsely selected coefficients is controlled at a PCER of $0.052$.

where $\mathbf{1}()$ denotes the indicator function. For a range of desired Type I error level the stability selection can be used to estimate stable sets. Applying the described algorithm, for each estimated stable set a sparse PC can be fitted and the corresponding BIC can be calculated. Similar to the approach of Lee et al. (2010), from the different sparse and 'stable' PC solutions along the range of the Type I error levels, the solution that minimizes the BIC is chosen. Figure 1, shows an example how the BIC relates to the per-comparison error rate (PCER) of different sparse PC solutions estimated by stability selection. Furthermore, the Type I error that can be controlled at the minimal BIC can be used as an indicator for the noise in the data matrix. With increasing noise level the stability selection will hardly identify stable variables and hence the Type I error at minimal BIC will increase. This property can be used to decide how many PCs might be relevant.

## 2 Outlook

A major drawback of conventional PCA is that each PC is a linear combination of all variables in the data and thus the interpretation of PCs might be difficult. Moreover, in high-dimensional data sets many variables may be noise variables that can cover the relevant structures in the data. In order to find PCs with many coefficients in the loadings vector being equal to zero, several sparse PCA methods have been proposed, i.e. Witten et al. (2009), Shen and Huang (2008) and Zou et al. (2004). Here we have presented a new sparse PCA approach that finds sparse PCs by applying stability selection. The stability selection is a variable selection approach that combines subsampling with high-dimensional selection algorithms, e.g. $l1$-penalized regression models. In addition, the stability selection provides Type I error control of falsely selecting variables. We face the problem of finding sparse models with good prediction accuracy by applying the BIC to a number of 'stable' models that have been precedingly selected by stability selection. So far we have implemented a parallelized version of the algorithm using the statistical programming language R (R Core Team, 2013). An R-package with additional visualization functions, such as biplots and screeplots is currently under development. We will compare the proposed method with other sparse PCA approaches in a simulation study. Several simulation scenarios and different criteria to assess the performance of each method will be investigated. For example, the number of true- and false-positive non-zero coefficients in the loadings vectors and the orthogonality of the resulting PCs will be considered. Furthermore, we will demonstrate the application of our method on a high-dimensional molecular data set.

## References

Lee, M., H. Shen, J. Z. Huang, and J. S. Marron (2010). Biclustering via sparse singular value decomposition. *Biometrics 66*(4), 1087–1095.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society - Series B: Statistical Methodology 72*(4), 417–473.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Shen, H. and J. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis 99*(6), 1015–1034.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics Oxford England 10*(3), 515–534.

Zou, H., T. Hastie, and R. Tibshirani (2004). Sparse Principal Component Analysis. pp. 1–30.