Package ‘glmpca’

September 27, 2019

Title  Dimension Reduction of Non-Normally Distributed Data
Version  0.1.0
Description  Implements a generalized version of principal components analysis
            (GLM-PCA) for dimension reduction of non-normally distributed data such as
            counts or binary matrices.
License  Artistic-2.0
Depends  R (&gt;= 3.6), stats
Imports
Suggests  knitr, MASS, testthat, covr, ggplot2
URL  https://github.com/willtownes/glmpca
BugReports  https://github.com/willtownes/glmpca/issues
VignetteBuilder  knitr
LazyData  false
RoxygenNote  6.1.1
Encoding  UTF-8
NeedsCompilation  no
Author  F. William Townes [aut, cre, cph],
       Kelly Street [aut],
       Jake Yeung [ctb]
Maintainer  F. William Townes <will.townes@gmail.com>
Repository  CRAN
Date/Publication  2019-09-27 10:50:05 UTC

R topics documented:

  glmpca ................................................................. 2

Index  4
Description

This function implements the GLM-PCA dimensionality reduction method for high-dimensional count data.

Usage

```r
glmpca(Y, L, fam = c("poi", "nb", "mult", "bern"), ctl = list(maxIter = 1000, eps = 1e-04), penalty = 1, verbose = FALSE, init = list(factors = NULL, loadings = NULL), nb_theta = 100, X = NULL, Z = NULL, sz = NULL)
```

Arguments

- **Y**: matrix of count data with features as rows and observations as columns.
- **L**: the desired number of latent dimensions (integer).
- **fam**: character describing the likelihood to use for the data (poisson, negative binomial, binomial approximation to multinomial, bernoulli).
- **ctl**: a list of control parameters for optimization.
- **penalty**: the L2 penalty for the latent factors (default = 1). Regression coefficients are not penalized.
- **verbose**: logical value indicating whether the current deviance should be printed after each iteration (default = FALSE).
- **init**: a list containing initial estimates for the factors (V) and loadings (V) matrices.
- **nb_theta**: see `negative.binomial` (nb_theta->∞ equivalent to Poisson).
- **X**: a matrix of column (observations) covariates. Any column with all same values (eg. 1 for intercept) will be removed. This is because we force the intercept and want to avoid collinearity.
- **Z**: a matrix of row (feature) covariates, usually not needed.
- **sz**: numeric vector of size factors to use in place of total counts.

Details

The basic model is \( R = AX' + ZG' + VU' \), where \( E[Y] = M = \text{linkinv}(R) \). Regression coefficients are \( A \) and \( G \), latent factors are \( U \) and loadings are \( V \). The objective function being optimized is the deviance between \( Y \) and \( M \), plus an L2 (ridge) penalty on \( U \) and \( V \).
Value

A list containing:

- **factors** a matrix $U$ whose rows match the columns (observations) of $Y$. It is analogous to the principal components in PCA. Each column of the factors matrix is a different latent dimension.

- **loadings** a matrix $V$ whose rows match the rows (features/dimensions) of $Y$. It is analogous to loadings in PCA. Each column of the loadings matrix is a different latent dimension.

- **coefX** a matrix $A$ of coefficients for the observation-specific covariates matrix $X$. Each row of coefX corresponds to a row of $Y$ and each column corresponds to a column of $X$. The first column of coefX contains feature-specific intercepts which are included by default.

- **coefZ** a matrix $G$ of coefficients for the feature-specific covariates matrix $Z$. Each row of coefZ corresponds to a column of $Y$ and each column corresponds to a column of $Z$. By default no such covariates are included and this is returned as NULL.

- **dev** a vector of deviance values. The length of the vector is the number of iterations it took for GLM-PCA’s optimizer to converge. The deviance should generally decrease over time. If it fluctuates wildly, this often indicates numerical instability, which can be improved by increasing the penalty parameter.

- **family** an S3 object of class glmpca_family. This is a minor extension to the family or negative.binomial object used by functions like glm and glm.nb. It is basically a list with various internal functions and parameters needed to optimize the GLM-PCA objective function. For the negative binomial case, it also contains the final estimated value of the dispersion parameter ($\text{nb_theta}$).

References


See Also

- [prcomp](https://cran.r-project.org/web/packages/prcomp/prcomp.pdf)

Examples

```r
# create a simple dataset with two clusters
mu<-rep(c(.5,3),each=10)
mu<-matrix(exp(rnorm(100*20)),nrow=100)
mu[,1:10]<-mu[,1:10]*exp(rnorm(100))
clust<-rep(c("red","black"),each=10)
Y<-matrix(rpois(prod(dim(mu)),mu),nrow=nrow(mu))
# visualize the latent structure
res<-glmpca(Y, 2)
Factors<-res$Factors
plot(Factors[,1],Factors[,2],col=clust,pch=19)
```
Index

family, 3

glm, 3
glm.nb, 3
glmpca, 2

negative.binomial, 2, 3

prcomp, 3