ALStructure Manual

February 20, 2018

`alstructure` *Main function for execution of the* `ALStructure` *algorithm*

Description

Computes global ancestry estimates under the admixture model given a SNP data matrix $\boldsymbol{X}$. This function is based on the `ALStrcture` algorithm from (Cabreros and Storey 2017).

Usage

```
alstructure(X, d_hat = NULL, svd_method = "base", tol = 1e-05,
  max_iters = 1000, order_method = "ave_admixture", P_init, Q_init)
```

Arguments

`X` The $m \times n$ SNP data matrix.

`d_hat` Estimate of the latent space dimension d. If left blank, this is estimated by the function `estimate_d()`

`svd_method` One of "base" or "truncated_svd." If "base" is chosen, the base `svd()` function is used. If "truncated_svd" is chosen, the truncated svd algorithm `propack.svd()` from the `svd` package is used.

`tol` The convergence criterion. If $RMSE(\hat{\boldsymbol{Q}}_t - \hat{\boldsymbol{Q}}_{t+1}) < tol$, then the algorithm halts

`max_iters` The maximum number of iterations (repetitions of steps (6) and (7) in Algorithm 1) to be executed

`order_method` One of "ave_admixture" or "var_explained." If "ave_admixture," the d populations are ordered by decreasing average admixture accross samples (i.e. $1/n \sum_j q_{ij}$). If "var_explained", the d populations are ordered be decreasing variation explained. Specifically, we compute a modified version of the eigen-R^2 statistic from (L. S. Chen and Storey 2008). The statistic is modified in the following ways: 1) we treat rows of $\boldsymbol{Q}$ as the response variables 2) we regress each row of Q on the eigenvectors of $\boldsymbol{G}$ rather than the eigenvectors of the data matrix itself 3) we take the weighted average only over the top d eigenvectors. and columns of $\boldsymbol{P}$ are ordered by amount of variation explained by each row of $\boldsymbol{Q}$ by the function `order_Q`

`P_init` Optional initialization of $\boldsymbol{P}$. Only available for `cALS` method.

`Q_init` Optional initializtion of $\boldsymbol{Q}$. Only available for `cALS` method.

Value

A list with the following elements:

P_hat : The estimated $\boldsymbol{P}$ matrix. Each column of $\boldsymbol{P}$ may be interpreted as a vector of allele frequencies for a specific ancestral population.

Q_hat : The estimated $\boldsymbol{Q}$ matrix. Each column of $\boldsymbol{Q}$ may be interpreted as the admixture proportions of a specific individual.

rowspace : a list with the following elements:

- **vectors** : The top d eigenvectors of the matrix $\boldsymbol{G}$ sorted by decreasing eigenvalue. These vectors approximate the subspace spanned by the rows of $\boldsymbol{Q}$.
- **values** : The top d eigenvalues of the matrix $\boldsymbol{G}$ sorted by decreasing eigenvalue.

References

Cabreros, I., and J. D. Storey. 2017. “A Nonparametric Estimator of Population Structure Unifying Admixture Models and Principal Components Analysis.” BioRxiv. Cold Spring Harbor Laboratory. doi:10.1101/240812.

Hao, W., M. Song and J. D. Storey. 2015. “lfa: Logistic Factor Analysis for Categorical Data.” R package version 1.8.0, https://github.com/StoreyLab/lfa.

`estimate_F` *Estimates the individual-specific allele frequency matrix* $\boldsymbol{F}$

Description

Estimates the $m \times n$ individual-specific allele frequency matrix $\boldsymbol{F}$ using the method of Latent Subspace Estimation (X. Chen and Storey 2015) as described in (Cabreros and Storey 2017) in Section 2.3.

Usage

```
estimate_F(X, d, svd_method = "base")
```

Arguments

`X`	The $m \times n$ SNP data matrix
`d`	The rank of $\boldsymbol{F}$. This can be estimated using the function `d_estimate()`.
`svd_method`	One of "base" or "truncated_svd." If "base" is chosen, the base `svd()` function is used. If "truncated_svd" is chosen, the truncated svd algorithm `propack.svd()` from the `svd` package is used.

Value

A list with the following elements:

F_hat : The $m \times n$ matrix $\hat{\boldsymbol{F}}$.

rowspace : a list with the following elements:

- **vectors** : The top d eigenvectors of the matrix $\boldsymbol{G}$ sorted by decreasing eigenvalue. These vectors approximate the subspace spanned by the rows of $\boldsymbol{Q}$.
- **values** : The top d eigenvalues of the matrix $\boldsymbol{G}$ sorted by decreasing eigenvalue.

References

Cabreros, I., and J. D. Storey. 2017. “A Nonparametric Estimator of Population Structure Unifying Admixture Models and Principal Components Analysis.” BioRxiv. Cold Spring Harbor Laboratory. doi:10.1101/240812.

Chen, X., and J. D. Storey. 2015. “Consistent Estimation of Low-Dimensional Latent Structure in High-Dimensional Data.” ArXiv E-Prints, October.

`estimate_d` *Estimate the latent space dimension*

Description

Estimates the dimension of the rowspace of $\boldsymbol{Q}$ (equivalently, the rank of the matrix $\boldsymbol{F}$). This estimate $\hat{d}$ is based on the estimator from (Leek 2011), page 6.

Usage

```
estimate_d(X)
```

Arguments

`X`	the $m \times n$ SNP data matrix

Value

an estimate $\hat{d}$ of the dimension of the latent space dimension.

References

Leek, J. T. 2011. “Asymptotic conditional singular value decomposition for high-dimensional genomic data.” Biometrics 67 (4): 344–52.

`factor_F` *A fast algorithm for factoring $\hat{\boldsymbol{F}}$.*

Description

An algorithm for finding approximate factors $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{Q}}$ of the matrix $\hat{\boldsymbol{F}}$ that obey constraints of the admixture model:

1. $p_{ij} \in [0,1] \, \forall (i,j)$
2. $q_{ij} \geq 0 \, \forall (i,j)$ and $\sum_i q_{ij} = 1 \, \forall j$

This algorithm is described in Algorithm 2 in (Cabreros and Storey 2017). While it lacks the theoretical guarantees of the `cALS` algorithm, it is much faster.

Usage

```
factor_F(F_hat, d, tol = 1e-05, max_iters = 1000)
```

Arguments

`F_hat`	The estimate of the matrix $\boldsymbol{F}$ to be factored
`d`	The dimension of the latent space. This can be estimated by the function `d_estimate`.
`tol`	The convergence criterion. If $RMSE(\hat{\boldsymbol{Q}}_t - \hat{\boldsymbol{Q}}_{t+1}) < tol$, then the algorithm halts
`max_iters`	The maximum number of iterations (repetitions of steps (6) and (7) in Algorithm 1) to be executed

Value

A list with the following elements:

P_hat : The estimated $\boldsymbol{P}$ matrix. Each column of $\boldsymbol{P}$ may be interpreted as a vector of allele frequencies for a specific ancestral population.

Q_hat : The estimated $\boldsymbol{Q}$ matrix. Each column of $\boldsymbol{Q}$ may be interpreted as the admixture proportions of a specific individual.

@references Cabreros, I., and J. D. Storey. 2017. "A Nonparametric Estimator of Population Structure Unifying Admixture Models and Principal Components Analysis." BioRxiv. Cold Spring Harbor Laboratory. doi:10.1101/240812.

`lse` *Estimates the latent subspace*

Description

Estimates the rowspace of Q using the method of latent subspace estimation. The function returns the top d eigenvalues and vectors of the matrix

$$\boldsymbol{G} = \frac{1}{m}\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{D}$$

where the matrix D is a diagonal matrix with each diagonal entry d_{ii} an estimate of the average of the variances of the random variables in the i column of $\boldsymbol{X}$. As is proven in (X. Chen and Storey 2015), the span of the top d eigenvectors of $\boldsymbol{G}$ span the same space as the rows of $\boldsymbol{Q}$. The eigenvectors are returned in order of decreasing eigenvalue.

Usage

```
lse(X, d, svd_method = "base")
```

Arguments

`X`	The $m \times n$ SNP data matrix
`d`	The rank of $\boldsymbol{F}$. This can be estimated using the function `d_estimate()`. When $d = n$, all eigenvectors of $\boldsymbol{G}$ are returned.
`svd_method`	One of "base" or "truncated_svd." If "base" is chosen, the base `svd()` function is used. If "truncated_svd" is chosen, the truncated svd algorithm `propack.svd()` from the `svd` package is used.

Value

A list with the following elements:

vectors : The top d eigenvectors of the matrix $\boldsymbol{G}$ sorted by decreasing eigenvalue. These vectors approximate the subspace spanned by the rows of $\boldsymbol{Q}$.

values : The top d eigenvalues of the matrix $\boldsymbol{G}$ sorted by decreasing eigenvalue.

References

Chen, X., and J. D. Storey. 2015. “Consistent Estimation of Low-Dimensional Latent Structure in High-Dimensional Data.” ArXiv E-Prints, October.

`order_pops` *Orders the d populations*

Description

Orders the d populations according to one of two methods: "ave_admixture" or "var_explained." Function returns matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ with permuted columns and rows according to the determined ordering of populations.

Usage

```
order_pops(P, Q, method = "ave_admixture", Q_space = NULL)
```

Arguments

`P` The $m \times d$ loadings matrix with columns to be ordered

`Q` The $d \times n$ admixture matrix with rows to be ordered

`method` One of "ave_admixture" or "var_explained." If "ave_admixture," the d populations are ordered by decreasing average admixture accross samples (i.e. $1/n \sum_j q_{ij}$). If "var_explained", the d populations are ordered be decreasing variation explained. Specifically, we compute a modified version of the eigen-R^2 statistic from (L. S. Chen and Storey 2008). The statistic is modified in the following ways: 1) we treat rows of $\boldsymbol{Q}$ as the response variables 2) we regress each row of Q on the eigenvectors of $\boldsymbol{G}$ rather than the eigenvectors of the data matrix itself 3) we take the weighted average only over the top d eigenvectors.

`Q_space` Only required for "var_explained" methodThe list containing the top d eigenvectors and their corresponding eigenvalues of the $\boldsymbol{G}$. These may be obtained through the function `lse`.

Value

A list with the following elements:

P_ordered : The permuted $\boldsymbol{P}$

Q_ordered : The permuted $\boldsymbol{Q}$

perm_mat : The permutation matrix $\boldsymbol{A}$ such that $\boldsymbol{P}\boldsymbol{A}^T = \boldsymbol{P}_{\text{ord}}$ and $\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}_{\text{ord}}$.

References

Chen, L. S., and J. D. Storey. 2008. “Eigen-R2 for dissecting variation in high-dimensional studies.” Bioinformatics 24 (19): 2260–2.

simulate_admixture *Simulates data from the PSD model*

Description

Creates a data frame that contains the parameters of the admixture model $(\boldsymbol{F}, \boldsymbol{P}, \boldsymbol{Q})$ as well as a single draw $\boldsymbol{X}$ such that $x_{ij} \sim \text{Bernoulli}(f_{ij})$. The $\boldsymbol{Q}$ matrix is drawn from the Dirichlet distribution with parameter α, and the $\boldsymbol{P}$ matrix is simulated from the Balding-Nichols model. The parameter α is supplied by the user. The $m \times 2$ matrix of Balding-Nichols parameters is optional. If `BN_params` is not supplied, the parameters are derived from a random sample of estimated Balding-Nichols parameters from the Human Genomes Diversity Project (HGDP) dataset. The Balding-Nichols parameter estimates are provided by (Gopalan et al. 2016), and included in this package in the object `hgdpBN`.

Usage

```
simulate_admixture(m, n, d, alpha, BN_params = NA, seed = NA)
```

Arguments

`m`	number of SNPs
`n`	number of individuals
`d`	number of groups
`alpha`	dirichlet parameter; `length(alpha) = d`
`BN_params`	a $m \times 2$ matrix of parameters. The first column contains $\mathrm{F_{ST}}$ for each SNP, while the second column contains the allele frequency.

Value

a list with the following elements:

P : the $m \times d$ matrix of loadings

Q : the $d \times n$ matrix of latent admixture components

F : the $m \times n$ matrix PQ

X : a random draw such that $x_{ij} \sim \text{Bernoulli}(f_{ij}, 2)$.

References

Gopalan, P., W. Hao, D. M. Blei, and J. D. Storey. 2016. “Scaling probabilistic models of genetic variation to millions of humans.” Nature Genetics 48 (12): 1587–90.

Index