

# Package ‘vsp’

December 5, 2022

**Type** Package

**Title** Vintage Sparse PCA for Semi-Parametric Factor Analysis

**Version** 0.1.1

**Description** Provides fast spectral estimation of latent factors in random dot product graphs using the vsp estimator. Under mild assumptions, the vsp estimator is consistent for (degree-corrected) stochastic blockmodels, (degree-corrected) mixed-membership stochastic blockmodels, and degree-corrected overlapping stochastic blockmodels.

**License** MIT + file LICENSE

**URL** <https://github.com/RoheLab/vsp>

**BugReports** <https://github.com/RoheLab/vsp/issues>

**Depends** R (>= 3.1)

**Imports** ggplot2, glue, invertiforms, LRMF3, magrittr, Matrix, rlang, RSpectra, stats, tibble, withr

**Suggests** covr, dplyr, GGally, igraph, igraphdata, knitr, purrr, rmarkdown, scales, testthat (>= 3.0.0), tidygraph, tidyr

**VignetteBuilder** knitr

**Config/testthat/edition** 3

**Encoding** UTF-8

**RoxygenNote** 7.1.2

**NeedsCompilation** no

**Author** Karl Rohe [aut],  
Muzhe Zeng [aut],  
Alex Hayes [aut, cre, cph] (<<https://orcid.org/0000-0002-4985-5160>>),  
Fan Chen [aut]

**Maintainer** Alex Hayes <[alexphayes@gmail.com](mailto:alexphayes@gmail.com)>

**Repository** CRAN

**Date/Publication** 2022-12-05 16:00:02 UTC

## R topics documented:

bff . . . . .	2
bind_varimax_z . . . . .	3
get_svd_u . . . . .	4
get_z_hubs . . . . .	4
plot_ipr_pairs . . . . .	5
plot_mixing_matrix . . . . .	6
plot_varimax_z_pairs . . . . .	6
screeplot.vsp_fa . . . . .	8
set_z_factor_names . . . . .	9
vsp . . . . .	9
vsp.svd_like . . . . .	12
vsp_fa . . . . .	13
<b>Index</b>	<b>15</b>

---

bff	<i>Find features most associated with cluster membership</i>
-----	--

---

### Description

Find features most associated with cluster membership

### Usage

```
bff(loadings, features, num_best)
```

### Arguments

loadings	An n by k matrix of weights that indicates how important that ith user is to the jth cluster, i.e., the Z or Y matrix calculated by <code>vsp()</code> .
features	An n by d matrix of features measured for each node in the network.
num_best	An integer indicating how many of the top features for differentiating between loadings you want.

### Details

See `vignette("bff")`.

### Value

An n by k matrix whose  $[i, j]$  entry is the ith "most important" feature for cluster j.

---

bind_varimax_z	<i>Add Z factor loadings to node table of tidygraph</i>
----------------	---

---

## Description

Add Z factor loadings to node table of tidygraph

## Usage

```
bind_varimax_z(graph, fa, ...)
```

```
bind_varimax_y(graph, fa, ...)
```

```
bind_svd_u(graph, fa, ...)
```

```
bind_svd_v(graph, fa, ...)
```

## Arguments

graph	A <a href="#">tidygraph::tbl_graph</a> object.
fa	Optionally, a <a href="#">vsp</a> object to extract varimax loadings from. If you do not passed a <a href="#">vsp</a> object, one will be created.
...	Arguments passed on to <a href="#">vsp</a>
x	Either a graph adjacency matrix, <a href="#">igraph::igraph</a> or <a href="#">tidygraph::tbl_graph</a> . If x is a <a href="#">matrix</a> or <a href="#">Matrix::Matrix</a> then x[i, j] should correspond to the edge going from node i to node j.
rank	The number of factors to calculate.

## Value

The same graph object with columns factor1, ..., factor{rank} in the table of node information.

## Functions

- `bind_varimax_y`: Add Y factor loadings to node table of tidygraph
- `bind_svd_u`: Add left singular vectors to node table of tidygraph
- `bind_svd_v`: Add right singular vectors to node table of tidygraph

---

`get_svd_u`*Get left singular vectors in a tibble*

---

**Description**

Get left singular vectors in a tibble

**Usage**

```
get_svd_u(fa, factors = 1:fa$rank)
```

```
get_svd_v(fa, factors = 1:fa$rank)
```

```
get_varimax_z(fa, factors = 1:fa$rank)
```

```
get_varimax_y(fa, factors = 1:fa$rank)
```

**Arguments**

`fa` A `vsp_fa()` object.

`factors` The specific columns to index into. The most reliable option here is to index with an integer vector of column indices, but you could also use a character vector if columns have been named. By default returns all factors/singular vectors.

**Value**

A `tibble::tibble()` with one row for each node, and one column containing each of the requested factor or singular vector, plus an additional `id` column.

**Functions**

- `get_svd_v`: Get right singular vectors in a tibble
- `get_varimax_z`: Get varimax Y factors in a tibble
- `get_varimax_y`: Get varimax Z factors in a tibble

---

`get_z_hubs`*Get most important hubs for each Z factor*

---

**Description**

Get most important hubs for each Z factor

**Usage**

```
get_z_hubs(fa, hubs_per_factor = 10, factors = 1:fa$rank)
```

```
get_y_hubs(fa, hubs_per_factor = 10, factors = 1:fa$rank)
```

**Arguments**

fa                    A `vsp_fa()` object.

hubs\_per\_factor

The number of important nodes to get per latent factor. Defaults to 10.

factors

The specific columns to index into. The most reliable option here is to index with an integer vector of column indices, but you could also use a character vector if columns have been named. By default returns all factors/singular vectors.

**Value**

A `tibble::tibble()` where each row corresponds to a single hub, and three columns:

- id: Node id of hub node
- factor: Which factor that node is a hub for. Nodes can be hubs of multiple factors.
- loading: The actual value of the hubs factor loading for that factor.

**Functions**

- `get_y_hubs`: Get most important hubs for each Y factor

---

plot\_ipr\_pairs

*Plot pairs of inverse participation ratios for singular vectors*

---

**Description**

When IPR for a given singular vector is  $O(1)$  rather than  $O(1 / \sqrt{n})$ , this can indicate that the singular vector is localizing on a small subset of nodes. Oftentimes this localization indicates overfitting. If you see IPR values that are not close to zero (where "close to zero" is something you sort of have to pick up over time), then you need to do some further investigation to see if you have localization and that localization corresponds to overfitting. Note, however, that not all localization is overfitting.

**Usage**

```
plot_ipr_pairs(fa)
```

**Arguments**

fa                    A `vsp_fa()` object.

**Value**

A `tibble::tibble()` with one row for each node, and one column containing each of the requested factor or singular vector, plus an additional `id` column.

---

`plot_mixing_matrix`     *Plot the mixing matrix B*

---

**Description**

Plot the mixing matrix B

**Usage**

```
plot_mixing_matrix(fa)
```

**Arguments**

`fa`                     A `vsp_fa()` object.

**Value**

A `tibble::tibble()` with one row for each node, and one column containing each of the requested factor or singular vector, plus an additional `id` column.

---

`plot_varimax_z_pairs`     *Create a pairs plot of select Y factors*

---

**Description**

To avoid overplotting, plots data for a maximum of 1000 nodes. If there are more than 1000 nodes, samples 1000 nodes randomly proportional to row norms (i.e. nodes with embeddings larger in magnitude are more likely to be sampled).

**Usage**

```
plot_varimax_z_pairs(fa, factors = 1:min(5, fa$rank), ...)
```

```
plot_varimax_y_pairs(fa, factors = 1:min(5, fa$rank), ...)
```

```
plot_svd_u(fa, factors = 1:min(5, fa$rank))
```

```
plot_svd_v(fa, factors = 1:min(5, fa$rank))
```

**Arguments**

fa	A <code>vsp_fa()</code> object.
factors	The specific columns to index into. The most reliable option here is to index with an integer vector of column indices, but you could also use a character vector if columns have been named. By default returns all factors/singular vectors.
...	Arguments passed on to <code>GGally::ggpairs</code>
data	data set using. Can have both numerical and categorical data.
mapping	aesthetic mapping (besides x and y). See <code>aes()</code> . If mapping is numeric, columns will be set to the mapping value and mapping will be set to NULL.
columns	which columns are used to make plots. Defaults to all columns.
title	title, x label, and y label for the graph
xlab	title, x label, and y label for the graph
ylab	title, x label, and y label for the graph
upper	see Details
lower	see Details
diag	see Details
params	deprecated. Please see <code>wrap_fn_with_param_arg</code>
axisLabels	either "show" to display axisLabels, "internal" for labels in the diagonal plots, or "none" for no axis labels
columnLabels	label names to be displayed. Defaults to names of columns being used.
labeller	labeller for facets. See <code>labellers</code> . Common values are "label_value" (default) and "label_parsed".
switch	switch parameter for <code>facet_grid</code> . See <code>ggplot2::facet_grid</code> . By default, the labels are displayed on the top and right of the plot. If "x", the top labels will be displayed to the bottom. If "y", the right-hand side labels will be displayed to the left. Can also be set to "both"
showStrips	boolean to determine if each plot's strips should be displayed. NULL will default to the top and right side plots only. TRUE or FALSE will turn all strips on or off respectively.
legend	May be the two objects described below or the default NULL value. The legend position can be moved by using <code>ggplot2's</code> theme element <code>pn</code> + <code>theme(legend.position = "bottom")</code> <ul style="list-style-type: none"> <li><b>a numeric vector of length 2</b> provides the location of the plot to use the legend for the plot matrix's legend. Such as <code>legend = c(3,5)</code> which will use the legend from the plot in the third row and fifth column</li> <li><b>a single numeric value</b> provides the location of a plot according to the display order. Such as <code>legend = 3</code> in a plot matrix with 2 rows and 5 columns displayed by column will return the plot in position <code>c(1,2)</code></li> <li><b>a object from <code>grab_legend()</code></b> a predetermined plot legend that will be displayed directly</li> </ul>
cardinality_threshold	maximum number of levels allowed in a character / factor column. Set this value to NULL to not check factor columns. Defaults to 15

`progress` NULL (default) for a progress bar in interactive sessions with more than 15 plots, TRUE for a progress bar, FALSE for no progress bar, or a function that accepts at least a plot matrix and returns a new `progress::progress_bar`. See `ggmatrix_progress`.

`proportions` Value to change how much area is given for each plot. Either NULL (default), numeric value matching respective length, `grid::unit` object with matching respective length or "auto" for automatic relative proportions based on the number of levels for categorical variables.

`legends` deprecated

## Value

A `ggplot2::ggplot()` plot or `GGally::ggpairs()` plot.

## Functions

- `plot_varimax_y_pairs`: Create a pairs plot of select Z factors
- `plot_svd_u`: Create a pairs plot of select left singular vectors
- `plot_svd_v`: Create a pairs plot of select right singular vectors

---

<code>screepLOT.vsp_fa</code>	<i>Create a screepLOT from a factor analysis object</i>
-------------------------------	---

---

## Description

Create a screepLOT from a factor analysis object

## Usage

```
## S3 method for class 'vsp_fa'
screepLOT(x, ...)
```

## Arguments

<code>x</code>	A <code>vsp_fa()</code> object.
<code>...</code>	Ignored, included only for consistency with S3 generic.

## Value

A `tibble::tibble()` with one row for each node, and one column containing each of the requested factor or singular vector, plus an additional `id` column.



---

set\_z\_factor\_names      *Give the dimensions of Z factors informative names*

---

### Description

Give the dimensions of Z factors informative names

### Usage

```
set_z_factor_names(fa, names)
```

```
set_y_factor_names(fa, names)
```

### Arguments

fa                      A `vsp_fa()` object.  
names                    Describe new names for Z/Y factors.

### Value

A new `vsp_fa()` object, but the columns names of Z and the row names of B have been set to names (for `set_z_factor_names`), and the column names of B and the column names of Y have been set to names (for `set_y_factor_names`).

### Functions

- `set_y_factor_names`: Give the dimensions of Y factors informative names

---

vsp                      *Semi-Parametric Factor Analysis via Vintage Sparse PCA*

---

### Description

This code implements TODO.

### Usage

```
vsp(x, rank, ...)
```

```
## Default S3 method:  
vsp(x, rank, ...)
```

```
## S3 method for class 'matrix'  
vsp(  
  x,
```

```
rank,
...,
center = FALSE,
recenter = FALSE,
degree_normalize = TRUE,
renormalize = FALSE,
tau_row = NULL,
tau_col = NULL,
kaiser_normalize_u = FALSE,
kaiser_normalize_v = FALSE,
rownames = NULL,
colnames = NULL
)

## S3 method for class 'Matrix'
vsp(
  x,
  rank,
  ...,
  center = FALSE,
  recenter = FALSE,
  degree_normalize = TRUE,
  renormalize = FALSE,
  tau_row = NULL,
  tau_col = NULL,
  kaiser_normalize_u = FALSE,
  kaiser_normalize_v = FALSE,
  rownames = NULL,
  colnames = NULL
)

## S3 method for class 'dgCMatrix'
vsp(
  x,
  rank,
  ...,
  center = FALSE,
  recenter = FALSE,
  degree_normalize = TRUE,
  renormalize = FALSE,
  tau_row = NULL,
  tau_col = NULL,
  kaiser_normalize_u = FALSE,
  kaiser_normalize_v = FALSE,
  rownames = NULL,
  colnames = NULL
)
```

```
## S3 method for class 'igraph'
vsp(x, rank, ..., edge_weights = NULL)
```

### Arguments

x	Either a graph adjacency matrix, <a href="#">igraph::igraph</a> or <a href="#">tidygraph::tbl_graph</a> . If x is a <a href="#">matrix</a> or <a href="#">Matrix::Matrix</a> then x[i, j] should correspond to the edge going from node i to node j.
rank	The number of factors to calculate.
...	These dots are for future extensions and must be empty.
center	Should the adjacency matrix be row <i>and</i> column centered? Defaults to FALSE.
recenter	Should the varimax factors be re-centered around the original factor means? Only used when center = TRUE, defaults to FALSE.
degree_normalize	Should the regularized graph laplacian be used instead of the raw adjacency matrix? Defaults to TRUE. If center = TRUE, A will first be centered and then normalized.
renormalize	Should the regularized graph laplacian be used instead of the raw adjacency matrix? Defaults to TRUE. If center = TRUE, A will first be centered and then normalized.
tau_row	Row regularization term. Default is NULL, in which case we use the row degree. Ignored when degree_normalize = FALSE.
tau_col	Column regularization term. Default is NULL, in which case we use the column degree. Ignored when degree_normalize = FALSE.
kaiser_normalize_u	Whether or not to use Kaiser normalization when rotating the left singular vectors U. Defaults to FALSE.
kaiser_normalize_v	Whether or not to use Kaiser normalization when rotating the right singular vectors V. Defaults to FALSE.
rownames	Character vector of row names of x. These row names are propagated into the row names of the U and Z. Defaults to NULL.
colnames	Character vector of column names of x. These column names are propagated into the row names of the V and Y. Defaults to NULL.
edge_weights	When x is an <a href="#">igraph::igraph</a> , an edge attribute to use to form a weighted adjacency matrix.

### Details

Sparse SVDs use [RSpectra](#) for performance.

### Value

An object of class vsp. TODO: Details

**Examples**

```
library(LRMF3)

vsp(m1100k, rank = 2)
```

---

vsp.svd\_like

---

*Perform varimax rotation on a low rank matrix factorization*


---

**Description**

Perform varimax rotation on a low rank matrix factorization

**Usage**

```
## S3 method for class 'svd_like'
vsp(
  x,
  rank,
  ...,
  centerer = NULL,
  scaler = NULL,
  recenter = FALSE,
  renormalize = FALSE,
  kaiser_normalize_u = FALSE,
  kaiser_normalize_v = FALSE,
  rownames = NULL,
  colnames = NULL
)
```

**Arguments**

x	Either a graph adjacency matrix, <a href="#">igraph::igraph</a> or <a href="#">tidygraph::tbl_graph</a> . If x is a <a href="#">matrix</a> or <a href="#">Matrix::Matrix</a> then x[i, j] should correspond to the edge going from node i to node j.
rank	The number of factors to calculate.
...	These dots are for future extensions and must be empty.
centerer	TODO
scaler	TODO
recenter	Should the varimax factors be re-centered around the original factor means? Only used when center = TRUE, defaults to FALSE.
renormalize	Should the regularized graph laplacian be used instead of the raw adjacency matrix? Defaults to TRUE. If center = TRUE, A will first be centered and then normalized.

kaiser_normalize_u	Whether or not to use Kaiser normalization when rotating the left singular vectors U. Defaults to FALSE.
kaiser_normalize_v	Whether or not to use Kaiser normalization when rotating the right singular vectors V. Defaults to FALSE.
rownames	Character vector of row names of x. These row names are propagated into the row names of the U and Z. Defaults to NULL.
colnames	Character vector of column names of x. These column names are propagated into the row names of the V and Y. Defaults to NULL.

### Examples

```
library(LRMF3)
library(RSpectra)

s <- svds(m1100k, k = 2)
mf <- as_svd_like(s)
fa <- vsp(mf, rank = 2)
```

---

vsp_fa	<i>Create a vintage sparse factor analysis object</i>
--------	---

---

### Description

vsp\_fa objects are a subclass of `LRMF3::fa_like()`, with additional fields `u`, `d`, `v`, `transformers`, `R_U`, and `R_V`

### Usage

```
vsp_fa(
  u,
  d,
  v,
  Z,
  B,
  Y,
  transformers,
  R_U,
  R_V,
  rownames = NULL,
  colnames = NULL
)
```

**Arguments**

u	A <code>matrix()</code> of "left singular-ish" vectors.
d	A <code>numeric()</code> vector of "singular-ish" values.
v	A <code>matrix()</code> of "right singular-ish" vectors.
Z	A <i>matrix</i> of embeddings for each observation.
B	A mixing <i>matrix</i> describing how observation embeddings and topics interact. Does not have to be diagonal!
Y	A <i>matrix</i> describing the compositions of various topics or factors.
transformers	A list of transformations from the <code>invertiforms</code> package.
R_U	Varimax rotation matrix use to transform u into Z.
R_V	Varimax rotation matrix use to transform v into Y.
rownames	Identifying names for each row of the original data. Defaults to NULL, in which cases each row is given a row number left-padded with zeros as a name.
colnames	Identifying names for each column of the original data. Defaults to NULL, in which cases each column is given a row column left-padded with zeros as a name.

**Value**

A `svd_fa` object.

# Index

aes, 7

bff, 2

bind\_svd\_u (bind\_varimax\_z), 3

bind\_svd\_v (bind\_varimax\_z), 3

bind\_varimax\_y (bind\_varimax\_z), 3

bind\_varimax\_z, 3

facet\_grid, 7

get\_svd\_u, 4

get\_svd\_v (get\_svd\_u), 4

get\_varimax\_y (get\_svd\_u), 4

get\_varimax\_z (get\_svd\_u), 4

get\_y\_hubs (get\_z\_hubs), 4

get\_z\_hubs, 4

GGally::ggpairs, 7

GGally::ggpairs(), 8

ggmatrix\_progress, 8

ggplot2::ggplot(), 8

grab\_legend, 7

igraph::igraph, 3, 11, 12

invertiforms, 14

labellers, 7

LRMF3::fa\_like(), 13

matrix, 3, 11, 12

matrix(), 14

Matrix::Matrix, 3, 11, 12

numeric(), 14

plot\_ipr\_pairs, 5

plot\_mixing\_matrix, 6

plot\_svd\_u (plot\_varimax\_z\_pairs), 6

plot\_svd\_v (plot\_varimax\_z\_pairs), 6

plot\_varimax\_y\_pairs  
(plot\_varimax\_z\_pairs), 6

plot\_varimax\_z\_pairs, 6

progress\_bar, 8

screepLOT.vsp\_fa, 8

set\_y\_factor\_names  
(set\_z\_factor\_names), 9

set\_z\_factor\_names, 9

tibble::tibble(), 4–6, 8

tidygraph::tbl\_graph, 3, 11, 12

unit, 8

vsp, 3, 9

vsp(), 2

vsp.svd\_like, 12

vsp\_fa, 13

vsp\_fa(), 4–9

wrap\_fn\_with\_param\_arg, 7