Package: sdim (via r-universe)

October 4, 2026

Type Package

Title An R Package for Supervised Dimension Reduction

Version 0.1.0

Description Implements five factor extraction methods for asset pricing and macroeconomic forecasting: principal component analysis (PCA), partial least squares (PLS), scaled PCA (sPCA) of Huang, Jiang, Li, Tong, and Zhou (2022) <doi:10.1287/mnsc.2021.4020>, the reduced-rank approach (RRA) of He, Huang, Li, and Zhou (2023) <doi:10.1287/mnsc.2022.4563>, and Instrumented PCA (IPCA) of Kelly, Pruitt, and Su (2019) <doi:10.1016/j.jfineco.2019.05.001>.

URL https://gabbocg.github.io/sdim/, https://github.com/gabbocg/sdim

BugReports https://github.com/gabbocg/sdim/issues

License MIT + file LICENSE

Encoding UTF-8

Language en-GB

LazyData true

Depends R (>= 4.1.0)

Imports stats, graphics, Rcpp

LinkingTo Rcpp, RcppArmadillo

Suggests frenchdata, knitr, readxl, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

RoxygenNote 7.3.2

Config/testthat/edition 3

NeedsCompilation yes

Author Gabriel Cabrera [aut, cre] (ORCID: <https://orcid.org/0000-0001-8803-2338>)

Maintainer Gabriel Cabrera <gabriel.cabrera.guz@gmail.com>

Repository https://cran.r-universe.dev

Date/Publication 2026-07-11 08:30:02 UTC

RemoteUrl https://github.com/cran/sdim

RemoteRef HEAD

RemoteSha 7e22bf9332d4f5697d0f2814032de9914073a651

Contents

estimate_ar_res . . . 2
estimate_ardl_multi . . . 3
eval_factors . . . 4
grunfeld . . . 5
he2023_dacheng202 . . . 6
he2023_factors . . . 7
he2023_ff17vw . . . 8
he2023_ff30vw . . . 8
he2023_ff48ew . . . 9
he2023_ff48vw . . . 10
he2023_ff5 . . . 11
huang2022_ip . . . 12
huang2022_macro . . . 13
ipca_est . . . 14
oos_standardize . . . 15
pca_est . . . 15
pls_est . . . 16
predict.sdim_fit . . . 17
predict.sdim_spca . . . 17
rra_est . . . 18
select_ar_lag_sic . . . 19
spca_est . . . 19

Index **22**

`estimate_ar_res` *Estimate AR(p) model*

Description

Fits an autoregressive model of order `p` for the horizon-`h` target and returns the OLS coefficients and residuals.

Usage

```
estimate_ar_res(y, h, p)
```

Arguments

y	Numeric vector of the target variable.
h	Positive integer; forecast horizon.
p	Non-negative integer; AR lag order.

Value

A list with components:

a_hat Coefficient vector (intercept first).

res Residual vector.

Examples

```
y <- arima.sim(list(ar = 0.7), n = 200)
ar_fit <- estimate_ar_res(y, h = 1, p = 1)
ar_fit$a_hat
```

---

estimate_ardl_multi *Estimate ARDL(p1, p2) model*

Description

Fits an autoregressive distributed lag model for the horizon-h target, with p1 lags of y and p2 lags of additional regressors z (e.g., extracted factors).

Usage

```
estimate_ardl_multi(y, z, h, p)
```

Arguments

y	Numeric vector of the target variable.
z	Numeric matrix of additional regressors (e.g., factor estimates).
h	Positive integer; forecast horizon.
p	Integer vector of length 2: c(p1, p2) where p1 is the number of AR lags and p2 the number of z lags.

Value

Coefficient vector (intercept, AR lags, then z lags).

Examples

```
y <- rnorm(200)
z <- matrix(rnorm(200 * 3), 200, 3)
coefs <- estimate_ardl_multi(y, z, h = 1, p = c(1, 1))
coefs
```

`eval_factors` *Evaluate extracted factors against target returns*

Description

Computes the two performance measures from He, Huang, Li, Zhou (2023), Section 2.4: Total adj-R^2 (Equation 19) and root-mean-squared pricing error (RMSPE, Equation 20).

Usage

```
eval_factors(ret, factors)
```

Arguments

`ret`	Numeric matrix or data frame (T x N) of excess returns for the target portfolios.
`factors`	Numeric matrix (T x K) of extracted factors, e.g. `fit$factors` from `pca_est`, `pls_est`, or `rra_est`.

Value

A named numeric vector with four elements:

RMSPE Root-mean-squared pricing error (percent). Average over assets of the per-asset RMSE of $R_{it} - \hat{\beta}_i' f_t$ (intercept excluded from the fitted value), as in Equation 20. Multiplied by 100 when `ret` is in decimal units.

TotalR2 Total adjusted R^2 (percent), as in Equation 19.

SR Mean absolute alpha-to-residual-volatility ratio (Sharpe ratio of pricing errors).

A2R Mean absolute alpha-to-mean-return ratio.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
set.seed(1)
ret <- matrix(rnorm(100 * 10) / 100, 100, 10)
X   <- matrix(rnorm(100 * 8), 100, 8)
fit <- pca_est(X = X, nfac = 3)
eval_factors(ret = ret, factors = fit$factors)
```

grunfeld *Grunfeld (1958) investment dataset*

Description

Panel data on gross investment for 11 US firms over 20 years (1935–1954), originally from Grunfeld (1958). This is a classic panel dataset used for validating the IPCA estimator against the Python `ipca` package (Kelly, Pruitt, Su, 2019).

Usage

```
grunfeld
```

Format

A data.frame with 220 rows and 5 variables:

firm Character; firm name (11 unique firms).

year Integer; year of observation (1935–1954).

invest Numeric; gross investment (millions of dollars).

value Numeric; market value of the firm (millions of dollars).

capital Numeric; stock of plant and equipment (millions of dollars).

Source

Grunfeld, Y. (1958). The Determinants of Corporate Investment. Ph.D. thesis, Department of Economics, University of Chicago. Loaded from the `statsmodels` Python package (`statsmodels.datasets.grunfeld`).

References

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are Covariances: A Unified Model of Risk and Return. *Journal of Financial Economics*, 134(3), 501–524. doi:10.1016/j.jfineco.2019.05.001

Examples

```
head(grunfeld)

# Reshape for ipca_est(): T x N matrix and T x N x L array
firms <- sort(unique(grunfeld$firm))
years <- sort(unique(grunfeld$year))
N <- length(firms)
TT <- length(years)

ret <- matrix(NA, TT, N)
Z   <- array(NA, dim = c(TT, N, 2))
for (i in seq_along(firms)) {
  idx <- grunfeld$firm == firms[i]
  ret[, i]  <- grunfeld$invest[idx]
```

```
  Z[, i, 1] <- grunfeld$value[idx]
  Z[, i, 2] <- grunfeld$capital[idx]
}

fit <- ipca_est(ret, Z, nfac = 1)
print(fit)
```

`he2023_dacheng202` *Dacheng 202-portfolio value-weighted returns from He, Huang, Li, Zhou (2023)*

Description

Monthly value-weighted returns on 202 portfolios (from Dacheng Xiu's replication data) from the replication package of He, Huang, Li, Zhou (2023). Used as the target return matrix (`target`) in the RRA, PLS, and PCA estimators. Columns are named sequentially `p001`–`p202`.

Usage

```
he2023_dacheng202
```

Format

A data.frame with 552 rows and 203 variables:

date First day of each month, class `Date`.

p001 Portfolio 1 return (percent).

p002 Portfolio 2 return (percent).

... Portfolios p003 through p202 (percent).

Source

He, Huang, Li, Zhou (2023) replication package, doi:10.1287/mnsc.2022.4563.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
head(he2023_dacheng202[, 1:5])
```

`he2023_factors` *Factor proxies from He, Huang, Li, Zhou (2023)*

Description

Monthly returns on 70 factor proxies from the replication package of He, Huang, Li, Zhou (2023): the five Fama-French factors (MKT, SMB, HML, RMW, CMA) plus 65 anomaly-based long-short portfolios. Used as factor proxies (`X`) in the RRA, PLS, and PCA estimators.

Usage

```
he2023_factors
```

Format

A data.frame with 516 rows and 71 variables:

date First day of each month, class `Date`.

MKT Market excess return (percent).

SMB Small-minus-big size factor (percent).

HML High-minus-low value factor (percent).

RMW Robust-minus-weak profitability factor (percent).

CMA Conservative-minus-aggressive investment factor (percent).

... 65 additional anomaly-based long-short factors (percent).

Note

The sample period ends 2016-12-01, twelve months earlier than the portfolio datasets (`he2023_ff48vw`, etc., which end 2017-12-01). Align dates before passing `he2023_factors` as `X` and any portfolio dataset as `target`.

Source

He, Huang, Li, Zhou (2023) replication package, doi:10.1287/mnsc.2022.4563.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
head(he2023_factors[, 1:5])
```

he2023_ff17vw — *Fama-French 17-industry value-weighted portfolios from He, Huang, Li, Zhou (2023)*

Description

Monthly value-weighted returns on the 17 Fama-French industry portfolios from the replication package of He, Huang, Li, Zhou (2023). Used as the target return matrix (`target`) in the RRA, PLS, and PCA estimators.

Usage

```
he2023_ff17vw
```

Format

A data.frame with 528 rows and 18 variables:

date First day of each month, class `Date`.

Food Food products portfolio return (percent).

... 16 additional industry portfolio returns (percent).

Source

He, Huang, Li, Zhou (2023) replication package, doi:10.1287/mnsc.2022.4563.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
head(he2023_ff17vw[, 1:5])
```

he2023_ff30vw — *Fama-French 30-industry value-weighted portfolios from He, Huang, Li, Zhou (2023)*

Description

Monthly value-weighted returns on the 30 Fama-French industry portfolios from the replication package of He, Huang, Li, Zhou (2023). Used as the target return matrix (`target`) in the RRA, PLS, and PCA estimators.

Usage

```
he2023_ff30vw
```

Format

A data.frame with 528 rows and 31 variables:

date First day of each month, class `Date`.

Food Food products portfolio return (percent).

... 29 additional industry portfolio returns (percent).

Source

He, Huang, Li, Zhou (2023) replication package, doi:10.1287/mnsc.2022.4563.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
head(he2023_ff30vw[, 1:5])
```

`he2023_ff48ew` *Fama-French 48-industry equal-weighted portfolios from He, Huang, Li, Zhou (2023)*

Description

Monthly equal-weighted returns on the 48 Fama-French industry portfolios from the replication package of He, Huang, Li, Zhou (2023). Used as the target return matrix (`target`) in the RRA, PLS, and PCA estimators.

Usage

```
he2023_ff48ew
```

Format

A data.frame with 528 rows and 49 variables:

date First day of each month, class `Date`.

Agric Agriculture portfolio return (percent).

Food Food products portfolio return (percent).

... 46 additional industry portfolio returns (percent).

Source

He, Huang, Li, Zhou (2023) replication package, doi:10.1287/mnsc.2022.4563.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
head(he2023_ff48ew[, 1:5])
```

---

he2023_ff48vw — *Fama-French 48-industry value-weighted portfolios from He, Huang, Li, Zhou (2023)*

Description

Monthly value-weighted returns on the 48 Fama-French industry portfolios from the replication package of He, Huang, Li, Zhou (2023). Used as the target return matrix (`target`) in the RRA, PLS, and PCA estimators.

Usage

```
he2023_ff48vw
```

Format

A data.frame with 528 rows and 49 variables:

date First day of each month, class `Date`.

Agric Agriculture portfolio return (percent).

Food Food products portfolio return (percent).

... 46 additional industry portfolio returns (percent).

Source

He, Huang, Li, Zhou (2023) replication package, doi:10.1287/mnsc.2022.4563.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
head(he2023_ff48vw[, 1:5])
```

`he2023_ff5` *Fama-French 5-factor data from He, Huang, Li, Zhou (2023)*

Description

Monthly Fama-French 5-factor returns plus momentum and liquidity, from the replication package of He, Huang, Li, Zhou (2023). The risk-free rate (`RF`) column is used in the replication scripts to convert gross returns to excess returns.

Usage

```
he2023_ff5
```

Format

A data.frame with 652 rows and 9 variables:

date First day of each month, class `Date`.

Mkt-RF Market excess return (percent).

SMB Small-minus-big size factor (percent).

HML High-minus-low value factor (percent).

RMW Robust-minus-weak profitability factor (percent).

CMA Conservative-minus-aggressive investment factor (percent).

RF Risk-free rate (percent).

FFMOM Fama-French momentum factor (percent).

Pastor_Liq Pastor-Stambaugh liquidity factor (percent).

Note

Sample period is 1963-07-01 to 2017-10-01 (652 months). Row 127 corresponds to 1974-01-01, aligning with the start of `he2023_factors`. To extract the matching RF series use `he2023_ff5$RF[127:642]`.

Source

He, Huang, Li, Zhou (2023) replication package, doi:10.1287/mnsc.2022.4563.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
head(he2023_ff5)
```

huang2022_ip *Industrial production growth from Huang, Jiang, Li, Tong, Zhou (2022)*

Description

Monthly growth rate of U.S. industrial production, computed as the first difference of the log IP index, spanning January 1960 to December 2019 (720 months). Used as the forecast target in the out-of-sample replication of Table 4 from Huang et al. (2022).

Usage

```
huang2022_ip
```

Format

A numeric vector of length 720 containing monthly log-differences of the Industrial Production Index. The `"dates"` attribute is an integer vector of dates in `YYYYMM` format (196001 to 201912).

Note

The IP growth target is constructed from raw IP levels provided in the authors' replication package, independently from the INDPRO series in `huang2022_macro` (which uses the FRED-MD transformation code). The two series are numerically identical for this variable.

See the replication script in `inst/replications/huang2022_table4.R`.

Source

Replication package of Huang, Jiang, Li, Tong, and Zhou (2022), doi:10.1287/mnsc.2021.4020.

References

Huang, D., Jiang, F., Li, K., Tong, G., and Zhou, G. (2022). Scaled PCA: A New Approach to Dimension Reduction. *Management Science*, 68(3), 1678–1695. doi:10.1287/mnsc.2021.4020

Examples

```
data(huang2022_ip)
length(huang2022_ip)   # 720
head(huang2022_ip)
```

huang2022_macro *FRED-MD macro predictors from Huang, Jiang, Li, Tong, Zhou (2022)*

Description

Monthly observations on 123 macroeconomic variables from the FRED-MD database (McCracken and Ng, 2016), spanning January 1960 to December 2019 (720 months). Variables are transformed for stationarity using the transformation codes listed in the online data appendix of the paper. Covers output and income, labour market, consumption, housing, inventories and orders, money and credit, interest rates, exchange rates, and prices.

Usage

```
huang2022_macro
```

Format

A numeric matrix with 720 rows and 123 columns. Each column is a macroeconomic time series transformed for stationarity following McCracken and Ng (2016). Column names correspond to FRED-MD mnemonics. The `"dates"` attribute is an integer vector of dates in `YYYYMM` format (196001 to 201912).

Note

This dataset is used together with `huang2022_ip` in the replication of Table 4 from Huang et al. (2022). See the replication script in `inst/replications/huang2022_table4.R`.

Source

Replication package of Huang, Jiang, Li, Tong, and Zhou (2022), doi:10.1287/mnsc.2021.4020.

References

Huang, D., Jiang, F., Li, K., Tong, G., and Zhou, G. (2022). Scaled PCA: A New Approach to Dimension Reduction. *Management Science*, 68(3), 1678–1695. doi:10.1287/mnsc.2021.4020

McCracken, M. W. and Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business and Economic Statistics*, 34(4), 574–589. doi:10.1080/07350015.2015.1086655

Examples

```
data(huang2022_macro)
dim(huang2022_macro)        # 720 x 123
head(colnames(huang2022_macro), 10)
```

`ipca_est` *IPCA factor extraction*

Description

IPCA factor extraction

Usage

```
ipca_est(ret, Z, nfac, max_iter = 100, tol = 1e-06, factor_mean = "zero")
```

Arguments

`ret`	Numeric matrix (T x N) of asset returns. Use `NA` for missing observations (unbalanced panel).
`Z`	Numeric array (T x N x L) of asset characteristics. `NA`s must mirror `ret` exactly.
`nfac`	Positive integer; number of latent factors K to extract.
`max_iter`	Maximum ALS iterations (default 100).
`tol`	Convergence tolerance on Frobenius norm of loading change (default 1e-6).
`factor_mean`	Character scalar specifying how the factor mean is modelled. One of `"zero"` (default, no mean adjustment), `"constant"` (time-series average), or `"VAR1"` (VAR(1) with intercept).

Value

An object of class `"sdim_fit"` with fields: `factors` (T x K), `lambda` (L x K characteristic loadings, i.e. Gamma in Kelly et al.), `eigvals` (factor variances), `factor_mean` (character scalar), `call`, `method = "ipca"`, `nfac`. If `factor_mean = "constant"`: also `mu` (length-K mean vector). If `factor_mean = "VAR1"`: also `var_coef` (K x K), `var_intercept` (length-K), `var_resid` ((T-1) x K).

References

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are Covariances: A Unified Model of Risk and Return. *Journal of Financial Economics*, 134(3), 501–524. doi:10.1016/j.jfineco.2019.05.001

Examples

```
set.seed(1)
ret <- matrix(rnorm(50 * 10) / 100, 50, 10)
Z   <- array(rnorm(50 * 10 * 4), dim = c(50, 10, 4))
fit <- ipca_est(ret, Z, nfac = 2)
print(fit)
```

oos_standardize *Standardize columns to zero mean and unit variance*

Description

Standardize columns to zero mean and unit variance

Usage

```
oos_standardize(X)
```

Arguments

X	A numeric matrix.

Value

A matrix with the same dimensions as X, where each column has been centred and scaled to unit variance.

Examples

```
X <- matrix(rnorm(100), 20, 5)
Xs <- oos_standardize(X)
round(colMeans(Xs), 10)
round(apply(Xs, 2, sd), 10)
```

pca_est *PCA factor extraction*

Description

PCA factor extraction

Usage

```
pca_est(target = NULL, X, nfac, gamma = -1)
```

Arguments

target	Ignored; accepted for API uniformity with other estimators.
X	Numeric matrix or data frame (T x L) of factor proxies.
nfac	Positive integer; number of factors to extract.
gamma	Numeric scalar controlling mean adjustment in the second-moment matrix. ‘gamma = -1‘ (default) gives the sample covariance (traditional PCA). ‘gamma = 10‘ and ‘gamma = 1‘ give the Lettau-Ludvigson variants from He et al. (2023).

Value

An object of class "sdim_fit".

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
set.seed(1)
X <- matrix(rnorm(100 * 8), 100, 8)
fit <- pca_est(X = X, nfac = 3)
print(fit)
```

---

pls_est *PLS factor extraction (Matlab-faithful NIPALS algorithm)*

Description

PLS factor extraction (Matlab-faithful NIPALS algorithm)

Usage

```
pls_est(target, X, nfac)
```

Arguments

target	Numeric matrix (T x N) of target variables (e.g., asset returns). A vector is coerced to a T x 1 matrix.
X	Numeric matrix or data frame (T x L) of factor proxies.
nfac	Positive integer; number of PLS components to extract.

Value

An object of class "sdim_fit".

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
set.seed(1)
X <- matrix(rnorm(100 * 8), 100, 8)
Y <- matrix(rnorm(100 * 5), 100, 5)
fit <- pls_est(target = Y, X = X, nfac = 3)
print(fit)
```

predict.sdim_fit *Project new data onto estimated factor loadings*

Description

Project new data onto estimated factor loadings

Usage

```
## S3 method for class 'sdim_fit'
predict(object, newdata, ...)
```

Arguments

object	An object of class "sdim_fit".
newdata	A numeric matrix or data frame with the same number of columns as the original predictor matrix.
...	Additional arguments (currently ignored).

Value

A numeric matrix of projected factors with nrow(newdata) rows and ncol(object$factors) columns.

predict.sdim_spca *Project new data onto estimated sPCA factor loadings*

Description

Standardizes newdata using the training column means and standard deviations, scales by the estimated (possibly winsorized) regression slopes, and projects onto the sPCA loadings.

Usage

```
## S3 method for class 'sdim_spca'
predict(object, newdata, ...)
```

Arguments

object	An object of class "sdim_spca".
newdata	A numeric matrix or data frame with the same number of columns as the original predictor matrix.
...	Additional arguments (currently ignored).

Value

A numeric matrix of projected factors with `nrow(newdata)` rows and `ncol(object$factors)` columns.

`rra_est` *Reduced-Rank Approach (RRA) factor extraction*

Description

Implements the two-step GMM estimator of He, Huang, Li, and Zhou (2023). Factor proxies `X` are rotated to maximise explanatory power for the target return matrix `target`, using diagonal GMM weighting matrices.

Usage

```
rra_est(target, X, nfac, compute_stat = FALSE)
```

Arguments

`target`	Numeric matrix (T x N) of target variables (e.g., asset returns). A vector is coerced to a T x 1 matrix.
`X`	Numeric matrix or data frame (T x L) of factor proxies.
`nfac`	Positive integer; number of RRA factors to extract.
`compute_stat`	Logical; if `TRUE`, compute the GMM J-test statistic for overidentifying restrictions. Returned as `NULL` when `FALSE` (default) or when degrees of freedom <= 0.

Value

An object of class `"sdim_fit"`.

References

He, J., Huang, J., Li, F., and Zhou, G. (2023). Shrinking Factor Dimension: A Reduced-Rank Approach. *Management Science*, 69(9). doi:10.1287/mnsc.2022.4563

Examples

```
set.seed(1)
X <- matrix(rnorm(100 * 8), 100, 8)
Y <- matrix(rnorm(100 * 5), 100, 5)
fit <- rra_est(target = Y, X = X, nfac = 3)
print(fit)
```

select_ar_lag_sic *Select AR lag order by SIC (BIC)*

Description

Selects the lag order for an autoregressive model of the horizon-h target $y_{t,h}$ by minimising the Schwarz Information Criterion.

Usage

```
select_ar_lag_sic(y, h, p_max)
```

Arguments

y	Numeric vector of the target variable.
h	Positive integer; forecast horizon. For h = 1 the target is simply y.
p_max	Maximum lag order to consider. The function evaluates p = 0, 1, ..., p_max.

Value

Integer: selected lag order. A value of 0 means the intercept-only model is preferred.

Examples

```
y <- rnorm(200)
select_ar_lag_sic(y, h = 1, p_max = 4)
```

spca_est *Scaled PCA factor extraction*

Description

Implements scaled principal component analysis (sPCA): predictors are first standardized, then each standardized predictor is scaled by its univariate predictive slope on the target, and finally principal components are extracted from the scaled predictors.

Usage

```
spca_est(target, X, nfac, winsorize = FALSE, winsor_probs = c(0, 99))
```

Arguments

target	A numeric vector of length T_reg (T_reg <= T).
X	A numeric matrix or data frame with T rows and N columns. When length(target) < nrow(X), the first length(target) rows of the standardized X are used for the scaling regression while all T rows are used for standardization and factor extraction. This matches the out-of-sample workflow in Huang et al. (2022), where the predictive regression y_{t+1} ~ X_t uses fewer rows than the full training window.
nfac	A positive integer giving the number of factors to extract.
winsorize	Logical; if TRUE, winsorize absolute slope estimates before scaling predictors.
winsor_probs	Numeric vector of length 2 giving winsorization percentiles. Used only when winsorize = TRUE.

Details

The function follows the MATLAB implementation of Huang, Jiang, Li, Tong, and Zhou (2022).

Value

An object of class "sdim_spca" with components:

factors A T x nfac matrix of extracted sPCA factors.

beta A numeric vector of predictor-specific predictive slopes.

beta_scaled A numeric vector of scaling coefficients actually used.

col_means Column means of X (used by predict).

col_sds Column standard deviations of X (used by predict).

Xs The standardized predictor matrix.

scaleXs The scaled standardized predictor matrix.

lambda The estimated loading matrix.

residuals Residual matrix from the PCA reconstruction step.

ve2 Average squared residual by row.

eigvals Singular values from the decomposition of scaleXs %*% t(scaleXs).

call The matched function call.

References

Huang, D., Jiang, F., Li, K., Tong, G., and Zhou, G. (2022). Scaled PCA: A New Approach to Dimension Reduction. *Management Science*, 68(3), 1678–1695. doi:10.1287/mnsc.2021.4020

Examples

```
set.seed(123)
X <- matrix(rnorm(200 * 10), nrow = 200, ncol = 10)
y <- rnorm(200)

fit <- spca_est(target = y, X = X, nfac = 3)
dim(fit$factors)
head(fit$beta)

# Predictive alignment: target has fewer rows than X
fit2 <- spca_est(target = y[1:199], X = X, nfac = 3)
dim(fit2$factors)  # 200 x 3 (factors for all T rows)
```

Index

∗ **datasets**
grunfeld, 5
he2023_dacheng202, 6
he2023_factors, 7
he2023_ff17vw, 8
he2023_ff30vw, 8
he2023_ff48ew, 9
he2023_ff48vw, 10
he2023_ff5, 11
huang2022_ip, 12
huang2022_macro, 13

estimate_ar_res, 2
estimate_ardl_multi, 3
eval_factors, 4

grunfeld, 5

he2023_dacheng202, 6
he2023_factors, 7
he2023_ff17vw, 8
he2023_ff30vw, 8
he2023_ff48ew, 9
he2023_ff48vw, 10
he2023_ff5, 11
huang2022_ip, 12, *13*
huang2022_macro, *12*, 13

ipca_est, 14

oos_standardize, 15

pca_est, *4*, 15
pls_est, *4*, 16
predict.sdim_fit, 17
predict.sdim_spca, 17

rra_est, *4*, 18

select_ar_lag_sic, 19
spca_est, 19