Package ‘hdtrd’

November 11, 2024

Type Package

Title Testing relevant difference in high-dimensional linear regression with applications to detect transferability

Version 1.0.1

Author Xu Liu [aut,cre]

Maintainer Xu Liu <liu.xu@sufe.edu.cn>

Description Provide the p-value of the test statistic for high-dimensional relevant difference in the generalized linear regression models and its application to transfer learning. In the paper Liu(2024), we propose novel statistics to test relevant difference of two high-dimensional coefficients in the generalized linear regression models. The proposed method can serve as a means for high-dimensional transfer learning the generalized linear regression models.

License GPL (>= 2)

Depends R (>= 3.2.0), Matrix, glmnet, limSolve

RoxygenNote 7.1.2

NeedsCompilation yes

Repository github

URL https://github.com/xliusufe/hdtrd

Encoding UTF-8

R topics documented:

hdtrd-package 2
bandmatrix 3
eigmax 4
predict_utr 5
projection 6
pvalclc 7
pvalgc 8
pvalrd 10
pvaltrans 11
pvaltrans_cv 13
simulData 16
translasso 17
translasso0 18
utrans 19

Index 21

`hdtrd-package` *Testing relevant difference in high-dimensional linear regression with applications to detect transferability*

Description

Provide the p-value of the test statistic for high-dimensional relevant difference in the generalized linear regression models and its application to transfer learning. In the paper Liu(2024), we propose novel statistics to test relevant difference of two high-dimensional coefficients in the generalized linear regression models. The proposed method can serve as a means for high-dimensional transfer learning the generalized linear regression models.

Details

Package:	hdtrd
Type:	Package
Version:	1.0.1
Date:	2024-06-08
License:	GPL (>= 2)

References

Cui, H., Guo, W. and Zhong, W. (2018). Test for high-dimensional regression coefficients using refitted cross-validation variance estimation. The Annals of Statistics, 46, 958-988.

Chen, Z., Cheng, V. X. and Liu, X. (2024). Hypothesis testing on high dimensional quantile regression. Journal of Econometrics.

Chen, J., Li, Q., and Chen, H. Y. (2022). Testing generalized linear models with highdimensional nuisance parameters. Biometrika, 110. 83-99.

Guo, B.and Chen, S. X. (2016). Tests for high dimensional generalized linear models. Journal of the Royal Statistical Society, Series B, 78, 1079-1102.

Karoui, N, E. (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. The Annals of Statistics, 36(6), 2757-2790.

Kong, W. and Valiant, G. (2017). Spectrum estimation from samples. Annals of Statistics. 45, 2218-2247.

Li, S., Cai, T. T., and Li, H. (2022a). Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control. Journal of the American Statistical Association, 118, 2171-2183.

Li, S., Cai, T. T., and Li, H. (2022b). Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality. Journal of the Royal Statistical Society Series B, 84, 149-173.

Liu, S. (2024). Unified Transfer Learning Models for High-Dimensional Linear Regression. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR. 238, 1036-1044.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Liu, X., Zheng, S. and Feng, X. (2020). Estimation of error variance via ridge regression. Biometrika. 107, 481-488.

Tian, Y. and Feng, Y. (2023) Transfer Learning Under High-Dimensional Generalized Linear Models. Journal of the American Statistical Association, 118, 2684-2697.

Yang, W., Guo, X. and Zhu, L. (2023). Score function-based tests for ultrahigh-dimensional linear models. arXiv:2212.08446.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. Journal of the American Statistical Association, 112, 757-768.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

`bandmatrix` *Construct a sparse banded matrix.*

Description

Provide a sparse banded matrix.

Usage

```
bandmatrix(rho, p, T = 5)
```

Arguments

`rho`	A vector with length T.
`p`	The dimension of the banded matrix.
`T`	The width of band. Default is `T = 5`.

Value

`sighalf`	The matrix $\Gamma \in \mathcal{R}^{(p+T)\times p}$ satisfying $\Sigma = \Gamma^T\Gamma$.
`sigma`	The sparse banded matrix $\Sigma \in \mathcal{R}^{p\times p}$.

References

Chen, Z., Cheng, V. X. and Liu, X. (2024). Hypothesis testing on high dimensional quantile regression. Journal of Econometrics.

Examples

```
p   <- 6
T   <- 3
rho <- seq(T)/(T+1)
fit <- bandmatrix(rho, p, T)
fit$sigma
```

`eigmax` *Estimation of the largest eigenvalue of covariance of a high-dimensional vector*

Description

Provide the estimator of the largest eigenvalue of covariance of a high-dimensional vector (Liu (2024)), as well as all estimated eigenvalues.

Usage

```
eigmax(X, zK = NULL, tJ = NULL, K = 1000, J = 1000, method = 'mpmo',
          nmoms = NULL, timeout = 0L)
```

Arguments

`X`	A data matrix with dimension $n \times p$.
`zK`	A matrix with dimension $K \times 2$, a given complex number, where the first column is the real part and the second column is the imaginary part. Default is `zK = NULL`, where `zK[,1] = rnorm(K)` is generated from standard normal distribution, and `zK[,2] = rep(1,K)/sqrt(n)`.
`tJ`	A J-vector. Default is `tJ = NULL`, where `tJ` is a grid of points in the interval $[\lambda_{\min}(\Sigma), \lambda_{\max}(\Sigma)]$.
`K`	A positive integer, which is the number of complex numbers `zK`. Default is `K = 1000`.
`J`	A positive integer, which is the length of `tJ`. Default is `J = 1000`.
`method`	There are three methods, `'mpmo'`, `'mplp'` and `'empi'`, to estimate the largest eigenvalue of Σ, see details in Liu (2024). Default is `method = 'mpmo'`.
`nmoms`	The number of moments. Default is `nmoms = NULL`, where `nmoms = 7` if `method = 'mpmo'`, `nmoms = 4` if `method = 'mplp'`, and `nmoms` is useless if `method = 'empi'`.
`timeout`	An integer: timeout variable in seconds, defaults to 0L which means no limit is set, see details in the function `linp` of R package "limSolve".

Details

See details in the paper Liu (2024).

Here, for the methods to estimate the largest eigenvalue of Σ, `'mpmo'` denotes the MPMO method; `'mplp'` denotes the MPLP method; and `'empi'` denotes the EMPI method.

Value

`lammax`	Estimator of the largest eigenvalue of Σ.
`lamest`	All estimated eigenvalues of Σ.

References

Karoui, N, E. (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. The Annals of Statistics, 36(6), 2757-2790.

Kong, W. and Valiant, G. (2017). Spectrum estimation from samples. Annals of Statistics. 45, 2218-2247.

Liu, X. (2024). High-dimensional test of relevant difference and its application totransfer learning. Manuscript.

Tian, X., Lu, Y., and Li, W. (2015). A robust test for sphericity of high-dimensional covariance matrices. Journal of Multivariate Analysis, 141, 217-227.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

Examples

```
p = 300
n = 200
sig = toeplitz(0.5^(c(1:p)-1))
sighalf = chol( sig )
X  = matrix(rnorm(n*p), nrow = n)
eigens = eigmax(X = X, method = 'mpmo')
eigens$lammax
```

---

predict_utr *Prediction of a new predictor*

Description

Provide the prediction for a new predictor.

Usage

```
predict_utr(fittrans, X, type = "response")
```

Arguments

`fittrans`	An object from fitting `utrans`.
`X`	A new predictor, a matrix with dimension $n \times p$.
`type`	The type of prediction, including "response" (Default) and "class". Here "response" provides the predicted probability when `family = "binomial"`. "class" predict `0/1` response for logistic regression. Applies only when `family = "binomial"`.

Details

See details in the paper Liu (2024).

Value

`yhat`	The new response $\hat{y}$ based on the new predictor x.

References

Liu, S. (2024). Unified Transfer Learning Models for High-Dimensional Linear Regression. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR. 238, 1036-1044.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Tian, X., Lu, Y., and Li, W. (2015). A robust test for sphericity of high-dimensional covariance matrices. Journal of Multivariate Analysis, 141, 217-227.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

Examples

```
data(simulData_trans_gauss)
fittrans <- utrans(target = dataset[[1]], source = dataset[-1], idtrans = seq(5))

p = ncol(dataset[[1]]$X)
n = 5
sig = toeplitz(0.5^(c(1:p)-1))
sighalf = chol( sig )
x = matrix(rnorm(n*p), nrow = n)

predict_utr(fittrans, x)
```

`projection` *Projection of y onto the closure of covariates x*

Description

Provide the projection of y onto the closure of covariates x.

Usage

```
projection(x, y, family = "gaussian", method = 'lasso', isresid = TRUE)
```

Arguments

`x`	Covariates, a $n \times p$-matrix.
`y`	Response, a n-vector.
`family`	Family for the generalized linear models, including 'gaussian', 'binomial', and 'poisson'. Default is `family = "gaussian"`.
`method`	There are two methods, "qfabs" and "lasso", to estimate the nuisance parameter $\boldsymbol{\alpha}$ in quantile regression. Default is `method = 'lasso'`.
`isresid`	logical. Projected residual $\hat{\boldsymbol{\eta}} = \boldsymbol{x} - \hat{H}\boldsymbol{z}$ is output if `isresid = TRUE`. Coefficient matrix $\hat{H}$ is calculated if `isresid = FALSE`. Default is `resids = TRUE`.

Details

High-dimensional test of relevant difference and its application to transferability test in the generalized Linear regression models (see details in the paper Liu (2024))

$$y_i = H\boldsymbol{X}_i^T.$$

Value

`proj` Projection.

References

Cheng, C., Feng, X., Huang, J. and Liu, X. (2022). Regularized projection score estimation of treatment effects in high-dimensional quantile regression. Statistica Sinica. 32, 23-41.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Examples

```
data(simulData_test_gauss)
x <- datahb$X
y <- datahb$Y
proj <- projection(x, y)
```

`pvalclc` *P-value for high-dimensional test in the generalized linear regression models*

Description

Provide p-value for high-dimensional test in the generalized linear regression models when the nuisance parameter is high-dimensional, see Chen et. al. (2022) for details.

Usage

```
pvalclc(data, family = 'gaussian', method = 'lasso', resids = NULL, psi = NULL)
```

Arguments

`data` A list, including Y (response), $\boldsymbol{X}$, $\boldsymbol{Z}$, where $\boldsymbol{Z}$ can be `NULL`.

`family` Family for the generalized linear models, including 'gaussian', 'binomial', and 'poisson'. Default is `family = "gaussian"`.

`method` There are two methods, "gfabs" and "lasso", to estimate the nuisance parameter $\boldsymbol{\alpha}$ in GLMs. Default is `method = 'lasso'`, which calls `glmnet`.

`resids` An n-vector, which is residual of the GLM. Default is `resids = NULL`. The canonical link function is used if `resids` and `psi` are `NULL`.

`psi` An n-vector, which is $\psi(X_i, \beta_0, \phi) = g'(X_i^\top \beta_0)/V(\mu_i(\beta_0); \phi)$, see Guo and Chen (2016) for the details. Default is `psi = NULL`. The canonical link function is used if both `resids` and `psi` are `NULL`. Here, `psi = rep(1,n)` if `psi = NULL`.

Details

The generalized linear regression models (see details in the paper Guo and Chen (2016))

$$\mu_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma},$$

where $\boldsymbol{Z}_i^T \boldsymbol{\gamma}$ is the control mean function.

The hypothesis test problem is

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \quad versus \quad H_1 : \boldsymbol{\beta} \neq \mathbf{0}.$$

One can input estimated residual `resids` = $y_i - \hat{\mu}_i$ and `psi = NULL` which produces the test statistic and p-value given by Chen et. al. (2022), where $\hat{\mu}_i$ is an estimator of μ_i according to Chen et. al. (2023).

Value

`pvals`	P-value of the corresponding test statistic.
`Tn`	Test statistic $\lvert\hat{U}_n\rvert/\sqrt{2\hat{R}_n}$. Reject H_0 if $\lvert\hat{U}_n\rvert/\sqrt{2\hat{R}_n} > z_{1-\alpha/2}$.

References

Guo, B.and Chen, S. X. (2016). Tests for high dimensional generalized linear models. Journal of the Royal Statistical Society, Series B, 78, 1079-1102.

Chen, J., Li, Q., and Chen, H. Y. (2023). Testing generalized linear models with highdimensional nuisance parameters. Biometrika. 110, 83-99.

Examples

```
data(simulData_test_gauss)
pvals <- pvalclc(data = datahb, family = "gaussian")
pvals$pvals
```

`pvalgc` *P-value for high-dimensional test in the generalized linear regression models*

Description

Provide p-value for high-dimensional test in the generalized linear regression models, see Guo and Chen (2016) for details.

Usage

```
pvalgc(data, family = "gaussian", resids = NULL, psi = NULL)
```

Arguments

`data` A list, including Y (response), $\boldsymbol{X}$, $\boldsymbol{Z}$, where $\boldsymbol{Z}$ can be `NULL`.

`family` Family for the generalized linear models, including 'gaussian', 'binomial', and 'poisson'. Default is `family = "gaussian"`.

`resids` An n-vector, which is residual of the GLM. Default is `resids = NULL`. The canonical link function is used if `resids` and `psi` are `NULL`.

`psi` An n-vector, which is $\psi(X_i, \beta_0, \phi) = g'(X_i^\top \beta_0)/V(\mu_i(\beta_0); \phi)$, see Guo and Chen (2016) for the details. Default is `psi = NULL`. The canonical link function is used if `resids` and `psi` are `NULL`. `psi = rep(1,n)` if `psi = NULL`.

Details

The generalized Linear regression models (see details in the paper Guo and Chen (2016))

$$\mu_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma},$$

where $\boldsymbol{Z}^T \boldsymbol{\gamma}$ is the control mean function.

The hypothesis test problem is

$$H_0 : \boldsymbol{\beta} = \boldsymbol{0} \quad versus \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{0}.$$

One can input estimated residual `resids =` $y_i - \hat{\mu}_i$ and `psi = NULL` which produces the test statistic and p-value given by Chen et. al. (2022), where $\hat{\mu}_i$ is an estimator of μ_i according to Chen et. al. (2023).

Value

`pvals` P-value of the corresponding test statistic.

`Tn` test statistic $\hat{U}_n/\sqrt{2\hat{R}_n}$. Reject H_0 if $\hat{U}_n/\sqrt{2\hat{R}_n} > z_{1-\alpha}$.

References

Guo, B.and Chen, S. X. (2016). Tests for high dimensional generalized linear models. Journal of the Royal Statistical Society, Series B, 78, 1079-1102.

Chen, J., Li, Q., and Chen, H. Y. (2023). Testing generalized linear models with highdimensional nuisance parameters. Biometrika. 110, 83-99.

Examples

```
data(simulData_test_gauss)
pvals <- pvalgc(data = datahb, family = "gaussian")
pvals$pvals
```

`pvalrd` *P-value for high-dimensional testing of relevant difference in the generalized linear regression models when the nuisance parameter is high-dimensional*

Description

Provide p-value for high-dimensional testing of relevant difference in generalized linear regression models (Liu (2024)) when the nuisance parameter is high-dimensional.

Usage

```
pvalrd(data, family = "gaussian", delta0 = 0.1, method = 'lasso',
          resids = NULL, sigma2 = NULL, lammax = NULL)
```

Arguments

`data`	A list, including Y (response), $\boldsymbol{X}$, $\boldsymbol{Z}$, where $\boldsymbol{Z}$ is high-dimensional.
`family`	Family for the generalized linear models, including 'gaussian', 'binomial', and 'poisson'. Default is `family = "gaussian"`.
`delta0`	Relevant difference, a given value by hypothesis test problem $H_0 : \|\beta\|_2 \leq \delta_0$. Default is `delta0 = 0.1`.
`method`	There are two methods, "qfabs" and "lasso", to estimate the nuisance parameter $\boldsymbol{\alpha}$ in quantile regression. Default is `method = 'lasso'`.
`resids`	An n-vector, which is residual of the GLM under H_0. Default is `resids = NULL`, where the canonical link function is used if `resids` and `psi` are `NULL`.
`sigma2`	Estimator of error's variance if family = "gaussian". Default is `sigma2 = NULL`, where `sigma2 = 1`.
`lammax`	Estimator of the largest eigenvalue $\sup_{\|\beta\|_2\leq\delta_0} \beta^T\Sigma^2\beta$. Default is `lammax = NULL`, which is estimated empirically by $\lambda_{\max}(S_n)/(1+\sqrt{p/n})$, see details in Liu (2024).

Details

High-dimensional test of relevant difference and its application to transferability test in the generalized Linear regression models (see details in the paper Liu (2024))

$$\mu_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \boldsymbol{Z}_i^T\boldsymbol{\gamma},$$

where $\boldsymbol{Z}^T\boldsymbol{\gamma}$ is the control mean function, and $\boldsymbol{X}$ is high-dimensional.

The hypothesis test problem is

$$H_0 : \|\boldsymbol{\beta}\| \leq \delta_0 \quad versus \quad H_1 : \|\boldsymbol{\beta}\| > \delta_0.$$

Value

`pvals`	P-value of the corresponding test statistic.
`Tn`	Standardized test statistic.

References

Karoui, N, E. (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. The Annals of Statistics, 36(6), 2757-2790.

Kong, W. and Valiant, G. (2017). Spectrum estimation from samples. Annals of Statistics. 45, 2218-2247.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Tian, X., Lu, Y., and Li, W. (2015). A robust test for sphericity of high-dimensional covariance matrices. Journal of Multivariate Analysis, 141, 217-227.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

Examples

```
data(simulData_test_gauss)
pvals <- pvalrd(data = datahb)
pvals$pvals
```

`pvaltrans` *P-value for high-dimensional testing of relevant difference in high-dimensional transfer learning in the generalized linear regression models.*

Description

Provide p-value for high-dimensional testing of relevant difference in high-dimensional transfer learning in the generalized linear regression models (Liu (2024)).

Usage

```
pvaltrans(target, source, family = "gaussian", delta0 = 0.1, nsource = 10,
          testmethd = 'pvalrd', method = 'lasso', resids = NULL,
          isproj = FALSE, proj = NULL,sigma2 = NULL, lammax = NULL,
          nmoms = NULL, zK = NULL, J = NULL, K = NULL, timeout = 0L)
```

Arguments

`target`	The target dataset, a list, including Y (response), $\boldsymbol{X}$ (covariates).
`source`	The source dataset, a list with sublist. Each sublist includes Y (response), $\boldsymbol{X}$ (covariates).
`family`	Family for generalized linear models, including ‘gaussian’, ‘binomial’, and ‘poisson’. Default is `family = "gaussian"`.
`delta0`	Relevant difference, a given value by hypothesis test problem $H_0 : \|\beta\|_2 \leq \delta_0$. Default is `delta0 = 0.1`.
`nsource`	The number of source datasets. Default is `nsource = 10`.
`testmethd`	There are two methods, "pvalrd" and "pvalclc", to calculate the p-value. Default is `testmethd = 'pvalrd'`, see details in Liu (2024).

`method`	There are two methods, "glm" and "lasso", to estimate the nuisance parameter $\boldsymbol{\alpha}$ under the null hypothesis in the generalized linear regression models, where "glm" method estimates nuisance parameter for classic low-dimensional setting, and "lasso" for high-dimensional setting. Default is `method = 'lasso'` for high-dimensional setting.
`resids`	An n-vector, which is residual of GLM under H_0. Default is `resids = NULL`, where the canonical link function is used if `resids` and `psi` are `NULL`.
`isproj`	logical. Projection score method is applied if `isproj = TRUE`. Default is `isproj = FALSE`, which means that no projection score is applied.
`proj`	The estimated residual of projection score, a list, where each element is a $n \times p$-matrix, $\hat{\eta} = x - \hat{H}z$. Default is `proj = NULL`, which means that projection score is calculated.
`sigma2`	Estimator of error's variance if family = "gaussian". Default is `sigma2 = NULL`, where `sigma2 = 1`.
`lammax`	Esimator of the largest eigenvalue $\sup_{\|\beta\|_2 \leq \delta_0} \beta^T\Sigma^2\beta$, see details in `eigmax`. Default is `lammax = NULL`, where `lammax` is estimated by EMPI method, see `eigmax`. If `testmethd = 'pvalrd'`, there are two choices `lammax = 'mpmo'` or `lammax = 'mplp'`. It is useless if `testmethd = 'pvalclc'`.
`nmoms`	The number of moments. Default is `nmoms = NULL`, where `nmoms = 7` if `method = 'mpmo'`, `nmoms = 4` if `method = 'mplp'`, and `nmoms` is useless if `method = 'empi'`.
`zK`	A matrix with dimension $K \times 2$, a given complex number, where the first column is the real part and the second column is the imaginary part. Default is `zK = NULL`, where `zK[,1] = rnorm(K)` is generated from standard normal distribution, and `zK[,2] = rep(1,K)/sqrt(n)`.
`J`	A positive integer, which is the length of `tJ`. Default is `J = NULL`, which means `J = max(500,3*n,2*p)+200`.
`K`	A positive integer, which is the number of complex numbers `zK`. Default is `K = NULL`, which means `K = max(500,3*n,2*p)+200`.
`timeout`	An integer: timeout variable in seconds, defaults to 0L which means no limit is set, see details in the function `linp` of R package "limSolve".

Details

High-dimensional test of relevant difference and its application to transferability test in the generalized Linear regression models (see details in the paper Liu (2024)).

Linear regression model for target data:

$$Y_{0i} = \boldsymbol{X}_{0i}^T\boldsymbol{\beta}_0 + \epsilon_{0i},$$

and

linear regression model for the kth source data:

$$Y_{ki} = \boldsymbol{X}_{ki}^T\boldsymbol{\beta}_k + \epsilon_{ki},$$

where $\boldsymbol{X}^T\boldsymbol{\beta}$ is a baseline mean function, and $\boldsymbol{X}$ is high-dimensional.

The hypothesis test problem is

$$H_0 : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \delta_0 \quad versus \quad H_1 : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| > \delta_0.$$

Here, for the methods to estimate the largest eigenvalue of Σ, `'mpmo'` denotes the MPMO method; `'mplp'` denotes the MPLP method; and `'empi'` denotes the EMPI method.

Value

pvals P-value of the corresponding test statistic, which is a vector with length nsource.

References

Chen, Z., Cheng, V. X. and Liu, X. (2024). Hypothesis testing on high dimensional quantile regression. Journal of Econometrics.

Karoui, N, E. (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. The Annals of Statistics, 36(6), 2757-2790.

Kong, W. and Valiant, G. (2017). Spectrum estimation from samples. Annals of Statistics. 45, 2218-2247.

Liu, S. (2024). Unified Transfer Learning Models for High-Dimensional Linear Regression. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR. 238, 1036-1044.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Liu, X., Zheng, S. and Feng, X. (2020). Estimation of error variance via ridge regression. Biometrika. 107, 481-488.

Tian, Y. and Feng, Y. (2023) Transfer Learning Under High-Dimensional Generalized Linear Models. Journal of the American Statistical Association, 118, 2684-2697.

Yang, W., Guo, X. and Zhu, L. (2023). Score function-based tests for ultrahigh-dimensional linear models. arXiv:2212.08446.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. Journal of the American Statistical Association, 112, 757-768.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

Examples

```
data(simulData_trans_gauss)
pvals <- pvaltrans(target = dataset[[1]], source = dataset[-1])
pvals
```

---

pvaltrans_cv *P-value for high-dimensional testing of relevant difference in high-dimensional transfer learning in the generalized linear regression models via cross validation method.*

Description

Provide p-value for high-dimensional testing of relevant difference in high-dimensional transfer learning in the generalized linear regression models via cross validation method (Liu (2024)).

Usage

```
pvaltrans_cv(target, source, family = "gaussian", delta0 = 0.1, nsource = 10,
             method = 'lasso', ncv = 10, alpha = 0.05, resids = NULL,
             isproj = FALSE, proj = NULL,sigma2 = NULL, lammax = NULL,
             nmoms = NULL, zK = NULL, J = NULL, K = NULL, timeout = 0)
```

Arguments

`target`	The target dataset, a list, including Y (response), $\boldsymbol{X}$ (covariates).
`source`	The source dataset, a list with sublist. Each sublist includes Y (response), $\boldsymbol{X}$ (covariates).
`family`	Family for generalized linear models, including 'gaussian', 'binomial', and 'poisson'. Default is `family = "gaussian"`.
`delta0`	Relevant difference, a given value by hypothesis test problem $H_0 : \|\beta\|_2 \le \delta_0$. Default is `delta0 = 0.1`.
`nsource`	The number of source datasets. Default is `nsource = 10`.
`method`	There are two methods, "glm" and "lasso", to estimate the nuisance parameter $\boldsymbol{\alpha}$ under the null hypothesis in the generalized linear regression models, where "glm" method estimates nuisance parameter for classic low-dimensional setting, and "lasso" for high-dimensional setting. Default is `method = 'lasso'` for high-dimensional setting.
`ncv`	Number of folds in the cross-validation, which is used to select transferable level δ_0. Default is `ncv = 10`.
`alpha`	logical. Projection score method is applied if `isproj = TRUE`. Default is `isproj = FALSE`, which means that no projection score is applied.
`resids`	An n-vector, which is residual of GLM under H_0. Default is `resids = NULL`, where the canonical link function is used if `resids` and `psi` are `NULL`.
`isproj`	logical. Projection score method is applied if `isproj = TRUE`. Default is `isproj = FALSE`, which means that no projection score is applied.
`proj`	The estimated residual of projection score, a list, where each element is a $n \times p$-matrix, $\hat{\eta} = x - \hat{H}z$. Default is `proj = NULL`, which means that projection score is calculated.
`sigma2`	Estimator of error's variance if family = "gaussian". Default is `sigma2 = NULL`, where `sigma2 = 1`.
`lammax`	Esimator of the largest eigenvalue $\sup_{\|\beta\|_2 \le \delta_0} \beta^T \Sigma^2 \beta$, see details in `eigmax`. Default is `lammax = NULL`, where `lammax` is estimated by EMPI method, see `eigmax`. If `testmethd = 'pvalrd'`, there are two choices `lammax = 'mpmo'` or `lammax = 'mplp'`. It is useless if `testmethd = 'pvalclc'`.
`nmoms`	The number of moments. Default is `nmoms = NULL`, where `nmoms = 7` if `method = 'mpmo'`, `nmoms = 4` if `method = 'mplp'`, and `nmoms` is useless if `method = 'empi'`.
`zK`	A matrix with dimension $K \times 2$, a given complex number, where the first column is the real part and the second column is the imaginary part. Default is `zK = NULL`, where `zK[,1] = rnorm(K)` is generated from standard normal distribution, and `zK[,2] = rep(1,K)/sqrt(n)`.
`J`	A positive integer, which is the length of `tJ`. Default is `J = NULL`, which means `J = max(500,3*n,2*p)+200`.
`K`	A positive integer, which is the number of complex numbers `zK`. Default is `K = NULL`, which means `K = max(500,3*n,2*p)+200`.
`timeout`	An integer: timeout variable in seconds, defaults to 0L which means no limit is set, see details in the function `linp` of R package "limSolve".

Details

High-dimensional test of relevant difference and its application to transferability test in the generalized Linear regression models (see details in the paper Liu (2024)).

Linear regression model for target data:

$$Y_{0i} = \boldsymbol{X}_{0i}^T \boldsymbol{\beta}_0 + \epsilon_{0i},$$

and

linear regression model for the kth source data:

$$Y_{ki} = \boldsymbol{X}_{ki}^T \boldsymbol{\beta}_k + \epsilon_{ki},$$

where $\boldsymbol{X}^T\boldsymbol{\beta}$ is a baseline mean function, and $\boldsymbol{X}$ is high-dimensional.

The hypothesis test problem is

$$H_0 : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \delta_0 \quad versus \quad H_1 : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| > \delta_0.$$

Here, for the methods to estimate the largest eigenvalue of Σ, `'mpmo'` denotes the MPMO method; `'mplp'` denotes the MPLP method; and `'empi'` denotes the EMPI method.

Value

`pvals`	P-value of the corresponding test statistic, which is a vector with length `nsource`.
`s_opt`	The `s_opt`th δ_0 is Selected.

References

Chen, Z., Cheng, V. X. and Liu, X. (2024). Hypothesis testing on high dimensional quantile regression. Journal of Econometrics.

Karoui, N, E. (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. The Annals of Statistics, 36(6), 2757-2790.

Kong, W. and Valiant, G. (2017). Spectrum estimation from samples. Annals of Statistics. 45, 2218-2247.

Liu, S. (2024). Unified Transfer Learning Models for High-Dimensional Linear Regression. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR. 238, 1036-1044.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Liu, X., Zheng, S. and Feng, X. (2020). Estimation of error variance via ridge regression. Biometrika. 107, 481-488.

Tian, Y. and Feng, Y. (2023) Transfer Learning Under High-Dimensional Generalized Linear Models. Journal of the American Statistical Association, 118, 2684-2697.

Yang, W., Guo, X. and Zhu, L. (2023). Score function-based tests for ultrahigh-dimensional linear models. arXiv:2212.08446.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. Journal of the American Statistical Association, 112, 757-768.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

Examples

```
data(simulData_trans_gauss)
np      <- dim(dataset[[1]]$X)
delta0  <- c(1:10)*log(np[1])/np[2]
## pvals <- pvaltrans_cv(target = dataset[[1]], source = dataset[-1], delta0 = delta0, nsource = 1)
```

`simulData` *Simulated data for generalized linear regression models*

Description

Simulated data for generalized linear regression models.

- 'Linear regression' for testing relevant difference (`simulData_test_gauss`),
- 'Poisson regression' for testing relevant difference (`simulData_test_poiss`),
- 'Logistic regression' for testing relevant difference (`simulData_test_binom`).
- 'Linear regression' for transfer learning (`simulData_trans_gauss`),
- 'Poisson regression' for transfer learning (`simulData_trans_poiss`), and
- 'Logistic regression' for transfer learning (`simulData_trans_binom`).

Each dataset includes a list entitled

- `datahb` in `simulData_test` for linear regression models,
- `data_binom` in `simulatedData_Binom` for logistic regression models,
- `data_poiss` in `simulatedData_Poiss` for Poisson regression models,
- `dataset` in `simulData_trans_gauss`, `simulData_trans_binom` and `simulData_trans_poiss` for linear regression, logistic regression and Poisson regression models, respectively. `dataset[[1]]` is the target dataset, and `dataset[-1]` is the 10 source datasets.

Usage

```
data(simulData_test_gauss)
```

Details

For `simulData_test_gauss`, we simulated data generated from linear regression models

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\alpha} + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{X}^T \boldsymbol{\alpha}$ is a baseline mean function.

- Y: the response, an n-vector,
- X: the baseline variable with dimension $n \times p$,
- Z: the interested variable with dimension $n \times q$.

For `simulData_trans_gauss`, we simulated data generated from linear regression models

$$Y_{0i} = \boldsymbol{X}_{0i}^T \boldsymbol{\beta}_0 + \epsilon_{0i},$$

and

Linear regression model for the kth source data:

$$Y_{ki} = \boldsymbol{X}_{ki}^T \boldsymbol{\beta}_k + \epsilon_{ki},$$

where $\boldsymbol{X}^T \boldsymbol{\beta}$ is a baseline mean function, and $\boldsymbol{X}$ is high-dimensional.

References

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Examples

```
data(simulData_test_gauss)
y <- datahb$Y[1:5]
dim(datahb$X)
dim(datahb$Z)

data(simulData_trans_gauss)
y <- dataset[[1]]$Y
dim(dataset[[1]]$X)

dim(dataset[[2]]$X)
```

`translasso` *Estimation of coefficient for the target data by Trans-Lasso from the source data*

Description

Provide the estimator of coefficient for the target data by Trans-Lasso from the source data (Li et al. (2022b)).

Usage

```
translasso(target, source = NULL, idtrans = NULL, nvec = NULL, Itil = NULL, l1 = TRUE)
```

Arguments

`target`	The target dataset, a list, including Y (response), $\boldsymbol{X}$ (covariates).
`source`	The source dataset, a list with sublist. Each sublist includes Y (response), $\boldsymbol{X}$ (covariates). `source` could be `NULL`, in which case `utrans` only fits the target data by `glmnet`.
`idtrans`	The transferable source indices. It can be either a subset of 1,..., length(source). Default is `idtrans = NULL`, which is `idtrans = seq(length(source))`.
`nvec`	A vector integers with length $K_s + 1$, each element of which is the number of indices of samples. Default is `idtrans = NULL`, which is `nvec = c(n0,n1,...,n_K)`.
`Itil`	A vector integers, which is the indices samples on target data.
`l1`	method to estimate the parameter in linear regression. Default is `l1 = TRUE`, which means that lasso is applied.

Details

See details in the paper Li et al. (2022b)

Value

`beta`	The coefficient (including intercept term) fitted target data by trans-lasso.

References

Li, S., Cai, T. T., and Li, H. (2022a). Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control. Journal of the American Statistical Association, 118, 2171-2183.

Li, S., Cai, T. T., and Li, H. (2022b). Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality. Journal of the Royal Statistical Society Series B, 84, 149-173.

Liu, S. (2024). Unified Transfer Learning Models for High-Dimensional Linear Regression. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR. 238, 1036-1044.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Tian, X., Lu, Y., and Li, W. (2015). A robust test for sphericity of high-dimensional covariance matrices. Journal of Multivariate Analysis, 141, 217-227.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

Examples

```
data(simulData_trans_gauss)
fit <- translasso(target = dataset[[1]], source = dataset[-1], idtrans = seq(5))
fit$beta[1:10]
```

---

translasso0 *Estimation of coefficient for the target data by oracle Trans-lasso from the source data*

Description

Provide the estimator of coefficient for the target data by oracle Trans-Lasso from the source data (Li et al. (2022b)).

Usage

```
translasso0(target, source = NULL, idtrans = NULL, nvec = NULL, lamconst = NULL, l1 = TRUE)
```

Arguments

`target`	The target dataset, a list, including Y (response), $\boldsymbol{X}$ (covariates).
`source`	The source dataset, a list with sublist. Each sublist includes Y (response), $\boldsymbol{X}$ (covariates). `source` could be `NULL`, in which case `utrans` only fits the target data by `glmnet`.
`idtrans`	The transferable source indices. It can be either a subset of 1,..., length(source). Default is `idtrans = NULL`, which is `idtrans = seq(length(source))`.
`nvec`	A vector integers with length $K_s + 1$, each element of which is the number of indices of samples. Default is `idtrans = NULL`, which is `nvec = c(n0,n1,...,n_K)`.
`lamconst`	A numeric number, which is a constant in the form `lamconst*sqrt(2*log(p)/n)`. Default is `lamconst = NULL`, which means that `lamconst` is selected by `glmnet`.
`l1`	method to estimate the parameter in linear regression. Default is `l1 = TRUE`, which means that lasso is applied.

Details

See details in the paper Li et al. (2022b)

Value

beta The coefficient (including intercept term) fitted target data by trans-lasso.

References

Li, S., Cai, T. T., and Li, H. (2022a). Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control. Journal of the American Statistical Association, 118, 2171-2183.

Li, S., Cai, T. T., and Li, H. (2022b). Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality. Journal of the Royal Statistical Society Series B, 84, 149-173.

Liu, S. (2024). Unified Transfer Learning Models for High-Dimensional Linear Regression. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR. 238, 1036-1044.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Tian, X., Lu, Y., and Li, W. (2015). A robust test for sphericity of high-dimensional covariance matrices. Journal of Multivariate Analysis, 141, 217-227.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

Examples

```
data(simulData_trans_gauss)
fit <- translasso0(target = dataset[[1]], source = dataset[-1], idtrans = seq(5))
fit$beta[1:10]
```

---

utrans *Estimation of coefficient for the target data by transfer learning from the source data*

Description

Provide the estimator of coefficient for the target data by transfer learning from the source data (Liu (2024)).

Usage

```
utrans(target, source, family = "gaussian", idtrans = NULL)
```

Arguments

`target`	The target dataset, a list, including Y (response), $\boldsymbol{X}$ (covariates).
`source`	The source dataset, a list with sublist. Each sublist includes Y (response), $\boldsymbol{X}$ (covariates). `source` could be `NULL`, in which case `utrans` only fits the target data by `glmnet`.
`family`	Family for generalized linear models, including 'gaussian', 'binomial', and 'poisson'. Default is `family = "gaussian"`.
`idtrans`	The transferable source indices. It can be either a subset of 1,..., length(source). Default is `idtrans = NULL`, which is `idtrans = seq(length(source))`.

Details

See details in the paper Liu (2024)

Value

`fitglmnet`	The object from fitting `cv.glmnet` by CV method, see details in R package "glmnet".
`beta`	The coefficient (including intercept term) of the GLMs to fit target data by transfer learning.
`family`	The response type.

References

Liu, S. (2024). Unified Transfer Learning Models for High-Dimensional Linear Regression. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR. 238, 1036-1044.

Liu, X. (2024). Testing relevant difference in high-dimensional linear regression with applications to detect transferability. Manuscript.

Tian, X., Lu, Y., and Li, W. (2015). A robust test for sphericity of high-dimensional covariance matrices. Journal of Multivariate Analysis, 141, 217-227.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. Journal of Statistical Software, Code Snippets, 30, 1-15.

Examples

```
data(simulData_trans_gauss)
fit <- utrans(target = dataset[[1]], source = dataset[-1], idtrans = seq(5))
fit$beta[1:10]
```

Index

∗ **datasets**
    simulData, 16
∗ **package**
    hdtrd-package, 2

bandmatrix, 3

eigmax, 4

hdtrd *(*hdtrd-package*)*, 2
hdtrd-package, 2

predict_utr, 5
projection, 6
pvalclc, 7
pvalgc, 8
pvalrd, 10
pvaltrans, 11
pvaltrans_cv, 13

simulData, 16

translasso, 17
translasso0, 18

utrans, 19