Package 'jafar'

September 27, 2025

Type Package

Title Joint Additive Factor Regression from Multi-View Data

Version 0.1.0

Description Fit supervised and unsupervised Bayesian integrative factor models on multi-view data.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Depends R (>= 3.5)

Imports Rcpp,
RcppArmadillo,
RcppParallel,
truncnorm,
Hmisc,
stats,
caret,
fields,
ggplot2,
cowplot,
latex2exp,
knitr,
svMisc,
MASS

LinkingTo Rcpp,
RcppArmadillo,
RcppParallel

SystemRequirements C++17

Contents

`bsfp.predict.oos` *Out-of-sample prediction for BSFP*

Description

Modified version of the function `bsfp.predict` from the GitHub repo BSFP for out-of-sample predictions.

Usage

```
bsfp.predict.oos(
  bsfp.fit,
  test_data,
  response_type = "continuous",
  model_params = NULL,
  nsample,
  progress = TRUE,
  starting_values = NULL
)
```

Arguments

`bsfp.fit`	Results from fitting `bsfp` on training data.
`test_data`	Matrix-list dataset of held-out test data.
`response_type`	Continuous or binary response. Must be one of 'continuous' (deafult) or 'binary'.
`model_params`	May be `NULL` if `model_params=NULL` in `bsfp` fit. Otherwise, specify as (`error_vars`, `joint_vars`, `indiv_vars`, `beta_vars`, `response_vars`).
`nsample`	Integer specifing number of Gibbs sampling iterations
`progress`	Boolean determining if progress of the sampler be displayed
`starting_values`	List of starting values for $\mathbf{V}, \mathbf{U}_s, \mathbf{W}_s, \mathbf{V}_s$ for $s = 1, \ldots, q$. If `NULL`, initialize from prior.

Details

Generate new scores for held-out test data based on a training fit of BSFP. Uses the estimated ranks and joint and individual loadings. Cannot be used if missing values are present in test data.

Value

Returns a list with the following parameters:

`test_data`	Test data provided by user
`EY.draw`	List of posterior samples for the E(Y|X), i.e. $\beta_0 + \mathbf{V}\boldsymbol{\beta}_{joint} + \sum_{s=1}^{q} \mathbf{V}_s\boldsymbol{\beta}_s$ for each Gibbs sampling iteration.
`V.draw`	List of posterior samples for joint scores, $\mathbf{V}$
`U.train`	List of posterior samples for joint loadings for each source, $\mathbf{U}_s$ for $s = 1, \ldots, q$ given by the training BSFP fit
`W.train`	List of posterior samples for individual loadings for each source, $\mathbf{W}_s$ for $s = 1, \ldots, q$ given by the training BSFP fit
`Vs.draw`	List of posterior samples for individual scores for each source, $\mathbf{V}_s$ for $s = 1, \ldots, q$
`ranks`	Vector with the estimated joint and individual ranks. `ranks[1]` is the estimated joint rank. `ranks[2:(q+1)]` correspond to the individual ranks for each source.
`tau2.train`	List of posterior samples for the response variance if the response was continuous given by training BSFP fit
`beta.train`	List of posterior samples for the regression coefficients used in the predictive model given by training BSFP fit
`Xm.draw`	List of posterior samples for missing predictors imputations

---

`features_reorder_HC` *predictors preprocess: reorder features via hierarchical clustering for better visualization*

Description

predictors preprocess: reorder features via hierarchical clustering for better visualization

Usage

```
features_reorder_HC(X_m, X_m_test = NULL, K0_HC = 15)
```

Arguments

`X_m`	Train set predictors
`X_m_test`	Test set predictors
`K0_HC`	Reference number of clusters for hierarchical clustering (default: 15)

Value

List of preprocessed features and rescaling factors

gibbs_jafar *Gibbs Sampler for JAFAR*

Description

Fits a Joint Additive FActor Regression (JAFAR) model using Gibbs sampling. Variation across multiple data-views is explained via shared and study-specific latent factors. Default and optional outputs include posterior means of the induced covariances, posterior samples of residual variances, latent factors, and factor loadings. Supports parallel computation and tempered updates to limit rank estimation in extreme large-p-small-n settings.

Usage

```
gibbs_jafar(
  X_m,
  y = NULL,
  yBinary = F,
  K0 = NULL,
  K0_m = NULL,
  tMCMC = 20000,
  tBurnIn = 15000,
  tThin = 10,
  hyperparams = list(),
  get_latent_vars = TRUE,
  get_last_sample = FALSE,
  parallel = TRUE,
  tempered = FALSE,
  rescale_pred = FALSE
)
```

Arguments

X_m	Multi-view input data. Rows should correspond to samples, columns to features. (list of length M; m-th element: matrix n x p_mm).
y	Vector responses (of length n). Set to NULL for unsupervised mode (default: NULL).
yBinary	Logical, indicating if the response(s) are binary (default: FALSE).
K0	Upper bound to numbers of shared latent factors (optional) If NULL, K0 is set to 3*log(max(p_m))
K0_m	Upper bounds to numbers of view-specific latent factors (optional) Length should equal length(X_m). If NULL, K0[m] is set to 3*log(max_(p_m[m]))
tMCMC	Total number of MCMC iterations (default: 20000).
tBurnIn	Number of burn-in iterations (default: 15000).
tThin	Thinning interval for saving samples (default: 10).
hyperparams	List of hyperparameters for the D-CUSP prior distributions. Missing hyperparameters are replaced by defaults encoded in jafar_set_hyperparameters.
get_latent_vars	Logical, whether to return latent factors and loading matrices (default: TRUE).

get_last_sample
: Logical, whether to return only the last sample of the MCMC chain (default: FALSE).

parallel
: Logical, whether to use parallel computation for the loadings' update (default: TRUE).

tempered
: Logical, temperature parameter for tempered sampling (default: FALSE, no tempering).

rescale_pred
: Logical, whether to rescale loadings when computing response predictions (default: FALSE).

Details

The number of samples in output is tEff=(tMCMC-tBurnIn)%/%tThin. The output list contains:

- KNumber shared latent factors (vector of length tEff).
- K_GmNumber view-specific latent factors (matrix tEff x M).
- K_Lm_effNumbers of shared factors active in each view (matrix tEff x M).
- K_Gm_effNumbers of specific factors active in each view (matrix tEff x M).
- active_LmBinary indicators of shared factors activity across views (binary array tEff x K x Ms).
- Cov_m_meanPosterior mean of the covariance matrix for each dataset (list of length M; m-th element: matrix p_m[m] x p_m[m]).
- Marg_Var_mMarginal variances of features (list of length M; m-th element: matrix tEff x p_m[m]).
- s2_inv_mInverse residual variances across views (list of length M; m-th element: matrix tEff x p_m[m]).
- mu_mFeatures intercepts across views (list of length M; m-th element: matrix tEff x p_m[m]).
- hyper_paramList of hyperparameters used for the model, including user-specified values and defaults ones were missing.

If is_supervised = TRUE:

- K_T_effNumbers of shared factors active in the response (vector of length tEff).
- K_Tm_effNumbers of specific factors active in the response (matrix tEff x M).
- active_TBinary indicators of shared factors activity in the response (binary matrix tEff x K).
- active_TmBinary indicators of specific factors activity in the response (list of length M; m-th element: matrix tEff x K_Gm[m]).
- s2_invResponse inverse residual variances (vector of length tEff).
- mu_yResponse intercept (vector of length tEff).
- ThetaResponse loadings on shared factors (matrix tEff x K).
- Theta_mResponse loadings on specific factors (list of length M; m-th element: matrix tEff x K_Gm[m]).
- y_MCLatent probit utilities (matrix tEff x n). (only if yBinary = TRUE).

If get_latent_vars = TRUE:

- Lambda_mLoadings matrices on shared factors (list of length M; m-th element: array tEff x p_m[m] x K).

- Gamma_mLoadings matrices on view-specific factors (list of length M; m-th element: array tEff x p_m[m] x K_Gm[m]).
- etaShared latent factors (array tEff x n x K).
- phi_mView-specific latent factors (list of length M; m-th element: array tEff x n x K_Gm[m]).

If the input matrices X_m contain missing values:

- Xm_MCPosterior samples of imputed values for missing entries. A list of length M; the m-th element is itself a list (one per feature with missingness), each containing an tEff × n_miss matrix of imputed values across MCMC iterations.
- na_idxList of length M; the m-th element gives the column indices of missing entries in X_m[[m]].
- na_row_idxList of length M; the m-th element gives the corresponding row indices of missing entries in X_m[[m]].

If get_last_sample = TRUE:

- last_sampleList of posterior values of all parameters at the last MCMC iteration, including latent factors, loadings, residual variances, and hyperparameters.

Value

A list containing posterior samples, latent variables (if requested), and other relevant model outputs.

Note

- Ensure that all matrices in X_m have the same number of rows (subjects).
- Missing data in X_m are allowed as NA and imputed in the MCMC.

---

gibbs_jfr *Gibbs Sampler for JFR*

Description

Fits a Joint Factor Regression (JFR) model using Gibbs sampling. The model can be fitted in both unsupervised (no response) and supervised (with response y) settings. Default and optional outputs include posterior means of the induced covariances, posterior samples of residual variances, latent factors, and factor loadings. Supports parallel computation and tempered updates to limit rank estimation in extreme large-p-small-n settings.

Usage

```
gibbs_jfr(
  X_m,
  y = NULL,
  yBinary = F,
  K0 = NULL,
  tMCMC = 20000,
  tBurnIn = 15000,
  tThin = 10,
  hyperparams = list(),
```

```
  get_latent_vars = TRUE,
  get_last_sample = FALSE,
  parallel = TRUE,
  tempered = FALSE,
  rescale_pred = FALSE
)
```

Arguments

X_m	Multi-view input data. Rows should correspond to samples, columns to features. (list of length M; m-th element: matrix n x p_mm).
y	Vector responses (of length n). Set to NULL for unsupervised mode (default: NULL).
yBinary	Logical, indicating if the response(s) are binary (default: FALSE).
K0	Upper bound to numbers of latent factors (optional) If NULL, K0 is set to 3*log(max(p_m))
tMCMC	Total number of MCMC iterations (default: 20000).
tBurnIn	Number of burn-in iterations (default: 15000).
tThin	Thinning interval for saving samples (default: 10).
hyperparams	List of hyperparameters for the D-CUSP prior distributions. Missing hyperparameters are replaced by defaults encoded in jafar_set_hyperparameters.
get_latent_vars	Logical, whether to return latent factors and loading matrices (default: TRUE).
get_last_sample	Logical, whether to return only the last sample of the MCMC chain (default: FALSE).
parallel	Logical, whether to use parallel computation for the loadings' update (default: TRUE).
tempered	Logical, temperature parameter for tempered sampling (default: FALSE, no tempering).
rescale_pred	Logical, whether to rescale loadings when computing response predictions (default: FALSE).

Details

The number of samples in output is tEff=(tMCMC-tBurnIn)%/%tThin. The output list contains:

- KNumber latent factors (vector of length tEff).
- K_Lm_effNumbers of latent factors active in each view (matrix tEff x M).
- active_LmBinary indicators of latent factors activity across views (binary array tEff x K x M).
- Cov_m_meanPosterior mean of the covariance matrix for each dataset (list of length M; m-th element: matrix p_m[m] x p_m[m]).
- Marg_Var_mMarginal variances of features (list of length M; m-th element: matrix tEff x p_m[m]).
- s2_inv_mInverse residual variances across views (list of length M; m-th element: matrix tEff x p_m[m]).
- mu_mFeatures intercepts across views (list of length M; m-th element: matrix tEff x p_m[m]).

- hyper_paramList of hyperparameters used for the model, including user-specified values and defaults ones were missing.

If is_supervised = TRUE:

- K_T_effNumbers of latent factors active in the response (vector of length tEff).
- active_TBinary indicators of latent factors activity in the response (binary matrix tEff x K).
- s2_invResponse inverse residual variances (vector of length tEff).
- mu_yResponse intercept (vector of length tEff).
- ThetaResponse loadings on latent factors (matrix tEff x K).
- y_MCLatent probit utilities (matrix tEff x n). (only if yBinary = TRUE).

If get_latent_vars = TRUE:

- Lambda_mLoadings matrices on latent factors (list of length M; m-th element: array tEff x p_m[m] x K).
- etaLatent factors (array tEff x n x K).

If the input matrices X_m contain missing values:

- Xm_MCPosterior samples of imputed values for missing entries. A list of length M; the m-th element is itself a list (one per feature with missingness), each containing an tEff × n_miss matrix of imputed values across MCMC iterations.
- na_idxList of length M; the m-th element gives the column indices of missing entries in X_m[[m]].
- na_row_idxList of length M; the m-th element gives the corresponding row indices of missing entries in X_m[[m]].

If get_last_sample = TRUE:

- last_sampleList of posterior values of all parameters at the last MCMC iteration, including latent factors, loadings, residual variances, and hyperparameters.

Value

A list containing posterior samples, latent variables (if requested), and other relevant model outputs.

Note

- Ensure that all matrices in X_m have the same number of rows (subjects).
- Missing data in X_m are allowed as NA and imputed in the MCMC.

`jafar_set_hyperparameters` *Set Hyperparameters for JAFAR and JFR models*

Description

Helper function to set hyperparameters for `gibbs_jfr` and `gibbs_jafar`. Missing hyperparameters are assigned default values. Supports both unsupervised and supervised (response-guided) settings.

Usage

```
jafar_set_hyperparameters(hyperparams_list, M, is_supervised = FALSE)
```

Arguments

`hyperparams_list`	A named list of hyperparameters to be used in the model.
`M`	Integer, number of data-views.
`is_supervised`	Logical, whether the model is supervised (default: FALSE).

Details

Default hyperparameters include:

- `seed`: random seed for reproducibility (default: 123).
- `t0, t1, t0_adapt`: adaptation parameters for MCMC (default: `t0=-1, t1=-5e-4, t0_adapt=200`).
- `a_m, b_m`: shape and rate of inverse-gamma prior for idiosyncratic noise in each view. Scalars of vectors of length `M` (default: `a_m[m]=3, b_m[m]=1`).
- `prec0m`: precision of normal prior on intercepts. Scalar of vector of length `M` (default: `prec0m[m]=2`).
- `var_spike`: variance of normal spike in cusps. Scalar of vector of length `M` (default: `var_spike[m]=0.005`).
- `a_chi, b_chi`: hyperparameters for slab inverse-gamma prior in cusps. Scalars of vectors of length `M` (default: `a_chi[m]=0.5, b_chi[m]=0.1`).
- `alpha_L, alpha_G`: DP concentration parameters giving the expected number of factors, shared and local. Scalars of vectors of length `M` (default: `alpha_L[m]=5, alpha_G[m]=5`).

If `is_supervised = TRUE`, additional hyperparameters for the response model are

- `a_sig, b_sig`: shape and rate of inverse-gamma prior for idiosyncratic noise (default: `a_sig=3, b_sig=1`).
- `prec0`: precision of normal prior on intercept (default: `prec0=2`).
- `var_spike_y`: variance of normal spike (default: `var_spike_y=0.005`).
- `a_theta, b_theta`: hyperparameters for slab inverse-gamma prior in the slab (default: `a_theta=0.5, b_theta=0.1`).
- `a_xi, b_xi`: shape parameters for beta prior on mixture weight in response loadings (default: `a_xi=3, b_xi=2`).

Value

A named list of hyperparameters with defaults filled in where missing. Scalar values are replicated `M` times where necessary.

multiviewMatchAlign *Perform rotational alignment using multi-view* MatchAlign.

Description

Perform rotational alignment using multi-view MatchAlign.

Usage

```
multiviewMatchAlign(ris_MCMC)
```

Arguments

risMCMC	Posterior samples, as returned by gibbs_jafar or gibbs_jfr.

Value

A modified version of the input risMCMC, with latent factors, loading matrices, and (if supervised) response loadings rotated according to multi-view MatchAlign.

plot_coefficients *Plot induced regression coefficients for y|X=x*

Description

Plot induced regression coefficients for y|X=x

Usage

```
plot_coefficients(yPred, out_path = "~/Desktop/", out_name = "coefficients")
```

Arguments

yPred	Response predictions, output of predict_y or predict_y_raw
out_path	Output path where the generated plot will be saved (default: "~/Desktop/")
out_name	Output file name (default: "coefficients")

plot_correlations *Plot the empirical and inferred within-view correlation matrices*

Description

Plot the empirical and inferred within-view correlation matrices

Usage

```
plot_correlations(
  risMCMC,
  X_m = NULL,
  out_path = "~/Desktop/",
  out_name = "cor_matrices"
)
```

Arguments

risMCMC	Posterior samples, output of gibbs_jafar or gibbs_jfr
X_m	Training set multi-view predictors (optional, default: NULL). If NULL, only inferred correlation matrices are visualized. If not NULL, the empirical correlation matrices are displayed besides the inferred ones
out_path	Output path where the generated plot will be saved (default: "~/Desktop/")
out_name	Output file name (default: "cor_matrices")

plot_loadings *Plot posterior means of factor loadings.*

Description

Rotational alignment must be performed in advanced through the function multiviewMatchAlign

Usage

```
plot_loadings(
  risMCMC,
  out_path = "~/Desktop/",
  out_name_shared = "shared_loadings",
  out_name_specific = "specific_loadings"
)
```

Arguments

risMCMC	Posterior samples, output of gibbs_jafar or gibbs_jfr
out_path	Output path where the generated plot will be saved
out_name_shared	Output file name for the shared component plot (default: "n_factors_shared")
out_name_specific	Output file name for the specific components plot (default: "n_factor_specific")

plot_n_factors *Plot MCMC samples of the inferred number of factors*

Description

Plot MCMC samples of the inferred number of factors

Usage

```
plot_n_factors(
  risMCMC,
  out_path = "~/Desktop",
  out_name_shared = "n_factors_shared",
  out_name_specific = "n_factor_specific"
)
```

Arguments

risMCMC	Posterior samples, output of gibbs_jafar or gibbs_jfr
out_path	Output path where the generated plot will be saved
out_name_shared	Output file name for the shared component plot (default: "n_factors_shared")
out_name_specific	Output file name for the specific components plot (default: "n_factor_specific")

plot_predictions *Plot response predictions against true values*

Description

Plot response predictions against true values

Usage

```
plot_predictions(
  yPred,
  yTrue,
  risMCMC,
  out_path = "~/Desktop/",
  out_name = "predictions"
)
```

Arguments

yPred	Response predictions, output of predict_y or predict_y_raw
yTrue	True values of the responses
risMCMC	Posterior samples, output of gibbs_jafar or gibbs_jfr
out_path	Output path where the generated plot will be saved (default: "~/Desktop/")
out_name	Output file name (default: "predictions")

predict_y — *Response predictions and induced regression coefficients for JAFAR and JFR*

Description

Response predictions and induced regression coefficients for JAFAR and JFR

Usage

```
predict_y(Xpred, risMCMC, rescale_pred = FALSE)
```

Arguments

Xpred	A list of M features' matrices, the m-th of dimension nPred x p_m[m] or possibly with missing (X_m[[m]]=NULL)).
risMCMC	Output of gibbs_jafar or gibbs_jfr containing posterior samples.
rescale_pred	Logical, whether to rescale loadings when computing response predictions (default: FALSE).

Value

A list containing posterior samples of the predicted responses (matrix tEff x nPred), and of the induced regression coefficients for each view (list of length M; m-th element: tEff x p_m[m]).

predict_y_raw — *Response predictions for JAFAR and JFR*

Description

Response predictions for JAFAR and JFR

Usage

```
predict_y_raw(Xpred, risMCMC, rescale_pred = FALSE)
```

Arguments

Xpred	A list of M features' matrices, the m-th of dimension nPred x p_m[m] or possibly with missing (X_m[[m]]=NULL)).
risMCMC	Output of gibbs_jafar or gibbs_jfr containing posterior samples.
rescale_pred	Logical, whether to rescale loadings when computing response predictions (default: FALSE).

Value

A list containing posterior samples of the predicted responses (matrix tEff x nPred).

preprocess_X *predictors preprocess: center & rescale + cdf transform (optional)*

Description

predictors preprocess: center & rescale + cdf transform (optional)

Usage

```
preprocess_X(X_m, X_m_test = NULL, copula = F)
```

Arguments

X_m	Train set predictors
X_m_test	Test set predictors
copula	Apply cdf transformation

Value

List of preprocessed features and rescaling factors

preprocess_y *response preprocess: center & rescale*

Description

response preprocess: center & rescale

Usage

```
preprocess_y(yTrain, yTest = NULL)
```

Arguments

yTrain	Train set responses
yTest	Test set responses

Value

List of preprocessed responses and rescaling factors

Index