Package 'Entropic.Scree'

October 29, 2026

Title Information-Theoretic Measurement Mapping and Rank Estimation

Version 1.0.2

Copyright Terrence J. Lee-St. John (Enli)

Description Maps relationships among measurements using a double-centered, joint-entropy-normalized mutual information matrix. Provides informational-axis loadings, primary and extended spectral-boundary diagnostics, and spectral-volume and informational-gravity summaries for discrete and discretized tabular data. The selected Primary Informational Rank is a candidate information-spectrum boundary, not an identified generating dimension. Optional, explicitly recorded missing-data preparation supports categorical missing states and numeric median fills with protected missingness indicators. For methodological details, see Lee-St. John (2026) <doi:10.5281/zenodo.22028086>.

URL https://doi.org/10.5281/zenodo.22028086

License Apache License 2.0

Encoding UTF-8

Imports data.table, ggplot2, infotheo, parallel, patchwork, Rcpp

LinkingTo Rcpp

Config/roxygen2/version 8.1.0

NeedsCompilation yes

Author Terrence J. Lee-St. John [aut, cre]

Maintainer Terrence J. Lee-St. John <terry@enli.com.au>

Repository CRAN

Date/Publication 2026-10-29 08:10:02 UTC

Contents

Entropic.Scree 2
Update.Entropic.Scree 8

Index **10**

`Entropic.Scree` *Entropic Scree Measurement Mapping and Rank Estimation*

Description

Maps relationships among measurements using a double-centered, joint-entropy-normalized mutual information matrix. Provides informational-axis loadings, primary and extended spectral-boundary diagnostics, and spectral-volume and informational-gravity summaries for discrete and discretized tabular data. The selected Primary Informational Rank is a candidate boundary in the information spectrum; a returned candidate does not establish a resolved elbow or correspondence with generating dimension. For the methods paper, see doi:10.5281/zenodo.22028086.

Usage

```
Entropic.Scree(
  data,
  low_entropy_thresh = 0.05,
  num_bins = NULL,
  bin_multiplier = 1,
  num_cores = parallel::detectCores() - 2,
  interactive_mode = TRUE,
  purge_constants = TRUE,
  check_collinearity = TRUE,
  triple_tap_window = 20,
  fwer_alpha = 0.01,
  extract_eigenvectors = TRUE,
  extract_bipolar_modules = TRUE,
  bipolar_top_n = NULL,
  return_processed_data = FALSE,
  categorical_missing = c("error", "category"),
  numeric_missing = c("error", "median_indicator")
)
```

Arguments

`data`	A data.table with numeric/integer, factor, character, or logical columns and unique names. Numeric category labels should be supplied as factors if they represent nominal states. Base data.frames are not supported.
`low_entropy_thresh`	Numeric. Minimum marginal entropy in nats for original measurements and unprotected indicators. Default 0.05. Zero-entropy original columns are always excluded; an indicator with a retained parent is protected.
`num_bins`	Integer from 2 through 46340. Numeric discretization target and native counting capacity. NULL uses max(2, ceiling(bin_multiplier * N^(1/3))). It must cover every retained categorical vocabulary, including an opted-in missing state; violations are rejected. Increasing it also changes the target for numeric measurements.

`bin_multiplier` Numeric. Multiplier for the automatic bin calculation. Default is 1.0.

`num_cores` Integer. Number of CPU cores for OpenMP parallelization.

`interactive_mode`
: Logical. If TRUE, displays diagnostics and, in an interactive R session, prompts for the primary rank and Extended Signal Tail boundary.

`purge_constants`
: Logical. If TRUE, removes constant measurements and identical duplicates before discretization.

`check_collinearity`
: Logical. If TRUE, checks numeric measurements for exact linear dependencies and removes dependent measurements when the check can be performed.

`triple_tap_window`
: Positive integer requested window w0 (default 20). The actual window is max(5, min(w0, floor(0.1 * m))).

`fwer_alpha` Numeric. Alpha used in the multiple-comparison-adjusted local t threshold. Default is 0.01; this does not establish whole-pipeline error control.

`extract_eigenvectors`
: Logical. Defaults to TRUE. Returns all retained measurements' loadings on every selected primary informational axis, plus the full eigenvector matrix. To omit both, set this and extract_bipolar_modules to FALSE.

`extract_bipolar_modules`
: Logical. Defaults to TRUE. Returns positive and negative loading lists for each primary informational axis; TRUE also enables eigenvector extraction.

`bipolar_top_n` NULL or numeric. NULL (default) keeps every nonzero loading in its corresponding pole. A number strictly between zero and one selects up to that proportion of the retained measurement count per pole; a positive integer selects up to that many per pole. This option never truncates measurement_loadings or eigenvectors.

`return_processed_data`
: Logical. If TRUE, returns the retained analysis table in original units only if measurements or values changed through filtering or opted-in missing-data handling. Internal MI binning/coding alone does not count as a data change. If unchanged, processed_data is NULL and processed_data_status explicitly says to use the original data. No file is written automatically.

`categorical_missing`
: Character. "error" (default) rejects categorical missing values. "category" explicitly represents missing factor, character, or logical values as a distinct, collision-free missing state; this is not imputation of an observed category.

`numeric_missing`
: Character. "error" (default) rejects numeric NA/NaN. "median_indicator" fills them with the column's observed median and adds a binary original-missingness companion. A wholly missing numeric column, or an infinite numeric value, is rejected under either policy.

Details

Rows of `measurement_loadings` identify retained measurements and columns `Axis_1`, ..., `Axis_K` identify the selected primary informational axes, where `K = K_roots`. Every loading is retained, including zeros. These are raw signed eigenvector coefficients (equal-axis coordinates), not eigenvalue-scaled coordinates, measurement-to-score correlations, or native observation-level factor scores. For eigenvalue-scaled coordinates on a chosen positive-axis set, multiply each eigenvector column by the square root of its corresponding `raw_eigenvalues` value. Do not use floored working weights to treat nonpositive raw directions as positive geometry.

The full `eigenvectors` matrix includes all spectral directions for later inspection or rank updates; its presence does not mean all directions were selected. Eigenvector signs are arbitrary, and repeated eigenvalues can admit different bases for the same subspace. Opposite poles express informational contrasts, not positive versus negative correlation between measurements.

The Extended Signal Tail lies after `K_roots` and ends inclusively at `K_extended`; it is empty when the boundaries coincide. Spectral-volume allocations are distinct from coordinate geometry and are not measured proportions of causal signal or measurement error. Factor-Specific Informational Gravity refers to an informational axis, not an identified latent generating factor.

Missing-data preparation. Both missing-data options are off by default. Numeric medians use observed finite values only. Each generated companion is 0 for an originally observed value and 1 for an originally missing value. Identical row-by-row missingness patterns share one companion with all associated parent names, avoiding duplicate indicator weighting. Names and missing-category labels never overwrite or merge existing values. The original input object is not modified. Imputation and explicit missing states change the represented data; results are not missingness-conditioned MI.

Original measurements undergo the usual constant, duplicate, numeric QR, and entropy checks first. Indicators do not enter the numeric QR check. An indicator is mandatory when any associated numeric parent survives, even below the entropy threshold or duplicating an encoded original measurement. Mandatory indicators precede orphaned indicators in filtering. When no parent survives, a companion faces normal binary constant, duplicate and entropy filtering against retained originals and mandatory companions. Constant required indicators are invalid. Complete columns generate no indicator. Protection depends on retaining at least one associated numeric parent, not on the indicator's marginal entropy or uniqueness. Numeric filling uses the ordinary sample median of observed values.

Metric definitions and paper notation. On common prepared records, marginal entropy is $H(X) = -\sum p(x) \ln p(x)$ in nats, and $I(X;Y) = H(X) + H(Y) - H(X,Y)$. Joint-entropy normalization gives $\mathcal{M}_{ij} = I(X_i;X_j)/H(X_i,X_j)$. Double-centering produces $\mathcal{M}_c = \mathbf{H}\mathcal{M}\mathbf{H}$, with $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/m$; its trace is the sum of the raw eigenvalues λ_i. The detector's mean reference is $\mu = \mathrm{Tr}(\mathcal{M}_c)/m$; its baseline count is $b = \#\{i : \eta_i > \mu\}$, not a comparison with the mean working weight.

Working weights are $\eta_i = \max(\varepsilon, \lambda_i)$, with $\varepsilon = 10^{-9}$, $S_\varepsilon = \sum_{i=1}^{m} \eta_i$, and $p_i = \eta_i/S_\varepsilon$ over all m directions. $R_{eff} = \exp(-\sum_i p_i \ln p_i)$. With inclusive boundaries $1 \le k = K_{roots} \le e = K_{extended} \le m$, define $P_{sig} = \sum_{i=1}^{K_{extended}} \eta_i/S_\varepsilon$. Redundant Signal Volume is $m - R_{eff}$, Unique Signal Volume is $P_{sig}R_{eff}$, and Idiosyncratic Informational Volume is $(1-P_{sig})R_{eff}$. Their sum is m. Total Shared Signal Volume is $TSV = (m-R_{eff}) + P_{sig}R_{eff}$. Average Informational Gravity is AIG $= TSV/K_{roots}$. Factor-Specific Informational Gravity is $FSIG_i = TSV\eta_i/\sum_{j=1}^{K_{roots}} \eta_j$, and the Structural Topology Profile is $T_i = FSIG_i/FSIG_1 = \eta_i/\eta_1$ over primary axes. Printed allocation percentages use m as denominator; the spectral fraction P_{sig} is different from TSV/m.

Raw positive mass is $m_+ = \sum_i \max(0, \lambda_i)$ and raw negative mass is $m_- = \sum_{\lambda_i<0} |\lambda_i|$. The Negative Spectral Mass Ratio is $\mathrm{NSMR} = m_-/m_+$ for $m_+ > 0$, before flooring; `NSMR_percent` is exactly $100\,\mathrm{NSMR}$. Raw mass m_+ is distinct from working-weight sum S_ε. Internally, `m_plus` denotes only raw positive mass and `S_epsilon` denotes the working sum.

R identifiers. `raw_eigenvalues` corresponds to λ_i, `eigenvalues` to η_i, and `similarity_matrix` to $\mathcal{M}_c$. `K_log_gap` and `K_triple_tap` correspond to the paper's automatic recommendations K_{log} and K_{tap}; `K_roots` and `K_extended` are the selected boundaries. `total_signal_volume` corresponds to TSV, `FSIG_final` to $FSIG_i$, and `structural_topology_profile` to T_i. `measurement_loadings` contains v_{ik} on the selected primary axes, with every retained measurement included. `Axis_1`, `Axis_2`, etc. name the corresponding informational-axis pole lists. Auxiliary allocations use `FSIG_auxiliary_bulk` and `FSIG_auxiliary_mean_reference`; neither is a unit-eigenvalue Kaiser rule.

The auxiliary bulk endpoint is the last axis before a detected macro gap; without such a gap it is b. Its allocation cutoff is max(e, endpoint). The mean-reference auxiliary cutoff is max(e, b). Each auxiliary applies the same TSV formula at its own cutoff and distributes that volume across the current k primary proportions. These are implementation diagnostics, not separately defined primary metrics in the paper. All cutoffs and their sources are returned and all auxiliary vectors refresh on a rank update.

Numerical geometry and interpretation. A raw axis is numerically positive when its eigenvalue exceeds $m\epsilon_{machine}\max(1, \max_i |\lambda_i|)$. An entirely degenerate spectrum is rejected before floor-based ranks or gravity can be presented. The tolerance is a numerical guard, not a spectral boundary estimator. Selected coefficients on null or negative raw directions may still be exported algebraically, with a warning and explicit status; they do not constitute a wholly positive-spectrum coordinate map. A working spectrum entirely at the floor also produces a warning. The working-weight and allocation formulas are not altered by this diagnostic. Local t thresholds are nominal heuristic thresholds, not established family-wise error control.

Data return. processed_data contains retained original-unit values, opted-in median fills or explicit categorical missing labels, and retained 0/1 companions. Internal discretization and category codes are solely MI-estimation inputs and do not trigger export. Integer versus double storage alone is not a change in numeric observations. The status records changed, added and dropped columns, or explains that the original input can be reused.

Value

A list containing the following components:

`eigenvalues` Working spectral weights: raw eigenvalues floored at 1e-9, in descending order.

`raw_eigenvalues`
Eigenvalues of the centered information matrix before flooring.

`similarity_matrix`
The double-centered, joint-entropy-normalized mutual information matrix.

`retained_features`
Names of measurements retained after preparation and entropy filtering.

`bin_distributions`
Per-measurement lists or arrays of category/bin counts.

`centered_trace` Sum of the diagonal of the centered information matrix.

NSMR
: Negative Spectral Mass Ratio: absolute negative raw eigenvalue mass divided by positive raw eigenvalue mass, as a ratio.

NSMR_percent
: Negative Spectral Mass Ratio in percentage units: exactly 100 * NSMR.

R_eff
: Total Unique Probabilistic Volume: entropy effective rank of the entire working spectrum.

K_log_gap
: Automatic Log-Gap recommendation for the Primary Informational Rank.

K_triple_tap
: Automatic Triple-Tap recommendation for the inclusive Extended Signal Tail boundary.

triple_tap_multiplier
: Nominal local t-multiplier at an accepted departure, or the full-window reference multiplier when the scan does not trigger. NA when the search is unavailable; rank_diagnostics records fallback status.

K_roots
: Selected Primary Informational Rank; not by definition the Intrinsic Generative Rank r.

K_extended
: Selected inclusive Extended Signal Tail boundary used for volume allocation.

total_signal_volume
: Total Shared Signal Volume, in variable equivalents.

unique_signal_volume
: Unique Signal Volume allocated to the selected signal region.

redundant_signal_volume
: Redundant Signal Volume: retained measurement count minus R_eff.

idiosyncratic_volume
: Idiosyncratic Informational Volume: remaining effective-volume allocation outside the selected signal region.

AIG
: Average Informational Gravity, in variable equivalents per primary informational axis.

FSIG_final
: Factor-Specific Informational Gravity allocated across the primary informational axes.

structural_topology_profile
: Structural Topology Profile: primary-axis FSIG divided by the leading-axis FSIG.

measurement_loadings
: Named matrix of every retained measurement's loading on every selected primary informational axis, returned by default. NULL only when both extraction options are FALSE.

eigenvectors
: Named full eigenvector matrix, including unselected directions, returned by default. NULL only when both extraction options are FALSE.

bipolar_modules
: Positive and negative loading lists for each selected primary axis, containing every nonzero loading per pole by default. Zero coefficients remain in measurement_loadings but belong to neither pole. List keys Axis_1, Axis_2, etc. match the measurement_loadings columns. Each contains Positive_Anchor and Negative_Anchor loading lists for that informational axis. NULL when extract_bipolar_modules is FALSE.

`processed_data` Retained original-unit analysis data.table only when requested and changed; otherwise NULL. Internal MI codes are not returned here.

`processed_data_status`
List with requested, changed (NA if not checked), returned, message, and, when requested, changed_columns, dropped_columns and added_columns. An unchanged input produces an explicit message and no duplicate table.

`preparation_info`
Policies, input/prepared row counts, bin target, exclusions, categorical missing labels/counts/codes, observed numeric medians and parent links. Each indicator receipt lists its parents, surviving parents, retained/protected status, ordinary filter findings, entropy in nats, and 0/1 meanings (internally coded 1/2).

`missingness_indicators`
Retained raw 0/1 companion columns when analysis-data return is requested; NULL when none or not requested. Identical missingness patterns share a column.

`positive_axis_diagnostics`
Numerical tolerance, positive raw-axis indices, whether all selected axes are positive, whether all working weights equal the floor, and floor_weight_fraction = sum(eta - pmax(lambda,0))/sum(eta). These diagnostics are unavailable when the input result has no raw eigenvalues.

`rank_diagnostics`
Selection status, primary detector path, automatic boundaries, extended fallback status, search endpoint, actual local window, nominal alpha, boundary eigenvalue ratio and percent drop, and macro-stitch status. These do not assert a resolved elbow.

`auxiliary_diagnostics`
Macro boundary index (first after the gap, NA if absent), inclusive macro endpoint and source, unfloored mean-eigenvalue reference and its working-weight count, and both current inclusive auxiliary allocation cutoffs.

`FSIG_auxiliary_bulk`
Current primary-axis allocation using the broader of K_extended and the explicit macro endpoint, or mean-reference count when no macro gap exists.

`FSIG_auxiliary_mean_reference`
Current primary-axis allocation using the broader of K_extended and the count of working weights above the mean raw eigenvalue.

Output structure

Axis-specific pole lists use `Axis_1`, `Axis_2`, and so on, matching the columns of `measurement_loadings`. Each pole list contains `Positive_Anchor` and `Negative_Anchor` coefficients. `NSMR` is a ratio; `NSMR_percent` is 100 times that ratio.

Within `auxiliary_diagnostics`, `macro_boundary_index` identifies the first axis after a detected macro gap (NA when no gap is detected). `macro_endpoint` is the inclusive endpoint, and `macro_endpoint_source` identifies whether it comes from a detected gap or the mean-reference count. `bulk_allocation_cutoff` is the inclusive auxiliary allocation endpoint. `Update.Entropic.Scree()` returns a result without modifying its input object in place.

Examples

```
# Generate a small random dataset
dummy_data <- data.table::data.table(
  V1 = rnorm(50),
  V2 = rnorm(50),
  V3 = rnorm(50)
)

# Run without interactive rank-selection prompts
res <- Entropic.Scree(dummy_data, interactive_mode = FALSE, num_cores = 1)
res$measurement_loadings
# All measurements on one selected axis, with no loading cutoff:
res$measurement_loadings[, 1, drop = FALSE]
# Pole summaries use the same informational-axis names:
res$bipolar_modules$Axis_1
res$NSMR_percent
```

Update.Entropic.Scree *Update Entropic Scree Spectral Boundaries*

Description

Changes the selected Primary Informational Rank and/or Extended Signal Tail boundary of an existing result. Recalculates spectral-volume and informational-gravity allocations and the primary-axis loading table without recomputing mutual information or the eigendecomposition.

Usage

```
Update.Entropic.Scree(
  scree_obj,
  new_K_roots = NULL,
  new_K_extended = NULL,
  bipolar_top_n = NULL,
  interactive_mode = TRUE
)
```

Arguments

scree_obj	A valid output list from Entropic.Scree().
new_K_roots	Integer. New Primary Informational Rank. NULL retains the current selection.
new_K_extended	Integer. New inclusive Extended Signal Tail boundary. NULL retains the current selection. The resulting pair must satisfy 1 <= K_roots <= K_extended <= m; invalid pairs are rejected, not silently clamped.
bipolar_top_n	NULL or numeric. NULL (default) keeps every nonzero loading per pole. A fraction strictly between zero and one limits each pole to that proportion of the retained measurement count; a positive integer sets a per-pole count. The complete measurement_loadings matrix is never truncated by this option.

`interactive_mode`
: Logical. If TRUE, prints updated metric dashboards to the console.

Details

The returned primary-axis loading matrix uses the same saved full eigenvectors. All retained measurements are included on each selected axis. If the input result has no saved eigenvectors, loading outputs remain NULL; this function does not silently repeat the eigendecomposition. Pole lists use Axis_1, Axis_2, etc., matching the primary loading-matrix columns. See `Entropic.Scree` for definitions and interpretation. Both explicitly named auxiliary allocations refresh using the current ranks. Their explicit inclusive cutoffs and mean raw-eigenvalue reference are recorded in auxiliary_diagnostics. Macro metadata are calculated from the saved spectrum without an eigensolve. Geometry status is unavailable when the input has no raw eigenvalues. Preparation receipts and any saved original-unit analysis data remain unchanged.

Value

The updated result list, returned invisibly, including the selected boundaries, primary measurement_loadings, complete pole lists by default, spectral-volume allocations, informational gravity, and Structural Topology Profile. The saved full eigenvectors and eigenvalues are unchanged.

Output structure

Axis-specific pole lists use `Axis_1`, `Axis_2`, and so on, matching the columns of `measurement_loadings`. Each pole list contains `Positive_Anchor` and `Negative_Anchor` coefficients. `NSMR` is a ratio; `NSMR_percent` is 100 times that ratio.

Within `auxiliary_diagnostics`, `macro_boundary_index` identifies the first axis after a detected macro gap (NA when no gap is detected). `macro_endpoint` is the inclusive endpoint, and `macro_endpoint_source` identifies whether it comes from a detected gap or the mean-reference count. `bulk_allocation_cutoff` is the inclusive auxiliary allocation endpoint. `Update.Entropic.Scree()` returns a result without modifying its input object in place.

Index

Entropic.Scree, 2, *9*

Update.Entropic.Scree, 8