**covSTATIS: a multi-table technique**

**for network neuroscience**

Giulia Baracchini1\*, Ju-Chi Yu2\*, Jenny Rieck3, Derek Beaton4, Cheryl Grady5, Hervé Abdi6, R. Nathan Spreng1

**Affiliations**

1 Montreal Neurological Institute, Department of Neurology and Neurosurgery, McGill University, Montreal, Canada

2 Campbell Family Mental Health Institute, Centre for Addiction and Mental Health, Toronto, Canada

3 Health Canada, Ottawa, Canada

4 Data Science & Advanced Analytics, St. Michael’s Hospital, Toronto, Canada

5 Rotman Research Institute, Baycrest Health Centre, University of Toronto, Toronto, Canada

6 School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, USA

\*theseauthors contributed equally to this work

**Correspondence**

Giulia Baracchini: [giulia.baracchini@mail.mcgill.ca](mailto:giulia.baracchini@mail.mcgill.ca)

**Abstract**

**Main**

Correlation, covariance and distance matrices are among the most commonly used data types in network neuroscience research (Kriegeskorte et al., 2008). They are typically built via pairwise comparisons of regional functional or structural MRI, genetic, PET, or electrophysiological values. Each matrix entry denotes the similarity in some neural measure between region pairs, ultimately reflecting the brain’s network organization. Such matrices, referred to hereafter also as data tables, are typically positive semi-definite, that is all their eigenvalues are non-negative, and symmetric.

In network neuroscience, data tables are often obtained from sets of variables collected on the same individuals (e.g., multiple scans or sessions, multiple imaging modalities), or from the same variables collected on different individuals (e.g., one imaging scan on several participants). Data tables are then compared with one another to answer a variety of questions: statistical reliability, temporal network structure, multi-modal network organization, individual differences, group or population effects (CITE). Network neuroscience thus relies on statistical methods able to align and compare multiple data tables at once. We refer to these methods as multi-table methods.

At their core, multi-table methods rely on decomposition and similarity techniques: data tables are first reduced to their lower dimensional structures, and similarity amongst them is then quantified via distance measures. While a variety of multi-table methods exist in the statistical literature (CITE), their adoption in network neuroscience remains limited, due in large part to researchers’ unfamiliarity with these techniques. To date, gradient-based analysis remains the most widely employed multi-table method in network neuroscience. This approach comes from the generalized Procrustes family of methods and involves matrix decomposition, realignment and similarity analyses (CITE). Single data tables are first separately decomposed into lower dimensional components that maximally explain their variance, via a non-linear technique called diffusion map embedding (CITE Margulies and Coifman). Resulting components from each table are then realigned to a common template and compared to one another (CITE). Gradient-based analysis has been, thus far, applied to a variety of research questions, both in health and disease (CITE), yet it can yield complex results that are challenging to interpret. Complexity primarily lies in the difficulty of reconstructing data patterns captured by each component and tracing individual data back to these components.

If on the one hand, gradient-based analysis provides a robust statistical framework to capture the integrated principles of brain organization, on the other hand advancing our understanding of brain network organization ultimately requires methodologies that preserve data fidelity while enhancing interpretability. Amalgamating multiple data tables in fact inherently amplifies data complexity and dimensionality, underscoring the need, in network neuroscience, for approaches capable of both reducing complexity and delivering interpretable outcomes. To this end, here we introduce an alternative multi-table method, covSTATIS, that sets its roots in similar statistical principles as gradient-based analysis, but allows for a dynamic interaction between input and output data therefore maximizing on interpretability.

covSTATIS sits at the intersection of gradient-based analysis, generalized Canonical Correlation Analysis and Similarity Network Fusion techniques. It is part of the generalized Procrustean family of methods and it is an extension of Principal Component Analysis. covSTATIS stands for “covariance STATIS” with STATIS being a French acronym for “structuring three-way statistical tables” (CITE). It also appears in the statistical literature as DISTATIS, when distance, instead of covariance, matrices are used (distance = 1-covariance). In covSTATIS, single data tables are first linearly combined into a group matrix, the *compromise matrix*, that best recapitulates common patterns across tables. The compromise matrix then undergoes eigenvalue decomposition (EVD) to obtain orthogonal components that maximally explain variance across tables. The distance between each data table and the compromise matrix is then calculated to assess similarity among the observations. Since components are derived from the optimal combination of all data tables, covSTATIS allows for the reconstruction of the patterns captured by each component with respect to the compromise matrix (*global factor scores*), and the projection of each table back onto the single components (*partial factor scores*). This step is what differentiates covSTATIS from gradient-based analysis: decomposition occurs on the combination of all data tables, not on each table singularly, aiding interpretation and examination of individual observations. INSERT FIG steps in txt

covSTATIS is a popular multi-table technique in adjacent scientific fields, such as food quality research (), chemistry (), ecology (), and molecular biology (). In neuroimaging to date, covSTATIS has been applied to a limited number of scenarios: to compare the effects of different preprocessing strategies on brain activation patterns across individuals (Churchill et al., 2012a, 2012b), to examine the similarity of spatial maps generated by different machine learning classifiers within individuals (Yourgnov et al., 2014), to combine and contrast different runs of MRI data within and across individuals (Sha et al., 2015), to integrate and contrast brain activity across different task conditions (St-Laurent et al., 2015) and across different stimuli within a task (Connolly et al., 2016; Mitchell et al., 2016; Rundel et al., 2018). In network neuroscience, thus far, covSTATIS has been used to compare spatial patterns of fMRI connectivity across rest and task states (Rieck), and to estimate temporal profiles of resting-state fMRI connectivity (Baracchini et al., 2023). Other potential applications of covSTATIS are investigations of group and/or individual differences in spatial and/or temporal network structure, deep phenotyping of network metrics, and multimodal assessments of network measures within and across individuals (**Figure 2**). covSTATIS outputs can then be integrated in secondary analyses, such as the examination of brain-behavior relationships.

To facilitate the adoption of covSTATIS by the network neuroscience community, we provide a step-by-step tutorial of covSTATIS using openly available data. In this tutorial, we use covSTATIS to examine how functional network structure varies spatially across task conditions in a healthy adult lifespan sample of 144 individuals. While in the MRI scanner, participants were administered an n-back task with three different conditions (0-back, 1-back, 2-back). Data used in this tutorial are available online (OSF link) and are described in detail in our previous publication (Rieck). The tutorial can be accessed at: <https://github.com/giuliabaracc/covSTATIS_netneuro/blob/main/Code/covstatis_vignette.html> (R markdown file also available).

While still in its early days, covSTATIS represents a theoretically and computationally simple tool for multi-table analyses, able to handle and explain high dimensional, large, complex data typical of network neuroscience research. <add current developments - discriminant, inferences (k1, kk Derek), spare STATIS - and limitations>

A screenshot of a computer

Description automatically generated

A screenshot of a game

Description automatically generated

**Methods**

**Notations**

A matrix is denoted by a bold, upper case letter (e.g., **X**), a vector is denoted by a bold, lower case letter (e.g., **x**), and a scalar is denoted by an italic letter (e.g., *X*). Given *I* data tables, we used the subscript to indicate individual data tables (e.g., **X*i***). **I** denotes the identity matrix, where the diagonal elements are 1s and elements off the diagonal are 0s, with comfortable dimensions. From a matrix **X**, the *j*th column is denoted by **x***j*, and the value on the *k*th row and the *j*th column is denoted by x*i*, *j*. For an *I* × *J* matrix, the minimum of *I* and *J* is the largest possible rank, denoted by *L*, of **X**. The trace(**X**) operator gives the sum of the diagonal elements of **X**.

**covSTATIS**

To generate the compromise space that best represents the common patterns across all data tables (e.g., correlation/covariance matrices), covSTATIS first derives weights from a pairwise similarity matrix, called the RV matrix, which quantifies the similarity between data tables, irrespective of their rotation or scaling, via the RV coefficient. Formally, given two *J* × *J* data tables **X***i* and **X***i’* (e.g., two connectivity matrices with *J* ROIs from the 2 observations *i* and *i’*, e.g., participants or tasks), the RV coefficient between these two matrices is computed as:

(1)

RV ranges between 0 and 1, akin to a squared Pearson’s correlation coefficient.

These RV coefficients are then stored in an *I* × *I* RV matrix, denoted by **C**, where c*i*,*i’* stores the RV coefficient between **X***i* and **X***i’*. As **C** gives the similarity between data tables, the first component of **C** best represents the common pattern across tables, and the first eigenvector of **C** (**u**1) quantifies how similar each table is to this common pattern. As a result, to build the compromise space, weights for each data table are derived by the scaled **u**1 which scaled to sum to 1. Formally, **C** undergoes the EVD:

such that ,(2)

where **Ω** is an *R* × *R* diagonal matrix of eigenvalues of **C** with *R* denoting the rank of **C**, and **U** is a *I* × *R* matrix of eigenvectors of **C**. Next the weights of **X***i* (denoted by *β*i) are obtained as:

, (3)

Where *u*i1 is the *i*thvalue of **u**1, whichcorresponds to **X***i*. The compromise (**X**+) is then computed as the weighted sum of all data matrices, where

, (4)

and decomposed by EVD:

such that , (5)

where **Λ** is an *L* × *L* diagonal matrix of eigenvalues of **X+** with *L* denoting the rank of **X+**, and **P** is a *J* × *L* matrix of eigenvectors of **X+**. From EVD, the *global* factor scores **F** (i.e., factor scores from the compromise) are computed as:

(6)

and the *partial* factor scores (i.e., the factor scores derived from the projection individual tables onto the compromise) are computed as:

.(7)

It is worth noted that these partial factor scores enjoy a barycentric property, where their weighted sums give the global factor scores:

. (8)

**Data availability**

**Code availability**

**Future applications of covSTATIS**

citations impo here: <https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.198?casa_token=XUQa7Dk4PToAAAAA%3ADnVsApr0Fjqh471gGtK8zzEW0QjHDzkxbumcKE0oD-zuH4oUufAmIWHOOI91D7-jnWVFKleTQuML3vc>

such as generalized Procrustes methods, PARAFAC, Tucker’s three-way decomposition, Kroencker-based methods, tensor decomposition, similarity network fusion analysis (CITE),

[1 paragraph on Applications of STATIS, Ju-Chi review + our applications in neuroimaging basically use cases] - Giu still

[Math] - Ju-Chi

[Software demo] - Jenny

[discussion & future applications] - Giu and Ju-Chi (we can use it to integrate across brain and behavior too - Sarah’s paper)

group comparisons task fMRI connectivity

spatial patterns rest-task fMRI connectivity

temporal patterns sliding window/rs dynamic connectivity

start with Asian Noodles

then neuroimaging data

**JR/DB ms:** [**https://github.com/jennyrieck/C-MARINeR/blob/master/manuscript/covstatis\_ms.pdf**](https://github.com/jennyrieck/C-MARINeR/blob/master/manuscript/covstatis_ms.pdf) **+ repo:** [**https://github.com/jennyrieck/C-MARINeR**](https://github.com/jennyrieck/C-MARINeR)

**MDS = single table**

**STATIS = multitable MDS**

* **STATIS:**
  1. **Shinkareva et al. (2008):**
     + **They use machine learning to identify the neural pattern associated with individual object and object categories**
     + **STATIS was used to combine region by region dissimilarity matrix from each participants**
       - **Regions were selected by machine learning methods**
       - **Build region-by-region dissimilarity matrix**
       - **Each dissimilarity matrix was transformed to a cross-product matrix**
       - **Normalized by the first eigenvalue**
       - **Create a subject-by-subject compromise by averaging the weighted cross-product matrices**
       - **Weights were from the first principal component of a subject-by-subject RV-coefficient matrix**
     + **Results:**
       - **Different regions were encoding different information, with clusters formed according to their functions**
       - **Brain activation that is used in identification differs qualitatively land systematically across regions**
       - **Bilaterally homologous regions were similar to each other, despite being physically distant from each other**
       - **For frontal cortex, activation in the two hemispheres was more distinct and left-lateralized**
       - **There is a cross-participant commonality in the neural signature at the level of semantic propery representation (and no just visual features).**

1. **ChurChill et al. (2012a): Standard Temporal Motion and Physiological Noise Correction Methods**
   * **Use DiSTATIS to test whether the effect of different pipelines correlate with significant common changes in activation pattern.**
   * **Other method such as NPAIRS, Friedman nonparametric rank-test, and PDA were used.**
     + **PDA is a multivariate testing adaptation of LDA; PDA is more sensitive to subject-dependent artifacts than GLM**
   * **Choosing pipeline that maximize performance for each subject, with the tradeoff of between-subject heterogeneity**
     + **Pipeline performance was evaluated by (Reproducibility, Prediction) metrics**
     + **Fixed pipeline order, with all possible combination of including/not including physiological noise correction (PNC), motion correction (MC), and motion parameter regression (MPR). Also, the combination includes different temporal detrending (DET) models with different polynomial orders.**
   * **Start with row and column centered distance matrix of rSPM(z) reproducibility between all possible pipeline.**
   * **Distances may be converted to correlation to measure reproducibility between pipeline.**
   * **DiSTATIS helps downweight outlier subjects, and extract the most common pattern of SPM similarity across subjects**
   * **Conslusion: (when pipeline is fixed)**
     + **When the highest detrending order is even, it offers a better (R,P).**
     + **Motion correction improves performance for individual-subject optimization, but might be limited to low motion in the original sample.**
     + **Along with detrending, PNC is important to optimize permorfance.**
     + **In general, MPR and DET dominant the effect of preprocessing steps, and induce similar spatial effect. MC effect is weaker, and with PDA, the PC denoising step can effectively correct for motion artifact. MC is also orthogonal to PNC and MPR. DISTATIS is rather insensitive to even/odd DET probably due to insensitivity to temporal patterns, and suggests that they are pretty robust to this order effect.**

1. **ChurChill et al. (2012b):**
   * **Integrated PCA and two ICA-based denoising methods into the preprocessing pipeline.**
   * **DiSTATIS was ued to evaluate if this is an important step, and if the choice of different techniques has a significant impact.**
     + **ICA-PESTICA**
     + **ICA-MELODIC**
     + **PDA using regularized PCA**
   * **Similar technique to the other paper**
   * **The optimal fixed pipelines vary significantly as a function of task contrast**
   * **Optimizing the ICA subspace using a data-driven step-up procedure optimized on (P,R), as with PCA, may potentially improve this model**
   * **Optimizing the subspace using PCA generally allows for more predictive and reproducible signal subspace estimation; this is beneficial, given the relative computational efficiency and consistency of the estimated subspace for**

**PCA, over ICA algorithms.**

1. **Yourgnov et al. (2014):**
   * **Examine the similarity of spatial maps created by different MVPA classifier.**
   * **DiSTATIS:**
     + **Within-subject distance matrices: a double centered 6 (methods) x 6 distance matrix for each participant**
     + **Compute a subject by subject Rv matrix from the within-subject distance matrices, then do an SVD**
     + **Compute compromise by the weighted sum of the double centered distance matrices of each subject**
       - **Weights are computed as the first factor scores/the sum of the first factor scores of all subject (the first singular value?)**
   * **DiSTATIS results:**
     + **3 overlapping pairs:**
       - **QD and LD-PC: both PCA-based regularization**
       - **SVM and LD-RR: both use L2 penalty for regularization**
       - **GNB-L and GNB-B: both are univariate Gaussian Naïve Bayes**
     + **The strength of the contrast interacts with the pattern similarities between the three pairs of methods**
     + **LD-RR and SVM are most sensitive to contrast effects such that their pattern may reflect either multivariate or univariate features**

1. **Sha et al. (2015):**
   * **Multivariate brain activation pattern analysis: images of animal species with different levels of animacy, and also two inanimate categories of tools**
   * **Use SVM for pattern classification with leave-one-run-out cross validation**
   * **STATIS was used to combine data of 7 runs for each participants:**
     + **7 Runs x 12 (stimuli categories) x n (size of ROIs in voxels)**
     + **First, they used STATIS to combine 7 tables of each run, and create a grand table.**
     + **Next, they penalize each column (voxel) with specific F criterion.**
     + **Third, they centered column-wise, then centered the rows. -> similar to calculating Pearson correlation distance between observations ~ in DSMs (dissimilarity matrix)**
     + **Finally, they generate the compromise, with each row being scaled by its norm (sqrt(SS)).**
   * **STATIS was then used to combine participants of each group:**
     + **Compute 11 compromise for each participants: the weighted linear sum of the seven 12 x n tables (for each run)**
     + **Then, compute a group-wise STATIS:**
       - **Stacked 11 compromise of each participant horizontally**
       - **PCA**
       - **Yields voxel loadings for each voxel in each participant**
   * **Volumetric loadings were mapped to the standard surface mesh grid**
   * **Use supplementary projection**
   * **Summary:**
     + **Multi-way data analysis**
     + **Combine all runs for each participants --> combine all participants**
     + **12 categories and n voxels**
     + **Loadings for different voxels are map to structure map for each individuals**
     + **Then are mapped onto the standard surface mesh grid**
     + **These loadings are being treated as betas in GLM, and the t-test was performed**
   * **The view that suggest dichotomous encoding of living-nonliving distinction exists in the ventral vision pathway tends to be wrong; instead, a graded dimension of animacy describe the representation space in the ventral pathway.**
   * **The representation for the low animacy animal share the same region as the representation for tools**
2. **St-Laurent et al. (2015):**
   * **Use distance matrix to represent individual's activation pattern**
   * **DiSTATIS was used to combine MDS pattern across participants**
   * **Based on results from searchlight analysis, called shrinkage discriminant anlysis (SDA). SDA is a regularized LDA.**

1. **Connolly et al. (2016):**
   * **Examine how predacity is represented in brain activation in visual perception.**
   * **Perform searchlight MVPA with SVM, the identified ROIs were then grouped according to their functions (the network they belong to) with cluster analysis. Finally, STATIS was used to visualize the representational structure within clusters.**
   * **"Cluster analysis was used to investigate how information is represented and transformed along the pathways identified by SVM. This is done by using clustering to divide the pathways into subregions based on intrinsic functional organization."**
   * **N x M matrix for each subject; N is number of stimulus categories, M is number of voxels. (Since it's defined by cluster analysis, it varies across participants.)**
   * **STATIS was used to combined this N x M matrix of different subjects.**
   * **Preprocessing:**
     + **Mean centered along columns**
     + **Mean centered along rows**
     + **Normalized along rows by dividing values by row norms**
     + **Yields the N x N cross-product matrix put into STATIS equivalent to Pearson's correlation matrices of rows**
   * **DiSTATIS? Used MDS calculated using STATIS**
   * **"MDS within subregions provided visual evidence for a representational space separating animal classes based on perceived threat along the first two PCs in STSa, and by the animacy continuum along the first PC in LOC."**
2. **Mitchell (2016):**
   * **Use MDS to represent individual's activation pattern**
   * **DiSTATIS was used to combine MDS pattern across participants**
   * **Strong separation between animate and inanimate item, and big versus small items**
3. **Rundel et al. (2018):**
   * **Use MVPA with STATIS to investigate whether processing of lexical word frequency is anatomically distinct from processing of semantic animacy category information.**
   * **Use searchlight to identify ROIs**
   * **STATIS is used as a way to group matrices of different participants.**
   * **Use STATIS with MDS --> DiSTATIS**

* **Using STATIS without referring to the name - Rv coefficient:**
  1. **Kherif et al. (2003): first apply Rv coefficient to fMRI analysis**
     1. **Use Rv coefficient to measure the homogeneity of subject sampled.**
     2. **Compute Rv coefficient of temporal and spatial data and transformed these Rv into distances.**
     3. **Use MDS to examine the relationship between different subjects**
  2. **Shinkareva et al. (2006): reference to Kherif et al. (2003)**
     1. **Presented a unified feature selection and classification procedure of classifying subjects into groups based on their spatio-temporal data. This approach accounts for and identifies intergroup spatial and temporal variability.**
     2. **This study used Rv coefficient to develop feature selection and classification methods.**
  3. **O'Toole (2007): introducing applying multivariate analysis to neuroimaging analysis**