




Data and text mining

# Signed Distance Correlation (SiDCo): an online implementation of distance correlation and partial distance correlation for data-driven network analysis

Francesco Monti<sup>1,2,‡</sup>, David Stewart<sup>1,2,‡</sup>, Anuradha Surendra<sup>1,2</sup>, Irina Alecu<sup>2,3</sup>,  
Thao Nguyen-Tran <sup>2,3,4</sup>, Steffany A.L Bennett <sup>2,3,4,\*</sup>,  
Miroslava Čuperlović-Culf <sup>1,2,3,\*</sup>

<sup>1</sup>National Research Council of Canada, Digital Technologies Research Centre, Ottawa, Ontario, Canada

<sup>2</sup>Neural Regeneration Laboratory and India Taylor Lipidomic Research Platform, Ottawa, Ontario, Canada

<sup>3</sup>Department of Biochemistry, Microbiology, and Immunology and Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada

<sup>4</sup>Department of Chemistry and Biomolecular Sciences, Centre for Catalysis Research and Innovation, University of Ottawa, Ottawa, Ontario, Canada

\*Corresponding author. Digital Technologies Research Centre, National Research Council of Canada, 1200 Montreal Road, Ottawa, ON, K1A 0R6, Canada; E-mail: cuperlovim@nrc.ca (M.C.-C.); Neural Regeneration Laboratory and India Taylor Lipidomic Research Platform, 451 Smyth Road, Ottawa, Ontario, K1H 8M5, Canada; E-mail: SteffanyAnn.Bennett@uottawa.ca (S.A.L.B.).

<sup>‡</sup>Equal first authors.

Associate Editor: Jonathan Wren

Received 4 January 2023; revised 9 March 2023; editorial decision 14 April 2023

## Abstract

**Motivation:** There is a need for easily accessible implementations that measure the strength of both linear and non-linear relationships between metabolites in biological systems as an approach for data-driven network development. While multiple tools implement linear Pearson and Spearman methods, there are no such tools that assess distance correlation.

**Results:** We present here Signed Distance COrrrelation (SiDCo). SiDCo is a GUI platform for calculation of distance correlation in omics data, measuring linear and non-linear dependencies between variables, as well as correlation between vectors of different lengths, e.g. different sample sizes. By combining the sign of the overall trend from Pearson's correlation with distance correlation values, we further provide a novel "signed distance correlation" of particular use in metabolomic and lipidomic analyses. Distance correlations can be selected as one-to-one or one-to-all correlations, showing relationships between each feature and all other features one at a time or in combination. Additionally, we implement "partial distance correlation," calculated using the Gaussian Graphical model approach adapted to distance covariance. Our platform provides an easy-to-use software implementation that can be applied to the investigation of any dataset.

**Availability and implementation:** The SiDCo software application is freely available at <https://complimet.ca/sidco>. Supplementary help pages are provided at <https://complimet.ca/sidco>. **Supplementary Material** shows an example of an application of SiDCo in metabolomics.

## 1 Introduction

The analysis of biological networks, as a parallel investigation to the study of individual feature characteristics, requires robust quantification of the interconnections between features within biological systems (Ma'ayan 2011). Several methods for data-driven network determination of feature interconnections have been used in the analysis of metabolomic data. Pearson or Spearman correlation-based

methods are arguably the most prevalent (Amara et al. 2022). While providing critical information about the direction of dependencies, both methods measure linear or monotonic correlations and cannot detect non-linear feature interactions (Rosato et al. 2018). Distance correlation, a non-parametric approach for correlation analysis, can measure various types of data relationships (linear and non-linear) as well as the correlations between vectors of different lengths (Gábor and Maria 2009; Székely and Rizzo 2013; Edelman et al.

2021). In metabolomic and lipidomic datasets, distance correlation can take into consideration the sparse coverage of feature data, the potential for determining non-linear relationships, and the possibly random network topologies associated with metabolism and inherent to lipidomic and metabolomic datasets with zero correlation only obtained for fully independent features.

Despite these advantages, few publications have used distance correlation to analyze metabolomic data (Oliveira et al. 2015; Tang et al. 2019; Cuperlovic-Culf et al. 2021). We suggest that this is, in part, due to the lack of easily accessible implementations. Moreover, no parallel implementation, to our knowledge, allows users to assess partial correlations, calculated as the measure of association between pairs of features while removing the confounding effects of other variables. To address this need and thereby provide new methods for the reconstruction of regulatory metabolomic and lipidomic networks, we present here Signed Distance COrelation (SiDCo), a web-based application of both signed distance correlation and partial distance correlation implemented using the Gaussian Graphical Model (GGM) previously implemented for other correlation approaches (Lauritzen 1996).

## 2 Implementation

SiDCo is implemented in Python with a RShiny front-end. It is compatible with all web browsers. Two analytical tabs allow users to perform either distance correlation or partial distance correlation. In both applications, users define their desired threshold values and  $P$  values. Data are automatically z-score normalized across all samples prior to analysis. Users are reminded that missing values must be imputed according to their specifications or data will be returned with the descriptive error message.

Distance correlations and  $P$  values are calculated and presented as described below and a correlation directionality sign is derived from Pearson correlation analysis as an indication of the overall linear trend in the data. Distance correlation calculations in SiDCo are provided in three forms: (i) “one-to-one,” calculating correlations between each pair of features, (ii) “one-to-all,” providing correlations for each feature with all other features combined, and (iii) partial correlation calculated for each pair of features while controlling for the contributions of other features, i.e. covariates.

Distance correlation,  $dCor(X, Y)$  between features  $X$  and  $Y$  and distance covariance,  $dCov(X, Y)$  are calculated as:

$$dCor(X, Y) = \frac{dCov(X, Y)}{\sqrt{dVar(X)dVar(Y)}}, \quad dCov(X, Y)^2 = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n A_{j,k} B_{j,k},$$

where  $A$  and  $B$  represent doubly centered distance matrices for variables  $X$  and  $Y$ , respectively, measured in  $n$  samples. Distance variances ( $dVar$ ) are:  $dVar(X) = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n A_{j,k}^2$  and  $dVar(Y) = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n B_{j,k}^2$

In a one-to-one correlation calculation, an array of values for each feature is compared with an array of values for all the other features one at a time. In this case, a doubly centered distance matrix is calculated as:

$A_{j,k} = a_{j,k} - a_{j.} - a_{.k} + \bar{a}$ , and  $B_{j,k} = b_{j,k} - b_{j.} - b_{.k} + \bar{b}$ ; where Euclidean distance is used to calculate  $x_j$  to  $x_k$  or  $y_j$  to  $y_k$  as  $a_{j,k} = \sqrt{(x_j - x_k)(x_j - x_k)}$  and  $b_{j,k} = \sqrt{(y_j - y_k)(y_j - y_k)}$ .  $a_{j.}$ ,  $b_{j.}$  and  $a_{.k}$ ,  $b_{.k}$  are the  $j$ -row and  $k$ -column mean values and  $\bar{a}$ ,  $\bar{b}$  are the overall mean of matrices  $A$  and  $B$ .

In a one-to-all case, the distance covariance for each feature out of  $m$  features in  $n$  dimensional sample space is compared to that of all the other features in  $n \times (m-1)$  dimensional space. The doubly centered distance matrix for variable  $Y$  used in the calculation of

$dCov$  is here:  $b_{j,k} = \sqrt{\sum_{s=1}^{m-1} (y_{js} - y_{ks})(y_{js} - y_{ks})}$  and equivalent for  $a_{j,k}$  for  $X$ .

The distance correlation  $P$  value is calculated using the Student's  $t$  cumulative distribution function with  $t$  value calculated as:

$t(X, Y) = dCor(X, Y) \sqrt{n-2} / \sqrt{1 - dCor(X, Y)^2}$  and corresponding two-sided  $P$  value for the  $t$ -distribution with  $n-2$  degrees of freedom. The sign of the distance correlation is given by the sign of the Pearson correlation following (Pardo-Diaz et al. 2021). The final output is provided as an .xlsx file and includes distance correlation and corresponding  $P$  values. Output of one-to-one analysis also includes the Pearson and Spearman correlations and their corresponding  $P$  values for completeness.

Partial correlation, the correlation between two features corrected for contributions of other features, is calculated as (following GGM):

$$\rho_{i,j} \cdot \forall k \neq i,j = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$$

Where matrix  $\omega = \Sigma^{-1}$  is inverse of  $\Sigma$  - distance covariance matrix. The inverse of the distance covariance matrix uses the Moore-Penrose method for pseudo-inverse which is equivalent to standard inversion for non-singular square matrices and multiplicative inverse for singular matrices where inverse is not possible. Partial distance correlation calculation should only be performed when number of features is smaller than number of samples. Here  $P$  values are calculated using the Fisher z-transformed correlation values:  $z_{ij} = 0.5 * \log \frac{1+\rho_{ij}}{1-\rho_{ij}}$  and cumulative standard normal distribution ( $cdf$ ) function:  $p_{i,j} = 2(1 - cdf(z_{ij} * \sqrt{N - M - 1}))$ . Results are provided as .xlsx downloads.

## 3 Conclusion

SiDCo is an open-access Web-based application for the calculation of signed and partial distance correlations between features available at <https://complimet.ca/SiDCo> where detailed instructions are provided.

## Supplementary data

Supplementary data are available at *Bioinformatics* online.

## Conflict of interest

None declared.

## Funding

This work was supported in part by RGPIN-2019-06796 to S.A.L.B. from the NSERC, as well as an operating grant [AI-4D-102-3] to S.A.L.B. and M.C.C. from the NRC AI4D Program.

## References

- Amara A, Frainay C, Jourdan F et al. Networks and graphs discovery in metabolomics data analysis and interpretation. *Front Mol Biosci* 2022;9: 841373. <https://doi.org/10.3389/fmolb.2022.841373>.
- Cuperlovic-Culf M, Cunningham EL, Teimoorinia H et al. Metabolomics and computational analysis of the role of monoamine oxidase activity in delirium and SARS-COV-2 infection. *Sci Rep* 2021;11:10629. <https://doi.org/10.1038/s41598-021-90243-1>.
- Edelmann D, Móri TF, Székely GJ. On relationships between the Pearson and the distance correlation coefficients. *Stat Probab Lett* 2021;169:108960. <https://doi.org/10.1016/j.spl.2020.108960>.
- Gábor JS, Maria LR. Brownian distance covariance. *Ann Appl Stat* 2009;3: 1236-65. <https://doi.org/10.1214/09-AOAS312F>.

- Lauritzen SL. *Graphical Models*. Oxford: Clarendon Press, 1996.
- Ma'ayan A. Introduction to network analysis in systems biology. *Sci Signal* 2011;4:tr5. <https://doi.org/10.1126/scisignal.2001965>.
- Oliveira AP, Dimopoulos S, Busetto AG *et al*. Inferring causal metabolic signals that regulate the dynamic TORC1-dependent transcriptome. *Mol Syst Biol* 2015;11:802. <https://doi.org/10.15252/msb.20145475>.
- Pardo-Diaz J, Bozhilova LV, Beguerisse-Diaz M *et al*. Robust gene coexpression networks using signed distance correlation. *Bioinformatics* 2021;37:1982–9. <https://doi.org/10.1093/bioinformatics/btab041>.
- Rosato A, Tenori L, Cascante M *et al*. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics* 2018;14:37. <https://doi.org/10.1007/s11306-018-1335-y>.
- Székely GJ, Rizzo ML. Partial distance correlation with methods for dissimilarities. *arXiv: Methodology* 2013;1310.2926.
- Tang ZZ, Chen G, Hong Q *et al*. Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. *Front Genet* 2019;10:454. <https://doi.org/10.3389/fgene.2019.00454>.