The C-SHIFT algorithm for normalizing covariances

Evgenia Chunikhina*, Paul Logan†, Yevgeniy Kovchegov‡, Anatoly Yambartsev§, Debashis Mondal¶, Andrey Morgun‖
*Data Science, Pacific University, Forest Grove, OR, USA Email: chunikhe.vazhno@gmail.com
†HP Inc., Corvallis, OR, USA
‡Department of Mathematics, Oregon State University, Corvallis, OR, USA
§IME, University of São Paulo, São Paulo, Brazil
¶Dept. of Mathematics and Statistics, Washington University in St. Louis, St. Louis, MO, USA
‖College of Pharmacy, Oregon State University, Corvallis, OR, USA

***Abstract*—Omics technologies are powerful tools for analyzing patterns in gene expression data for thousands of genes. Due to a number of systematic variations in experiments, the raw gene expression data is often obfuscated by undesirable technical noises. Various normalization techniques were designed in an attempt to remove these non-biological errors prior to any statistical analysis. One of the reasons for normalizing data is the need for recovering the covariance matrix used in gene network analysis. In this paper, we introduce a novel normalization technique, called the covariance shift (C-SHIFT) method. This normalization algorithm uses optimization techniques together with the blessing of dimensionality philosophy and energy minimization hypothesis for covariance matrix recovery under additive noise (in biology, known as the bias). Thus, it is perfectly suited for the analysis of logarithmic gene expression data. Numerical experiments on synthetic data demonstrate the method's advantage over the classical normalization techniques. Namely, the comparison is made with Rank, Quantile, cyclic LOESS (locally estimated scatterplot smoothing), and MAD (median absolute deviation) normalization methods. We also evaluate the performance of C-SHIFT algorithm on real biological data.**

Gene expression analysis plays an important role in genomic research. Several omics technologies such as RNAseq and microarrays allow for the collection of massive amounts of simultaneous measurements of gene expression levels of thousands to tens of thousands of genes. Analyzing different patterns of gene expressions helps to gain insight into complex biological phenomena such as development, aging, onset and progression of diseases, and cellular response/reaction to drugs/treatments. Although new technologies are constantly developing, it is well known that all of them generate some technical noise which affects the measured gene expression levels [13, 24, 31]. To extract accurate biological information it becomes necessary to normalize the data to filter out/compensate for these non-biological noises/errors. Normalization is a crucial pre-processing step in the gene expression data analysis. The gene expression data will vary significantly after different normalization methods. Thus, the results of further data analysis (e.g. gene expression network) will be critically dependent on a choice of a normalization technique. A variety of normalization procedures have been used on gene expression data sets. See [4, 5, 18, 21, 23, 26, 28, 30] and reference therein for a review and comparison of current normalization strategies. In this paper we develop a novel normalization technique, called the covariance shift (C-SHIFT) method, and compare it to the following well known normalization methods used in large scale data analysis: Rank, Quantile, cyclic LOESS (locally estimated scatterplot smoothing), and MAD (median absolute deviation). See [1, 5, 25, 26] and references therein for more details on the above listed normalization methods. There is an important distinction: while Rank, Quantile, LOESS and other normalizations normalize the data, C-SHIFT algorithm normalizes the covariances. The need to normalize the covariances is caused by the presence of bias.

A. Bias.

Consider a situation where the gene expression data is subjected to multiplicative noise (aka bias). Let M be the number of genes and N be the number of measurements. Next, we let $X_n^{(i)}$ denote the true gene expression, where subscript index n stands for the n-th gene in the network and the superscript index i stands for the i-th measurement.

The observed gene expression, denoted by $\widetilde{X}_n^{(i)}$, is different from $X_n^{(i)}$ due to all gene expressions in the i-th measurement being distorted by a multiplicative noise $W^{(i)}$, i.e.,

$$\widetilde{X}_n^{(i)} = W^{(i)} X_n^{(i)}, \tag{1}$$

where random variables $X_n^{(i)}$ are independent of the variable $W^{(i)}$. Additionally random variables $W^{(i)}$ $(i = 1, \ldots, N)$ are assumed to be independent and identically distributed (i.i.d.). Here, both the observed and the true gene expressions are positive, i.e., $X_n^{(i)} > 0$ and $W^{(i)} > 0$.

In biology, the multiplicative noise $W^{(i)}$ is referred to as the bias. The bias is prompted by random events causing an error in the measurement of the total amount of RNA. Such random events are often related to different levels of tissue preservation in different samples that leads to variability of RNA degradation. Consequently, this leads to an RNA detection problem. Additionally, there are other technical reasons for an error in the measurement of the total amount of RNA in a given sample that may lead to a bias in (1). All other noise (e.g. misreading parts of RNA) goes into the variable $X_n^{(i)}$.

The multiplicative noise in (1) implies the corresponding additive noise (bias) in the logarithimic gene expression data:

$$\widetilde{Y}_n^{(i)} = Y_n^{(i)} + V^{(i)}, \tag{2}$$

where we let $\widetilde{Y}_n^{(i)} := \log \widetilde{X}_n^{(i)}$, $Y_n^{(i)} := \log X_n^{(i)}$, and $V^{(i)} := \log W^{(i)}$.

B. Impact of bias on covariances and correlations.

While the bias may not appear critical, they are known to cause significant problems in the analyses of gene correlation structure. Specifically, this phenomenon is known [24] to cause the disappearance of the large magnitude negative correlations in the observed biological data, $\widetilde{X}_n$ and $\widetilde{Y}_n$, which hampers the ability to perform certain types of statistical data analysis, such as the false discovery rate (FDR) method.

The bias, whether multiplicative as in (1) or additive as in (2), causes the correlations to be shifted away from -1. In particular, the independent additive noise in (2) implies an increase of theoretical covariance as

$$Cov(\widetilde{Y}_n, \widetilde{Y}_m) = Cov(Y_n, Y_m) + \omega, \tag{3}$$

where $\omega = Var(V) > 0$. Consequently, the correlations in the logarithmic data are equal to

$$corr(\widetilde{Y}_n, \widetilde{Y}_m) = \frac{Cov(Y_n, Y_m) + \omega}{\sqrt{\Big(Var(Y_n) + \omega\Big)\Big(Var(Y_m) + \omega\Big)}}. \tag{4}$$

If $Cov(Y_n, Y_m)$ is negative, by adding $\omega > 0$ in the numerator and the denominator, we obtain

$$corr(\widetilde{Y}_n, \widetilde{Y}_m) > corr(Y_n, Y_m).$$

Hence, the disappearance of large magnitude negative correlations.

The purpose of the covariance shift (C-SHIFT) algorithm developed in this current manuscript is to normalize covariances in the logarithmic data and restore the correlations, thus offsetting the impact of the additive bias in (2). Consequently, the comparison of C-SHIFT covariance normalization algorithm with methods of normalizing data such as Rank, Quantile, or LOESS can only be done in terms of the effectiveness of recovering true empirical correlations. This comparison will be implemented on synthetic data in Section II-B and on real biological data in Section III.

The problem of improving the existing and developing new normalization methods is very important for scientists working with biological data. The fact that normalization alters the data-correlation structure was stated in Saccenti [30]. Besides [30] gives a comprehensive overview of normalization methods. In Bolstad *et al.* [5] the authors compare three complete data normalization methods (cyclic LOESS, contrast based method, and quantile), that make use of data from all arrays in an experiment, with two methods that make use of a baseline array. The comparison was done on two publicly available datasets with the results favoring the complete data methods. For more on the normalization methods, see [1, 6, 7, 9, 13, 14, 26, 29, 32].

C. Paper structure and workflow diagram.

The paper is organized as follows. In Section I, we formulate C-SHIFT method from the underlying theoretical considerations. The pseudocode for the C-SHIFT algorithm is given in Section II-A. Section II-B contains numerical experiments on two synthetic datasets. Section III evaluates the outcomes of correlation recovery using six real biological datasets from GEO depository. The results and future directions are discussed in Section IV. Finally, Section V contains the proofs.

Workflow diagram can be found in Fig. 1.

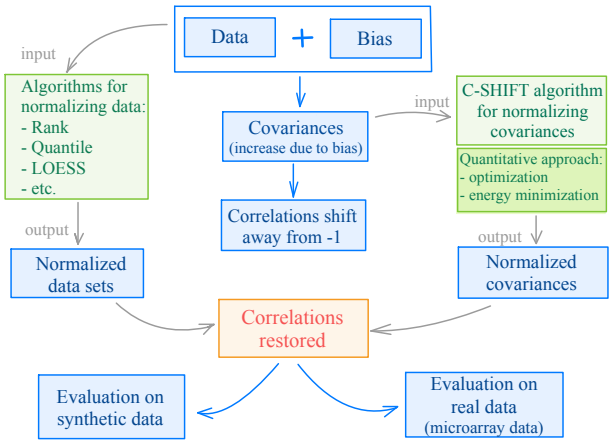

Fig. 1. Workflow diagram.

I. Theoretical derivations

Denote by $\widehat{Cov}$ the empirical covariances taken over N samples for each of $\binom{M}{2}$ pairs of genes. Similarly, let $\widehat{Var}$ denote the empirical variance. Then, equation (2) yields the observed empirical covariance

$$\widehat{Cov}(\widetilde{Y}_n, \widetilde{Y}_m) = \widehat{Cov}(Y_n, Y_m) - \hat{a}_n - \hat{a}_m + \hat{\omega} \quad (5)$$

for all pairs of gene indices n and m, where $\hat{a}_n = -\widehat{Cov}(Y_n, V)$ for all $n = 1, \ldots, M$, and $\hat{\omega} = \widehat{Var}(V) > 0$. As is often the case, $\hat{\omega}$ can be very large relative to the values of $\hat{a}_n$, causing the disappearance of the large magnitude negative correlations in empirical data.

The goal of the covariance shift (C-SHIFT) normalization method introduced here is the recovery of the true empirical covariances $\widehat{Cov}(Y_n, Y_m)$ and the respective true empirical correlations in the case of the logarithmic gene expression data or any other situations with additive noise as in (2).

Let $\widetilde{C} = \left(\widehat{Cov}(\widetilde{Y}_n, \widetilde{Y}_m)\right)_{n,m}$ be the empirical covariance matrix of the observed data $\widetilde{Y}_n^{(i)}$, and let $C = \left(\widehat{Cov}(Y_n, Y_m)\right)_{n,m}$ be the empirical covariance matrix of the cleaned data $Y_n^{(i)}$ (i.e., the true empirical covariance) that we desire to recover. Formula (5) rewritten in the matrix form states

$$C = \widetilde{C} + \hat{a}\mathbf{1}^T + \mathbf{1}\hat{a}^T - \hat{\omega}\,\mathbf{1}\mathbf{1}^T, \quad (6)$$

where $\hat{a} = \left(\hat{a}_1, \ldots, \hat{a}_M\right)^T$, and $\mathbf{1}$ denotes the column vector of 1's, hence $\mathbf{1}\mathbf{1}^T$ is a square matrix of 1's.

Our goal here is to estimate $\hat{a}$ and $\hat{\omega}$ in (6), and thus recover the true empirical covariance matrix C. We assume large dimension M. There will be two cases.

Case I: If $\det(\widetilde{C}) = 0$ (e.g. $N < M$), we make a small perturbation of the diagonal entries of $\widetilde{C}$ (the variances) resulting in a new covariance matrix being positive definite whose smallest eigenvalue is still very close to zero. Next, we use energy minimization to estimate $\hat{a}_n$ and $\hat{\omega}$ in (6).

Case II: If $\widetilde{C}$ is positive definite (full rank), our approach exploits the phenomenon sometimes referred to as the *curse of dimensionality* [3, 27] and sometimes as the *blessing of dimensionality* [8, 12, 16], postulating that in higher dimensions almost all data points are located near extrema (i.e., in the outer shell)*. In other words, for large M, we anticipate the smallest eigenvalue of C to be near zero. As a rigorous bound, we observe that if some of the correlations $corr(Y_n, Y_m)$ are located in $[-1, \delta - 1]$ interval, then the smallest eigenvalue of C is located within $\left[0, \delta \min_n \widehat{Var}(Y_n)\right]$ interval. Thus, as in Case I, under the blessing of dimensionality assumption, we again use energy minimization for estimating $\hat{a}_n$ and $\hat{\omega}$.

Next, we will need the following result.

Proposition 1. *Suppose $\mathcal{M}$ is a symmetric positive definite square matrix, and let*

$$v^* := \max\left\{v : \mathcal{M} - v\,\mathbf{1}\mathbf{1}^T \text{ is positive semidefinite}\right\}.$$

Then,

$$v^* = \frac{1}{\mathbf{1}^T \mathcal{M}^{-1} \mathbf{1}}.$$

The proof of Proposition 1 is given in Section V.

Suppose the empirical covariance matrix $\widetilde{C}$ is positive definite, i.e., $\widetilde{C}$ is of full rank. Consider values of a column vector $\alpha = (\alpha_1, \ldots, \alpha_M)^T$ such that

$$\widetilde{C} + \alpha\mathbf{1}^T + \mathbf{1}^T\alpha$$

is positive definite. If we let

$$v(\alpha) := \frac{1}{\mathbf{1}^T \left(\widetilde{C} + \alpha\mathbf{1}^T + \mathbf{1}\alpha^T\right)^{-1} \mathbf{1}}, \quad (7)$$

then Prop. 1 implies

$$C_\alpha := \widetilde{C} + \alpha\mathbf{1}^T + \mathbf{1}\alpha^T - v(\alpha)\mathbf{1}\mathbf{1}^T \quad (8)$$

is positive semidefinite with $\det(C_\alpha) = 0$.

Next, recall the quantities $\hat{a}$ and $\hat{w}$ in (6). If $\widetilde{C}$ is rank deficient, we perturb its diagonal entries by adding small positive (random or deterministic) values, and if $\widetilde{C}$ has full rank, we assume the blessing of dimensionality phenomenon holds. Thus, in either case, we

*In this paper we will refer to the phenomenon as the blessing of dimensionality rather than the curse of dimensionality.

work under the assumption that $\widetilde{C}$ is positive definite with its smallest eigenvalue located near zero. Then, Prop. 1 implies $\hat{w} \approx v(\hat{a})$, where $v(\alpha)$ is as defined in (7). Therefore, letting $\alpha = \hat{a}$ in (8), we will have $C_{\hat{a}}$ approximating C expressed as in (6).

Now, for a matrix X, let $\|X\|_F$ denote the Frobenius norm of X and let $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2$ be the energy function. Our next assumption states that $\hat{a}$ can be estimated by the minimizer α^* of the energy function $\mathcal{E}(C_\alpha)$, i.e., we estimate $\hat{a}$ with

$$\alpha^* = \operatorname{argmin} \|C_\alpha\|_F.$$

The assumption is additionally justified by the observation that a random adjustment of the covariance via an additive noise (bias) as in (5) will result in an energy increase, i.e., $\mathcal{E}(\widetilde{C}) > \mathcal{E}(C)$.

Matrix C_{α^*} will approximate $C_{\hat{a}}$, which, in turn, approximates the desired true empirical covariance matrix C. The covariance shift (C-SHIFT) algorithm works as follows: it uses optimization algorithms to estimate α^* and outputs C_{α^*} as an estimate for C. See Algorithm 1 in Section II-A.

The following lemma yields a close form expression for the gradient $\nabla\|C_\alpha\|_F^2$ that will be used to estimate α^* which minimizes $\|C_\alpha\|_F$.

Lemma 1. *Suppose the empirical covariance matrix $\widetilde{C}$ is of full rank, and the quantities C_α and $v(\alpha)$ are as in (8) and (7). Then, the gradient of the Frobenius norm squared is given by*

$$\begin{aligned}\frac{1}{4}\nabla\|C_\alpha\|_F^2 = &M\alpha + \widetilde{C}\mathbf{1} + [a - M\,v(\alpha)]\mathbf{1} \\ &+ [M^2\,v^2(\alpha) - c\,v(\alpha) - 2Ma\,v(\alpha)]A_\alpha^{-1}\mathbf{1},\end{aligned} \tag{9}$$

where $\|\cdot\|_F$ denotes the Forbenius norm, and we let

$$A_\alpha := \widetilde{C} + \alpha\mathbf{1}^T + \mathbf{1}\alpha^T, \quad c := \mathbf{1}^T\widetilde{C}\mathbf{1}, \quad a := \mathbf{1}^T\alpha.$$

The proof of Lemma 1 is given in Section V.

First, we observe that C_α is invariant under the addition of multiples of $\mathbf{1}$. Thus, without loss of generality, we restrict the domain to a hyperplane $a = \text{Const}$. Next, we notice that $\mathbf{1}^T\nabla\|C_\alpha\|_F^2 = 0$ in (9). Thus, in the gradient descent method, the value of a remains constant, i.e., throughout the algorithm, vector α remains on the same hyperplane $\mathbf{1}^T\alpha = \text{Const}$.

In our next lemma, we find the Hessian of $\|C_\alpha\|_F^2$. When minimizing $\|C_\alpha\|_F$ (equivalently, $\|C_\alpha\|_F^2$) both the gradient and the Hessian of $\|C_\alpha\|_F^2$ are inputted in the optimization algorithm such as trust-region or gradient descent.

Lemma 2. *Suppose the empirical covariance matrix $\widetilde{C}$ is of full rank, and the quantities C_α, $v(\alpha)$, and A_α are as in (8), (7), and (10) respectively. Then, the Hessian of $\|C_\alpha\|_F^2$, denoted by $H_\alpha := \operatorname{Hess}\big(\|C_\alpha\|_F^2\big)$ is expressed as follows*

$$\begin{aligned}\frac{1}{4}H_\alpha = &MI + \mathbf{1}\mathbf{1}^T - 2M\,v(\alpha)\left(A_\alpha^{-1}\mathbf{1}\mathbf{1}^T + \mathbf{1}\mathbf{1}^TA_\alpha^{-1}\right) \\ &+ \left(3M^2\,v(\alpha) - c - 2Ma\right)v(\alpha)A_\alpha^{-1}\mathbf{1}\mathbf{1}^TA_\alpha^{-1} \\ &- \left(M^2\,v(\alpha) - c - 2Ma\right)A_\alpha^{-1},\end{aligned} \tag{10}$$

where I is the identity matrix, $c = \mathbf{1}^T\widetilde{C}\mathbf{1}$, $a := \mathbf{1}^T\alpha$.

The proof of Lemma 2 is given in Section V.

Next, we show the convexity of $\|C_\alpha\|_F^2$. This is needed for the validity of optimization algorithms such as trust-region or gradient descent.

Theorem 1. *Suppose the empirical covariance matrix $\widetilde{C}$ is of full rank, and the quantities C_α and $v(\alpha)$ are as in (8) and (7). Then, the Frobenius norm squared $\|C_\alpha\|_F^2$ is convex, i.e.,*

$$\triangle\|C_\alpha\|_F^2 \geqslant 0 \quad \forall\alpha. \tag{11}$$

The proof of Theorem 1 is given in Section V.

II. C-SHIFT algorithm and experiments

In this section we provide the covariance shift (C-SHIFT) algorithm and evaluate its performance on synthetic datasets. Moreover, we compare the C-SHIFT algorithm with the well-known and frequently used normalization methods: Quantile, Rank, LOESS, and Median absolute deviation (MAD). Our empirical results demonstrate that the C-SHIFT algorithm outperforms other methods.

A. C-SHIFT algorithm

The pseudocode for the C-SHIFT algorithm is given in Algorithm 1. Note that the algorithms takes into account both cases: when $\widetilde{C}$ has full rank and when $\widetilde{C}$ is rank deficient (i.e., $\widetilde{C}$ is positive semi-definite but not positive definite). When $\widetilde{C}$ is rank deficient the rank of $\widetilde{C} + \alpha\mathbf{1}^T + \mathbf{1}\alpha^T$ may exceed the rank $\widetilde{C}$ by no more than 2, and therefore may also be rank deficient. Therefore, to make $\widetilde{C}$ a full rank we add to it a diagonal matrix $\mathsf{diag}(f)$, where f is a vector of i.i.d. random variables from $\mathsf{Unif}[0, 1]$.

To find the optimal $\alpha^* = \arg\min_\alpha \|C_\alpha\|_F^2$, we use gradient and Hessian, provided in equations (9) and (10), in the trust-region algorithm to minimize $\|C_\alpha\|_F^2$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2022.3151840, IEEE/ACM Transactions on Computational Biology and Bioinformatics

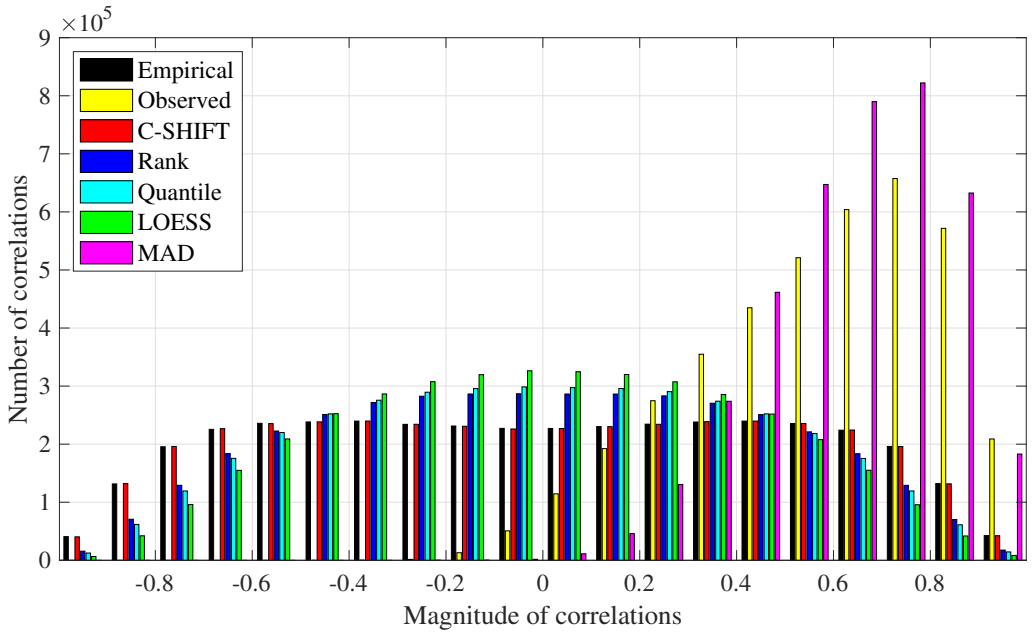

Fig. 2. A bar graph of correlations for the RCM dataset. On the x-axis we display the range of correlations, partitioned into intervals of length 0.1. The height of each bar describes the number of correlations that belong to the corresponding interval. Bars of different colors correspond to different correlation matrices, indicated in the legend.

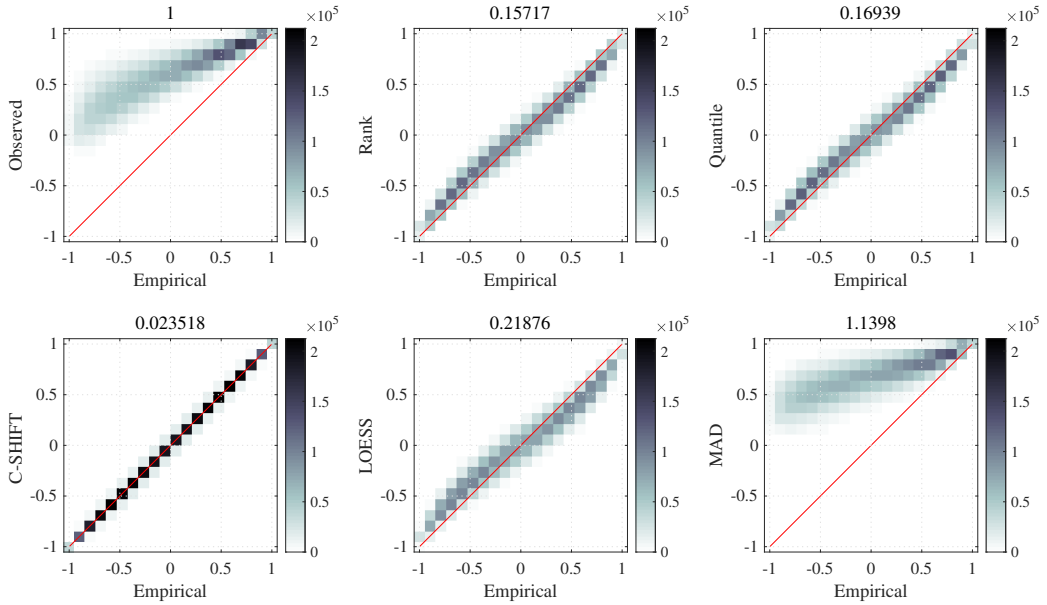

Fig. 3. The heat maps for the RCM dataset. Each heat map illustrates the transformation of the true empirical correlations $corr(Y_n, Y_m)$ (horizontal axis) after adding bias and applying the corresponding normalization method. In the top left plot the vertical axis represents the observed correlations $corr(\widetilde{Y}_n, \widetilde{Y}_m)$. In the remaining five heat maps, the vertical coordinates represent the correlations after normalization. Going clockwise, these five heat maps are Rank, Quantile, MAD, LOESS, and C-SHIFT. The darker the color, the higher the density. The number on top of each heat map indicates the relative leftover error after normalization. Smaller numbers indicate better recovery performance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2022.3151840, IEEE/ACM Transactions on Computational Biology and Bioinformatics

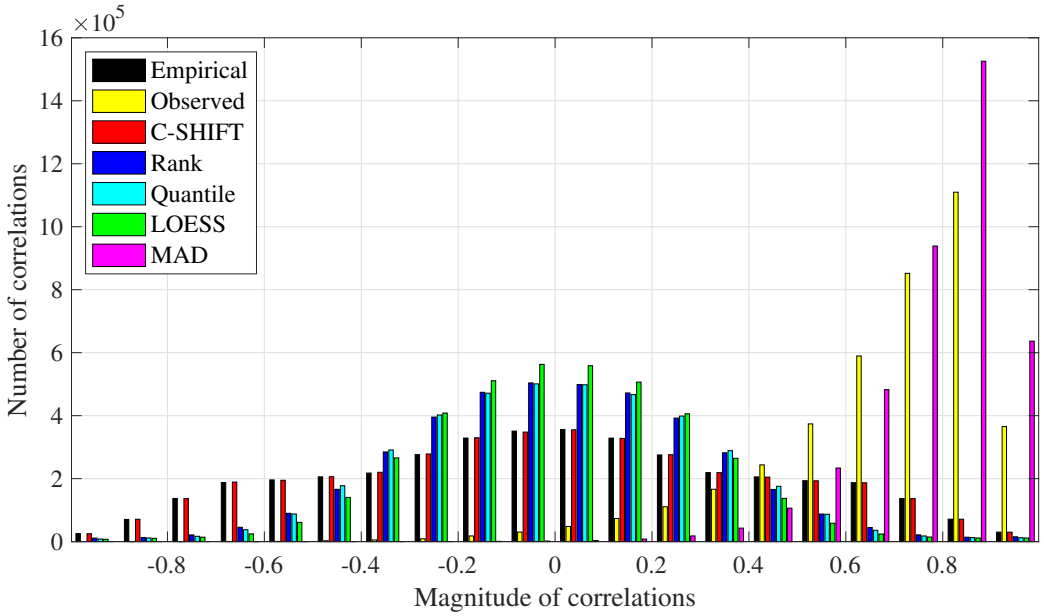

Fig. 4. A bar graph of correlations for the Cascade dataset. On the x-axis we display the range of correlations, partitioned into intervals of length 0.1. The height of each bar describes the number of correlations that belong to the corresponding interval. Bars of different colors correspond to different correlation matrices, indicated in the legend.

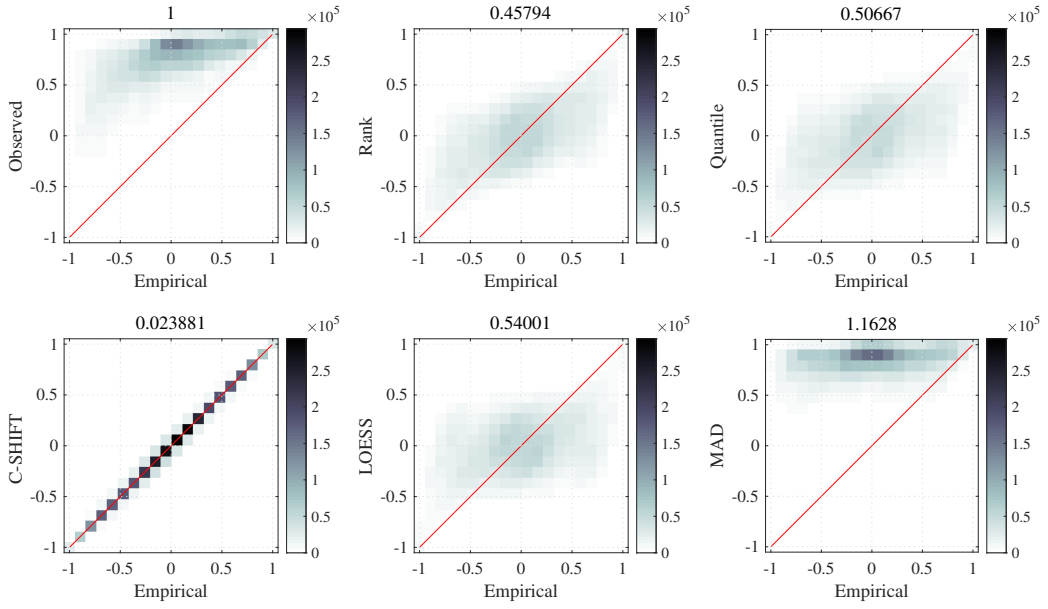

Fig. 5. The heat maps for the Cascade dataset. Each heat map illustrates the transformation of the true empirical correlations $corr(Y_n, Y_m)$ (horizontal axis) after adding bias and applying the corresponding normalization method. In the top left plot the vertical axis represents the observed correlations $corr(\widetilde{Y}_n, \widetilde{Y}_m)$. In the remaining five heat maps, the vertical coordinates represent the correlations after normalization. Going clockwise, these five heat maps are Rank, Quantile, MAD, LOESS, and C-SHIFT. The darker the color, the higher the density. The number on top of each heat map indicates the relative leftover error after normalization. Smaller numbers indicate better recovery performance.

Input: observed covariance matrix $\widetilde{C}$
Output: recovered empirical covariance matrix C
if $\widetilde{C}$ is rank deficient **then**
  $f \leftarrow$ i.i.d. $\mathsf{Unif}[0,1]$
  $\widetilde{C} \leftarrow \widetilde{C} + \mathsf{diag}(f)$
end if
$v(\alpha) \leftarrow \left[\mathbf{1}^T \left(\widetilde{C} + \alpha\mathbf{1}^T + \mathbf{1}\alpha^T\right)^{-1} \mathbf{1}\right]^{-1}$
$C_\alpha \leftarrow \widetilde{C} + \alpha\mathbf{1}^T + \mathbf{1}\alpha^T - v(\alpha)\mathbf{1}\mathbf{1}^T$
$\alpha^* \leftarrow \arg\min_\alpha \|C_\alpha\|_F^2$
$C \leftarrow C_{\alpha^*}$
if $\widetilde{C}$ is rank deficient **then**
  $C \leftarrow C - \mathsf{diag}(f)$
end if
return C

Algorithm 1: C-SHIFT

B. Numerical experiments

In this section we conduct experiments on two synthetic datasets that we generate using random covariance method (RCM) and cascade method. We start by describing both methods.

1) Data generation:

a) Random covariance method (RCM): We generate a synthetic dataset with $M = 2000$ genes and $N = 50$ measurements (samples) using RCM. For that we first generate an auxiliary matrix $H \in \mathbb{R}^{M\times m}$ ($m = 2$) whose entries are independent random variables, uniformly distributed over the interval $I = [-10, 10]$. Next, we sample a diagonal matrix $D \in \mathbb{R}^{M\times M}$ with diagonal entries being i.i.d. exponential random variables with parameter $\lambda_D = 30$. We let $\Sigma = HH^T + D$ be the population (parameter) covariance matrix. Then we generate the true empirical logarithmic data $Y^{(i)} = \left(Y_n^{(i)}\right) \sim \mathcal{N}\big(0, \Sigma\big)$ for each $i = 1, \ldots, N$. Finally, we set the observed logarithmic data be $\widetilde{Y}_n^{(i)} = Y_n^{(i)} + V^{(i)}$, where vector $V^{(i)}$ are $\mathcal{N}\big(-0.01, 100\big)$ random variables.

b) Cascade method: The *cascade* datasets were generated according by a directed acyclic weighted network $G = (V, E)$ aka directed acyclic graph (DAG). The graph was randomly generated via a recurrent *cascade* model. The parent-offspring relation is represented by the direction of edges $E = \{(u, v)\}$ of the graph G, i.e., u is the parent vertex and v is its offspring. For any vertex v let $pa(v)$ be the set of its parents, $pa(v) = \{u \in V : (u, v) \in E\}$. Next, for each edge $(u, v) \in E$ an independent random weight c_{uv} is assigned, with c.d.f.

$$p\, U_{[a_-,b_-]}(x) + (1-p)\, U_{[a_+,b_+]}(x),$$

where the parameters $a_- < b_- \leqslant 0,\ 0 \leqslant a_+ < b_+$, and $p \in (0, 1)$ are fixed, and $U_A(x)$ denotes the uniform c.d.f. on an interval A. We generated a random weighted DAG with the nodes $v \in V$ representing the genes. The random variables $\{Y_v\}_{v\in V}$ representing the logarithmic gene expressions are generated as a noisy multiplicative cascade via the following structural linear recursive equations:

$$Y_v = \sum_{u\in pa(v)} c_{uv} Y_u + \varepsilon_v,$$

where the recursion begins with $Y_0 = y_0$, and proceeds from generation to generation. The noise variables $(\varepsilon_v, v \in V)$ are i.i.d. $\mathcal{N}\big(0, \sigma^2\big)$, sampled independently from the random weights c_{uv}. For simulation of $(Y_v, v \in V)$ the following values of parameters were chosen:

p	$[a_-, b_-]$	$[a_+, b_+]$	σ^2	y_0	$\lvert V\rvert$
1/3	$[-1.2, -0.5]$	$[0.5, 1.3]$	1	4.5	2000

2) Simulation results: We generate two datasets (RCM and Cascade) using the methods described in section II-B1. Each date set consists of a matrix with the empirical data $\left(Y_n^{(i)}\right) \in \mathbb{R}^{M\times N}$ and a matrix with the observed data $\left(\widetilde{Y}_n^{(i)}\right) \in \mathbb{R}^{M\times N}$. In both, RCM and Cascade datasets, we let $M = 2000$ genes and $N = 50$ measurements (samples). For each dataset, we normalize the covariance matrix $\widetilde{C}$, obtained from the observed data, by using C-SHIFT, Rank, Quantile, LOESS, and MAD methods. We compare the performance of the algorithms using the results presented in Figures 2-5.

In Figures 2 and 4 we depict the bar graphs of correlations for RCM and Cascade datasets, respectively. As we can see in both datasets, the correlations of the observed data (yellow) are shifted away from -1 so that there are no large magnitude negative correlations. The aim of the normalization algorithms is to shift the correlations back into correct positions, i.e., ideally, the correlations of the normalized data should match the empirical correlations. Note that for both datasets, the C-SHIFT method correctly recovers the number of correlations in each interval: the red bars almost perfectly match the black bars. In contrast, other normalization methods could not recover the correct numbers of correlations, especially for the correlations of larger magnitudes. Specifically, Rank, Quantile and LOESS normalization techniques tend to shift correlations mostly to the center of the bar plot, each forming a bell shape. Predictably, the MAD method has the worst performance in correlation recovery. Finally, among the other three normalization techniques (Quantile,

Rank, and LOESS), the latter method has the poorest performance.

Figures 3 and 5 contain six heat maps each, for RCM and Cascade datasets, respectively. Each heat map illustrates the transformation of the true empirical correlations $corr(Y_n, Y_m)$ (horizontal axis) after adding bias and applying the corresponding normalization method. We consider 2,001,000 correlations corresponding to all pairs of genes. For each point, representing a pair of genes (n, m), the horizontal coordinate equals the true empirical correlation $corr(Y_n, Y_m)$ in all six plots. The vertical coordinate in the top left heat map is the correlation in the observed data, $corr(\widetilde{Y}_n, \widetilde{Y}_m)$. Importantly, it shows the shift of correlations rightward in the observed data. In the remaining five heat maps, the vertical coordinates represent the correlations after normalization. Going clockwise, these five heat maps are Rank, Quantile, MAD, LOESS, and C-SHIFT. The darker the color, the higher the density. Notice that the heat map for C-SHIFT is almost perfectly diagonal, which demonstrates how well C-SHIFT recovers the correlations. Thus, in addition to correctly recovering the right numbers of correlations in each interval (which was demonstrated in Figures 2 and 4), the proposed C-SHIFT algorithm also returns (shifts back) the correlations to the correct margins. Hence, the heat map is a diagonal line. The number on top of each heat map indicates the relative leftover error after normalization, i.e., the ℓ^2-norm of the vector of differences between the horizontal and vertical coordinates, scaled by the Frobenius norm of the difference between the empirical and the observed correlation matrices. Thus, the left top heat map is assigned the value 1, and for each normalization method, the smaller the number the better it recovers the empirical correlation matrix. Any such number smaller than one is an improvement. The number for C-SHIFT is by far the smallest in each dataset (0.023518 and 0.023881), while in the case of MAD normalization, the corresponding number even exceeds 1.

III. Evaluation of C-SHIFT algorithm on real data

In this section we apply C-SHIFT algorithm to real biological data, and compare the resulting correlations to the correlations obtained by normalizing the same data with Rank, Quantile, and LOESS. In the analysis, we used scaled ℓ^1-norm to measure the distance between correlation matrices. Specifically, for two correlation matrices, $R = \big(r_{i,j}\big)$ and $R' = \big(r'_{i,j}\big)$, the norm

$$d(R, R') = \frac{1}{M(M-1)} \sum_{1 \leqslant i < j \leqslant M} |r_{i,j} - r'_{i,j}| \qquad (12)$$

measures the distance between R and R' on the scale from 0 to 1. We considered the following microarray datasets from GEO depository.

Two datasets come from GSE7803 in GEO depository [34]. Dataset **GSE7803[Carcinoma]** looks into 21 samples of invasive squamous cell carcinomas and 4,152 genes. Dataset **GSE7803[Normal]** has 10 normal cervical samples and 4,709 genes.

Dataset **GSE152738** from GEO depository [20] consists of 58 liver specimens from adult liver donors and 12,164 genes.

Dataset **GSE86858** obtained in [17] has 15,312 genes and 8 samples from obese diabetic mice treated with γ-oryzanol-encapsulated nanoparticles, of which, 4 were taken from liver and 4 from hypothalamus.

Two datasets come from GSE59412 [33], where it was discovered that the ectopic expression of miR-K12-11 differentially affected gene expression in BJAB cells of lymphoid origin and TIVE cells of endothelial origin. Dataset **GSE59412[TIVE]** consists of 8 samples of TIVE cells and 16,700 genes. Dataset **GSE59412[BJAB]** consists of 24 samples of BJAB cells and 19,296 genes.

All six datasets considered were not normalized prior to the analysis. Affy R package and MAS-5 method [2, 11, 15, 22] was used for reading and preliminary data analysis at the probe-level of affymetrix CEL files. We calculated Abesnt/Present Call for each probe set and subselected only the genes that are expressed in all samples.

Next, we summarize our observations. First, we notice that in all real and synthetic datasets considered in this analysis, the correlations produced by Rank, Quantile, and LOESS are close to each other. In the synthetic data, where the desired true empirical correlations are known, one easily encounters a situation where under a strong bias the correlations produced by C-SHIFT are significantly different from the correlations produced by Rank, Quantile, or LOESS. See Fig. 5.

Recall that $d(R, R')$ defined in (12) measures the distance between correlation matrices on the scale from 0 to 1. In the six real datasets considered in this work, we notice that the distances between the correlations obtained from C-SHIFT and either one of

Dataset	Rank vs. C-SHIFT	Quantile vs. C-SHIFT	LOESS vs. C-SHIFT
GSE7803[Carcinoma]	0.02824	0.017624	0.017557
GSE7803[Normal]	0.03425	0.023438	0.025578
GSE152738	0.019763	0.014113	0.015207
GSE86858	0.096963	0.095617	0.10688
GSE59412[TIVE]	0.046627	0.041095	0.043859
GSE59412[BJAB]	0.041824	0.038841	0.04086

TABLE I
DISTANCE (12) BETWEEN PAIRS OF CORRELATION MATRICES RECOVERED BY NORMALIZATION METHODS IN THE SIX DATASETS.

the three normalization methods used in comparison (Rank, Quantile, and LOESS) range between 0.01 and 0.1. See Table I and Figures 6 and 7. In five out of six datasets, the distance between C-SHIFT and any of the three normalization approaches does not exceed 0.05. A small but sizable mismatch of ≈ 0.1 between C-SHIFT and each of the three normalization methods is observed in dataset GSE86858.

IV. Discussion

In systems biology, the gene co-expression networks (GCN) are reconstructed from the correlations between the genes. GCN recovery relies on removing the bias with a normalization method, and thus improving the estimation of correlations between the pairs of genes. However, the standard normalization techniques such as Rank, Quantile, LOESS, and MAD are known to be insufficient at recovering true empirical correlations while the C-SHIFT algorithm is specifically designed to recover the true empirical correlations. The multiple experiments with synthetic datasets demonstrate the algorithm's superior performance at recovering true empirical correlations in comparison to the standard normalization techniques.

Working with the synthetic data, we noticed that the C-SHIFT algorithm's precision at removing the bias and recovering true empirical correlations improves as the number of genes M is increased. Additionally, C-SHIFT was observed to outperform the standard normalization methods when the variance of the bias $\omega = Var(V)$ was taken sufficiently large. Therefore, a typical situation when C-SHIFT outperforms Rank, Quantile, and LOESS at recovering true empirical correlations would be when either or both of the following scenerios holds: (i) the number of genes M is large with a relatively small number of measurements N; (ii) large value of the variance of the bias $\omega = Var(V)$. In all other observed situations, C-SHIFT, Rank, Quantile, and LOESS would show similar efficiency while MAD would perform significantly worth. For instance, in both simulated examples (RCM and Cascade) considered in Sect. II-B, we set $M = 2000$, $N = 50$, and $\omega = Var(V) = 100$.

Also, we observed that the correlations recovered by C-SHIFT, Rank, Quantile, and LOESS would essentially match in five out of six real datasets considered in this paper. One dataset (GSE86858) demonstrated small but sizable difference. In the case of GSE86858 dataset, $M = 15,312$ and $N = 8$. Moreover, as 4 of the samples were taken from liver and 4 from hypothalamus, this would suggest a greater variance of the bias $\omega = Var(V)$. This case appears to match both of the above described scenarios, (i) and (ii), that as observed in simulated data would cause C-SHIFT to outperform the standard normalization techniques at recovering true empirical correlations.

In perspective, C-SHIFT algorithm could benefit GCN reconstruction studies that consider much larger sample numbers N by combining together multiple publicly available datasets (e.g. see Feltus et al. [10] for Arabidopsis case study) since the combinations of datasets are likely to increase the variance of the bias $\omega = Var(V)$ as in scenario (ii) above, shown to favor C-SHIFT.

Importantly, we notice that the C-SHIFT algorithm corrects the positive shift of covariances (and correlations) observed when $\hat{\omega} = \widehat{Var}(V)$ is larger than $\hat{a}_n = -\widehat{Cov}(Y_n, V)$ $(n = 1, \ldots, M)$ in (5). Hence, the independence of V from Y_n assumption can be replaced with a weaker assumption stating that $Cov(Y_n, V) \ll Var(V)$. This will be explored in a follow-up publication.

An alternative version of the C-SHIFT algorithm is based on trace minimization approach instead of energy minimization. In this alternative C-SHIFT algorithm, the positive semi-definite matrix $C_{\alpha *}$ with

$$\alpha^* = \operatorname{argmin} \operatorname{Tr}\big(C_\alpha\big)$$

is used to approximate the true empirical covariance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2022.3151840, IEEE/ACM Transactions on Computational Biology and Bioinformatics

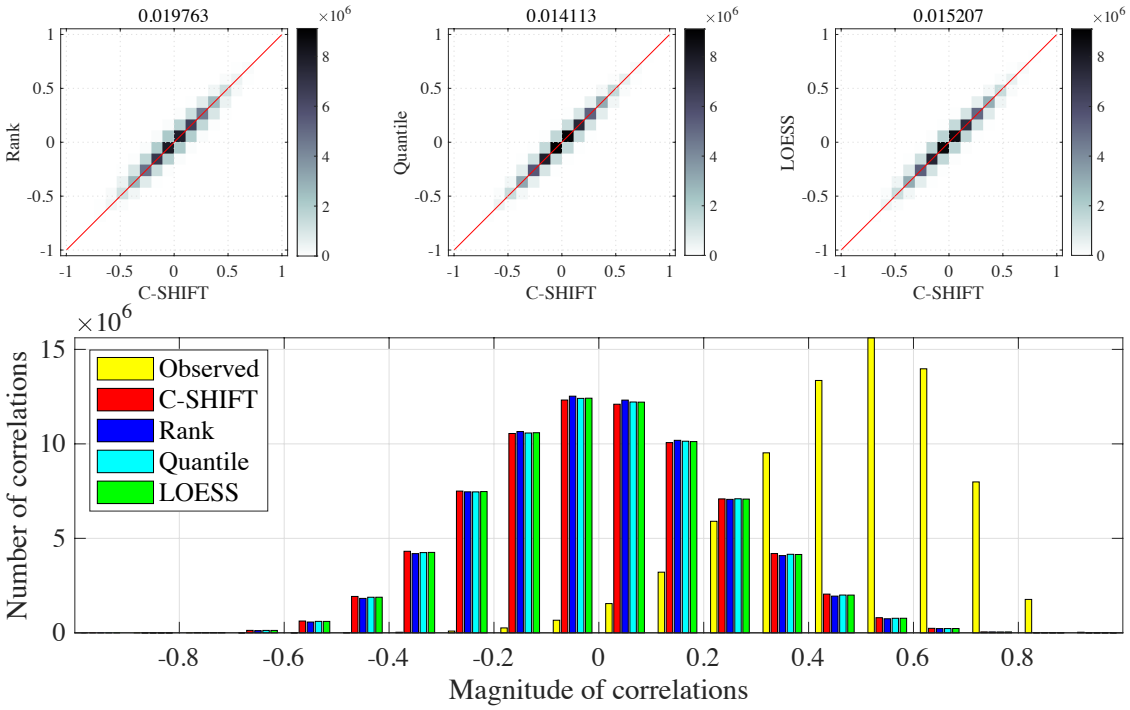

Fig. 6. Top row: heat maps for **GSE152738** dataset that compare the correlations obtained by Rank, Quantile, and LOESS to C-SHIFT. The number on top of each heat map represents the distance between correlation matrices as defined in (12). Bottom row: numbers of correlations for **GSE152738** dataset. Different colors correspond to different correlation matrices, indicated in the legend. Horizontal axis is partitioned into intervals of length 0.1.

matrix C. The analogs of Lemmas 1 and 2 and the convexity result in Theorem 1 are also established for $\mathrm{Tr}(C_\alpha)$ in the trace minimization approach. See [19]. Empirically it appears that this alternative approach produces the same α^* as the original C-SHIFT algorithm based on energy minimization as presented in this paper, and therefore it recovers the empirical covariance C with the same accuracy. Thus, the alternative, trace minimizing C-SHIFT algorithm can be used instead of Algorithm 1. This approach will be analyzed in a follow-up paper.

Finally, the C-SHIFT algorithm was deposited on GitHub at https://github.com/evcphd/C-SHIFT

V. Proofs

Proof of Proposition 1. Observe that

$$x^T\big(\mathcal{M} - v\,\mathbf{1}\mathbf{1}^T\big)x = x^T\mathcal{M}x - v\left(\sum x_i\right)^2 \geqslant 0$$

for all $x \in \mathbb{R}^M$ if and only if $v \leqslant v^*$, where v^* minimizes $x^T\mathcal{M}x$ under the condition $\sum x_i = Const.$ Next, applying the Lagrange multipliers method, we obtain $2\mathcal{M}x = \lambda\mathbf{1}$, and therefore,

$$v^* = \frac{x^T\mathcal{M}x}{(\sum x_i)^2} = \frac{\frac{\lambda}{2}x^T\mathbf{1}}{(\sum x_i)^2} = \frac{\lambda/2}{\mathbf{1}^T x} = \frac{1}{\mathbf{1}^T\,\mathcal{M}^{-1}\,\mathbf{1}}$$

as $x = \frac{\lambda}{2}\mathcal{M}^{-1}\mathbf{1}$. □

Proof of Lemma 1. By (8), we have

$$\|C_\alpha\|_F^2 = \|\widetilde{C}\|_F^2 + 2M\sum_{i=1}^{M}\alpha_i^2 + M^2\,v^2(\alpha) + 4\left(\mathbf{1}^T\widetilde{C}\alpha\right) + 2a^2 - 2c\,v(\alpha) - 4Ma\,v(\alpha) \tag{13}$$

Notice that

$$\frac{\partial}{\partial\alpha_i}A_\alpha = \bar{e}_i\mathbf{1}^T + \mathbf{1}\bar{e}_i^T \quad \text{and}$$
$$\frac{\partial}{\partial\alpha_i}A_\alpha^{-1} = -A_\alpha^{-1}\left(\bar{e}_i\mathbf{1}^T + \mathbf{1}\bar{e}_i^T\right)A_\alpha^{-1}, \tag{14}$$

where $\bar{e}_i$ is the i-th coordinate vector. Therefore, we have

$$\frac{\partial}{\partial\alpha_i}v(\alpha) = v^2(\alpha)\mathbf{1}^T A_\alpha^{-1}\left(\bar{e}_i\mathbf{1}^T + \mathbf{1}\bar{e}_i^T\right)A_\alpha^{-1}\,\mathbf{1}$$
$$= 2v(\alpha)\,\mathbf{1}^T A_\alpha^{-1}\bar{e}_i \tag{15}$$

implying

$$\nabla v(\alpha) = 2v(\alpha)A_\alpha^{-1}\mathbf{1}. \tag{16}$$

Next, the gradient $\nabla\|C_\alpha\|_F^2$ in (9) is found via the equations (13) and (16). □

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2022.3151840, IEEE/ACM Transactions on Computational Biology and Bioinformatics

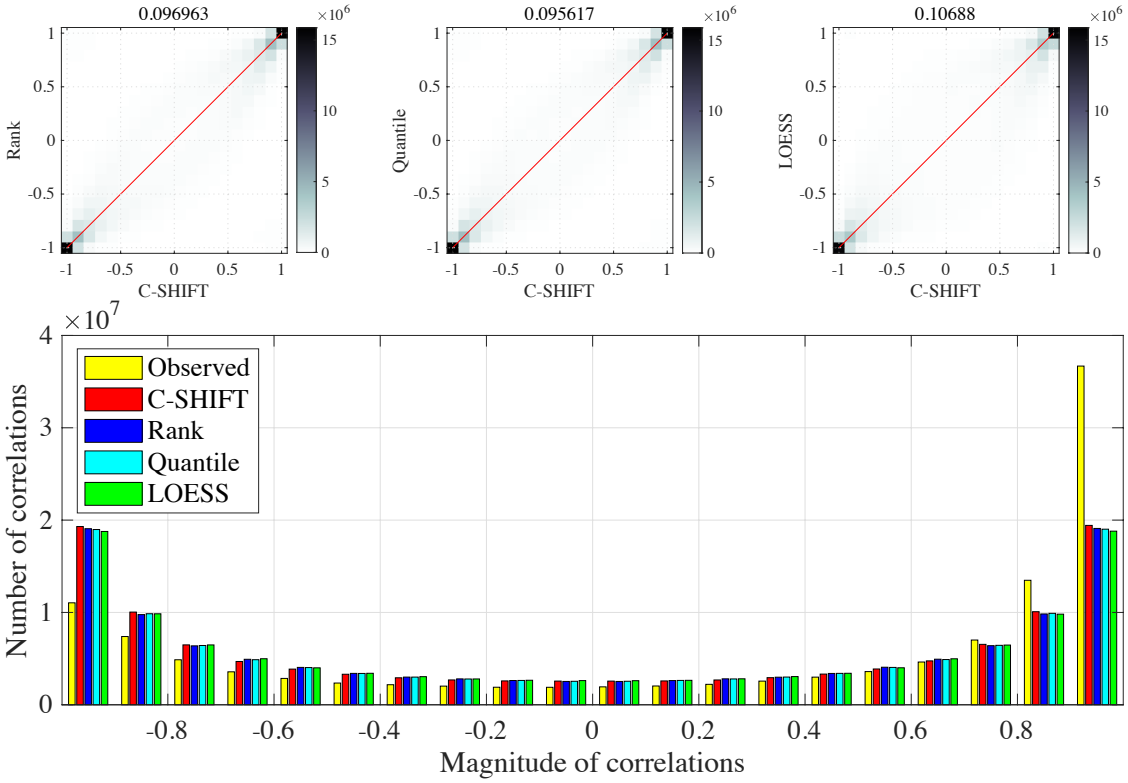

Fig. 7. Top row: heat maps for **GSE86858** dataset that compare the correlations obtained by Rank, Quantile, and LOESS to C-SHIFT. The number on top of each heat map represents the distance between correlation matrices as defined in (12). Bottom row: numbers of correlations for **GSE86858** dataset. Different colors correspond to different correlation matrices, indicated in the legend. Horizontal axis is partitioned into intervals of length 0.1.

Proof of Lemma 2. By (9), we have

$$\begin{aligned}\frac{1}{4}H_\alpha &= \frac{1}{4}\nabla\left(\nabla\|C_\alpha\|_F^2\right)^T \\ &= M\nabla\alpha^T + \nabla\mathbf{1}^T\big(a - M\,v(\alpha)\big) \\ &+ \Big(\nabla\big(M^2\,v^2(\alpha) - c\,v(\alpha) - 2Ma\,v(\alpha)\big)\Big)\mathbf{1}^T A_\alpha^{-1} \\ &+ \big(M^2\,v^2(\alpha) - c\,v(\alpha) - 2Ma\,v(\alpha)\big)\nabla\mathbf{1}^T A_\alpha^{-1},\end{aligned} \tag{17}$$

where $\nabla = \left(\frac{\partial}{\partial\alpha_1},\ldots,\frac{\partial}{\partial\alpha_M}\right)^T$ was used as the column vector of the partial derivative operators. The summation parts in (17) are calculated as follows. First,

$$M\nabla\alpha^T = MI. \tag{18}$$

Next, (16) implies

$$\begin{aligned}&\nabla\big(M^2\,v^2(\alpha) - c\,v(\alpha) - 2Ma\,v(\alpha)\big) \\ &= 2\big(2M^2\,v(\alpha) - c - 2Ma\big)v(\alpha)A_\alpha^{-1}\mathbf{1} - 2Mv(\alpha)\mathbf{1}.\end{aligned} \tag{19}$$

Equation (14) yields

$$\begin{aligned}\nabla\mathbf{1}^T A_\alpha^{-1} &= \sum_{i=1}^{M}\bar{e}_i\mathbf{1}^T\frac{\partial}{\partial\alpha_i}A_\alpha^{-1} \\ &= -A_\alpha^{-1}\mathbf{1}\mathbf{1}^T A_\alpha^{-1} - \frac{1}{v(\alpha)}A_\alpha^{-1}.\end{aligned} \tag{20}$$

Finally, (16) is used to derive

$$\nabla\mathbf{1}^T\big(a - M\,v(\alpha)\big) = \mathbf{1}\mathbf{1}^T - 2M\,v(\alpha)A_\alpha^{-1}\mathbf{1}\mathbf{1}^T. \tag{21}$$

Combining together equations (18)-(21) and substituting them into (17) we obtain (10). □

Proof of Theorem 1. We will use the notations from Lemmas 1 and 2 such as $c := \mathbf{1}^T\widetilde{C}\mathbf{1}$ and $a := \sum_{i=1}^{M}\alpha_i$. Without loss of generality we consider α on the hyperplane $a = 0$.

Here, $A_\alpha = \widetilde{C} + \alpha\mathbf{1}^T + \mathbf{1}\alpha^T$ is a positive definite symmetric matrix with eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_M > 0$ counted with respect to algebraic multiplicity, and let $\{v_i\}_{i=1,\ldots,M}$ be the corresponding orthonormal basis of eigenvectors.

As $\triangle\|C_\alpha\|_F^2 = \mathrm{Tr}\big(H_\alpha\big)$, equation (10) implies

$$\begin{aligned}\frac{1}{4}\triangle\|C_\alpha\|_F^2 &= M^2\Big(1 - v(\alpha)\mathrm{Tr}\big(A_\alpha^{-1}\big)\Big)\\ &\quad + c\Big(\mathrm{Tr}\big(A_\alpha^{-1}\big) - v(\alpha)\mathbf{1}^T A_\alpha^{-2}\mathbf{1}\Big)\\ &\quad + 3M\Big(Mv^2(\alpha)\mathbf{1}^T A_\alpha^{-2}\mathbf{1} - 1\Big). \qquad (22)\end{aligned}$$

The Laplacian in (22) is shown to be strictly positive in the following three steps. First, by the Cauchy-Bunyakovsky-Schwarz inequality, we have

$$\begin{aligned}&Mv^2(\alpha)\mathbf{1}^T A_\alpha^{-2}\mathbf{1} - 1\\ &\quad = v^2(\alpha)\Big(\|\mathbf{1}\|_2^2\|A_\alpha^{-1}\mathbf{1}\|_2^2 - \big(\mathbf{1}^T A_\alpha^{-1}\mathbf{1}\big)^2\Big) \geqslant 0. \qquad (23)\end{aligned}$$

Next, observe that $Mx + (1-x)^2 \geqslant 1$ for $M \geqslant 2$, and all $x \in [0,1]$. Thus, for a given probability mass function $\{p_k\}_{k=1,\ldots,M}$ such that $p_k < 1$ for all k, and a given index $i \in \{1,\ldots,M\}$, Jensen's inequality implies

$$\begin{aligned}&Mp_i + \left(\sum_{j:j\neq i}\lambda_j^{-1}p_j\right)\left(\sum_{j:j\neq i}\lambda_j p_j\right)\\ &\quad = Mp_i + (1-p_i)^2\left(\sum_{j:j\neq i}\lambda_j^{-1}q_j\right)\left(\sum_{j:j\neq i}\lambda_j q_j\right)\\ &\quad \geqslant Mp_i + (1-p_i)^2 \ \geqslant 1 \qquad (24)\end{aligned}$$

where we let $q_j = \frac{p_j}{1-p_i}$ for all $j \neq i$. Summing over all i in (24), we obtain,

$$\begin{aligned}&\sum_i \lambda_i^{-1}p_i + \frac{1}{M}\sum_i \lambda_i^{-1}\left(\sum_{j:j\neq i}\lambda_j^{-1}p_j\right)\left(\sum_{j:j\neq i}\lambda_j p_j\right)\\ &\qquad \geqslant \frac{1}{M}\sum_i \lambda_i^{-1}. \qquad (25)\end{aligned}$$

Eqn. (25) implies

$$\begin{aligned}&\sum_i \lambda_i^{-1}p_i + \frac{1}{M}\sum_i \lambda_i^{-1}p_i\left(\sum_{j:j\neq i}\lambda_j^{-1}\right)\left(\sum_k \lambda_k p_k\right)\\ &\qquad \geqslant \frac{1}{M}\sum_i \lambda_i^{-1}. \qquad (26)\end{aligned}$$

which rewrites as

$$\begin{aligned}&\sum_i \lambda_i^{-1}p_i + \frac{1}{M}\left(\sum_i \lambda_i^{-1}p_i\right)\left(\sum_j \lambda_j^{-1}\right)\left(\sum_k \lambda_k p_k\right)\\ &\quad \geqslant \frac{1}{M}\left(\sum_i \lambda_i^{-2}p_i\right)\left(\sum_k \lambda_k p_k\right) + \frac{1}{M}\sum_i \lambda_i^{-1}. \qquad (27)\end{aligned}$$

Finally, we let $p_i = \frac{1}{M}\big(\mathbf{1}^T v_i\big)^2$ and substitute the following expressions into (27):

$$\sum_i \lambda_i p_i = \frac{1}{M}\mathbf{1}^T A_\alpha \mathbf{1} = \frac{1}{M}\mathbf{1}^T\widetilde{C}\mathbf{1} = \frac{c}{M} \quad \text{as} \quad a = 0,$$

$$\sum_i \lambda_i^{-1}p_i = \frac{1}{M}\mathbf{1}^T A_\alpha^{-1}\mathbf{1} = \frac{1}{M\,v(\alpha)},$$

$$\sum_i \lambda_i^{-1} = \mathrm{Tr}\big(A_\alpha^{-1}\big), \quad \text{and} \quad \sum_i \lambda_i^{-2}p_i = \frac{1}{M}\mathbf{1}^T A_\alpha^{-2}\mathbf{1}.$$

Consequently, (27) rewrites as

$$\begin{aligned}&M^2\Big(1 - v(\alpha)\mathrm{Tr}\big(A_\alpha^{-1}\big)\Big)\\ &\quad + c\Big(\mathrm{Tr}\big(A_\alpha^{-1}\big) - v(\alpha)\mathbf{1}^T A_\alpha^{-2}\mathbf{1}\Big) \ \geqslant 0. \qquad (28)\end{aligned}$$

Substituting (23) and (28) into (22), we then obtain $\triangle\|C_\alpha\|_F^2 \geqslant 0$. □

Acknowledgments

We would like to thank the anonymous referees for encouraging remarks, valuable feedback, and suggesting ways to improve the paper. This research was supported by the FAPESP awards 2017/10555-0, 2018/14952-7 and 2018/07826-5, and by the NSF award DMS-1412557.

References

[1] Dhammika Amaratunga and Javier Cabrera. Analysis of data from viral dna microchips. *Journal of the American Statistical Association*, 96(456):1161–1170, 2001.

[2] Jose M Arteaga-Salas, Harry Zuzan, William B Langdon, Graham JG Upton, and Andrew P Harrison. An overview of image-processing methods for affymetrix genechips. *Briefings in Bioinformatics*, 9(1):25–33, 2008.

[3] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[4] Martin Bilban, Lukas K Buehler, Steven Head, Gernot Desoye, and Vito Quaranta. Normalizing dna microarray data. *Current Issues in Molecular Biology*, 4:57–64, 2002.

[5] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[6] Lixin Cheng, Leung-Yau Lo, Nelson LS Tang, Dong Wang, and Kwong-Sak Leung. Crossnorm: a novel normalization strategy for microarray data in cancers. *Scientific reports*, 6(1):1–11, 2016.

[7] Lixin Cheng, Xuan Wang, Pak-Kan Wong, Kwan-Yeung Lee, Le Li, Bin Xu, Dong Wang, and Kwong-Sak Leung. Icn: a normalization

method for gene expression data considering the over-expression of informative genes. *Molecular BioSystems*, 12(10):3057–3066, 2016.

[8] David L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS conference on Mathematical Challenges of the 21st Century*. Citeseer, 2000.

[9] Jianqing Fan, Heng Peng, and Tao Huang. Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. *Journal of the American Statistical Association*, 100(471):781–796, 2005.

[10] F Alex Feltus, Stephen P Ficklin, Scott M Gibson, and Melissa C Smith. Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an arabidopsis case study. *BMC systems biology*, 7(1):1–12, 2013.

[11] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[12] Alexander N Gorban and Ivan Yu Tyukin. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2118):20170237, 2018.

[13] Alexander J Hartemink, David K Gifford, Tommi S Jaakkola, and Richard A Young. Maximum-likelihood estimation of optimal scaling factors for expression array normalization. In *Microarrays: Optical Technologies and Informatics*, volume 4266, pages 132–140. International Society for Optics and Photonics, 2001.

[14] Jianhua Hu and Xuming He. Enhanced quantile normalization of microarray data to reduce loss of information in gene expression profiles. *Biometrics*, 63(1):50–59, 2007.

[15] Rafael A Irizarry, Laurent Gautier, and Leslie M Cope. An r package for analyses of affymetrix oligonucleotide arrays. In *The analysis of gene expression data*, pages 102–119. Springer, 2003.

[16] Paul C Kainen. Utilizing geometric anomalies of high dimension: When complexity makes computation easier. In *Computer Intensive Methods in Control and Signal Processing*, pages 283–294. Springer, 1997.

[17] Chisayo Kozuka, Chigusa Shimizu-Okabe, Chitoshi Takayama, Kaku Nakano, Hidetaka Morinaga, Ayano Kinjo, Kotaro Fukuda, Asuka Kamei, Akihito Yasuoka, Takashi Kondo, et al. Marked augmentation of plga nanoparticle-induced metabolically beneficial impact of γ-oryzanol on fuel dyshomeostasis in genetically obese-diabetic ob/ob mice. *Drug delivery*, 24(1):558–568, 2017.

[18] Xueyan Liu, Nan Li, Sheng Liu, Jun Wang, Ning Zhang, Xubin Zheng, Kwong-Sak Leung, and Lixin Cheng. Normalization methods for the analysis of unbalanced transcriptome data: a review. *Frontiers in bioengineering and biotechnology*, 7:358, 2019.

[19] Paul Logan. *C-SHIFT, Quantile Theory, and Assessing Monotonicity*. PhD thesis, Oregon State University, 2020.

[20] Norihisa Nishimura, Davide De Battista, David R McGivern, Ronald E Engle, Ashley Tice, Rafaelle Fares-Gusmao, Juraj Kabat, Anna Pomerenke, Hanh Nguyen, Shinya Sato, et al. Chitinase 3-like 1 is a profibrogenic factor overexpressed in the aging liver and in patients with liver cirrhosis. *Proceedings of the National Academy of Sciences*, 118(17), 2021.

[21] Taesung Park, Sung-Gon Yi, Sung-Hyun Kang, SeungYeoun Lee, Yong-Sung Lee, and Richard Simon. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4(1):33, 2003.

[22] Stuart D Pepper, Emma K Saunders, Laura E Edwards, Claire L Wilson, and Crispin J Miller. The utility of mas5 expression summary and detection call algorithms. *BMC bioinformatics*, 8(1):1–12, 2007.

[23] Sylvain Pradervand, Johann Weber, Jérôme Thomas, Manuel Bueno, Pratyaksha Wirapati, Karine Lefort, G Paolo Dotto, and Keith Harshman. Impact of normalization on mirna microarray expression profiling. *RNA*, 15(3):493–501, 2009.

[24] Xing Qiu, Andrew I Brooks, Lev Klebanov, and Andrei Yakovlev. The effects of normalization on the correlation structure of microarray data. *BMC bioinformatics*, 6(1):1–11, 2005.

[25] Xing Qiu, Hulin Wu, and Rui Hu. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics*, 14(1):124, 2013.

[26] John Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32(4):496–501, 2002.

[27] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531, 2010.

[28] Youlan Rao, Yoonkyung Lee, David Jarjoura, Amy S Ruppert, Chang-gong Liu, Jason C Hsu, and John P Hagan. A comparison of normalization techniques for microrna microarray data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

[29] Cavan Reilly, Changchun Wang, and Mark Rutherford. A method for normalizing microarrays using genes that are not differentially expressed. *Journal of the American Statistical Association*, 98(464):868–878, 2003.

[30] Edoardo Saccenti. Correlation patterns in experimental data are affected by normalization procedures: consequences for data analysis and network inference. *Journal of Proteome Research*, 16(2):619–634, 2017.

[31] Andreas Scherer. *Batch effects and noise in microarray experiments: sources and solutions*. John Wiley & Sons, 2009.

[32] Gordon K Smyth and Terry Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–273, 2003.

[33] Yajie Yang, Isaac W Boss, Lauren M McIntyre, and Rolf Renne. A systems biology approach identified different regulatory networks targeted by kshv mir-k12-11 in b cells and endothelial cells. *BMC genomics*, 15(1):1–17, 2014.

[34] Yali Zhai, Rork Kuick, Bin Nan, Ichiro Ota, Stephen J Weiss, Cornelia L Trimble, Eric R Fearon, and Kathleen R Cho. Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies hoxc10 as a key mediator of invasion. *Cancer research*, 67(21):10163–10172, 2007.

Evgenia Chunikhina is an Assistant Professor of Data Science at Pacific University, Oregon, USA. She received her B.S. in Applied Mathematics, Computer Science and Mechanics from Voronezh State University, Russia. She received her M.S. in Mathematics, M.Eng. in Computer Science, and Ph.D. in Computer Science from Oregon State University in 2014, 2015, and 2018 respectively. Dr. Chunikhina completed a one-year postdoc at the University of São Paulo, São Paulo, Brazil. Her research interests include machine learning, artificial intelligence, statistical signal processing, information theory, compressed sensing, networks, random graphs, and biostatistics. Dr. Chunikhina is a member of the IEEE.

Paul R. Logan received his B.S. in Physics from the South Dakota School of Mines & Technology and his M.S. in Physics from Arizona State University. He acquired his M.S. in Statistics from Oregon State University in 2016. He earned his Ph.D. in Statistics from Oregon State University in 2020. After a 14-month internship with HP Inc., Dr. Logan permanently joined the company as a Statistician in 2020. His areas of focus include statistical tool app development, functional data analysis, multiple hypothesis testing, machine learning, and interpreting kappa studies.

Yevgeniy Kovchegov works in the field of probability and stochastic processes. His research interests include mathematics of data science, mathematical models of statistical mechanics, models of mathematical biology, stochastic self-similarity, random networks, and mathematical statistics. He graduated from Stanford University in 2002 with a Ph.D. in mathematics. From 2002 to 2005 he was VIGRE assistant professor at UCLA department of mathematics. Starting in 2005 he works at Oregon State University department of mathematics, where in 2017 he became a full professor. Dr. Kovchegov is currently an associate editor for Stochastics and Dynamics, and for Rocky Mountain Journal of Mathematics.

Anatoly Yambartsev is an associate professor at Department of Statistics, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil. He earned his Ph.D. in Mathematics and Applied Mathematics from Moscow State University, Russia in 1999. Dr. Yambartsev completed a two-year postdoc at the University of São Paulo, Brazil, and in 2005, became an assistant professor at the same university. His research interests include Markov processes, statistical mechanics, computational statistics, and biostatistics.

Debashis Mondal is an associate professor at the Department of Mathematics and Statistics, Washington University in St. Louis. From 2014 to 2021 Dr. Mondal worked at the Department of Statistics, Oregon State University. He received his Ph.D. in statistics from the University of Washington in 2007 and both his bachelor and master's degrees in statistics from the Indian Statistical Institute in Kolkata, India, in 2000 and 2002. Dr. Mondal's research interests include spatial statistics, MCMC, anomaly detection, ecology, and time series. He is a recipient of the NSF career award, the young researcher award by the International Indian Statistical Association, and is an elected member of the International Statistical Institute. Dr. Mondal is an associate editor of Journal of American Statistical Association.

Andrey Morgun is an Associate Professor at the College of Pharmacy, Oregon State University. He received his MD degree from Kharkiv National Medical University in 1995. He earned his M.S. in immunogenetics and transplantation and Ph.D. in clinical immunology, immunogenetics and transplant immunology from Federal University of São Paulo, Brazil in 1998 and 2002, respectively. Dr. Morgun completed his postdoctoral training at NIH from 2005 to 2011. Dr. Morgun's research interests include System Biology, Immunology, Microbiome, and Pharmacogenomics.