BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 17 2007, pages 2298–2305
doi:10.1093/bioinformatics/btm328

Systems biology

Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data

Dongxiao Zhu[1,*], Youjuan Li[2] and Hua Li[1]

[1]Stowers Institute for Medical Research, 1000 E 50th Street, Kansas City, MO 64110 and [2]Department of Statistics, University of Michigan, Ann Arbor, MI 48105, USA

Received on January 23, 2007; revised on May 31, 2007; accepted on June 17, 2007

Advance Access publication June 22, 2007

Associate Editor: Golan Yona

ABSTRACT

Summary: Estimating pairwise correlation from replicated genome-scale (a.k.a. OMICS) data is fundamental to cluster functionally relevant biomolecules to a cellular pathway. The popular Pearson correlation coefficient estimates bivariate correlation by averaging over replicates. It is not completely satisfactory since it introduces strong bias while reducing variance. We propose a new multivariate correlation estimator that models all replicates as independent and identically distributed (i.i.d.) samples from the multivariate normal distribution. We derive the estimator by maximizing the likelihood function. For small sample data, we provide a resampling-based statistical inference procedure, and for moderate to large sample data, we provide an asymptotic statistical inference procedure based on the Likelihood Ratio Test (LRT). We demonstrate advantages of the new multivariate correlation estimator over Pearson bivariate correlation estimator using simulations and real-world data analysis examples.

Availability: The estimator and statistical inference procedures have been implemented in an R package 'CORREP' that is available from CRAN [http://cran.r-project.org] and Bioconductor [http://www.bioconductor.org/].

Contact: doz@stowers-institute.org or dongxiaozhu@yahoo.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The ever-increasing use of high-throughput data acquisition technologies, such as gene expression arrays and mass spectrometry, has generated an enormous amount of genome-scale data, in which every gene or gene product is associated with a series of numeric measurements. Analysis of these data provide many opportunities to understand the underlying cellular processes. Some of the familiar tasks in analysis of genome-wide measurements include pre-processing (Irizarry *et al.,* 2003; Li and Wong, 2001; Zhou and Rocke, 2005), detecting differentially expressed genes (Cui and Churchill, 2003; Pavelka *et al.,* 2004; Sartor *et al.,* 2006), gene clustering (Medvedovic and Sivaganesan, 2002; Medvedovic *et al.*, 2004; Liu *et al.,* 2006; Yeung *et al.,* 2001), sample classification and biomarker discovery (Nguyen and Rocke, 2002a, Nguyen and Rocke, 2002b; Yeung and Bumgarner, 2005) and gene network reconstruction (Lee *et al.*, 2004; Zhu *et al.*, 2005a). These analyses are useful in understanding the complex web of bio-molecule interaction and regulation in the living cell.

Many types of genome-wide data are noisy, and have to be replicated to account for the inherent variability. In this work, we discuss replication in two contexts, i.e. profiling gene expression using expression arrays, and profiling protein abundance in biological samples using mass-spectrometry.

Various forms of data replication are employed in both cDNA and Affymetrix microarray experiments. In cDNA microarray experiments, the differences between replication strategies are in the extent to which the replicates can be treated as statistically independent random samples from a population. The more popular replication methods can be divided into within-slide and between-slide replications (Speed, 2003). The former is generated by printing the same cDNA on a slide more than once for quality control purposes. The latter is generated by using different mRNA aliquot for each slide hybridization that can be further divided into technical replicates and biological replicates depends on whether the mRNA aliquots are from a single mouse or strain or not. In Affymetrix experiments, each probe set consists of about 11 probe pairs. The probe-level intensities can be treated as replicated measures of the gene expression scores.

In proteomics research, quantification experiments of the proteins or the peptides present in a sample are usually performed in replicates. For example, the recently introduced iTRAQ protein quantification technique (Ross *et al.*, 2004) derivatizes peptides using multiple (e.g. four) isobaric mass tags i.e. mass-to-charge ratios (m/z) 114, 115, 116 and 117 Da, and it also imparts identical reversed-phase retention properties to differentially labeled peptides. Therefore, four separate peptide mixtures can be individually labeled with different mass tags, combined, and separated by reversed-phase high performance liquid chromatography (HPLC). The relative abundance of a protein is thus represented by a number of ratios of peptides abundance to the internal standard. These ratios can be viewed as replicated measurements of the protein abundance (Hardt *et al.,* 2005).

Many data analysis approaches were able to draw reliable inference from noisy data by taking good advantage

*To whom correspondence should be addressed.

of replicates. For example, analysis of variance (ANOVA) have been applied to screen differentially expressed genes between different physiological/genetic conditions (Cui and Churchill, 2003). Bayesian mixture models have been developed and applied to replicated microarray data in order to infer gene clusters and classify tissues (Medvedovic and Sivaganesan, 2002; Medvedovic *et al*., 2004; Yeung and Bumgarner, 2005). However, many popular approaches to multivariate pattern discovery, such as hierarchical clustering or co-expression network construction, are based on estimating the statistical correlation between a pair of replicated biomolecule expression profiles, and the widespread bivariate correlation estimators used for that purpose (e.g. Pearson correlation coefficient) employ average expression profiles over replicates. The averaging, while on one hand reduces variance, on the other hand, tends to give rise to more biased correlation estimation. Thus the simple averaging over replicates is not completely satisfactory especially for noisy omics data that do not have good within-replicate correlation.

Here, we propose a new multivariate correlation estimator that exploits all replicates of a data set by assuming they are i.i.d. samples from the multivariate normal distribution. Based on a covariance structure that explicitly models within-replicate and between-replicate correlations of the data, we derive a maximum likelihood (ML)-based correlation estimator by maximizing the likelihood function of the replicated data. We establish a theoretical connection between the two correlation estimators when there is no replicate, and we provide a suite of statistical inference procedures for small sample size and large sample size data. Our work was strongly influenced by the previous works in estimating interclass and intraclass correlation from familial data (Chu and kshirsagar, 1994; Keen and Srivastava, 1991; Konishi, 1985; Konishi *et al.,* 1991; Srivastava, 1984).

2 METHODS

2.1 Pearson bivariate correlation estimator

Suppose the biomolecule abundance levels are simultaneously measured over n independent samples. There are m replicated measures available for each sample. Let x_{ij}, y_{ij} denote the abundance levels for arbitrary biomolecule X and biomolecule Y in the ith replicate in the jth sample. The application of Pearson correlation coefficient requires averaging biomolecule abundance profiles, i.e. $\bar{x}_j$ and $\bar{y}_j$, $j = 1, 2,\ldots, n$. The sample Pearson bivariate correlation coefficient between biomolecules X and Y is defined as:

$$cor(X, Y) = \frac{\sum_{j=1}^{n}(\bar{x}_j - \bar{x})(\bar{y}_j - \bar{y})}{\sqrt{\sum_{j=1}^{n}(\bar{x}_j - \bar{x})^2 \sum_{j=1}^{n}(\bar{y}_j - \bar{y})^2}}, \quad (1)$$

where $\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}, \bar{y}_j = \frac{1}{m}\sum_{i=1}^{m} y_{ij}$ and $\bar{x}$ and $\bar{y}$ are the grand means for x_{ij}, y_{ij}, $i = 1, 2,\ldots, m$, $j = 1, 2,\ldots, n$, respectively.

2.2 Multivariate correlation estimator

Instead of averaging, we exploit all the replicated observations by assuming the data are i.i.d. samples from a multivariate normal distribution with a specified correlation matrix and a mean vector, i.e. $Z_j = (x_{j1}, \ldots, x_{jm}, y_{j1}, \ldots, y_{jm})^T$, $j = 1, 2, \ldots, n$, follows a $2m$-variate normal distribution $N(\mu, \sum)$, where $\mu = \begin{bmatrix} \mu_x e_m \\ \mu_y e_m \end{bmatrix}$, $e_m = (1, \ldots, 1)^T$ is an $m \times 1$ vector, the correlation matrix $\sum$ is a $2m \times 2m$ matrix:

$$\Sigma = \begin{pmatrix} 1 & \ldots & \rho_x & \rho & \ldots & \rho \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_x & \ldots & 1 & \rho & \ldots & \rho \\ \rho & \ldots & \rho & 1 & \ldots & \rho_y \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho & \ldots & \rho & \rho_y & \ldots & 1 \end{pmatrix} = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{bmatrix}, \quad (2)$$

where the inter-molecule correlation ρ is the parameter of interest, and the intra-molecule correlation ρ_x or ρ_y are nuisance parameters. The ρ_x and ρ_y indicate the quality of replicates that high quality replicates tend to have high value, and vice versa. We employ three parameters: ρ, ρ_x and ρ_y to model the correlation structure of replicated data.

Assuming a multivariate normal model, the maximum likelihood estimate (MLE) of ρ can be derived as follows (see Appendix for more details):

$$\hat{\mu}_x = \frac{1}{n}\frac{1}{m}\sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij} \quad (3)$$

Similarly,

$$\hat{\mu}_y = \frac{1}{n}\frac{1}{m}\sum_{j=1}^{n}\sum_{i=1}^{m} y_{ij} \quad (4)$$

therefore, $\hat{\mu} = \begin{bmatrix} \hat{\mu}_x e_m \\ \hat{\mu}_y e_m \end{bmatrix}$

The MLE of $\sum$ is

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(Z_j - \hat{\mu})(Z_j - \hat{\mu})^T \quad (5)$$

To derive the MLE of ρ, it would be preferable to obtain the likelihood explicitly as a function of ρ. However, this is intractable in practice (see Appendix for a detailed discussion). Our approach is to use the average of the elements of $\hat{\Sigma}_{xy}$ estimated via Equation (5):

$$\hat{\rho} = \mathrm{Avg}(\hat{\Sigma}_{xy}) \quad (6)$$

It follows that the $\hat{\rho}$ remains invariant for data with unequal numbers of replicates (m) since it is not a function of m.

2.3 Theoretical connection between the two estimators

The sample Pearson correlation coefficient [Equation (1)] can be also written into the following form:

$$cor(X, Y) = \frac{\sum_{j=1}^{n}(\bar{x}_j - \bar{x})(\bar{y}_j - \bar{y})}{(n-1)S_X S_Y}, \quad (7)$$

where S_X and S_Y are SDs of X and Y, respectively. When there is no replicate ($m = 1$), the correlation matrix $\sum$ is reduced to a 2 by 2 matrix with diagonal elements equal to 1 and off-diagonal elements equal to ρ. It is easy to show from Equation (5) that

$$\hat{\rho} = \frac{\sum_{j=1}^{n}(\bar{x}_j - \bar{x})(\bar{y}_j - \bar{y})}{nS_X S_Y} \quad (8)$$

Hence, we derive the connection between the two estimators when there is no replicate as follows:

$$\hat{\rho} = \frac{n-1}{n} cor. \quad (9)$$

Algorithm 1 Permutation Test

1: Estimate the multivariate correlation ($\hat{\rho}_0$) of the original data using Eqs. 5, 6.
2: **if** sample size $n \leq 10$ **then**
3: **for all** $n!$ permutations of $(X_{.1}, Y_{.1}),\ldots,(X_{.n}, Y_{.n})$ **do**
4: keep $X_{.j}$ constant, permute the $Y_{.j}$, and re-estimate the multivariate correlation of the permutated data (i.e.$\hat{\rho}_k$, $k = 1, 2, \ldots, n!$)
5: **end for**
6: **end if**
7: **if** sample size $n > 10$ **then**
8: **for** P random permutations of $(X_{.1}, Y_{.1}),\ldots,(X_{.n}, Y_{.n})$ **do**
9: keep $X_{.j}$ constant, permute the $Y_{.j}$, and recalculate the multivariate correlation of the permutated data (i.e. $\hat{\rho}_k$, $k = 1, 2, \ldots, P$, $P \leq n!$)
10: **end for**
11: **end if**
12: $\hat{\rho}_k$ forms empirical null distribution of ρ
13: The empirical two-sided p-value is calculated as:

$$p = \frac{\sum_{k=1}^{K} I_k}{K}, I_k = \begin{cases} 1 & \text{if} |\hat{\rho}_k| \geq |\hat{\rho}_0|, \\ 0 & \text{if Otherwise}, \end{cases} \quad (11)$$

where $K = n!$ or $K = P$

Algorithm 2 Bootstrap Confidence Interval (Hollander and Wolfe 1999)

1: **for all** B bootstrap trials **do**
2: Make n random draws with replacement from the bivariate sample $Z_1, Z_2, \ldots, Z_n$. i.e. each data vector Z_i, $i = 1,\ldots,n$ is sampled with equivalent probability $\frac{1}{n}$
3: Compute $\hat{\rho}$. Denote B values of $\hat{\rho}$ as $\hat{\rho}^{*1}, \hat{\rho}^{*2}, \ldots, \hat{\rho}^{*B}$
4: **end for**
5: An asymptotically distribution-free confidence interval for ρ, with approximate confidence coefficient 100(1-α)%, is $(\rho_L^{'}, \rho_U^{'})$ where

$$\rho_L^{'} = \hat{\rho}^{*(k)}, \rho_U^{'} = \hat{\rho}^{*(B+1-k)}$$

and

$$k = B(\alpha/2). \quad (12)$$

6: **if** k=$B(\alpha/2)$ is an integer **then**
7: $\rho_L^{'}$ is the kth-largest bootstrap replication and $\rho_U^{'}$ is the $(B+1-k)$th-largest replication
8: **end if**
9: **if** k=$B(\alpha/2)$ is not an integer **then**
10: k=$(B+1)(\alpha/2)>$, the largest integer that is less than or equal to $(B+1)(\alpha/2)$
11: **end if**

2.4 Small sample inference procedure

For small sample size data, we provide a resampling-based statistical inference procedure. For n replicated bivariate observations $Z_1 = (X_{.1}, Y_{.1})$, $Z_2 = (X_{.2}, Y_{.2})$, …, $Z_n = (X_{.n}, Y_{.n})$, where . represents row index, and 1,…,n represent column index, we test the following hypothesis:

$$H_0 : \rho_{X,Y} = 0 \text{ versus } H_\alpha : \rho_{X,Y} \neq 0. \quad (10)$$

The steps for performing the permutation test are described in Algorithm 1, and the steps for deriving the asymptotically distribution-free bootstrap confidence interval is described in Algorithm 2 (Efron and Tibshirani, 1993, Hollander and Wolfe, 1999).

2.5 Large sample inference procedure

For moderate to large sample data, we provide an LRT procedure for testing the hypothesis that the multivariate correlation ρ vanishes. Under the multivariate normal distribution assumption stated in Section 2.2, $Z_j \sim N(\mu, \sum)$, and we test the following hypothesis:

$$H_0 : Z \in N(\mu, \Sigma_0) \text{ versus } H_\alpha : Z \in N(\mu, \Sigma_1). \quad (13)$$

Here, both $\sum_0$ and $\sum_1$ are $(m_1 + m_2) \times (m_1 + m_2)$ matrices, where m_1 and m_2 are number of replicates for biomolecule X and Y (in the case of equal numbers of replicates for X and Y, we have $m_1 = m_2 = m$) and $\Sigma_0 = \begin{pmatrix} \Sigma_x & \mathbf{0}_{m_{12}} \\ \mathbf{0}_{m_{21}}^T & \Sigma_y \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{pmatrix}$, where $\sum_x$ and $\sum_y$, with diagonal elements identity and all the other entries being ρ_x and ρ_y, respectively. $\mathbf{0}_{m_{12}}$ is an $m_1 \times m_2$ zero matrix and $\mathbf{0}_{m_{21}}$ is an $m_2 \times m_1$ zero matrix, i.e. under the null hypothesis, the intermolecule correlation ρ vanishes. $\sum_{xy}$ is an $m_1 \times m_2$ matrix with all entries equal to ρ. Likewise, $\sum_{xy}^T$ is an $m_2 \times m_1$ matrix with all entries equal to ρ. The likelihood ratio (LR) statistic for testing the two different correlation structures can be derived as follows:

$$\wedge = \frac{|\hat{\Sigma}_0|^{-n/2} e^{-\frac{1}{2}\sum_{j=1}^{n}(Z_j - \hat{\mu})'(\hat{\Sigma}_0)^{-1}(Z_j - \hat{\mu})}}{|\hat{\Sigma}_1|^{-n/2} e^{-\frac{1}{2}\sum_{j=1}^{n}(Z_j - \hat{\mu})'(\hat{\Sigma}_1)^{-1}(Z_j - \hat{\mu})}}. \quad (14)$$

Note that for the test to be a true LRT, all the estimated quantities $\hat{(\cdot)}$ in the above formula should be MLE's. In Section 2.2, Equations (3 – 5) give the formula of MLE's of the mean vector and the correlation matrix under H_α. The MLE of the correlation matrix under H_0 can be determined as $\hat{\Sigma}_0 = \begin{pmatrix} \hat{\Sigma}_x & O \\ O & \hat{\Sigma}_y \end{pmatrix}$, where

$$\hat{\Sigma}_x = \frac{1}{n}\sum_{j=1}^{n}(X_j - \hat{\mu}_x)(X_j - \hat{\mu}_x)', \quad (15)$$

$$\hat{\Sigma}_y = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{\mu}_y)(Y_j - \hat{\mu}_y)'. \quad (16)$$

The LR statistic, denoted by G^2, is therefore:

$$G^2 = -2\log\wedge = n[trM - \log|M| - (m_1 + m_2)], \quad (17)$$

where $M = (\hat{\Sigma}_0)^{-1}\hat{\Sigma}_1$. The LR statistic is asymptotically χ^2 distributed with $2(m_1 \times m_2)$ degrees of freedom under H_0 (Anderson, 1958).

3 RESULTS

3.1 Simulation setup

We used mean squared error (MSE) as an objective criterion for comparison. It is defined as

$$\text{MSE} = \text{E}_\rho(\hat{\rho}_l - \rho)^2 = 1/S\sum_{l=1}^{S}(\hat{\rho}_l - \rho)^2, \quad (18)$$

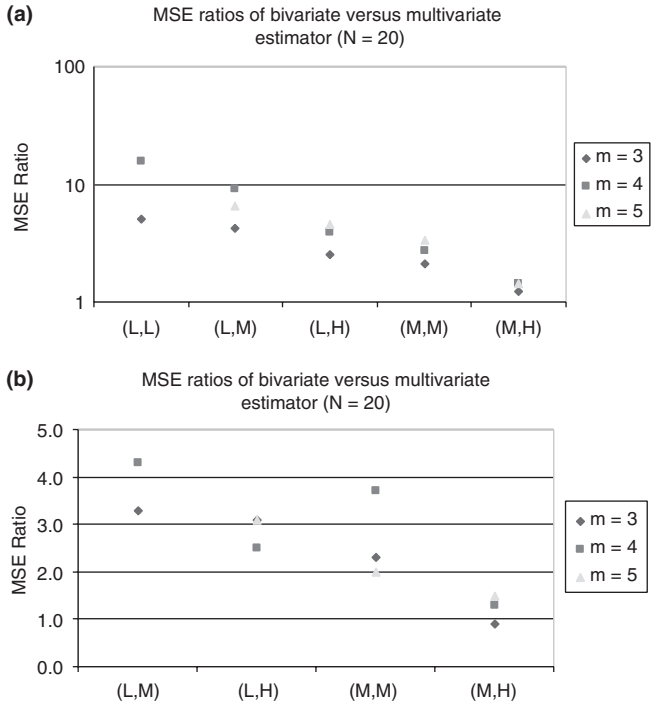

Fig. 1. MSE ratios of the bivariate versus multivariate correlation estimators. Sample size is fixed ($N = 20$), and replicates vary from $m = 3$; 4 and 5. (**a**) Low true correlation cut-off (in logarithmic scale). (**b**) High true correlation cut-off. Colour version of this figure is available as Supplementary material online.

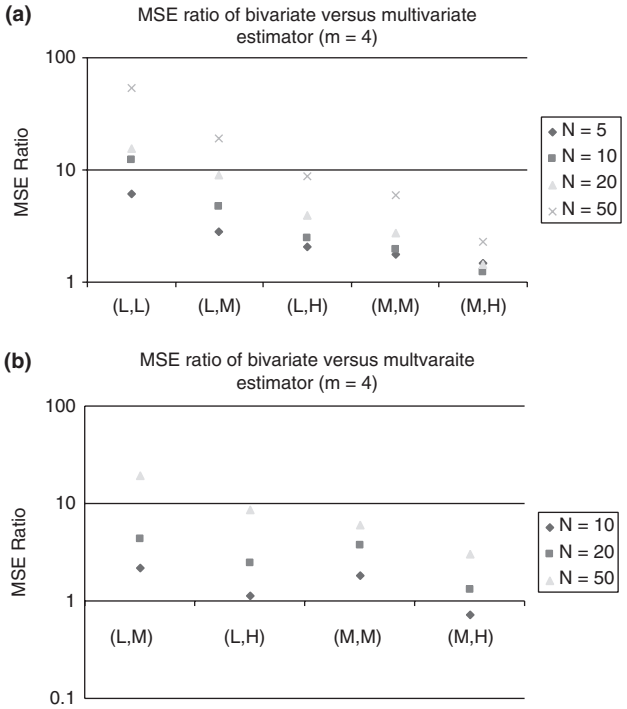

Fig. 2. MSE ratios of the bivariate versus the multivariate correlation estimators. Replicates are fixed ($m = 4$), and sample sizes vary (in logarithmic scale). (**a**) Low true correlation cut-off. (**b**) High true correlation cut-off. Colour version of this figure is available as Supplementary material online.

where l is the index, and S is the total number of simulation runs, ρ is the true correlation coefficient and $\hat{\rho}_l$ is the lth (either multivariate or bivariate) estimation of the true correlation. An estimator with smaller MSE is considered to be better.

Genome-wide data may have different sample sizes (n), different number of replicates (m) and different, often not well-studied, correlation structure ($\sum$). We aim to show that our multivariate estimator is consistently superior to Pearson bivariate correlation estimator in almost all the realistic combinations of the above parameters. In particular, we set the parameters to the following values:

- The sample size n: sample size represents the number of independent biological samples of interest. We set n to be 5, 10, 20 and 50, corresponding to small/median/large samples.
- The number of replicates m: most data only have a few replicates due to the cost of the genome-wide experiments. Therefore, we set m to be 3, 4 and 5.
- The within-replicate correlations ρ_x or ρ_y (see Methods Section): it is an indicator of data quality. We set it at three levels: very noisy (L) (0.1–0.3), noisy (M) (0.3–0.5) and clean (H) (0.6–0.8). Evaluation based on the first two levels are of particular interest to us, since genome-wide profiling is typically noisy.
- The between-replicate correlation ρ (see Methods Section): similarly, ρ can be chosen as low (0.2), median (0.4) and high (0.6). The high between-replicate correlation (typically 0.6) is of particular interest since it more reliably predict functional similarity (Zhu *et al.*, 2005b).

A comprehensive comparisons of the two estimators can thus be done by exhaustively exploring all possible combinations of the above parameters.

3.2 Simulation results

In Figures 1 and 2, vertical axis represents the MSE ratio of Pearson estimator over multivariate estimator. Ratios greater than 1 indicate that the multivariate correlation estimator outperforms the traditional bivariate estimator. The horizontal axis represents the different combinations of categorical within-replicate correlations. We compared the two estimators under five within-replicate correlation structures: i.e. (L, L), (L, M), (L, H), (M, M) and (M, H) and three between-replicate correlation cut-offs, 0.2, 0.4 and 0.6. Note that under few combination, simulations could not be preformed because the corresponding correlation matrices were not positive definite (PD).

In Figure 1, we fixed the sample size at $n = 20$ to examine the performance of the multivariate estimator with varying number of replicates versus the bivariate estimator by averaging over replicates. In Figure 1a and b, almost all examined MSE ratio are greater than 1 indicating superior performance of the multivariate estimator to the bivariate estimator. In particular, Figure 1a (lower between-replicate correlation cut-off $\rho = 0.4$)

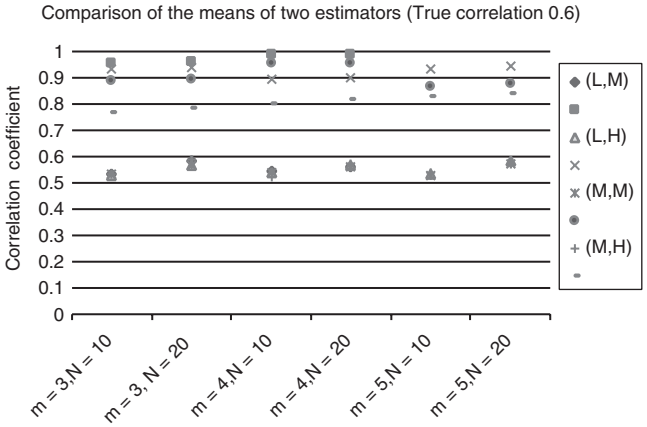

Fig. 3. Comparison of the means of two estimations. The upper cluster of curves (pink) represents the bivariate estimator for omics data with different replication quality. The lower cluster of curves (blue) represents the multivariate estimator for omics data with different replication quality. Colour version of this figure is available as Supplementary material online.

has a much higher overall MSE ratio. In Figure 2, we fixed the number of replicates at $m = 4$, which is typical in omics experiments to examine the performance over different number of samples. Similarly, almost all examined MSE ratio are greater than 1. In general, the superior performance of the multivariate estimator to the bivariate estimator is an increasing function of the sample size (n), number of replicates (m), but it is a decreasing function of data quality (ρ_x and ρ_y) and correlation cut-off. For real-world data, we expect that our correlation estimator work even better than Pearson estimator as sample size, number of replicates increase and data quality decrease. The set of comprehensive comparison results are summarized in the Supplementary Table S1.

Checking into the results more carefully, we found that the mean bivariate estimates are greatly biased upward and the mean multivariate estimates are slightly biased downward (Figure 3, Supplementary Table S1). This can be translated into a significantly reduced false positive rate when applying the multivariate estimator. We also examined variances of the two estimators, our estimator has smaller variance for noisy data, and larger variance otherwise (Supplementary Table S1). Our multivariate estimator, while a little conservative and sometime having larger variance, is much more accurate than the Pearson bivariate correlation estimator (Figure 3).

3.3 Microarray data analysis examples

Many real-world data analysis tasks rely on accurate estimation of correlation matrix. Affymetrix made two spike-in data sets publicly available as benchmark to compare different probe set expression summarization methods, such as RMA (Irizarry *et al.,* 2003) and GCRMA (Wu and Irizarry, 2005) [downloadable from http://affycomp.biostat.jhsph.edu/]. We instead use the two spike-in data sets as benchmark to compare correlation estimation methods: Pearson correlation coefficient using either RMA or GCRMA expression scores (GR and GG in Table 1). Multivariate correlation estimator using probe-level data where perfect match (PM) intensities of the same probe set are treated as replicated measures of the gene expression score (PR and PG in Table 1). For each spike-in data set, both nominal values of gene expression and observed probe-level intensities are available. The correlation matrix estimated by the four methods summarized in Table 1 are then compared to the nominal correlation matrix in terms of MSE. The smaller the MSE is, the closer the estimated correlation matrix to the nominal one.

Either array or probe set can be treated as multivariate random variable, correspondingly we calculated MSE of both array correlation matrix and probe set correlation matrix. Note that in the second data set, only probe set correlation matrix is computable due to unequal probe levels within array. In Table 2, we observe that MSE of the multivariate correlation estimators (PR and PG) are uniformly smaller than that

Table 1. Summary of the four methods to estimate correlation from Affymetrix microarray data

Method	Description
PR	**P**robe-level, **R**MA background correction, normalization and multivariate correlation estimation
PG	**P**robe-level, **G**CRMA background correction, normalization and multivariate correlation estimation
GR	**G**ene-level, **R**MA background correction, normalization, summarization and Pearson correlation estimation
GG	**G**ene-level, **G**CRMA background correction, normalization, summarization and Pearson correlation estimation

PR and PG are probe-level methods using multivariate correlation estimator. GR and GG are gene-level methods using Pearson correlation estimator.

Table 2. Comparing the four correlation estimation method listed in Table 1 using two spike-in data sets

	Method	Exp1	Exp2	Exp3
ArrayCorr (spike-in1)	PR	0.033	0.033	0.034
	PG	0.022	0.021	0.022
	GR	0.080	0.081	0.082
	GG	0.056	0.054	0.052
GeneCorr (spike-in1)	PR	0.117	0.120	0.131
	PG	0.107	0.112	0.126
	GR	0.196	0.193	0.201
	GG	0.193	0.188	0.200
GeneCorr (spike-in2)	PR	0.147	0.146	0.146
	PG	0.142	0.141	0.142
	GR	0.251	0.251	0.254
	GG	0.272	0.273	0.272

Mean squared error is reported to access average squared deviation of correlation matrix estimated using PR, PG, GR or GG method from the nominal one.

Table 3. Comparing the multivariate correlation estimator with the Pearson bivariate correlation estimator over different numbers of replicates (m) using adjusted RAND index

	$m=2$	$m=3$	$m=4$
MulVar;Diana	0.80	0.85	**0.95**
MulVar;Complete	0.71	0.86	0.66
MulVar;Average	0.85	**0.90**	0.87
BiVar;Diana	0.72	0.65	0.88
BiVar;Complete	0.78	0.68	0.68
BiVar;Average	**0.87**	0.88	0.87

Highest RAND indices at each number of replicates are in bold face.

of Pearson correlation estimator (GR and GG). It strongly indicates that pattern discovery methods based on multivariate correlation estimator are more capable to discover the true patterns of the data.

We then compared our multivariate correlation estimator with Pearson correlation estimator in the context of hierarchical gene clustering. We analyzed a subset of 205 genes whose expression were profiled using four replicates under 20 physiological/genetic conditions (Ideker *et al.*, 2000). The 205 genes were previously classified into four functional groups (Yeung *et al.,* 2003) that we employed it as the external knowledge for comparing performance of two correlation estimators through hierarchical gene clustering approaches. Similar to Medvedovic *et al.* (2004), we constructed six data sets of two replicates and four data sets of three replicates. The (average) adjusted RAND index (Hubert and Arabie, 1985) was used to measure the consistency between clustering results and the external knowledge.

As shown in Table 3, overall clustering algorithms based on the multivariate correlation estimator are better consistent to the external biological knowledge (higher adjusted RAND index). Note that we are interested in identifying four large clusters instead of many small clusters, therefore, we expect the divisive hierarchical clustering more suitable for this task (Speed, 2003). In Table 3, the (average) adjusted RAND index of divisive hierarchical clustering (represented by Diana) monotonically increases with the number of replicates increases and is consistent best to the external knowledge with four replicates (adjusted RAND index 0.95).

We further compared the sensitivity and specificity of the two test procedures over a wide biologically relevant range. For fair comparisons, asymptotic Fisher Z-transform was used as test procedure for Pearson correlation estimator in parallel with the asymptotic LRT test procedure for the multivariate correlation estimator. Sensitivity is calculated as the ratio between the number of gene pairs called significant by the test procedure and the number of genes pairs that does fall into the same one of four previously defined functional categories. Specificity is calculated as the ratio between the number of gene pairs NOT called significant and the number of gene pairs that does NOT fall into the same one of four previously defined functional categories.

We made comparisons using top correlated gene pairs (1–15%) and the selection is based on the empirical results that only top correlated gene pairs are likely to be functionally relevant (Griffith *et al.,* 2005; Lee *et al.*, 2004). In the context of co-expression network, it corresponds to the well-observed phenomenon that gene co-expression network is typically very sparse (Lee *et al.*, 2004; Zhu *et al.*, 2005a). Our LRT procedure performs slightly better than Fisher's Z-transform test procedure in terms of both sensitivity and specificity (Supplementary Table S2). Note that there is much overlap of significant calls between the two test procedures (Supplementary Table S2).

We claim that the better performance of our correlation estimator and test procedure was achieved under an adverse condition of high within-replicate correlation (75% quantile: 0.89, 50% quantile: 0.73, 25% quantile: 0.34, see Supplementary Fig. S1). As demonstrated in simulations, we expect our estimator to significantly outperform Pearson bivariate estimator with noisy data (low within-replicate correlation) and more replicates.

4 CONCLUSION

In this article, we proposed a new multivariate correlation estimator that exploits all replicated data points in genome-scale data instead of averaging over replicates when using bivariate correlation estimator. Our simulation studies showed that the new estimator possesses low MSE, high accuracy and comparable variance to the Pearson correlation estimator. The analysis of a real-world data set further demonstrated the potential application of our method to pattern discovery problems from noisy genome-wide profiling data. In particular, our approach provide a new way of exploiting Affymetrix probe-level data to pattern discovery problems. Using two Affymetrix spike-in data sets, we have shown that our estimated correlation matrix is closer to the nominal correlation matrix than the competing methods. It seems to indicate that pattern discovery methods based on multivariate correlation estimator are better able to discover the true patterns from probe-level data. Since the multivariate correlation estimator was presented in a closed form, its computational cost is moderate compared to Pearson correlation estimator, therefore, pattern discovery algorithms based on the multivariate correlation estimator is very scalable to large omics data set.

ACKNOWLEDGEMENTS

We thank Arcady Mushegian, Norman Pavelka and Frank Emmert-Streib for helpful discussion. We would to thank anonymous reviewers for their helps in improving the quality of this manuscript.

REFERENCES

Anderson,T.W. (1958) *An introduction to Multivariate Analysis.* Wiley, New York.

Chu,S.K. and Kshirsagar,A.M. (1994) Correlation coefficient between two variables when the data set consists of observations on twins. *Parisankhyan Samikkha: Int. J. Stat.*, **1**, 1.

Cui,X.Q. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 201.

Efron,B. and Tibshirani,R.J. (1993) *An Introduction to the Boostrap*. Chapman & Hall, New York, USA.
Griffith,O.L. *et al.* (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. *Genomics*, **86**, 476–488.
Hardt,M. *et al.* (2005) Assessing the effects of diurnal variation on the composition of human parotid saliva: quantitative analysis of native peptides using iTRAQ reagents. *Anal. Chem.*, **77**, 4947–4954.
Hollander,A. and Wolfe,D. (1999) *Nonparametric Statistical Methods*. Wiley-Interscience, Hoboken NJ, USA.
Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.
Ideker,T. *et al.* (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.
Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
Keen,K.J. and Srivastava,M. S. (1991) The asymptotic variance of the interclass correlation coefficient. *Biometrika*, **78**, 225–228.
Konishi,S. (1985) Normalizing and variance stabilizing transformations for intraclass correlations. *Ann. Inst. Stat. Math.*, **37**, 87–94.
Konishi,S. *et al.* (1991) Inferences on multivariate measures of interclass and intraclass correlations in familial data. *J. R. Stat. Soc. B*, **53**, 649–659.
Lee,H.K. *et al.* (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.
Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression score computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
Liu,X. *et al.* (2006) Bayesian context-specific infinite mixture model for clustering of gene expression profiles across diverse microarray datasets. *Bioinformatics*, **22**, 1737–1744.
Medvedovic,M. and Sivaganesan,S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.
Medvedovic,M. *et al.* (2004) Bayesian mixtures for clustering replicated microarray data. *Bioinformatics*, **20**, 1222–1232.
Nguyen,D.V. and Rocke,D.M. (2002a) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.
Nguyen,D.V. and Rocke,D.M. (2002b) Multi-class cancer classification via partial least squares using gene expression profiles. *Bioinformatics*, **18**, 1216–1226.
Pavelka,N. *et al.* (2004) A power law global error model for the identification of differentially expressed genes in microarray data. *BMC Bioinformatics*, **5**, 203.
Ross,P.L. *et al.* (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics*, **3**, 1154–1169.
Sartor,M. A. *et al.* (2006) Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics*, **7**, 538.
Speed,T. (ed.) Statistical analysis of gene expression microarray data. CRC Press , Chapman & Hall Boca Raton FL USA.
Srivastava,M.A. (1984) Estimation of interclass correlations in familial data. *Biometrika*, **71**, 177–185.
Wu,Z. and Irizarry,R.A. (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J. Comput. Biol.*, **12**, 882–893.
Yeung,K.Y. and Bumgarner,R. (2005) Multi-class classification of microarray data with repeated measurements: application to cancer. *Genome Biol.*, **4**, R83.
Yeung,K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.
Yeung,K.Y., Medvedovic,M and Bumgarner,R. (2003) Clustering gene expression data with repeated measurements. *Genome Biol.*, **4**, R34.
Zhou,L. and Rocke,D. (2005) An expression index for Affymetrix GeneChips based on the generalized algorithm. *Bioinformatics*, **21**, 3983–3989.
Zhu,D. *et al.* (2005a) High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS). *J. Comput. Biol.*, **12**, 1027–1043.
Zhu,D. *et al.* (2005b) Network constrined clustering for gene microarray data. *Bioinformatics*, **21**, 4014–4021.

APPENDIX

The likelihood function of a $2m$-variate normal family is as the following:

$$L(\mu, \Sigma) = \frac{1}{(2\pi)^{nm}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n}(Z_j-\mu)'\Sigma^{-1}(Z_j-\mu)}, \tag{19}$$

where $\mu = \begin{bmatrix}\mu_x e_m \\ \mu_y e_m\end{bmatrix}$, $\mu \in \Re^{2m}$ with $e_m = (1, \ldots, 1)^T$. $\sum$ is the $2m \times 2m$ covariance matrix with the structure specified in Equation (2). To obtain the MLE, equivalently, we wish to minimize:

$$l(\mu, \Sigma) = n\log|\Sigma| + \sum_{j=1}^{n}(Z_j - \mu)'\Sigma^{-1}(Z_j - \mu). \tag{20}$$

By taking derivatives on μ_x and μ_y, we have:

$$\hat{\mu}_x = \frac{1}{n}\frac{1}{m}\sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij}, \tag{21}$$

$$\hat{\mu}_y = \frac{1}{n}\frac{1}{m}\sum_{j=1}^{n}\sum_{i=1}^{m} y_{ij}. \tag{22}$$

Now, in order to find the MLE of $\sum$, we do a slight transformation on the log-likelihood:

$$\begin{aligned} & l(\mu, \Sigma) \\ &= -\frac{nm}{2}ln2\pi - \frac{n}{2}ln|\Sigma| - \frac{1}{2}\sum_{j=1}^{n}(Z_j-\mu)^T\Sigma^{-1}(Z_j-\mu) \\ &= -\frac{nm}{2}ln2\pi - \frac{n}{2}ln|\Sigma| - \frac{1}{2}\sum_{j=1}^{n} tr(Z_j-\mu)^T\Sigma^{-1}(Z_j-\mu) \\ &= -\frac{nm}{2}ln2\pi - \frac{n}{2}ln|\Sigma| - \frac{1}{2}\sum_{j=1}^{n} tr\Sigma^{-1}(Z_j-\mu)(Z_j-\mu)^T, \end{aligned} \tag{23}$$

therefore,

$$\begin{aligned} \frac{\partial l(\mu, \Sigma)}{\partial \Sigma} &= -\frac{n}{2}\frac{\partial ln|\Sigma|}{\partial \Sigma} - \frac{1}{2}\frac{\partial}{\partial \Sigma}\sum_{j=1}^{n} tr\Sigma^{-1}(Z_j-\mu)(Z_j-\mu)^T \\ &= -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\sum_{j=1}^{n}(Z_j-\mu)(Z_j-\mu)^T. \end{aligned} \tag{24}$$

This gives:

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(Z_j - \hat{\mu})(Z_j - \hat{\mu})^T. \tag{25}$$

The ideal approach to obtain MLE of ρ is to get the likelihood explicitly expressed as a function of the unknown parameters $\theta = (\mu_x, \mu_y, \rho_x, \rho_y, \rho)^T$. In order to do that, we can introduce the transformation used by Srivastava (1984) for interclass correlations in familial data.

Let $e_m = (1, \ldots, 1)^T$, $0_m = (0, \ldots, 0)^T$ be constant $m \times 1$ vectors. I_m denotes the $m \times m$ identity matrix and 0_m is an $m \times m$ zero matrix. The canonical transformation of the original random vector Z_j, $Z_j \in \Re^{2m}$, is given by:

$$T_j = \begin{pmatrix} A_j & 0_m \\ 0_m^T & A_j \end{pmatrix}(x_{j1}, \ldots, x_{jm}, y_{j1}, \ldots, y_{jm})^T \tag{26}$$

where $A_j = (m^{-1}e_m, C_j^T)^T$ is an $m \times m$ matrix such that $C_j e_m = 0_m$ and $C_j C_j^T = I_{m-1}$. The likelihood function of the original data is independent of the choice of C_j (Keen and Srivastava, 1991).

Then the expectation and covariance matrix for the transformed random vector T_j can be expressed as:

$$\mu = E(T_j) = (\mu_x, 0_{m-1}^T, \mu_y, 0_{m-1}^T)^T, \tag{27}$$

$$\Sigma = cov(T_j) = \begin{bmatrix} \mathcal{X} & \Psi \\ \Psi^T & \mathcal{Y} \end{bmatrix}, \tag{28}$$

where $\mathcal{X} = [\rho_x + m^{-1}(1-\rho_x)] \oplus (1-\rho_x) I_{m-1}$, $\mathcal{Y} = [\rho_y + m^{-1}(1-\rho_y)] \oplus (1-\rho_y) I_{m-1}$ and $\Psi = \rho \oplus 0_{m-1, m-1}$.

Therefore, the $\sum^{-1}$ in the log likelihood function can be expressed as:

$$\Sigma^{-1} = \begin{bmatrix} \mathcal{G} & \mathcal{K} \\ \mathcal{K} & \mathcal{H}, \end{bmatrix} \tag{29}$$

with $\mathcal{G} = (\mathcal{X} - \Psi \mathcal{Y}^{-1} \Psi^T)^{-1}$, $\mathcal{H} = (\mathcal{Y} - \Psi \mathcal{X}^{-1} \Psi^T)^{-1}$, $\mathcal{K} = -\mathcal{X}^{-1} \Psi \mathcal{H}$.

We can also express $|\sum|$ in terms of the parameters ρ_x, ρ_y and ρ explicitly:

$$\begin{aligned} |\Sigma| &= |\mathcal{X}||\mathcal{Y} - \Psi^T \mathcal{X}^{-1} \Psi| \\ &= [(1-\rho_x)(1-\rho_y)]^{m-1} \frac{[(m-1)\rho_x + 1][(m-1)\rho_y + 1] - m^2\rho^2}{m^2} \end{aligned} \tag{30}$$

The MLE can therefore be derived by minimizing:

$$\begin{aligned} l = n(m-1)[log(1-\rho_x) + log(1-\rho_y)] \\ + n\log([(m-1)\rho_x + 1][(m-1)\rho_y + 1] - m^2\rho^2) - 2n\log m \\ + m \sum_{j=1}^{n} [\frac{b}{h}(\bar{x}_j - \mu_x)^2 + \frac{a}{h}(\bar{y}_j - \mu_y)^2] + \sum_{j=1}^{n} [\frac{SS_{x_j}}{(1-\rho_x)} + \frac{SS_{y_j}}{(1-\rho_y)}, \\ - 2\frac{m^2}{h}(\bar{x}_j - \mu_x)(\bar{y}_j - \mu_y)\rho], \end{aligned} \tag{31}$$

where $a = (m-1)\rho_x + 1$, $b = (m-1)\rho_y + 1$ and $h = [(m-1)\rho_x + 1][(m-1)\rho_y + 1] - m^2\rho^2$, and SS_{x_j}, SS_{y_j} denote the sample sum of squares for gene X and Y under the j_{th} condition, respectively:

$$SS_{x_j} = \sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2 \quad \text{with} \quad \bar{x}_j = \sum_{i=1}^{m} x_{ij}/m \tag{32}$$

Similarly,

$$SS_{y_j} = \sum_{i=1}^{m} (y_{ij} - \bar{y}_j)^2 \quad \text{with} \quad \bar{y}_j = \sum_{i=1}^{m} y_{ij}/m \tag{33}$$

By taking differentiation of this with respect to the unknown parameters $\theta = (\mu_x, \mu_y, \rho_x, \rho_y, \rho)^T$, we can get the true MLE estimates. However, we cannot get explicit solutions for any of the unknown parameters. The calculations were proved to be intractable in implementation, and we chose to use the estimation specified in Section 2 rather that the MLE solution for ρ.