Estimation of variances and covariances for high-dimensional data: a selective review

Tiejun Tong,[1] Cheng Wang[1] and Yuedong Wang[2]*

Estimation of variances and covariances is required for many statistical methods such as *t*-test, principal component analysis and linear discriminant analysis. High-dimensional data such as gene expression microarray data and financial data pose challenges to traditional statistical and computational methods. In this paper, we review some recent developments in the estimation of variances, covariance matrix, and precision matrix, with emphasis on the applications to microarray data analysis. © 2014 Wiley Periodicals, Inc.

How to cite this article:
WIREs Comput Stat 2014, 6:255–264. doi: 10.1002/wics.1308

Keywords: covariance matrix; high-dimensional data; microarray data; precision matrix; shrinkage estimation; sparse covariance matrix

INTRODUCTION

Variances and covariances are involved in the construction of many statistical methods including *t*-test, Hotelling's T^2-test, principal component analysis, and linear discriminant analysis. Therefore, the estimation of these quantities is of critical importance and has been well studied over the years. The recent flood of high-dimensional data, however, poses new challenges to traditional statistical and computational methods. For example, the microarray technology allows simultaneous monitoring of the whole genome. Due to the cost and other experimental difficulties such as the availabilities of biological materials, microarray data are usually collected in a limited number of samples. These kinds of data are often referred to as high-dimensional small sample size data, or 'large p small n' data, where p is the number of genes and n is the number of samples. Due to the small sample size, there is a large amount of uncertainty associated with standard estimates of parameters such as the sample mean and covariance. As a consequence, statistical analyses based on such estimates are usually unreliable.

Let $Y_i = (Y_{i1}, \dots, Y_{ip})^{\mathrm{T}}$ be independent random samples from a multivariate normal distribution,[1,2]

$$Y_i = \Sigma^{1/2} X_i + \mu, \quad i = 1, \dots, n, \tag{1}$$

where $\mu = (\mu_1, \dots, \mu_p)^{\mathrm{T}}$ is a p-dimensional mean vector, Σ is a $p \times p$ positive definite covariance matrix, $X_i = (X_{i1}, \dots, X_{ip})^{\mathrm{T}}$, and X_{ij} are independent and identically distributed (i.i.d.) random variables from the standard normal distribution. For microarray data, Y_{ij} represents the normalized gene expression level of gene j in the ith sample. In two-sample cDNA arrays, Y_{ij} may also represent the normalized log ratio of two-channel intensities.

In multivariate statistical analysis, one often needs to estimate the covariance matrix Σ or the inverse covariance matrix Σ^{-1}. The inverse covariance matrix is also called the precision matrix $\Omega = \Sigma^{-1}$. The estimation of the covariance matrix or its inverse has applications in many statistical problems including linear discriminant analysis,[3] Hotelling's T^2-test,[4] and Markowitz mean-variance analysis.[5] We write the

*Corresponding to: yuedong@pstat.ucsb.edu

[1]Department of Mathematics, Hong Kong Baptist University, Hong Kong, Hong Kong

[2]Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA

Conflict of interest: The authors have declared no conflicts of interest for this article.

sample covariance matrix as

$$S_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)^{\mathrm{T}},$$

where $\overline{\mathbf{Y}} = \sum_{i=1}^{n} \mathbf{Y}_i/n$ is the sample mean. When $p < n$, $(n-1)S_n$ follows a Wishart distribution and $S_n^{-1}/(n-1)$ follows an inverse Wishart distribution. In addition, $E\left(S_n^{-1}\right) = (n-1)\,\Omega/(n-p-2)$. A common practice is to estimate Σ by the sample covariance matrix S_n and estimate Ω by the scaled inverse covariance matrix $(n-p-2)\,S_n^{-1}/(n-1)$. These two estimators are consistent estimators of Σ and Ω when p is fixed and n goes to infinity.

For high-dimensional data such as microarray data, however, p can be as large as or even larger than n. As a consequence, the sample covariance matrix S_n is close to or is a singular matrix. This brings new challenges to the estimation of the covariance matrix and the precision matrix. In this paper, we review some recent developments in the estimation of variances and covariances. Specifically, we review (1) the estimation of variances, i.e., the diagonal matrix of $\boldsymbol{\Sigma}$, (2) the estimation of the covariance matrix $\boldsymbol{\Sigma}$, and (3) the estimation of the precision matrix Ω.

ESTIMATION OF VARIANCES

As reviewed in Cui and Churchill[6] and Ayroles and Gibson,[7] one commonly used method to identify differentially expressed genes is the analysis of variance (ANOVA). ANOVA is a very flexible approach for microarray experiments to compare more than two conditions. When there are only two conditions, the t-test may be used for detecting differential expression. Throughout the paper, for simplicity of illustration we consider only the two-color arrays with one factor at two levels, in which a paired t-test may be employed. Let $\boldsymbol{D} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$, where σ_j^2 are gene-specific variances for $j = 1, \ldots, p$, respectively. When the factor has more than two levels or the experiment involves more than one factor, the variances σ_j^2 correspond to residual variances in ANOVA or regression models.

In microarray data analysis, rather than the whole covariance matrix $\boldsymbol{\Sigma}$, there are many situations where only the estimation of gene-specific variances is required. We now provide several examples of these situations. The first example is a multiple testing problem in microarray data analysis. To identify differentially expressed genes, we test the hypotheses $H_{j0}: \mu_j = 0$ against $H_{j1}: \mu_j \neq 0$ for each gene j. Consider the test statistic $T_j = \sqrt{n}\,\overline{Y}_j/s_j$, where $\overline{Y}_j$ is the gene-specific sample mean and s_j^2 is the gene-specific sample variance. Then, only an estimate of D is needed rather than the whole covariance matrix Σ. The second example is the class prediction (or classification) problem. If we use Diagonal Linear Discriminant Analysis (DLDA) for class prediction,[8] then again we need to estimate $\boldsymbol{D}$ rather than Σ. For more details about DLDA and its variants, see Bickel and Levina,[9] Lee et al.,[10] Pang et al.,[11] and Huang et al.[12] The third example is the multivariate testing problem. To overcome the singularity problem, several researchers proposed diagonal Hotelling's T^2-tests where only an estimate of $\boldsymbol{D}$ is required. For more details, see, for example, Wu et al.,[13] Srivastava and Du,[14] Srivastava,[15] Park and Ayyala,[16] and Srivastava et al.[17]

Due to the small sample size n, however, the standard gene-specific sample variance s_j^2 is usually unstable. Consequently, the standard t-tests in the first example, the diagonal discriminant rules in the second example, and the diagonal Hotelling tests in the third example may not be reliable in practice. Various methods have been proposed for improving the estimation of gene-specific variances. Some of these methods are reviewed in the remainder of this section.

Shrinkage Estimators

A key to improving the variance estimation is to borrow information across genes, implicitly or explicitly, locally or globally. One of the earliest methods to stabilize the variance estimation was proposed by Tusher et al.[18] in 2001. In order to avoid the undue influence of the small variance estimates, Tusher et al.[18] proposed to estimate the standard deviation σ_j by $(s_j + c)/2$ in their SAM test, where c is a constant acting as a shrinkage factor. For the choice of the constant c, Efron et al.[19] suggested to use the 90th percentile of all estimated standard deviations, whereas Cui and Churchill[6] suggested to use the pooled sample variance.

In 2005, Cui et al.[20] proposed a James–Stein shrinkage estimator for the variances. For microarray data with $Y_{ij} \overset{\text{i.i.d.}}{\sim} N\left(\mu_j, \sigma_j^2\right)$, we have $s_j^2 = \sigma_j^2 \chi_{j,v}^2/v$, where for ease of notation, $\chi_{j,v}^2$ denote i.i.d. random variables which have a chi-squared distribution with $v = n - 1$ degrees of freedom. Taking the log transformation leads to

$$Z_j = \ln \sigma_j^2 + \varepsilon_j, \qquad (2)$$

where $Z_j = \ln\left(s_j^2\right) - m$, $\varepsilon_j = \ln\left(\chi_{j,v}^2/v\right) - m$, and $m = E\left\{\ln\left(\chi_{j,v}^2/v\right)\right\}$. Treating Z_j in Eq. (2) as normal random variables, the James–Stein shrinkage method[21] can be applied to derive a shrinkage estimate for $\ln \sigma_j^2$. Transforming back to the original scale, the final estimates of the variances are as follows:

$$\widetilde{\sigma}_j^2 = B\left(\prod_{j=1}^{p}\left(s_j^2\right)^{1/p}\right)\exp\left[\left(1 - \frac{(p-3)\,V}{\sum\left(\ln s_j^2 - \overline{\ln s_j^2}\right)^2}\right)_+ \times\left(\ln s_j^2 - \overline{\ln s_j^2}\right)\right], \qquad (3)$$

where $V = \mathrm{var}(\varepsilon_j)$, $\overline{\ln s_j^2} = \sum_{j=1}^{p} \ln\left(s_j^2\right)/p$, and $B = \exp(-m)$ is the bias correction factor such that $B\prod_{j=1}^{p}\left(s_j^2\right)^{1/p}$ gives an unbiased estimator of σ^2 when $\sigma_j^2 = \sigma^2$ for all j.

Note that Z_j in Eq. (2) can be far from normal when v is small. Therefore, the shrinkage variance estimates (Eq. (3)) can be suboptimal. Note also that the variance estimates appear in the denominator of the t-tests. Tong and Wang[22] showed that using direct estimates of $1/\sigma_j$ leads to a more powerful and robust test than using the reciprocal of the estimates of σ_j. Consequently, they considered the general estimation of $\left(\sigma_j^2\right)^t$ for any power $t \neq 0$. Note that σ_j and $1/\sigma_j$ are special cases with $t = 1/2$ and $t = -1/2$. Let $s_j^{2t} = \left(s_j^2\right)^t$, $s_{pool}^{2t} = \prod_{j=1}^{p}\left(s_j^2\right)^{t/p}$, and

$$h_n(t) = \left(\frac{v}{2}\right)^t \left(\frac{\Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v}{2} + \frac{t}{n}\right)}\right)^n, \qquad (4)$$

where $\Gamma(\cdot)$ is the Gamma function. Tong and Wang[22] proposed the following family of shrinkage estimators for $\left(\sigma_j^2\right)^t$:

$$\widehat{\sigma}_j^{2t} = \left(h_p(t)\, s_{\mathrm{pool}}^{2t}\right)^{\alpha}\left(h_1(t)\, s_j^{2t}\right)^{1-\alpha}, \quad 0 \le \alpha \le 1, \quad (5)$$

where $h_1(t)\, s_j^{2t}$ is an unbiased estimator of σ_j^{2t}, and $h_p(t)\, s_{pool}^{2t}$ is an unbiased estimator of σ^{2t} when $\sigma_j^2 = \sigma^2$ for all j. When $t = 1$, $\widehat{\sigma}_j^2$ is a simple modification of the estimator in Cui et al.[20] The shrinkage parameter α controls the degree of shrinkage from the gene-specific variance estimate $h_1(t)\, S_j^{2t}$ toward the bias-corrected geometric mean $h_p(t)\, s_{\mathrm{pool}}^{2t}$. There is no shrinkage when $\alpha = 0$, and all variance estimates are shrunken to the pooled variance when $\alpha = 1$. More recently, Tong et al.[23] proposed another James–Stein shrinkage estimator for the variances that shrunk the individual sample variance toward the arithmetic mean. For both shrinkage to the geometric mean and shrinkage to the arithmetic mean estimators, optimal shrinkage parameters were derived under both the Stein and squared loss functions. Asymptotic properties were investigated under the two schemes when either the number of degrees of freedom of each individual estimate or the number of individuals approaches infinity.

Bayesian Estimators

Baldi and Long[24] applied a Bayesian method to improve the estimation of variances. Specifically, they assumed the following conjugate prior for $\left(\mu_j, \sigma_j^2\right)$,

$$p\left(\mu_j, \sigma_j^2 | \alpha\right) = N\left(\mu_j; \mu_0, \sigma_j^2/\lambda_0\right) \mathcal{I}\left(\sigma_j^2; v_0, \sigma_0^2\right),$$

where $\boldsymbol{\alpha} = \left(\mu_0, \lambda_0, v_0, \sigma_0^2\right)$ are unknown hyperparameters, $N(x; a, b)$ represents the normal density function with mean a and variance b, and $\mathcal{I}(x; a, b)$ represents the scaled inverse gamma density with degrees of freedom a and scale b. The posterior density for $\left(\mu_j, \sigma_j^2\right)$ has the same functional form as the prior density.

$$p\left(\mu_j, \sigma_j^2 | Y_{1j}, \ldots, Y_{nj}\right) = N\left(\mu_j; \mu_n, \sigma_j^2/\lambda_n\right)\ \mathcal{I}\left(\sigma_j^2; v_n, \sigma_n^2\right),$$

where $\lambda_n = \lambda_0 + n$, $v_n = v_0 + n$, and

$$\mu_n = \frac{\lambda_0}{\lambda_0 + n}\mu_0 + \frac{n}{\lambda_0 + n}\overline{Y}_j,$$

$$v_n\sigma_n^2 = v_0\sigma_0^2 + (n-1)\, s_j^2 + \frac{\lambda_0 n}{\lambda_0 + n}\left(\overline{Y}_j - \mu_0\right)^2.$$

Note that the posterior mean μ_n is a weighted average of the prior mean μ_0 and the sample mean $\overline{Y}_j$. Baldi and Long[24] suggested to use $\mu_0 = \overline{Y}_j$. This leads to the posterior means of μ_i and σ_i^2 as

$$\widehat{\mu}_j = \overline{Y}_j \quad \text{and} \quad \widehat{\sigma}_j^2 = \frac{v_0\sigma_0^2 + (n-2)\, s_j^2}{v_0 + n - 2}.$$

The posterior modes have the same form as above with $n-2$ replaced by $n-1$. It is clear that both posterior mean and mode of σ_i^2 are shrinkage estimators. The background variance σ_0^2 is estimated by pooling together all the neighboring genes contained in a window of a certain size. The parameter

v_0 represents the degree of confidence in the background variance σ_0^2 versus the gene-specific sample variance.

Other methods under the Bayesian framework are summarized as follows. Lonnstedt and Speed[25] proposed a posterior odds of differential expression in a replicated two-color experiment using an empirical Bayes approach that combines information across genes. Kendziorski et al.[26] extended the empirical Bayes method using the hierarchical gamma–gamma and lognormal–normal models. Smyth[27] developed hierarchical models in the context of general linear models, see also Wright and Simon.[28] Hwang and Liu[29] and Zhao[30] applied some empirical Bayes approaches that shrunk both means and variances. Ji et al.[31] developed an empirical Bayes estimator for the variances by borrowing information across both genes and experiments.

Regression Estimators

It has been observed for microarray data that the variance increases proportionally with the intensity level.[32–35] One possible remedy to this problem is to transform the data and eliminate the dependence of the variance on the mean. See, for example, Durbin et al.,[36] Huber et al.,[37] Rocke,[38] Rocke and Durbin,[39] and Durbin and Rocke.[40]

Another remedy is to apply the regression method. Specifically, we assume a functional relationship between the mean and the variance: $\sigma_i^2 = g\left(\mu_i\right)$. The goal of the regression approach is then to estimate the variance–mean function g. Depending on prior knowledge, the function g may be modeled parametrically or non-parametrically. When modeled parametrically, we denote $g(\mu, \theta)$ as the variance function with parameter θ. Parametric models for microarray data include the constant coefficient of variation model,[32] $g(\mu) = \theta\mu^2$, and the quadratic model,[33,41] $g(\mu) = \theta_1 + \theta_2\mu^2$. Often it is difficult, if not impossible, to specify a parametric model for g. A non-parametric regression approach may be used in these situations. Any one of the non-parametric regression approaches such as smoothing splines and local polynomials could be used to model g non-parametrically.

Estimation of the parameter θ or the non-parametric function g needs to take several subtle issues into account. Note that the means μ_j represent a large number of unknown nuisance parameters in the estimation of the variance function. This is a Neyman–Scott type problem where care needs to be taken to derive consistent estimates. It is usually not difficult to construct consistent estimators for θ or g if μ_j is known. Denote $\widehat{\theta}\left(\boldsymbol{\mu}\right)$ and $\widehat{g}\left(\boldsymbol{\mu}\right)$ as consistent estimators for θ and g, respectively, where the dependence on $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ is expressed explicitly. In practice $\boldsymbol{\mu}$ is unknown. The sample mean $\overline{\boldsymbol{Y}} = \left(\overline{Y}_1, \ldots, \overline{Y}_p\right)^{\mathrm{T}}$ is a natural estimate of $\boldsymbol{\mu}$. Then, a direct approach is to replace $\boldsymbol{\mu}$ by $\overline{\boldsymbol{Y}}$ which leads to the estimates $\widehat{\theta}(\overline{\boldsymbol{Y}})$ and $\widehat{g}(\overline{\boldsymbol{Y}})$ for θ and g. Unfortunately, these naive estimates $\widehat{\theta}(\overline{\boldsymbol{Y}})$ and $\widehat{g}(\overline{\boldsymbol{Y}})$ are in general inconsistent as the sampling error in $\overline{\boldsymbol{Y}}$ is ignored.[42,43]

Regarding $\overline{\boldsymbol{Y}}$ as an error-prone unbiased measure of $\boldsymbol{\mu}$, the problem can be cast in a general framework of heteroscedastic measurement error. Therefore, the SIMEX (simulation extrapolation) method in Carroll et al.[44] may be applied to derive estimates of θ and g. However, due to correlation between the measurement error and the response, the naive application of the SIMEX method still does not lead to consistent estimates.[42,43] To overcome this problem, Carroll and Wang[42] and Wang et al.[43] proposed permutation SIMEX methods that lead to consistent estimates of θ and g. Other regression methods include Fan et al.[45] and Fang and Zhu.[46]

ESTIMATION OF THE COVARIANCE MATRIX

When p is fixed and n is large, the sample covariance matrix S_n is an unbiased and consistent estimator of the covariance matrix $\boldsymbol{\Sigma}$. However, for high-dimensional data in which p is close to or even larger than n, S_n may no longer be a good estimator of $\boldsymbol{\Sigma}$. In particular, S_n will be a singular matrix, or close to it, in such settings.

Many methods have been proposed in the literature to improve the estimation of $\boldsymbol{\Sigma}$. In essence, these methods can be classified into three categories corresponding to (1) $p < n$, (2) $p \geq n$, and (3) $p \gg n$, respectively. For category (1), as S_n is invertible with the eigenvalues being non-zero, attempts are often made to stabilize the estimation of eigenvalues. This was first proposed by Stein[47] and will be referred to as *Stein-type estimators*. For category (2), S_n is a singular matrix. To overcome this problem, one may consider an estimator like $\lambda S_n + (1-\lambda)\boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is an identity matrix of size $p \times p$ and $\lambda \in [0, 1)$ is a shrinkage parameter. Note that this type of method can be derived under the Bayes or empirical Bayes framework. We refer to the estimators of this type as *the ridge-type estimators*. Note that the covariance matrix associated with high-dimensional data such as microarrays can be sparse. In such settings, the above-mentioned shrinkage estimators for non-sparse $\boldsymbol{\Sigma}$ may no longer be applicable or the improvement may be negligible. This motivates researchers to consider new estimation methods that are specifically for sparse covariance matrices. We classify these estimators into category (3) and refer to them as *the sparse estimators*.

Stein-Type Estimators

When $p<n$, $\boldsymbol{S}_n$ is an unbiased estimator of $\boldsymbol{\Sigma}$. However, it is known that the eigenvalues of $\boldsymbol{S}_n$ tend to be more spread out than the eigenvalues of $\boldsymbol{\Sigma}$, especially when p is close to n. As a consequence, $\boldsymbol{S}_n$ may be unstable with the smallest estimated eigenvalues being too small and the largest too large.[48] For ease of exposition, we write

$$\boldsymbol{S}_n = \boldsymbol{U}_n \boldsymbol{\Lambda}_n \boldsymbol{U}_n^{\mathrm{T}},$$

where $\boldsymbol{U}_n$ is an orthogonal matrix, $\boldsymbol{\Lambda}_n = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ and $\lambda_1 \geq \ldots \geq \lambda_p$ are the eigenvalues of $\boldsymbol{S}_n$. Stein[47] proposed to shrink the eigenvalues of the sample covariance matrix to avoid the extreme eigenvalues. Specifically, he suggested to estimate the eigenvalues by

$$\widehat{\lambda}_j = \frac{n}{n-p+1+2\sum\limits_{i\neq j}\frac{\lambda_j}{\lambda_j-\lambda_i}}\lambda_j, \quad j=1,\ldots,p.$$

Letting $\widehat{\boldsymbol{\Lambda}}_n = \mathrm{diag}\left\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p\right\}$, the resulting estimator of $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{U}_n \widehat{\boldsymbol{\Lambda}}_n \boldsymbol{U}_n^{\mathrm{T}}. \tag{6}$$

The estimator 6 was derived by minimizing an unbiased estimate of the Stein loss function.[48] We refer to this estimator as the Stein estimator. Note that the Stein estimator does not preserve the order of the eigenvalues and the resulting eigenvalues can even be negative. Much research has been devoted to improve the Stein estimator. In particular, Haff[49] derived an estimator of $\boldsymbol{\Sigma}$ under the constraint that the order of the sample eigenvalues is maintained. Other Stein-type estimators can be found, for example, in Efron and Morris,[50] Dey and Srinivasan,[51] Yang and Berger,[52] Daniels and Kass,[53] Daniels and Kass,[48] and references therein.

Ridge-Type Estimators

When $p \geq n$, $\boldsymbol{S}_n$ is a singular matrix with the smallest eigenvalues being zero. In such settings, the Stein-type estimators are no longer applicable. To achieve an invertible estimate for the covariance matrix, Ledoit and Wolf[54] proposed to estimate $\boldsymbol{\Sigma}$ by the following ridge-type estimator,

$$\widetilde{\boldsymbol{\Sigma}} = \lambda_1 \boldsymbol{S}_n + \lambda_2 \boldsymbol{I}_p, \tag{7}$$

where $\boldsymbol{I}_p$ is the identity matrix of size p and λ_1 and λ_2 are shrinkage parameters. Under the squared loss function, they derived the optimal coefficients λ_1 and λ_2 and also proposed data-driven estimators for these coefficients. More recently, Fisher and Sun[55] considered a general convex combination of $\boldsymbol{S}_n$ and some target matrix $\boldsymbol{T}$,

$$\breve{\boldsymbol{\Sigma}} = \lambda \boldsymbol{S}_n + (1-\lambda)\boldsymbol{T}, \tag{8}$$

where $\lambda \in (0, 1)$ is the shrinkage parameter. The target matrix $\boldsymbol{T}$ is often chosen to be positive definite (and therefore non-singular) and well-conditioned. Consequently, the final estimator is also positive definite and well-conditioned for any dimensionality. Similar approaches can be found, for example, in Schäfer and Strimmer,[56] Warton,[57] Chen et al.,[58] Warton,[59] and references therein.

Sparse Estimators

For high-dimensional data with $p \gg n$, to have a good estimate of $\boldsymbol{\Sigma}$ one may have to rely on some sparsity assumptions about the covariance matrix. In such settings, the above-mentioned Stein-type and the ridge-type estimators are either no longer applicable or the improvement is nearly negligible. This suggests that new estimation methods are required for very large p. Let the covariance matrix be $\boldsymbol{\Sigma} = \{\sigma_{ij}\}_{p\times p}$ and the sample covariance matrix be $\boldsymbol{S}_n = \{s_{ij}\}_{p\times p}$. Under the sparsity conditions that most of the σ_{ij} are zero or close to zero, Bickel and Levina[60] proposed to estimate $\boldsymbol{\Sigma}$ by a thresholding method. Specifically, their estimator is

$$\boldsymbol{T}_s\left(\boldsymbol{S}_n\right) = \{s_{ij} I(|s_{ij}| \geq s)\}_{p\times p}, \tag{9}$$

where s is a tuning parameter serving as a threshold. The asymptotic properties of the proposed threshold estimator were established under some regularity conditions. Note that the threshold estimator in Bickel and Levina[60] can be regarded as a hard-thresholding estimator. Rothman et al.[61] considered a generalized thresholding rule that include hard and soft thresholding as in Donoho and Johnstone,[62] the Smoothly Clipped Absolute Deviation (SCAD) method in Fan and Li,[63] and the adaptive Least Absolute Shrinkage and Selection Operator (LASSO) method in Zou.[64] We note that a single threshold level was used for all the entries of the sample covariance matrix in the above approaches. More recently, Cai and Liu[65] proposed an adaptive thresholding method where the threshold level is entry-based and so the resulting estimator is more flexible. Other thresholding methods in the literature include, for example, Bickel and Levina,[66] Karoui,[67] Lam and Fan,[68] Cai and Zhou,[69] Cai and Yuan,[70] and references therein.

ESTIMATION OF THE PRECISION MATRIX

In many statistical analyses, we need an estimate of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ rather than an estimate of the covariance matrix. Examples include linear discriminant analysis,[3] Hotelling's T^2 test,[4] and Markowitz mean-variance analysis.[5]

In the special case when $\mathbf{\Sigma} = \text{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$, we have the diagonal precision matrix as $\Omega = \text{diag}\left(\sigma_1^{-2}, \ldots, \sigma_p^{-2}\right)$. Some methods for estimating the diagonal precision matrix have been proposed in the literature, e.g., the shrinkage estimators in Tong and Wang[22] and Tong et al.[23] For a general non-diagonal Ω, we can accordingly classify the existing estimators into three categories: (1) the Stein-type estimators, (2) the ridge-type estimators, and (3) the sparse estimators. The Stein-type estimators can be found, for example, in Dey[71] and Tsukuma and Konno,[72] and references therein. Due to space limitations, we will only provide a brief review of the ridge-type and sparse estimators.

Ridge-Type Estimators

Recall that an unbiased estimator of Ω is given by $\widehat{\Omega} = (n-p-2)\, S_n^{-1}/(n-1)$. Efron and Morris[50] proposed an empirical Bayesian estimator as

$$\widehat{\Omega}_{\text{EM}} = \frac{n-p-2}{n-1} S_n^{-1} + \frac{p^2+p-2}{(n-1)\,\text{tr}\left(S_n\right)} I_p. \qquad (10)$$

Similar methods can be found, for example, in Haff,[73] Haff,[74] Krishnamoorthy and Gupta,[75] Bodnar et al.,[76] and references therein. Note that all these estimators involve the term S_n^{-1} and so they apply to the situation when $p < n$ only.

When $p \geq n$, to overcome the singularity problem, Kubokawa and Srivastava[77] considered the following ridge-type estimator for the precision matrix,

$$\widehat{\Omega}_{\text{ridge}} = \alpha \left(S_n + \beta I_p\right)^{-1}, \qquad (11)$$

where α and β are two shrinkage coefficients. In their paper, an empirical Bayes approach was applied to estimate α and β. More recently, Wang et al.[78] proposed a data-driven estimator for the shrinkage coefficients using random matrix theory. Note that their proposed method is distribution-free. They further demonstrated in numerical studies that the proposed estimator performs better than the existing competitors in a wide range of settings.

Sparse Estimators

Let $a = (a_1, \ldots, a_p)^{\text{T}}$ be a vector and $A = \{a_{ij}\}_{p \times q}$ be a matrix. We define the element-wise l_1 norms as $|a|_1 = \sum_j |a_j|$ and $|A|_1 = \sum_{ij} |a_{ij}|$, and the l_∞ norms as $|a|_\infty = \max_j |a_j|$ and $|A|_\infty = \max_{ij} |a_{ij}|$. Various sparse estimators for Ω have been proposed in the recent literature. Most of them are based on a regularization approach. Banerjee et al.[79] proposed an l_1 penalized likelihood method:

$$\widehat{\Omega} = \text{argmin}_{\Omega > 0} \left\{\text{tr}\left(S_n \Omega\right) - \log|\Omega| + \lambda_n |\Omega|_1\right\}. \qquad (12)$$

Fan et al.[80] replaced the l_1 penalty in Eq. 12 by the SCAD penalty.[63] More recently, Cai et al.[81] considered another regularization method:

$$\text{Minimize } |\Omega|_1 \text{ subject to } |S_n\Omega - I_p| \leq \lambda_n, \qquad (13)$$

where λ_n is a tuning parameter. Other regularization methods include, for example, Yuan and Lin,[82] d'Aspremont et al.,[83] Friedman et al.,[84] Ravikumar et al.,[85] and references therein.

CONCLUSION

With the advent of high-throughput data such as microarrays, we are in an era of biotechnology innovation. Instead of working on a gene-by-gene basis, the microarray technology allows simultaneous monitoring of the whole genome. These data have motivated the development of reliable biomarkers for disease subtype classification and diagnosis, and for the identification of novel targets for drug treatment. Due to the cost and other experimental difficulties such as the availabilities of biological materials, microarray data are usually collected on a limited number of samples.

High-dimensional data such as microarray gene expression pose great challenges to traditional statistical and computational methods. In particular, the standard estimates of variances and covariances are usually unreliable. In this paper, we review some recent developments in the estimation of variances and covariances for high-dimensional data. The estimation of variances and covariances plays an important

© 2014 Wiley Periodicals, Inc.

role in statistical analysis including t-test, Hotelling's T^2-test, principal component analysis, and discriminant analysis. For instance, to test whether two gene sets are equal, Chen et al.[86] proposed a regularized Hotelling's test for both scenarios of $p < n$ and $p \geq n$; and Cai et al.[87] proposed another test statistic based on a linear transformation of the data by the precision matrix. In discriminant analysis, Friedman[88] proposed to use the ridge-type estimators of the covariance matrix. More recent works in this area include Guo et al.,[89] Shao et al.,[90] Fan et al.,[91] and among others. We have emphasized the applications to microarray data analysis. Nevertheless, the methods reviewed this paper have a wide variety of applications. For example, the estimation of precision matrix may be applied to graphical models.[45,65,81] We note that the review in this paper is selective and many other important approaches are not included due to space limitations.

There are many remaining challenges in the estimation of variances and covariances for high-dimensional data. For instance, the estimation usually involves unknown tuning parameters. Cross-validation and bootstrap methods have been proposed to select the tuning parameters. Most guidelines are based on simulation studies without theoretical justification.[92] Assumptions and evaluation criteria used in the estimation are somewhat arbitrary in the existing literature. This makes it difficult to have a fair comparison among estimation methods.

ACKNOWLEDGMENTS

Tiejun Tong's research was supported in part by Hong Kong Research Grant HKBU202711 and Hong Kong Baptist University FRG Grants FRG2/11-12/110 and FRG1/13-14/018. Cheng Wang's research was supported by NSF of China Grants 11101397, 71001095, and 11271347. Yuedong Wang's research was supported by NSF Grant DMS-07-06886. The authors thank the editor, the associate editor, and two referees for their constructive comments that led to a substantial improvement of this review article.

REFERENCES

1. Bai Z, Saranadasa H. Effect of high dimension: by an example of a two sample problem. *Stat Sin* 1996, 6:311–330.
2. Chen S, Zhang L, Zhong P. Tests for high-dimensional covariance matrices. *J Am Stat Assoc* 2010, 105:810–819.
3. Anderson TW. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. New York: Wiley; 2003.
4. Hotelling H. The generalization of student's ratio. *Ann Math Stat* 1931, 2:360–378.
5. Markowitz H. Portfolio selection. *J Finance* 1952, 7:77–91.
6. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 2003, 4:210.
7. Ayroles JF, Gibson G. Analysis of variance of microarray data. *Methods Enzymol* 2006, 411:214–233.
8. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 2002, 97: 77–87.
9. Bickel PJ, Levina E. Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 2004, 10:989–1010.
10. Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. *Comput Stat Data Anal* 2005, 48:869–885.
11. Pang H, Tong T, Zhao H. Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics* 2009, 65: 1021–1029.
12. Huang S, Tong T, Zhao H. Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics* 2010, 66:1096–1106.
13. Wu Y, Genton MG, Stefanski LA. A multivariate two-sample mean test for small sample size and missing data. *Biometrics* 2006, 62:877–885.
14. Srivastava MS, Du M. A test for the mean vector with fewer observations than the dimension. *J Multivariate Anal* 2008, 99:386–402.
15. Srivastava MS. A test for the mean vector with fewer observations than the dimension under non-normality. *J Multivariate Anal* 2009, 100:518–532.
16. Park J, Ayyala DN. A test for the mean vector in large dimension and small samples. *J Stat Plann Inference* 2013, 143:929–943.
17. Srivastava MS, Katayama S, Kano Y. A two sample test in high dimensional data. *J Multivariate Anal* 2013, 114:349–358.

© 2014 Wiley Periodicals, Inc.

18. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001, 98: 5116–5121.

19. Efron B, Tibshirani R, Storey JD, Tusher VG. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 2001, 96:1151–1160.

20. Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill GA. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 2005, 6:59–75.

21. James W, Stein C. Estimation with quadratic loss. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley, CA: University of California Press; 1961, 361–379.

22. Tong T, Wang Y. Optimal shrinkage estimation of variances with applications to microarray data analysis. *J Am Stat Assoc* 2007, 102:113–122.

23. Tong T, Jang H, Wang Y. James-Stein type estimators of variances. *J Multivariate Anal* 2012, 107:232–243.

24. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 2001, 17:509–519.

25. Lönnstedt I, Speed T. Replicated microarray data. *Stat Sin* 2002, 12:31–46.

26. Kendziorski CM, Newton MA, Lan H, Could MN. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med* 2003, 22:3899–3914.

27. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiment. *Stat Appl Genet Mol Biol* 2004, 3:1.

28. Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 2003, 19:2448–2455.

29. Hwang JTG, Liu P. Optimal tests shrinking both means and variances applicable to microarray data analysis. *Stat Appl Genet Mol Biol* 2010, 9:36.

30. Zhao Z. Double shrinkage empirical Bayesian estimation for unknown and unequal variances. *Stat Interface* 2010, 3:533–541.

31. Ji T, Liu P, Nettleton D. Borrowing information across genes and experiments for improved error variance estimation in microarray data analysis. *Stat Appl Genet Mol Biol* 2012, 11:12.

32. Chen Y, Dougherty ER, Bittner ML. Ratio-biased decisions and the quantitative analysis of cDNA microarray images. *J Biomed Opt* 1997, 2:364–374.

33. Rocke DM, Durbin B. A model for measurement error for gene expression arrays. *J Comput Biol* 2001, 8:557–569.

34. Strimmer K. Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics* 2003, 4:10.

35. Weng L, Dai H, Zhan Y, He Y, Stepaniants SB, Bassett DE. Rosetta error model for gene expression analysis. *Bioinformatics* 2006, 22:1111–1121.

36. Durbin B, Hardin J, Hawkins D, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 2002, 18:S105–S110.

37. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002, 18:S96–S104.

38. Rocke DM. Design and analysis of experiments with high throughput biological assay data. *Semin Cell Dev Biol* 2004, 15:708–713.

39. Rocke DM, Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* 2003, 19:966–972.

40. Durbin B, Rocke DM. Estimation of transformation parameters for microarray data. *Bioinformatics* 2003, 19:1360–1367.

41. Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* 2002, 18:1207–1215.

42. Carroll RJ, Wang Y. Nonparametric variance estimation in the analysis of microarray data: a measurement error approach. *Biometrika* 2008, 95:437–449.

43. Wang Y, Ma Y, Carroll RJ. Variance estimation in the analysis of microarray data. *J R Stat Soc B* 2009, 71:425–445.

44. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. New York: Chapman & Hall; 2006.

45. Fan J, Feng Y, Niu YS. Nonparametric estimation of genewise variance for microarray data. *Ann Stat* 2010, 38:2723–2750.

46. Fang Y, Zhu LX. Asymptotics of SIMEX-based variance estimation. *Metrika* 2012, 75:329–345.

47. Stein C. Estimation of a covariance matrix. Rietz lecture, 39th IMS Annual Meeting, Atlanta, Georgia 1975.

48. Daniels MJ, Kass RE. Shrinkage estimators for covariance matrices. *Biometrics* 2001, 57:1173–1184.

49. Haff LR. The variational form of certain Bayes estimators. *Ann Stat* 1991, 19:1163–1190.

50. Efron B, Morris C. Multivariate empirical Bayes and estimation of covariance matrices. *Ann Stat* 1976, 4:22–32.

51. Dey DK, Srinivasan C. Estimation of a covariance matrix under Stein's loss. *Ann Stat* 1985, 13:1581–1591.

© 2014 Wiley Periodicals, Inc.

52. Yang R, Berger JO. Estimation of a covariance matrix using the reference prior. *Ann Stat* 1994, 22:1195–1211.

53. Daniels MJ, Kass RE. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J Am Stat Assoc* 1999, 94:1254–1263.

54. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *J Multivariate Anal* 2004, 88:365–411.

55. Fisher TJ, Sun X. Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Comput Stat Data Anal* 2011, 55:1909–1918.

56. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol* 2005, 4:32.

57. Warton DI. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J Am Stat Assoc* 2008, 103:340–349.

58. Chen Y, Wiesel A, Hero AO. Shrinkage estimation of high dimensional covariance matrices. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009)*. Taipei, Taiwan: IEEE; 2009, 2937–2940.

59. Warton DI. Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics* 2011, 67:116–123.

60. Bickel PJ, Levina E. Covariance regularization by thresholding. *Ann Stat* 2008, 36:2577–2604.

61. Rothman AJ, Levina E, Zhu J. Generalized thresholding of large covariance matrices. *J Am Stat Assoc* 2009, 104:177–186.

62. Donoho DL, Johnstone JM. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 1994, 81:425–455.

63. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001, 96:1348–1360.

64. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006, 101:1418–1429.

65. Cai T, Liu W. Adaptive thresholding for sparse covariance matrix estimation. *J Am Stat Assoc* 2011, 106:672–684.

66. Bickel PJ, Levina E. Regularized estimation of large covariance matrices. *Ann Stat* 2008, 36:199–227.

67. Karoui NE. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann Stat* 2008, 36:2717–2756.

68. Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann Stat* 2009, 37:42–54.

69. Cai T, Zhou H. Optimal rates of convergence for sparse covariance matrix estimation. *Ann Stat* 2012, 40:2389–2420.

70. Cai T, Yuan M. Adaptive covariance matrix estimation through block thresholding. *Ann Stat* 2012, 40:2014–2042.

71. Dey DK. Improved estimation of a multinormal precision matrix. *Stat Probab Lett* 1987, 6:125–128.

72. Tsukuma H, Konno Y. On improved estimation of normal precision matrix and discriminant coefficients. *J Multivariate Anal* 2006, 97:1477–1500.

73. Haff LR. Minimax estimators for a multinormal precision matrix. *J Multivariate Anal* 1977, 7:374–385.

74. Haff LR. An identity for the Wishart distribution with applications. *J Multivariate Anal* 1979, 9:531–544.

75. Krishnamoorthy K, Gupta A. Improved minimax estimation of a normal precision matrix. *Can J Stat* 1989, 17:91–102.

76. Bodnar T, Gupta AK, Parolya N. Optimal linear shrinkage estimator for large dimensional precision matrix. arXiv preprint arXiv:1308.0931.

77. Kubokawa T, Srivastava MS. Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data. *J Multivariate Anal* 2008, 99:1906–1928.

78. Wang C, Pan G, Tong T, Zhu L. Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Stat Sin*. in press.

79. Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J Mach Learn Res* 2008, 9:485–516.

80. Fan J, Feng Y, Wu Y. Network exploration via the adaptive LASSO and SCAD penalties. *Ann Appl Stat* 2009, 3:521–541.

81. Cai T, Liu W, Luo X. A constrained L_1 minimization approach to sparse precision matrix estimation. *J Am Stat Assoc* 2011, 106:594–607.

82. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika* 2007, 94:19–35.

83. d'Aspremont A, Banerjee O, El Ghaoui L. First-order methods for sparse covariance selection. *SIAM J Matrix Anal Appl* 2008, 30:56–66.

84. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2008, 9:432–441.

85. Ravikumar P, Wainwright MJ, Raskutti G, Yu B. High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electron J Stat* 2011, 5:935–980.

86. Chen LS, Paul D, Prentice RL, Wang P. A regularized hotelling's T^2 test for pathway analysis in proteomic studies. *J Am Stat Assoc* 2011, 106:1345–1360.

87. Cai T, Liu W, Xia Y. Two-sample test of high dimensional means under dependence. *J R Stat Soc B* 2014, 76:349–372.

88. Friedman J. Regularized discriminant analysis. *J Am Stat Assoc* 1989, 84:165–175.

89. Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 2007, 8:86–100.

90. Shao J, Wang Y, Deng X, Wang S. Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann Stat* 2011, 39:1241–1265.

91. Fan Y, Jin J, Yao Z. Optimal classification in sparse Gaussian graphic model. *Ann Stat* 2013, 41:2537–2571.

92. Fang Y, Wang B, Feng Y. Tuning parameter selection in regularized estimations of large covariance matrices. arXiv preprint arXiv:1308.3416, 2013.