8-2014

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

Tiejun Tong
Hong Kong Baptist University, tongt@hkbu.edu.hk

Cheng Wang
Hong Kong Baptist University

Yuedong Wang
University of California

Link to published article: http://dx.doi.org/10.1002/wics.1308

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

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

Keywords

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

Abstract

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.

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. This kind of data is 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 $Y_i = (Y_{i1}, \ldots, Y_{ip})^T$ be independent random samples from a multivariate normal distribution[3,18],

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

*Corresponding to: yuedong@pstat.ucsb.edu
[1]Department of Mathematics, Hong Kong Baptist University, Hong Kong
[2]Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA

where $\mu = (\mu_1, \ldots, \mu_p)^T$ is a p-dimensional mean vector, Σ is a $p \times p$ positive definite covariance matrix, $X_i = (X_{i1}, \ldots, X_{ip})^T$, 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[1], Hotelling's T^2 test[48], and Markowitz mean-variance analysis[63]. We write the sample covariance matrix as

$$S_n = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})^T,$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ 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, $E(S_n^{-1}) = (n-1)\Omega/(n-p-2)$. 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 Σ, (2) the estimation of the covariance matrix Σ, and (3) the estimation of the precision matrix Ω.

Estimation of Variances

As reviewed in Cui and Churchill[22] and Ayroles and Gibson[2], 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 $\mathrm{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, 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 Σ, 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 $T_j = \sqrt{n}\bar{Y}_j/s_j$ where $\bar{Y}_j$ is the gene-specific sample mean and s_j^2 is the gene-specific sample variance. Then all we need is an estimate of D rather than the whole covariance matrix Σ. The second example is the class prediction (or classification) problem. If we use DLDA for class prediction[30], then again we need to estimate D rather than Σ. For more details about DLDA and its variants, see Bickel and Levina[6], Lee et al.[60], Pang et al.[64] and Huang et al.[49]. 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 D is required. For more details, see for example Wu et al.[89], Srivastava and Du[75], Srivastava[74], Park and Ayyala[65], and Srivastava et al.[76].

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.[82] in 2001. In order to avoid the undue influence of the small variance estimates, Tusher et al.[82] 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.[33] suggested to use the 90th percentile of all estimated standard deviations while Cui and Churchill[22] suggested to use the pooled sample variance.

In 2005, Cui et al.[23] proposed a James-Stein shrinkage estimator for the variances. For microarray data with $Y_{ij} \overset{iid}{\sim} N(\mu_j, \sigma_j^2)$, we have $s_j^2 = \sigma_j^2 \chi_{j,\nu}^2/\nu$, where for ease of notation, $\chi_{j,\nu}^2$ denote i.i.d. random variables which have a chi-squared distribution with $\nu = n - 1$ degrees of freedom. Taking the log-transformation leads to

$$Z_j = \ln \sigma_j^2 + \epsilon_j, \tag{2}$$

where $Z_j = \ln(s_j^2) - m$, $\epsilon_j = \ln(\chi_{j,\nu}^2/\nu) - m$ and $m = \mathrm{E}\{\ln(\chi_{j,\nu}^2/\nu)\}$. Treating Z_j in (2) as normal random variables, the James-Stein shrinkage method[52] 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

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

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

Note that Z_j in (2) can be far from normal when ν is small. Therefore, the shrinkage variance estimates (3) can be suboptimal. Note also that the variance estimates appear in the denominator of the t tests. Tong and Wang[80] 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 $(\sigma_j^2)^t$ 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 $s_j^{2t} = (s_j^2)^t$, $s_{pool}^{2t} = \prod_{j=1}^{p}(s_j^2)^{t/p}$ and

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

where $\Gamma(\cdot)$ is the Gamma function. Tong and Wang[80] proposed the following family of shrinkage estimators for $(\sigma_j^2)^t$,

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

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 $t = 1$, $\hat{\sigma}_j^2$ is a simple modification of the estimator in Cui et al.[23]. 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 $h_p(t)s_{pool}^{2t}$. 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.[79] proposed another James-Stein shrinkage estimator for the variances that shrunk the individual sample variance towards 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[4] applied a Bayesian method to improve the estimation of variances. Specifically, they assumed the following conjugate prior for (μ_j, σ_j^2),

$$p(\mu_j, \sigma_j^2|\alpha) = N(\mu_j; \mu_0, \sigma_j^2/\lambda_0)\mathcal{I}(\sigma_j^2; \nu_0, \sigma_0^2),$$

where $\alpha = (\mu_0, \lambda_0, \nu_0, \sigma_0^2)$ 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 (μ_j, σ_j^2) has the same functional form as the prior density

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

where $\lambda_n = \lambda_0 + n$, $\nu_n = \nu_0 + n$ and

$$\mu_n = \frac{\lambda_0}{\lambda_0 + n}\mu_0 + \frac{n}{\lambda_0 + n}\bar{Y}_j,$$
$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n-1)s_j^2 + \frac{\lambda_0 n}{\lambda_0 + n}(\bar{Y}_j - \mu_0)^2.$$

Note that the posterior mean μ_n is a weighted average of the prior mean μ_0 and the sample mean $\bar{Y}_j$. Baldi and Long[4] suggested to use $\mu_0 = \bar{Y}_j$. This leads to the posterior means of μ_i and σ_i^2 as

$$\hat{\mu}_j = \bar{Y}_j \quad \text{and} \quad \hat{\sigma}_j^2 = \frac{\nu_0 \sigma_0^2 + (n-2)s_j^2}{\nu_0 + n - 2}.$$

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 ν_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[61] 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.[55] extended the empirical Bayes method using the hierarchical gamma-gamma and lognormal-normal models. Smyth[73] developed hierarchical models in the context of general linear models; see also Wright and Simon[88]. Hwang and Liu[51] and Zhao[92] applied some empirical Bayes approaches that shrunk both means and variances. Ji et al.[53] 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[19,67,78,87]. 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.[32], Huber et al.[50], Rocke[69], Rocke and Durbin[68], and Durbin and Rocke[31].

Another remedy is to apply the regression method. Specifically, we assume a functional relationship between the mean and the variance: $\sigma_i^2 = g(\mu_i)$. 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 nonparametrically. 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[19], $g(\mu) = \theta\mu^2$, and the quadratic model[67,20], $g(\mu) = \theta_1 + \theta_2\mu^2$. Often it is difficult, if not impossible, to specify a parametric model for g. A nonparametric regression approach may be used in these situations. Any one of the nonparametric regression approaches such as smoothing splines and local polynomials could be used to model g nonparametrically.

Estimation of the parameter θ or the nonparametric 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 $\hat{\theta}(\mu)$ and $\hat{g}(\mu)$ as consistent estimators for θ and g respectively, where the dependence on $\mu = (\mu_1, \cdots, \mu_p)^T$ is expressed explicitly. In practice μ is unknown. The sample mean $\bar{Y} = (\bar{Y}_1, \cdots, \bar{Y}_p)^T$ is a natural estimate of μ. Then a direct approach is to replace μ by $\bar{Y}$ which leads to the estimates $\hat{\theta}(\bar{Y})$ and $\hat{g}(\bar{Y})$ for θ and g. Unfortunately, these naive estimates $\hat{\theta}(\bar{Y})$ and $\hat{g}(\bar{Y})$ are in general inconsistent since the sampling error in $\bar{Y}$ is ignored[16,84].

Regarding $\bar{Y}$ as an error-prone unbiased measure of μ, the problem can be cast in a general framework of heteroscedastic measurement error. Therefore, the SIMEX (simulation extrapolation) method in Carroll et al.[15] 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[16,84]. To overcome this problem, Carroll and Wang[16] and Wang et al.[84] proposed permutation SIMEX methods which lead to consistent estimates of θ and g. Other regression methods include Fan et al.[35] and Fang and Zhu[38].

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 Σ. 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 Σ. 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 Σ. 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), since 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[77] 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 $\lambda S_n + (1-\lambda) I_p$, where 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 shrinkage estimators for non-sparse Σ 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 $p < n$, S_n is an unbiased estimator of Σ. However, it is known that the eigenvalues of S_n tend to be more spread out than the eigenvalues of Σ, especially when p is close to n. As a consequence, S_n may be unstable with the smallest estimated eigenvalues being too small and the largest too large[25]. For ease of exposition, we write

$$S_n = U_n \Lambda_n U_n^T,$$

where U_n is an orthogonal matrix, $\Lambda_n = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ and $\lambda_1 \geq \cdots \geq \lambda_p$ are the eigenvalues of S_n. Stein[77] proposed to shrink the eigenvalues of the sample covariance matrix to avoid the extreme eigenvalues. Specifically, he suggested to estimate the eigenvalues by

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

Letting $\hat{\Lambda}_n = \mathrm{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_p\}$, the resulting estimator of Σ is

$$\hat{\Sigma} = U_n \hat{\Lambda}_n U_n^T. \tag{6}$$

The estimator (6) was derived by minimizing an unbiased estimate of the Stein loss function[25]. 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[47] derived an estimator of Σ 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[34], Dey and Srinivasan[27], Yang and Berger[90], Daniels and Kass[24], Daniels and Kass[25], and references therein.

Ridge-type estimators

When $p \geq n$, 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[59] proposed to estimate Σ by the following ridge-type estimator,

$$\tilde{\Sigma} = \lambda_1 S_n + \lambda_2 I_p, \tag{7}$$

where 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[41] considered a general convex combination of S_n and some target matrix T,

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

where $\lambda \in (0,1)$ is the shrinkage parameter. The target matrix T is often chosen to be positive definite (and therefore nonsingular) 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[71], Warton[85], Chen et al.[21], Warton[86], and references therein.

Sparse estimators

For high-dimensional data with $p \gg n$, to have a good estimate of Σ one may have to rely on some sparsity assumptions about the covariance matrix. In such settings, the above 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 $\Sigma = \{\sigma_{ij}\}_{p\times p}$ and the sample covariance matrix be $S_n = \{s_{ij}\}_{p\times p}$. Under the sparsity conditions that most of the σ_{ij} are zero or close to zero, Bickel and Levina[7] proposed to estimate Σ by a thresholding method. Specifically, their estimator is

$$T_s(S_n) = \{s_{ij} I(|s_{ij}| \geq s)\}_{p\times p}, \tag{9}$$

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[7] can be regarded as a hard-thresholding estimator. Rothman et al.[70] considered a generalized thresholding rule that include hard- and soft-thresholding as in Donoho and Johnstone[29], the SCAD method in Fan and Li[37], and the adaptive LASSO method in Zou[93]. 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[10] 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[8], Karoui[54], Lam and Fan[58], Cai and Zhou[14], Cai and Yuan[13], and references therein.

Estimation of the Precision Matrix

In many statistical analyses, we need an estimate of the precision matrix $\Omega = \Sigma^{-1}$ rather than an estimate of the covariance matrix. Examples include linear discriminant analysis[1], Hotelling's T^2 test[48], and Markowitz mean-variance analysis[63].

In the special case when $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, we have the diagonal precision matrix as $\Omega = \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_p^{-2})$. Some methods for estimating the diagonal precision matrix have been proposed in the literature, e.g., the shrinkage estimators in Tong and Wang[80] and Tong et al.[79]. 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[28] and Tsukuma and Konno[81], 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 $\hat{\Omega} = (n-p-2)S_n^{-1}/(n-1)$. Efron and Morris[34] proposed an empirical Bayesian estimator as

$$\hat{\Omega}_{\mathrm{EM}} = \frac{n-p-2}{n-1} S_n^{-1} + \frac{p^2+p-2}{(n-1)\mathrm{tr}(S_n)} I_p. \tag{10}$$

Similar methods can be found, for example, in Haff[45], Haff[46], Krishnamoorthy and Gupta[56], Bodnar et al.[9], 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[57] considered the following ridge-type estimator for the precision matrix,

$$\hat{\Omega}_{\mathrm{ridge}} = \alpha(S_n + \beta I_p)^{-1}, \tag{11}$$

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.[83] 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 $a = (a_1, \ldots, a_p)^T$ 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.[5] proposed an ℓ_1 penalized likelihood method:

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

Fan et al.[36] replaced the ℓ_1 penalty in (12) by the SCAD penalty[37]. More recently, Cai et al.[11] considered another regularization method:

$$\text{minimize } |\Omega|_1 \text{ subject to } |S_n\Omega - I_p| \leq \lambda_n, \tag{13}$$

where λ_n is a tuning parameter. Other regularization methods include, for example, Yuan and Lin[91], d'Aspremont et al.[26], Friedman et al.[43], Ravikumar et al.[66], 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 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.[17] proposed a regularized Hotelling's test for both scenarios of $p < n$ and $p \geq n$; and Cai et al.[12] proposed another test statistic based on a linear transformation of the data by the precision matrix. In discriminant analysis, Friedman[42] proposed to use the ridge-type estimators of the covariance matrix. More recent works in this area include Guo et al.[44], Shao et al.[72], Fan et al.[39], 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[35,10,11]. 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[40]. 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.

Acknowledgements

add the acknowledgements here.

References

1. Anderson TW. An Introduction to Multivariate Statistical Analysis, 3rd ed., 2003, New York: Wiley.

2. Ayroles JF, Gibson G. Analysis of Variance of Microarray Data *Methods in Enzymology* 2006, 411:214C233.

3. Bai Z, Saranadasa H. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* 1996, 6:311-330.

4. 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.

5. Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research* 2008, 9:485-516.

6. 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.

7. Bickel PJ, Levina E. Covariance regularization by thresholding. *Annals of Statistics* 2008, 36:2577-2604.

8. Bickel PJ, Levina E. Regularized estimation of large covariance matrices. *Annals of Statistics* 2008, 36:199-227.

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

10. Cai T, Liu W. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 2011, 106:672-684.

11. Cai T, Liu W, Luo X. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* 2011, 106:594-607.

12. Cai T, Liu W, Xia Y. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B* 2014, 76:349-372.

13. Cai T, Yuan M. Adaptive covariance matrix estimation through block thresholding. *Annals of Statistics* 2012, 40:2014-2042.

14. Cai T, Zhou H. Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics* 2012, 40:2389-2420.

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

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

17. Chen LS, Paul D, Prentice RL, Wang P. A Regularized Hotelling's T^2 Test for Pathway Analysis in Proteomic Studies. *Journal of the American Statistical Association* 2011, 106:1345-1360.

18. Chen S, Zhang L, Zhong P. Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association* 2010, 105:810-819.

19. Chen Y, Dougherty ER, Bittner ML. Ratio-biased decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* 1997, 2:364-374.

20. 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.

21. Chen Y, Wiesel A, Hero AO. Shrinkage estimation of high dimensional covariance matrices. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, 2937-2940. IEEE.

22. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* 2003, 4:210.

23. 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.

24. Daniels MJ, Kass RE. Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association* 1999, 94:1254-1263.

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

26. d'Aspremont A, Banerjee O, El Ghaoui L. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications* 2008, 30:56-66.

27. Dey DK, Srinivasan C. Estimation of a covariance matrix under Stein's loss. *Annals of Statistics* 1985, 13:1581-1591.

28. Dey DK, Srinivasan C. Improved estimation of a multinormal precision matrix. *Statistics & Probability Letters* 1987, 6:125-128.

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

30. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 2002, 97:77-87.

31. Durbin B, Rocke DM. Estimation of Transformation Parameters for Microarray Data. *Bioinformatics* 2003, 19:1360-1367.

32. Durbin B, Hardin J, Hawkins D, Rocke DM. A Variance-Stabilizing Transformation for Gene-Expression Microarray Data. *Bioinformatics* 2002, 18:S105-S110.

33. Efron B, Tibshirani R, Storey JD, Tusher VG. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001, 96:1151-1160.

34. Efron B, Morris C. Multivariate empirical bayes and estimation of covariance matrices. *Annals of Statistics* 1976, 4:22-32.

35. Fan J, Feng Y, Niu YS. Nonparametric estimation of genewise variance for microarray data. *Annals of Statistics* 2010, 38:2723-2750.

36. Fan J, Feng Y, Wu Y. Network exploration via the adaptive lasso and scad penalties. *Annals of Applied Statistics* 2009, 3:521-541.

37. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001, 96:1348-1360.

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

39. Fan Y, Jin J, Yao Z. Optimal classification in sparse Gaussian graphic model. *Annals of Statistics* 2013, 41:2537-2571.

40. Fang Y, Wang B, Feng Y. Tuning Parameter Selection in Regularized Estimations of Large Covariance Matrices. *arXiv preprint arXiv:1308.3416*, 2013.

41. Fisher TJ, Sun X. Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics & Data Analysis* 2011, 55:1909-1918.

42. Friedman J. Regularized discriminant analysis. *Journal of the American Statistical Association* 1989, 84:165-175.

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

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

45. Haff LR. Minimax estimators for a multinormal precision matrix. *Journal of Multivariate Analysis* 1977, 7:374-385.

46. Haff LR. An identity for the Wishart distribution with applications. *Journal of Multivariate Analysis* 1979, 9:531-544.

47. Haff LR. The variational form of certain Bayes estimators. *Annals of Statistics* 1991, 19:1163-1190.

48. Hotelling H. The generalization of student's ratio. *Annals of Mathematical Statistics* 1931, 2:360-378.

49. Huang S, Tong T, Zhao H. Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics* 2010, 66:1096-1106.

50. 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.

51. Hwang JTG, Liu P. Optimal Tests Shrinking Both Means and Variances Applicable to Microarray Data Analysis. emphStatistical Applications in Genetics and Molecular Biology 2010, 9:36.

52. James W, Stein C. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1961, 1:361-379.

53. Ji T, Liu P, Nettleton D. Borrowing information across genes and experiments for improved error variance estimation in microarray data analysis. *Statistical Applications in Genetics and Molecular Biology* 2012, 11:12.

54. Karoui NE. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics* 2008, 36:2717-2756.

55. Kendziorski CM, Newton MA, Lan H, Could MN. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* 2003, 22:3899-3914.

56. Krishnamoorthy K, Gupta A. Improved minimax estimation of a normal precision matrix. *Canadian Journal of Statistics* 1989, 17:91-102.

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

58. Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics* 2009, 37:42-54.

59. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis* 2004, 88:365-411.

60. Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis* 2005, 48:869-885.

61. Lönnstedt I, Speed T. Replicated microarray data. *Statistica Sinica* 2002, 12:31-46.

62. Maatta JM, Casella G. Developments in decision-theoretic variance estimation. *Statistical Science* 1990, 5:90-101.

63. Markowitz H. Portfolio selection. *The Journal of Finance* 1952, 7:77-91.

64. Pang H, Tong T, Zhao H. Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics* 2009, 65:1021-1029.

65. Park J, Nag Ayyala D. A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning and Inference* 2013, 143:929-943.

66. Ravikumar P, Wainwright MJ, Raskutti G, Yu B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* 2011, 5:935-980.

67. Rocke DM, Durbin B. A model for measurement error for gene expression arrays. *Journal of Computational Biology* 2001, 8:557-569.

68. Rocke DM, Durbin B. Approximate Variance-Stabilizing Transformations for Gene-Expression Microarray Data. *Bioinformatics* 2003, 19:966-972.

69. Rocke DM. Design and Analysis of Experiments with High Throughput Biological Assay Data. *Seminars in Cell and Developmental Biology* 2004, 15:708-713.

70. Rothman AJ, Levina E, Zhu J. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* 2009, 104:177-186.

71. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 2005, 4:32.

72. Shao J, Wang Y, Deng X, Wang S. Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics* 2011, 39:1241-1265.

73. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiment. *Statistical Applications in Genetics and Molecular Biology* 2004, 3:1.

74. Srivastava MS. A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis* 2009, 100:518-532.

75. Srivastava MS, Du M. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* 2008, 99:386-402.

76. Srivastava MS, Katayama S, Kano Y. A two sample test in high dimensional data. *Journal of Multivariate Analysis* 2013, 114:349-358.

77. Stein C. Reitz lecture. In *39th Annual Meeting IMS. Atlanta, Georgia* 1975.

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

79. Tong T, Jang H, Wang Y. James-Stein Type Estimators of Variances. *Journal of Multivariate Analysis* 2012, 107:232-243.

80. Tong T, Wang Y. Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association* 2007, 102:113-122.

81. Tsukuma H, Konno Y. On improved estimation of normal precision matrix and discriminant coefficients. *Journal of Multivariate Analysis* 2006, 97:1477-1500.

82. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of National Academic Science* 2001, 98:5116-5121.

83. Wang C, Pan G, Tong T, Zhu L. Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Statistica Sinica*, in revision.

84. Wang Y, Ma Y, Carroll RJ. Variance estimation in the analysis of microarray data. *Journal of the Royal Statistical Society, Series B* 2009, 71:425-445.

85. Warton DI. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association* 2008, 103:340-349.

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

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

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

89. Wu Y, Genton MG, Stefanski LA. A multivariate two-sample mean test for small sample size and missing data. *Biometrics* 2006, 62:877-885.

90. Yang R, Berger JO. Estimation of a covariance matrix using the reference prior. *Annals of Statistics* 1994, 22:1195-1211.

91. Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. *Biometrika* 2007, 94:19-35.

92. Zhao Z. Double shrinkage empirical Bayesian estimation for unknown and unequal variances. *Statistics and Its Interface* 2010, 3:533-541.

93. Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical association* 2006, 101:1418-1429.