Journal of Statistical Planning and Inference 139 (2009) 2284 – 2292

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Regularized parameter estimation of high dimensional *t* distribution

Ming Yuan[a,*], Jianhua Z. Huang[b]

[a]*School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive NW, Atlanta, GA 30332, USA*
[b]*Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843, USA*

ARTICLE INFO

Article history:
Received 31 August 2008
Received in revised form
22 October 2008
Accepted 23 October 2008
Available online 31 October 2008

Keywords:
EM algorithm
Empirical Bayes
Multivariate *t* distribution
Penalized likelihood

ABSTRACT

We propose penalized-likelihood methods for parameter estimation of high dimensional *t* distribution. First, we show that a general class of commonly used shrinkage covariance matrix estimators for multivariate normal can be obtained as penalized-likelihood estimator with a penalty that is proportional to the entropy loss between the estimate and an appropriately chosen shrinkage target. Motivated by this fact, we then consider applying this penalty to multivariate *t* distribution. The penalized estimate can be computed efficiently using EM algorithm for given tuning parameters. It can also be viewed as an empirical Bayes estimator. Taking advantage of its Bayesian interpretation, we propose a variant of the method of moments to effectively elicit the tuning parameters. Simulations and real data analysis demonstrate the competitive performance of the new methods.

1. Introduction

Multivariate *t* distribution is widely used in practice to model features such as fat tails, and excess kurtosis that are often observed in real applications such as internet traffic and finance, but cannot be captured by multivariate normal. Although useful from modeling perspective, the practical use of multivariate *t* distribution is often limited by the difficulty in parameter estimation, particularly so for high dimensional data.

Let $y=(y^{(1)},\ldots,y^{(p)})'$ be a p-dimensional random vector following a multivariate t distribution with ν degrees of freedom, i.e., with density function given by

$$f(y|\mu,\Psi)=\frac{\Gamma\left(\frac{\nu+p}{2}\right)|\Psi|^{-1/2}}{(\pi\nu)^{p/2}\Gamma(\nu/2)\left[1+\frac{1}{\nu}(y-\mu)'\Psi^{-1}(y-\mu)\right]^{(\nu+p)/2}}, \tag{1}$$

where the parameters μ and Ψ are often referred to as the location vector and scale matrix, respectively. Given a random sample $y_1,\ldots,y_n$ of y, we wish to estimate μ and Ψ. The most commonly used estimator is the maximum-likelihood estimate (MLE), which can be computed efficiently thanks to the EM algorithm first proposed by Liu and Rubin (1995). The performance of MLE, however, quickly deteriorates as the dimensionality p increases since the number of unknowns grows quadratically in p. Even worse, the MLE of Ψ may not be guaranteed to be positive definite for moderate or large p's.

In this paper, we propose a penalized-likelihood estimator for multivariate *t* distribution. We introduce a novel penalty that is proportional to the entropy loss between the estimate and an appropriately chosen shrinkage target. In the limiting case of multivariate normal which corresponds to multivariate *t* distribution with the degree of freedom approaches infinity, we show

* Corresponding author.
E-mail address: myuan@isye.gatech.edu (M. Yuan).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.10.014

that the proposed estimator has several popular covariance matrix estimators such as those of Haff (1980) and Ledoit and Wolf (2004) as special cases. In more general situations when the degree of freedom is finite, we show that similar to MLE, the proposed estimator can be computed efficiently using an EM algorithm. Furthermore, we demonstrate that the proposed estimate has a natural Bayesian interpretation. We propose an empirical Bayes strategy that is similar in spirit to the method of moments to select the tuning parameters. We demonstrate the effectiveness of the new estimator using simulations and real data analysis.

2. Method

2.1. Shrinkage estimator of the multivariate normal covariance matrix

To motivate our approach of regularized parameter estimation for multivariate t distribution, we first consider the limiting case of multivariate normal which amounts to letting the degree of freedom ν go to infinity. Suppose $y_1, \ldots, y_n \sim \mathrm{N}_p(\mu, \Sigma)$. The most popular estimate of the mean and covariance matrix is the MLE $\mu = \bar{y}$ and $S = (1/n)\sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})'$.

Although successful for lower dimensional problems, the performance of S quickly deteriorates as the dimension increases. In particular when $p \geqslant n - 1$, S is not positive definite. A number of approaches (Ledoit and Wolf, 2004; Huang et al., 2006; Yuan and Lin, 2007; Bickel and Levina, 2008) have been introduced in recent years to overcome the difficulties associated with high dimensions. A general strategy that is proven useful in this setup is through regularization:

$$(\hat{\mu}, \widehat{\Sigma}) = \min_{\mu, \Sigma \succ 0} \left\{ -\frac{2}{n} \log \text{ likelihood} + \lambda \text{ pen}(\Sigma) \right\}, \tag{2}$$

where $\Sigma \succ 0$ means that Σ is positive definite, and the penalty function $\text{pen}(\cdot)$ is chosen so that the covariance matrix estimate is shrunken towards a well-conditioned target. Denote Ω an appropriately chosen shrinking target matrix. We propose the following penalty function:

$$\text{pen}(\Sigma) = \ln |\Sigma| + \text{trace}(\Sigma^{-1}\Omega). \tag{3}$$

Note that this penalty function differs from the entropy loss between Σ and Ω only by a constant not depending on Σ. The penalty has several desirable properties. It is strictly convex over the cone of positive definite matrices, which ensures that the objective function of (2) has a unique minimizer and the optimization is computationally tractable. The penalty itself has a unique minimizer Ω, which induces shrinkage towards the pre-specified target.

Note that

$$-\frac{2}{n} \log \text{ likelihood} + \lambda\{\ln |\Sigma| + \text{trace}(\Sigma^{-1}\Omega)\} = \ln |\Sigma| + \frac{1}{n}\sum_{i=1} (y_i - \mu)'\Sigma^{-1}(y_i - \mu) + \lambda\{\ln |\Sigma| + \text{trace}(\Sigma^{-1}\Omega)\}. \tag{4}$$

The minimizing μ is $\hat{\mu} = \bar{y}$. The penalized-likelihood estimator of Σ therefore minimizes

$$\ln |\Sigma| + \text{trace}(\Sigma^{-1}S) + \lambda\{\ln |\Sigma| + \text{trace}(\Sigma^{-1}\Omega)\} = -(1 + \lambda) \ln |C| + \text{trace}\{C(S + \lambda\Omega)\}, \tag{5}$$

where $C = \Sigma^{-1}$ is the so-called concentration or precision matrix. Using the fact that $\partial \ln |C|/\partial C = C^{-1}$ and $\partial \,\text{trace}(CA)/\partial C = A$ for a comformable matrix A, it is easy to show that the penalized-likelihood estimate (PLE) of Σ is

$$\widehat{\Sigma} = \frac{1}{1 + \lambda} S + \frac{\lambda}{1 + \lambda} \Omega. \tag{6}$$

A particularly popular class of shrinkage estimator for the covariance matrix takes the form $\widehat{\Sigma} = aS + bI$ for some $a, b > 0$ where I is an identity matrix of appropriate dimension (Haff, 1980; Ledoit and Wolf, 2004). These estimates can also be cast as penalized-likelihood estimators within the proposed framework. To elaborate, consider the shrinking target being an appropriately scaled identity matrix, i.e., $\Omega = \gamma I$, where $\gamma > 0$ is an additional tuning parameter. It is clear that now (6) becomes a linear combination of S and I:

$$\widehat{\Sigma} = \frac{1}{1 + \lambda} S + \frac{\lambda\gamma}{1 + \lambda} I. \tag{7}$$

Haff (1980) introduced a family of empirical Bayes covariance matrix estimator. Under the assumption that $p < n$, the estimator is given by

$$\widehat{\Sigma}_{\text{Haff}} = \frac{pn - 2n - 2}{pn^2 |S|^{1/p}} I + \frac{n}{n + 1} S. \tag{8}$$

It is not hard to see that our PLE reduces to (8) with $\lambda = 1/n$ and

$$\gamma = \frac{(n + 1)(pn - 2n - 2)}{pn^2 |S|^{1/p}}. \tag{9}$$

Another special case is the one proposed by Ledoit and Wolf (2004), which is given by

$$\widehat{\Sigma}_{\mathrm{LW}} = \frac{b_n^2}{d_n^2} m_n I + \frac{a_n^2}{d_n^2} S, \tag{10}$$

where

$$m_n = \mathrm{trace}(S)/p, \quad d_n = \|S - m_n I\|_F^2,$$
$$b_n^2 = \min\left(d_n^2, \frac{1}{n}\sum_{i=1}^n \|(y_i - \bar{y})(y_i - \bar{y})' - S\|_F^2 \right),$$

$a_n^2 = d_n^2 - b_n^2$, and $\|\cdot\|_F$ stands for the Frobenius norm. This estimator amounts to (6) with $\lambda = b_n^2/a_n^2$ and $\gamma = m_n$.

To fix ideas, in what follows, we shall assume that $\Omega = \gamma I$. The methodological developments presented, however, can also be applied to other choices of Ω.

2.2. Penalized-likelihood estimation for multivariate t distribution

Since the multivariate normal distribution is the limit of multivariate t distribution as the degree of freedom $v \to \infty$, it is reasonable to extend the penalized likelihood (2) to the multivariate t distributions. We thus consider the following PLE:

$$\min_{\mu,\Psi} -\frac{2}{n}\sum_{i=1}^n l(y_i; \mu, \Psi) + \lambda\ \mathrm{pen}(\Psi) \quad \text{subject to } \Psi \text{ being positive definite}, \tag{11}$$

where $l(y_i; \mu, \Psi) = \ln f(y_i|\mu, \Psi)$ is the contribution of y_i to the log likelihood and $f(y_i|\mu, \Psi)$ is defined as in (1). Hereafter we write $\mathrm{pen}_\gamma(\cdot)$ instead of $\mathrm{pen}(\cdot)$ to emphasize its dependence on parameter γ through the choice of $\Omega = \gamma I$.

Similar to MLE, the proposed PLE does not have a closed form. To develop a computational algorithm, we take advantage of the fact that multivariate t distribution belongs to the family of scale mixture of normals. In particular, if $y|\tau \sim \mathrm{N}_p(\mu, \Psi/\tau)$ and $\tau \sim \chi_v^2/v$, then $y \sim t_p(\mu, \Psi, v)$, i.e., y's density is given by (1). This property suggests that we can augment the data and treat the problem as a missing data problem, then the EM algorithm (Dempster et al., 1977) can be applied for easy computation of the estimator. More precisely, for observed data $y_{\mathrm{obs}} = \{y_1, \ldots, y_n\}$, where $y_i \sim t_p(\mu, \Psi, v)$, independently, we augment y_{obs} to $y_{\mathrm{com}} = \{(y_1, \tau_1), \ldots, (y_n, \tau_n)\}$. If the weights $\tau_1, \ldots, \tau_n$ were observable, then the MLE and our PLE could be easily computed. Since τs' are not observable, we treat them as missing values and apply the EM algorithm to compute estimates of μ and Ψ. The EM algorithm alternates between an expectation (E) step and a maximization (M) step. In computing the MLE, the E step computes the expected value of the complete-data log likelihood with respect to the conditional distribution of complete-data given the observed data. The M step maximizes the resulting function. Under mild conditions each iteration of EM decreases the log likelihood of the observed data. Our adding a regularization penalty to the log likelihood does not change the convergence property of EM.

Let $\theta^{(t)} = (\mu^{(t)}, \Psi^{(t)})$ be the parameter estimate at the t th iteration of the EM algorithm. Then, at the $(t+1)$th iteration:

E step: Assuming $\theta = \theta^{(t)}$, compute the expected value of the weights τ_i given the observed data:

$$\tau_i^{(t+1)} = E(\tau_i|\theta^{(t)}, y_{obs}) = \frac{v + p}{v + d(y_i, \mu^{(t)}, \Psi^{(t)})},$$

where p is the dimension of y_i, v is the degrees of freedom, and $d(y, \mu, \Psi) = (y-\mu)'\Psi^{-1}(y-\mu)$ is the Mahalanobis distance between y and μ.

M step: Minimizing the expected penalized negative log likelihood, giving

$$\mu^{(t+1)} = \sum_{i=1}^n \tau_i^{(t+1)} y_i \Big/ \sum_{i=1}^n \tau_i^{(t+1)},$$

and

$$\Psi^{(t+1)} = \frac{1}{\lambda + 1}\left\{ \frac{1}{n}\sum_{i=1}^n \tau_i^{(t+1)}(y_i - \mu^{(t+1)})(y_i - \mu^{(t+1)})' + \lambda\gamma I \right\}.$$

2.3. Bayesian interpretation and tuning

The proposed estimator can be naturally interpreted as the maximum posterior estimate with a prior $\Psi^{-1} \sim W((1/n\eta)I, p + 1 + n\lambda)$, i.e.,

$$p(\Psi^{-1}|\lambda, \eta) \propto |\Psi|^{-n\lambda/2} \exp\{-n\eta\,\mathrm{trace}(\Psi^{-1})/2\}, \tag{12}$$

where $\eta = \lambda\gamma$. Therefore, the marginal density of $y_1, \ldots, y_n$ can be written as

$$p(y_1, \ldots, y_n|\mu, \lambda, \eta) = \int \left\{ \prod_{i=1}^{n} f(y_i|\mu, \Psi) \right\} p(\Psi^{-1}|\lambda, \eta)\, \mathrm{d}\Psi^{-1}. \tag{13}$$

Taking the empirical Bayes perspective, η and λ can be determined by maximizing the marginal likelihood. However, the lack of a closed form marginal density makes the optimization intractable. To overcome this problem, here we introduce a variant of the method of moments as an alternative to maximizing the marginal likelihood.

Denote

$$\tilde{S} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})', \tag{14}$$

and let $\tilde{\tilde{S}}$ be a $n \times n$ matrix whose (j, k) entry is given by

$$\tilde{\tilde{S}}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \{(y_i - \bar{y})(y_i - \bar{y})' - S\}_{jk}^2. \tag{15}$$

In Appendix A, we derived that

Theorem 1. *The marginal expectations of* $\mathrm{trace}(\tilde{S})$ *and* $\mathrm{trace}(\tilde{\tilde{S}})$ *with prior* $\Psi^{-1} \sim W((1/n\eta)I, p + 1 + n\lambda)$ *are given by*

$$E\{\mathrm{trace}(\tilde{S})\} = \frac{vp\gamma}{v-2}, \tag{16}$$

$$E\{\mathrm{trace}(\tilde{\tilde{S}})\} = \left\{ \left(1 - \frac{4}{n} + \frac{5}{n^2} - \frac{1}{n^3} - \frac{1}{n^4}\right) \frac{3v^2}{(v-4)(v-2)} - \left(1 - \frac{9}{n} + \frac{32}{n^2} - \frac{24}{n^3}\right) \frac{v^2}{(v-2)^2} \right\} \frac{p\gamma^2}{1 - 2/n\lambda}. \tag{17}$$

Using this result, an estimator of (λ, γ) can be obtained by matching $\mathrm{trace}(\tilde{S})$ and $\mathrm{trace}(\tilde{\tilde{S}})$ with their respective expectations.

Note that $E\{\mathrm{trace}(\tilde{\tilde{S}})\}$ involves the fourth moment of a multivariate t random variable and therefore is only valid when $v > 4$. This is not restrictive in practice though since in most applications, t distribution is used to model excess kurtosis which is only finite if $v > 4$. Nevertheless, when $v = 3$ or 4, a combination of moment matching and cross validation (CV) can be employed. Note that (16) is still valid. Hence we can match $\mathrm{trace}(\tilde{S})$ with its expectation to obtain $\hat{\gamma} = \mathrm{trace}(\tilde{S})(v-2)/pv$. Now that (17) no longer holds, we can choose λ using K-fold CV.

In CV, the full data set $\mathscr{D}$ is randomly split into K subsets of about the same size, denoted by $\mathscr{D}^{(1)}, \ldots, \mathscr{D}^{(K)}$. For each $k = 1, \ldots, K$, we use the data in $\mathscr{D} - \mathscr{D}^{(k)}$ to estimate the model parameters and the data in $\mathscr{D}^{(k)}$ to validate. The log likelihood could be used as the performance measure in our case. For each candidate value of λ, the K-fold cross-validated log likelihood criterion is defined as

$$\mathrm{CV}(\lambda, \hat{\gamma}) = -2 \sum_{k=1}^{K} \sum_{i \in I_k} l(y_i, \mu_{(-k)}, \Psi_{-k}),$$

where I_k is the index set of the data in $\mathscr{D}^{(k)}$, and $\hat{\mu}_{-k}$ and $\widehat{\Psi}_{-k}$ are the estimated μ and Ψ using the training data set $\mathscr{D} - \mathscr{D}^{(k)}$. Typically, K is set to be 5 or 10 and $\mathrm{CV}(\lambda, \hat{\gamma})$ is minimized over a grid of values of λ. Let $\hat{\lambda}$ be the minimizer of $\mathrm{CV}(\lambda, \hat{\gamma})$. Our final estimate of Σ is based on $\hat{\lambda}$, $\hat{\gamma}$ and the full data set.

3. Numerical example

To illustrate the merits of the proposed penalized-likelihood estimator, we consider here a financial application. Covariance matrix estimation is of great importance in portfolio construction. We compare the proposed estimator with several commonly used alternative covariance matrix estimators using a real data set. We consider four estimates: the sample covariance matrix (`Sample`), the covariance matrix estimator of Ledoit and Wolfe (2004; `LWE`), the `MLE`, and the `PLE`. For the latter two estimators, we estimate the covariance matrix by $\widehat{\Sigma} = v/(v-2)\widehat{\Psi}$. Clearly there are many other alternative covariance matrix estimators. We choose the sample covariance matrix for its simplicity and popularity, and `LWE` because it has been demonstrated to enjoy good performance when compared with other popular alternatives (Ledoit and Wolf, 2004).

We consider an analysis of the joint distribution of Fama and French's (1993) 25 asset returns in the past 10 years from January 1997 to December 2006. The multivariate normal assumption is rejected by Anscombe–Glynn (1983) test of kurtosis

Table 1
Estimated mean and standard deviation for each of the 25 asset returns from Fama and French (1993).

Asset	t_8 PLE		Sample		t_8 MLE		LWE
	Mean	Std	Mean	Std	Mean	Std	Std
S1B1	0.59	11.18	0.61	10.17	0.60	8.97	10.00
S1B2	1.37	8.84	1.46	8.27	1.38	7.09	8.15
S1B3	1.54	6.81	1.60	5.98	1.55	5.49	5.95
S1B4	1.74	6.47	1.80	5.45	1.75	5.24	5.44
S1B5	1.99	6.98	1.79	5.56	2.04	5.69	5.55
S2B1	0.68	9.96	0.65	8.47	0.70	8.06	8.35
S2B2	1.09	7.46	1.06	6.06	1.10	6.07	6.03
S2B3	1.46	6.45	1.36	5.05	1.49	5.26	5.07
S2B4	1.49	6.64	1.33	5.17	1.52	5.43	5.18
S2B5	1.53	7.12	1.37	5.61	1.57	5.81	5.60
S3B1	0.69	9.04	0.61	7.76	0.71	7.29	7.66
S3B2	1.14	6.83	1.06	5.42	1.15	5.55	5.41
S3B3	1.22	6.01	1.09	4.73	1.24	4.90	4.77
S3B4	1.29	6.16	1.09	4.79	1.32	5.03	4.82
S3B5	1.68	6.55	1.52	5.17	1.72	5.35	5.18
S4B1	0.86	8.06	0.95	7.02	0.85	6.47	6.95
S4B2	1.23	6.15	1.13	4.88	1.26	5.00	4.90
S4B3	1.42	5.87	1.25	4.77	1.44	4.77	4.80
S4B4	1.45	5.96	1.32	4.72	1.47	4.83	4.75
S4B5	1.40	6.00	1.21	4.94	1.43	4.85	4.96
S5B1	0.84	6.15	0.70	4.94	0.84	4.95	4.96
S5B2	1.19	5.53	0.96	4.51	1.20	4.47	4.55
S5B3	1.21	5.53	0.97	4.52	1.23	4.48	4.56
S5B4	1.20	5.13	0.92	4.34	1.23	4.13	4.40
S5B5	1.29	6.36	0.82	5.13	1.33	5.13	5.14

The sample mean is also used for calculating the LWE covariance matrix.

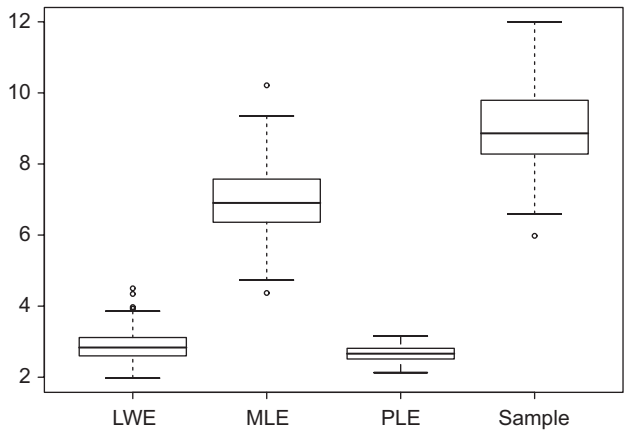

Fig. 1. Comparisons of four methods on simulated data sets. The boxplots of the losses from 100 simulation runs.

with p-value smaller than 0.01%. The averaged kurtosis of the 25 asset returns is 4.53 which indicates a multivariate t distribution with 8 degrees of freedom might be a more reasonable alternative. We estimate the location vector and scale matrix using the proposed method. The estimated mean and standard deviation of each asset are reported in Table 1.

To further evaluate the accuracy of the proposed method, we simulated data with the same characteristics as the observed asset returns. We generate $n = 120$ observations from a $p = 25$ dimensional multivariate t distribution with 8 degrees of freedom, and the location vector and scale matrix being those estimated from the above real data using PLE. Then we apply each of the four estimators to the simulated data. To evaluate the performance of the estimates, we use the following loss function for a covariance matrix estimate $\widehat{\Sigma}$:

$$L(\widehat{\Psi}) = \ln|\widehat{\Sigma}| + \operatorname{trace}(\Sigma\widehat{\Sigma}^{-1}) - \ln|\Sigma| - p. \tag{18}$$

The experiment was repeated 100 times. The boxplot of the losses for each of the method is given in Fig. 1, from which we can see that the PLE performs favorably over the other methods.

4. Discussions

In this paper, we proposed a novel regularization technique for parameter estimation in multivariate t and normal distribution. The estimator can be computed efficiently using an EM algorithm and the tuning parameters can be set using a variant of the method of moments or CV.

The proposed methods can be easily extended in several directions. First, in many applications, it might be desirable to shrink the scale matrix estimate towards a target other than γI. For example, Ledoit and Wolf (2003) demonstrated that assuming normality, in estimating the covariance matrix of stock returns, it is natural to consider shrinking the covariance matrix towards a factor model where the covariance matrix can be expressed as a linear combination of a diagonal matrix and a matrix with small ranks. The entropy type penalty (3) can be easily adopted for such purpose by setting Ω as a covariance matrix estimate obtained from the factor model.

Moreover, the proposed method can be modified to handle unknown degrees of freedom. If v is unknown, we need to add to the M step an updating formula of v by maximizing the actual observed data log likelihood over v given $(\mu, \Psi) = (\mu^{(t)}, \Psi^{(t)})$. Maximizing the actual log likelihood is preferred to maximizing the expected log likelihood (Liu and Rubin, 1995). The log likelihood, ignoring irrelevant constants, is

$$\left\{\ln\left(\Gamma\left(\frac{v+p}{2}\right)\right) - \ln\left(\Gamma\left(\frac{v}{2}\right)\right) + \frac{v}{2}\ln v\right\} - \frac{v+p}{2}\sum_{i=1}^{n}\ln\{v + (y_i-\mu)'\Psi^{-1}(y_i-\mu)\},$$

which can be maximized using a one-dimensional search such as Newton's method.

The proposed technique can also be generalized to a rich family of distributions that can be expressed as the scale mixture of multivariate normal distribution (Andrews and Mallows, 1974), i.e.,

$$y|\tau \sim MVN(\mu, \Psi/\tau),$$
$$\tau \sim \pi(\tau).$$

Multivariate t is a special case of scale mixture of multivariate normal with π being set as χ^2_v/v. Other popular examples include finite mixture of normals with shared covariance matrix where π is a discrete distribution; multivariate exponential power distribution (Haro-Lopez and Smith, 1999) where π is a positive weighted stable distribution; and multivariate stable distribution (Buckle, 1995) where π takes a nonstandard form. It is clear that the methodology developed in Section 2 can be modified to estimate parameters μ and Ψ in these models as well.

Acknowledgments

Yuan's work was partially supported by grants from the National Science Foundation (DMS-0624841 and DMS-0706724). Huang's work was partially supported by grants from the National Science Foundation (DMS-0606580) and the National Cancer Institute (CA57030).

Appendix A. Derivation of marginal moments

Simple algebraic manipulations yield

$$\mathrm{trace}(\tilde{S}) = \frac{1}{n-1}\sum_{i=1}^{n}\|y_i-\bar{y}\|^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}\|y_i-\mu\|^2 - n\|\bar{y}-\mu\|^2\right). \tag{A.1}$$

Therefore,

$$\begin{aligned} E\{\mathrm{trace}(\tilde{S})\} &= \frac{1}{n-1}\left[\sum_{i=1}^{n}E\{\|y_i-\mu\|^2\} - nE\{\|\bar{y}-\mu\|^2\}\right] \\ &= E(\|y_1-\mu\|^2) \\ &= E\{E(\|y_1-\mu\|^2|\Psi)\} \\ &= E\left\{\frac{v}{v-2}\mathrm{trace}(\Psi)\right\} \\ &= \frac{vp\gamma}{v-2}, \end{aligned}$$

where the last equation holds because $E(\Psi) = (n\eta/n\lambda)I = \gamma I$.

The derivation of (17) is somewhat tedious and we decompose it into several lemmas.

Lemma 1.

$$E\left[\left\{\sum_{i=1}^{n}(y_{ij}-\mu_j)^2\right\}^2\middle|\Psi\right]=nE\{(y_{1j}-\mu_j)^4|\Psi\}+n(n-1)E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.$$

Lemma 2.

$$E\left[(\bar{y}_{\cdot j}-\mu_j)^2\sum_{i=1}^{n}(y_{ij}-\mu_j)^2|\Psi\right]=\frac{1}{n}E\{(y_{1j}-\mu_j)^4|\Psi\}+\frac{n-1}{n}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.$$

Proof.

$$\begin{aligned}E\left[(\bar{y}_{\cdot j}-\mu_j)^2\sum_{i=1}^{n}(y_{ij}-\mu_j)^2|\Psi\right]&=E\left[\left\{\frac{1}{n}\sum_{k=1}^{n}(y_{kj}-\mu_j)\right\}^2\sum_{i=1}^{n}(y_{ij}-\mu_j)^2|\Psi\right]\\&=\frac{1}{n^2}E\left[\sum_{k=1}^{n}(y_{kj}-\mu_j)^2\sum_{i=1}^{n}(y_{ij}-\mu_j)^2|\Psi\right]\\&=\frac{1}{n}E\{(y_{1j}-\mu_j)^4|\Psi\}+\frac{n-1}{n}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\},\end{aligned}$$

where the last equality holds because of Lemma 1. □

Lemma 3.

$$E[(\bar{y}_{\cdot j}-\mu_j)^4|\Psi]=\frac{1}{n^3}E\{(y_{1j}-\mu_j)^4|\Psi\}+\frac{6(n-1)}{n^3}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.$$

Proof.

$$\begin{aligned}E[(\bar{y}_{\cdot j}-\mu_j)^4|\Psi]&=E\left[\left\{\frac{1}{n}\sum_{i=1}^{n}(y_{ij}-\mu_j)\right\}^4\middle|\Psi\right]\\&=\frac{1}{n^4}E\left[\sum_{i=1}^{n}(y_{ij}-\mu_j)^4|\Psi\right]+\frac{6}{n^4}E\left[\sum_{i\neq k}(y_{ij}-\mu_j)^2(y_{kj}-\mu_j)^2|\Psi\right]\\&=\frac{1}{n^3}E\{(y_{1j}-\mu_j)^4|\Psi\}+\frac{6(n-1)}{n^3}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.\end{aligned}$$ □

Lemma 4.

$$E\left\{\sum_{i=1}^{n}(y_{ij}-\bar{y}_{\cdot j})^4|\Psi\right\}=\left(n-4+\frac{6}{n}-\frac{3}{n^2}\right)E\{(y_{1j}-\mu_j)^4|\Psi\}+\left(6-\frac{24}{n}+\frac{18}{n^2}\right)E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.$$

Proof.

$$\begin{aligned}E\left\{\sum_{i=1}^{n}(y_{ij}-\bar{y}_{\cdot j})^4|\Psi\right\}&=E\left[\sum_{i=1}^{n}\{(y_{ij}-\mu_j)+(\mu_j-\bar{y}_{\cdot j})\}^4|\Psi\right]\\&=E\left\{\sum_{i=1}^{n}(y_{ij}-\mu_j)^4|\Psi\right\}+nE\{(y_{\cdot j}-\mu_j)^4|\Psi\}-4E\left\{\sum_{i=1}^{n}(y_{ij}-\mu_j)^3(y_{\cdot j}-\mu_j)|\Psi\right\}\\&\quad-4E\left\{\sum_{i=1}^{n}(y_{ij}-\mu_j)(y_{\cdot j}-\mu_j)^3|\Psi\right\}+6E\left\{\sum_{i=1}^{n}(y_{ij}-\mu_j)^2(y_{\cdot j}-\mu_j)^2|\Psi\right\}\\&=E\left\{\sum_{i=1}^{n}(y_{ij}-\mu_j)^4|\Psi\right\}+nE\{(y_{\cdot j}-\mu_j)^4|\Psi\}-\frac{4}{n}E\left\{\sum_{i=1}^{n}(y_{ij}-\mu_j)^4|\Psi\right\}-4nE\{(y_{\cdot j}-\mu_j)^4|\Psi\}\\&\quad+\frac{6}{n}E\{(y_{1j}-\mu_j)^4|\Psi\}+\frac{6(n-1)}{n}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}\\&=\left(n-4+\frac{6}{n}-\frac{3}{n^2}\right)E\{(y_{1j}-\mu_j)^4|\Psi\}+\left(6-\frac{24}{n}+\frac{18}{n^2}\right)E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.\end{aligned}$$ □

Lemma 5.

$$E(\tilde{S}_{jj}^2|\Psi) = \frac{1}{n}E\{(y_{1j}-\mu_j)^4|\Psi\} + \frac{n^2-2n+6}{n(n-1)}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.$$

Proof. Note that

$$\begin{aligned}\tilde{S}_{jj}^2 &= \left\{\frac{1}{n-1}\sum_{i=1}^n (y_{ij}-\bar{y}_{\cdot j})^2\right\}^2 \\ &= \frac{1}{(n-1)^2}\left\{\sum_{i=1}^n (y_{ij}-\mu_j)^2 - n(\bar{y}_{\cdot j}-\mu_j)^2\right\}^2 \\ &= \frac{1}{(n-1)^2}\left[\left\{\sum_{i=1}^n (y_{ij}-\mu_j)^2\right\}^2 + n^2(\bar{y}_{\cdot j}-\mu_j)^4 - 2n(\bar{y}_{\cdot j}-\mu_j)^2\sum_{i=1}^n (y_{ij}-\mu_j)^2\right].\end{aligned}$$

Now from Lemma 1,

$$E\left[\left\{\sum_{i=1}^n (y_{ij}-\mu_j)^2\right\}^2 \middle| \Psi\right] = nE(y_{1j}-\mu_j)^4 + n(n-1)E(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2, \tag{A.2}$$

from Lemma 3,

$$E\{n^2(\bar{y}_{\cdot j}-\mu_j)^4|\Psi\} = \frac{1}{n}E\{(y_{1j}-\mu_j)^4|\Psi\} + \frac{6(n-1)}{n}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}, \tag{A.3}$$

and from Lemma 2,

$$E\left\{2n(\bar{y}_{\cdot j}-\mu_j)^2\sum_{i=1}^n (y_{ij}-\mu_j)^2|\Psi\right\} = 2E\{(y_{1j}-\mu_j)^4|\Psi\} + 2(n-1)E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}. \tag{A.4}$$

To sum up, we have

$$E(\tilde{S}_{jj}^2|\Psi) = \frac{1}{n}E\{(y_{1j}-\mu_j)^4|\Psi\} + \frac{n^2-2n+6}{n(n-1)}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}. \qquad \square$$

Lemma 6.

$$E(\tilde{\tilde{S}}_{jj}|\Psi) = \left(1-\frac{4}{n}+\frac{5}{n^2}-\frac{1}{n^3}-\frac{1}{n^4}\right)\frac{3v^2\Psi_{jj}^2}{(v-4)(v-2)} - \left(1-\frac{9}{n}+\frac{32}{n^2}-\frac{24}{n^3}\right)\frac{v^2\Psi_{jj}^2}{(v-2)^2}.$$

Proof.

$$\begin{aligned}\tilde{\tilde{S}}_{jj} &= \frac{1}{n}\sum_{i=1}^n \{(y_{ij}-\bar{y}_{\cdot j})^2 - S_{jj}\}^2 \\ &= \frac{1}{n}\left\{\sum_{i=1}^n (y_{ij}-\bar{y}_{\cdot j})^4 + nS_{jj}^2 - 2S_{jj}\sum_{i=1}^n (y_{ij}-\bar{y}_{\cdot j})^2\right\} \\ &= \frac{1}{n}\sum_{i=1}^n (y_{ij}-\bar{y}_{\cdot j})^4 - S_{jj}^2.\end{aligned}$$

Recall that $S = (n-1)\tilde{S}/n$. Lemma 5 implies that

$$E(S_{jj}^2|\Psi) = \frac{(n-1)^2}{n^3}E\{(y_{1j}-\mu_j)^4|\Psi\} + \frac{n^3-3n^2+8n-6}{n^3}E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.$$

Together with Lemma 4,

$$E(\tilde{\tilde{S}}_{jj}|\Psi) = \left(1-\frac{5}{n}+\frac{8}{n^2}-\frac{4}{n^3}\right)E\{(y_{1j}-\mu_j)^4|\Psi\} - \left(1-\frac{9}{n}+\frac{32}{n^2}-\frac{24}{n^3}\right)E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}.$$

The proof is now completed by the fact that

$$E\{(y_{1j}-\mu_j)^4|\Psi\}=\frac{3\nu^2\Psi_{jj}^2}{(\nu-4)(\nu-2)}, \tag{A.5}$$

$$E\{(y_{1j}-\mu_j)^2(y_{2j}-\mu_j)^2|\Psi\}=\frac{\nu^2\Psi_{jj}^2}{(\nu-2)^2}. \quad \square \tag{A.6}$$

Now (17) can be obtained because $E(\Psi_{jj}^2)=\gamma^2/(1-2/n\lambda)$ (Siskind, 1972).

References

Andrews, D.F., Mallows, C.L., 1974. Scale mixtures of normal distributions. J. Roy. Statist. Soc. B 36, 99–102.
Anscombe, F.J., Glynn, W.J., 1983. Distribution of kurtosis statistic for normal statistics. Biometrika 70, 227–234.
Bickel, P.J., Levina, E., 2008. Regularized estimation of large covariance matrices. Ann. Statist. 36(1), 199–227.
Buckle, D., 1995. Bayesian inference for stable distributions. J. Amer. Statist. Assoc. 90, 605–613.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B 39, 1–38.
Fama, E., French, K., 1993. Common risk factors in the returns on stocks and bonds. J. Finance Econom. 33, 3–56.
Haff, L., 1980. Empirical Bayes estimation of the multivariate normal covariance matrix. Ann. Statist. 8, 586–597.
Haro-Lopez, R., Smith, A., 1999. On robust Bayesian analysis for location and scale parameters. J. Multivariate Anal. 70, 30–56.
Huang, J., Liu, N., Pourahmadi, M., Liu, L., 2006. Covariance selection and estimation via penalised normal likelihood. Biometrika 93, 85–98.
Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Empirical Finance 10, 603–621.
Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88, 365–411.
Liu, C., Rubin, D.B., 1995. ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statist. Sinica 5, 19–39.
Siskind, V., 1972. Second moments of inverse Wishart-matrix elements. Biometrika 59, 690–691.
Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94, 19–35.