Journal of Statistical Planning and Inference 177 (2016) 50–63

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

# Robust modeling using non-elliptically contoured multivariate $t$ distributions

Zhichao Jiang[a], Peng Ding[b,*]

[a] School of Mathematical Sciences, Peking University, Bejing 100871, China

[b] Department of Statistics, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

*Article history:*
Received 19 October 2015
Received in revised form 15 April 2016
Accepted 16 April 2016
Available online 4 May 2016

*MSC:*
00-01
99-00

*Keywords:*
Sample selection
Heavy-tailedness
Heckman selection model
Robit model
Linear mixed-effects model
Data augmentation
Parameter expansion

ABSTRACT

Models based on multivariate $t$ distributions are widely applied to analyze data with heavy tails. However, all the marginal distributions of the multivariate $t$ distributions are restricted to have the same degrees of freedom, making these models unable to describe different marginal heavy-tailedness. We generalize the traditional multivariate $t$ distributions to non-elliptically contoured multivariate $t$ distributions, allowing for different marginal degrees of freedom. We apply the non-elliptically contoured multivariate $t$ distributions to three widely-used models: the Heckman selection model with different degrees of freedom for selection and outcome equations, the multivariate Robit model with different degrees of freedom for marginal responses, and the linear mixed-effects model with different degrees of freedom for random effects and within-subject errors. Based on the normal mixture representation of our $t$ distribution, we propose efficient Bayesian inferential procedures for the model parameters based on data augmentation and parameter expansion. We show via simulation studies and real data examples that the conclusions are sensitive to the existence of different marginal heavy-tailedness.



## 1. Introduction

Normal distributions are widely used for statistical modeling due to their simplicity and interpretability. Many results and methods, such as ordinary least squares, can be derived analytically when the relevant variables are normally distributed. However, in practice, data may have heavy tails, which are difficult to deal with using normal models.

Models based on $t$ distributions are frequently applied for robust analysis (Zellner, 1976; Lange et al., 1989; Geweke, 1994; Liu and Wu, 1999; Liu, 2004; Gelman et al., 2014; Zhang et al., 2014), and they are attractive generalizations of the models based on normal distributions such as linear and Probit models. Student (1908) proposes the classical univariate $t$ distribution, which is symmetric and bell-shaped, but has heavier tails than the standard normal distribution. A multivariate $t$ distribution (MTD) is a multivariate generalization of the one-dimensional Student $t$ distribution. Because the MTD is elliptically contoured, its linear transformations follow MTDs with the same number of degrees of freedom. However, it is sometimes too restrictive to require all marginal degrees of freedom be the same. Previous literature generalizes the MTD through different ways. For a recent review, see Nadarajah and Dey (2005). Arellano-Valle and Bolfarine (1995) discuss three characterizations of the MTD within the class of elliptical contoured distributions (Cambanis et al., 1981). Fang et al. (2002) propose the meta-elliptical distributions using copula. Jones (2002) develops a dependent bivariate $t$

* Corresponding author.
*E-mail address:* pengdingpku@berkeley.edu (P. Ding).

http://dx.doi.org/10.1016/j.jspi.2016.04.004

distribution with different marginal degrees of freedom. However, none of their work allows the marginal distributions to be independent, which is a limitation for modeling. In this paper, we propose a non-elliptically contoured multivariate $t$ distribution (NECTD), allowing for different marginal degrees of freedom and independent marginal distributions. The bivariate case of the NECTD is similar to the formulation of Shaw and Lee (2008). Our NECTD, based on scale mixtures of the components of the multivariate normal distribution, are flexible enough to be incorporated into various models and enjoy easy Bayesian computation using data augmentation (Tanner and Wong, 1987) and parameter expansion (Liu et al., 1998; Meng and Van Dyk, 1999; Van Dyk and Meng, 2001). We further illustrate its potential applications by generalizing the Heckman selection model, multivariate Robit model, and linear mixed-effects model.

Sample selection (Heckman, 1979) or missing data (Little and Rubin, 2002) problems are common in applied research. The Heckman selection model (Heckman, 1979) is the most famous model dealing with sample selection, which consists of a Probit selection equation and a linear outcome equation. To deal with heavy-tailed data with sample selection, Marchenko and Genton (2012) propose a Heckman selection-$t$ model, modeling the error terms of the selection and outcome equations as a bivariate $t$ distribution. However, in the Heckman selection-$t$ model, the error terms are constrained to have the same number of degrees of freedom, which cannot handle cases with different heavy-tailedness in the selection and outcome equations. Ignoring the heterogeneity of the marginal numbers of degrees of freedom may lead to biased inference. In order to overcome this limitation, we propose a generalized selection-$t$ model based on the NECTD, allowing for different heavy-tailedness in the selection and outcome equations.

The Logistic or Probit model for binary data can be represented by a latent linear model with a Logistic or normal error distribution (Albert and Chib, 1993). To make such commonly-used models more robust to outliers, Liu (2004) proposes a Robit regression model, replacing the error in the latent linear model by a $t$ distribution. When generalizing the Robit model to multivariate settings, it may be restrictive to have all the marginal distributions sharing the same number of degrees of freedom. Fortunately, we can generalize the multivariate Robit model by assuming NECTD error terms.

The linear mixed-effects model is frequently used for analyzing repeatedly measured data (Hartley and Rao, 1967; Laird and Ware, 1982). It assumes normal distributions for both the random effects and the within-subject errors. Pinheiro et al. (2001) propose a robust linear mixed-effects model, in which the random effects and the within-subject errors follow a MTD. This model is widely used in practice (Lin and Lee, 2006, 2007). However, their model restricts the numbers of degrees of freedom of the random effects and the within-subject errors to be the same. Based on the NECTD, we propose a generalized linear $t$ mixed-effects model, allowing for different heavy-tailedness in the two sources of variations.

The paper proceeds as follows. We introduce the NECTD and discuss its statistical properties in Section 2. In Sections 3–5, we propose the generalized selection-$t$, Robit, and linear $t$ mixed-effects models, respectively. For each model, we propose a Bayesian inferential procedure for the parameters, give a numerical example, and show its application on a real dataset. We conclude with a discussion in Section 6. In Appendices A and B, we present the properties of the NECTD and provide the details of Bayesian inference for NECTD. In the online Supplementary Materials, we provide the details for Bayesian posterior computation, sensitivity analysis, the datasets, and R code to replicate the results in this paper.

## 2. Non-elliptically contoured multivariate $t$ distribution

The traditional $p$-dimensional MTD, $\boldsymbol{t}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, has probability density function:

$$f(\boldsymbol{x}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{p/2}\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\left\{1+\nu^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}^{-(\nu+p)/2}, \tag{1}$$

where $\boldsymbol{\mu}$ is the location parameter, $\boldsymbol{\Sigma}$ is the scale matrix, and $\nu$ is the number of degrees of freedom.

Let $\boldsymbol{I}_p$ denote a $p \times p$ identity matrix. We can represent the MTD as a ratio between a multivariate normal random vector and the square root of an independent Gamma random variable:

$$\boldsymbol{X} \mid q \sim \boldsymbol{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/q), \quad q \sim \chi^2_\nu/\nu,$$

or equivalently,

$$\boldsymbol{X} = \boldsymbol{\mu} + q^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}, \quad \boldsymbol{Z} \sim \boldsymbol{N}_p(\boldsymbol{0}, \boldsymbol{I}_p),\ q \sim \chi^2_\nu/\nu,\ q \perp\!\!\!\perp \boldsymbol{Z}. \tag{2}$$

The additional factor $q$ with $E(q) = 1$ does not change the location but amplifies the variability of the multivariate normal distribution $\boldsymbol{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. When $q$ falls close to zero, the MTD produces extreme values. Representation (2) implies that each marginal distribution of $\boldsymbol{X}$ follows a univariate $t$ distribution with the same number of degrees of freedom $\nu$, namely, $X_j \sim t_1(\mu_j, \sigma_j^2, \nu)$. Moreover, the traditional MTD is an elliptically contoured distribution, which enjoys nice mathematical properties (Fang et al., 1990; Anderson, 2003; Kotz and Nadarajah, 2004).

However, the constraint of a common number of degrees of freedom prevents modeling multivariate data with different heavy-tailedness in different dimensions. We tackle this problem by generalizing the traditional elliptically contoured MTD. Let $\boldsymbol{Q} = \mathrm{diag}\{q_1\boldsymbol{I}_{p_1}, \dots, q_s\boldsymbol{I}_{p_s}\}$ be a block diagonal matrix with $\sum_{j=1}^s p_j = p$ and $\{q_j \sim \chi^2_{\nu_j}/\nu_j : j = 1, \dots, s\}$. Instead of using the probability density function, we define NECTD using a scale mixture of a normal random vector:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{Q}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}, \quad \boldsymbol{Z} \sim \boldsymbol{N}_p(\boldsymbol{0}, \boldsymbol{I}_p), \tag{3}$$