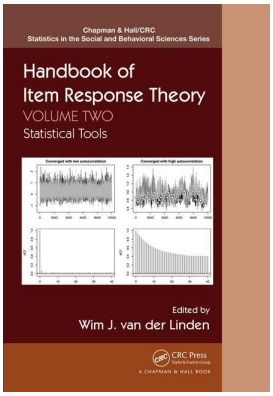

Handbook of Item Response Theory Volume Two Statistical Tools

Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series

J. van der Linden Wim

Multivariate Normal Distribution

Publication details
https://test.routledgehandbooks.com/doi/10.1201/b19166-5
Jodi M. Casabianca, Brian W. Junker
Published online on: 11 Feb 2016

3

Multivariate Normal Distribution

Jodi M. Casabianca and Brian W. Junker

CONTENTS

3.1 Introduction

In this chapter, we review several basic features of the multivariate normal distribution. Section 3.2 considers general properties of the multivariate normal density and Section 3.3 considers the sampling distribution of the maximum likelihood estimators (MLEs) of the mean vector and variance–covariance matrix based on iid (independent, identically distributed) random sampling of a multivariate normal distribution. Section 3.4 reviews the standard conjugate distributions for Bayesian inference with the multivariate normal distribution and Section 3.5 considers various generalizations and robustifications of the normal model. The properties of the multivariate normal distribution are well known and available in many places; our primary sources are the texts by Johnson and Wichern (1998) and Morrison (2005). A classic and comprehensive treatment is given by Anderson's (2003) text.

In item response theory (IRT), the multivariate normal and its generalizations are most often used as the underlying variables distribution for a data-augmentation version of the normal-ogive model (Bartholomew and Knott, 1999; Fox, 2010), as a population distribution for the proficiency parameters θ_p, for persons $p = 1, \ldots, P$, and as a prior distribution for other model parameters (e.g., difficulty parameters b_i, log-discrimination parameters $\log a_i$, etc.) whose domain is the entire real line or Euclidean space. Especially, in the latter two cases, the multivariate normal distribution serves to link standard IRT modeling with hierarchical linear models (HLMs) and other structures that incorporate various group structures and other dependence on covariates, to better model item responses in terms of the contexts in which they are situated (e.g., Fox, 2003, 2005a,b, 2010).

Because of the wide variety of applications of the normal distribution in IRT, we will depart somewhat from the notation used in the rest of the book in this chapter. We use $\mathbf{X} = (X_1, X_2, \ldots, X_K)^T$ to represent a generic K-dimensional random vector and X to represent a generic random variable. Their observed values will be denoted $\mathbf{x}$ and x, respectively. Also, because very many densities will be discussed, we will use the generic notation $f()$ to denote a density. The particular role or nature of $f()$ will be clear from context.

3.2 Multivariate Normal Density

The univariate normal density with mean μ and variance σ^2 for a random variable X is as follows:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu/\sigma)^2} \quad -\infty < x < \infty \tag{3.1}$$

As it is well known, $E[X] = \mu$, and $\text{Var}(X) = E[(X-\mu)^2] = \sigma^2$. We often write $X \sim N(\mu, \sigma^2)$ to convey that the random variable X has this density. The quantity in the exponent of the normal density thus measures the square of the distance between realizations of the random variable, X, and the mean μ, scaled in units of the standard deviation σ.

The multivariate normal density, for $\mathbf{X} \in \Re^K$, is as follows:

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{3.2}$$

where again $E[\mathbf{X}] = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T$ is the mean vector and $\text{Var}(\mathbf{X}) = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$ is the $K \times K$ symmetric nonnegative-definite variance–covariance matrix. We often write $\mathbf{X} \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or just $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if the dimension K is clear from context, to indicate that $\mathbf{X}$ follows the multivariate normal density. The quantity $(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ in the exponent is the squared distance from $\mathbf{x}$ to $\boldsymbol{\mu}$, again scaled by the variance–covariance matrix $\boldsymbol{\Sigma}$. In other contexts, this is called the *Mahalanobis distance* between $\mathbf{x}$ and $\boldsymbol{\mu}$ (Morrison, 2005).

The following theorem gives some properties of the multivariate normal random variables.

Theorem 3.1

If $\mathbf{X} \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

1. $E[\mathbf{X}] = \boldsymbol{\mu}$, and $\text{Var}(\mathbf{X}) = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$ (Johnson and Wichern, 1998).
2. If $\mathbf{W} = \mathbf{A}\mathbf{X} + \mathbf{b}$ for a constant matrix $\mathbf{A}$ and a constant vector $\mathbf{b}$, then $\mathbf{W} \sim N_K(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$ (Johnson and Wichern, 1998).
3. *Cholesky decomposition*: There exists a lower-triangular matrix $\mathbf{L}$ with nonnegative diagonal entries, such that $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ (Rencher, 2002). If $\mathbf{X} \sim N_K(\mathbf{0}, \mathbf{I}_{K\times K})$ then $\mathbf{W} = \mathbf{L}\mathbf{X} + \boldsymbol{\mu} \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
4. $\boldsymbol{\Sigma}$ is diagonal if and only if the components of $\mathbf{X}$ are mutually independent (Johnson and Wichern, 1998).

5. If $\mathbf{X}$ is partitioned into disjoint subvectors $\mathbf{X}_1$ and $\mathbf{X}_2$, and we write the following equation:

$$\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix} \sim N\left[\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right] \tag{3.3}$$

then the conditional distribution of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$, is also multivariate normal, with mean $\boldsymbol{\mu}_{1|2}$ and variance–covariance matrix $\boldsymbol{\Sigma}_{11|2}$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \tag{3.4}$$

$$\boldsymbol{\Sigma}_{11|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12} \tag{3.5}$$

(Johnson and Wichern, 1998).

3.2.1 Geometry of the Multivariate Normal Density

A useful geometric property of the multivariate normal distribution is that it is log quadratic: $\log f(\mathbf{x})$ is a simple linear function of the symmetric nonnegative definite quadratic form $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. This means that the level sets $\{\mathbf{x} : f(\mathbf{x}) = K\}$, or equivalently (after taking logs and omitting irrelevant additive constants) the level sets $\{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2\}$, will be ellipsoids. Figure 3.1 depicts bivariate normal density plots and the same densities using contour plots for two sets of variables with equal variance; variables in subplot (a) are uncorrelated and variables in subplot (b) are highly correlated. The contours are exactly the level sets $\{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2\}$ for various values of c^2.

Finding the principal axes of the ellipsoids is straightforward with Lagrange multipliers. For example, to find the first principal axis, we want to find the point $\mathbf{x}$ that is a maximum distance from $\boldsymbol{\mu}$ (maximize the squared distance $(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})$) that is still on the contour (satisfies the constraint $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$). Differentiating the Lagrange-multiplier objective function

$$g_\lambda(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}) - \lambda\left[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - c^2\right] \tag{3.6}$$

with respect to $\mathbf{x}$ and setting these derivatives equal to zero leads to the eigenvalue/eigenvector problem

$$(\boldsymbol{\Sigma} - \lambda I)(\mathbf{x}-\boldsymbol{\mu}) = 0 \tag{3.7}$$

It is now a matter of calculation to observe that the eigenvalues can be ordered so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \tag{3.8}$$

with corresponding mutually orthogonal eigenvectors $\mathbf{y}_k = (\mathbf{x}_k - \boldsymbol{\mu})$ lying along the principal axes of the ellipsoid with half-lengths $c\sqrt{\lambda_k}$.

Let $\mathbf{A}$ be the $K \times K$ matrix with columns $\mathbf{a}_k = \mathbf{y}_k/\|\mathbf{y}_k\|$. Then, $\mathbf{A}^T\mathbf{A} = I$, the identity matrix, and $\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A} = \text{diag}(\lambda_1, \ldots, \lambda_K)$, the $K \times K$ diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_K$. Now, consider the random vector $\mathbf{W} = \mathbf{A}^T(\mathbf{X}-\boldsymbol{\mu})$. It is easy to verify that $\mathbf{W} \sim N(\mathbf{0}, \text{diag}(\lambda_1, \ldots, \lambda_K))$. The components W_k of $\mathbf{W}$ are the (population) *principal components* of $\mathbf{X}$.

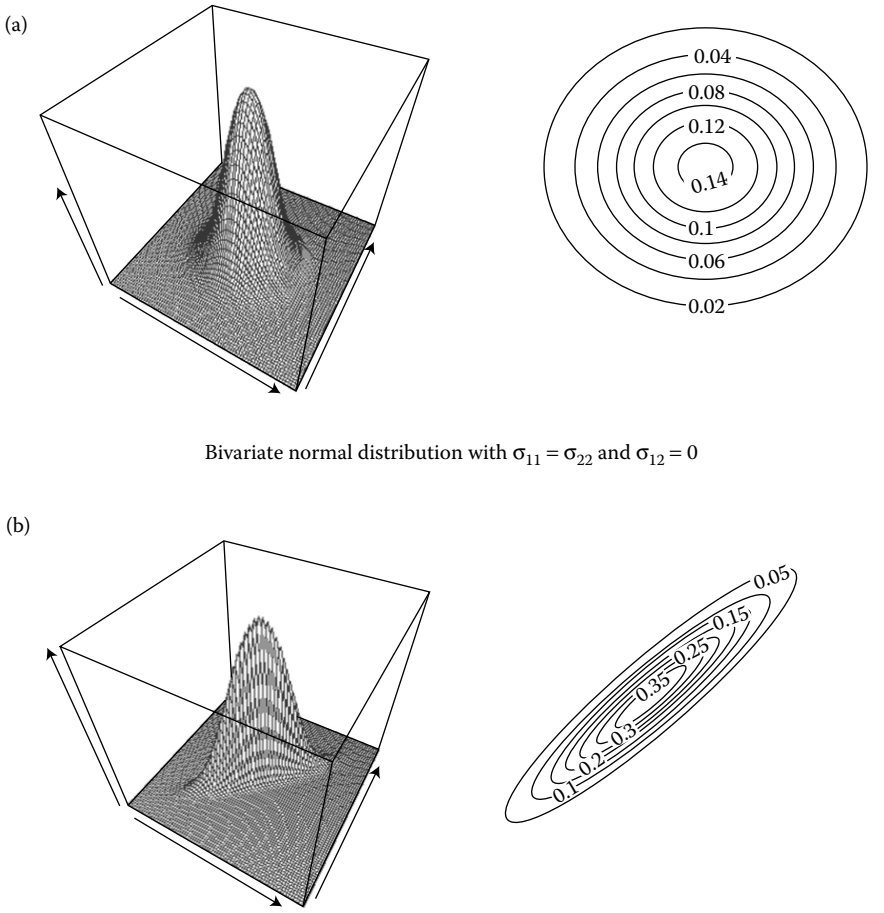

FIGURE 3.1
Bivariate density plots. Plots in (a) are the bivariate density and contour plots for two uncorrelated variables $\sigma_{12} = 0$ and plots in (b) are the bivariate density and contour plots for two perfectly correlated variables $\sigma_{12} = 1$.

3.3 Sampling from a Multivariate Normal Distribution

3.3.1 Multivariate Normal Likelihood

Let a set of $NK \times 1$ vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ represent an iid random sample from a multivariate normal population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Since densities of independent random variables multiply, it is easy to observe that the joint density will be as follows:

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_N | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{NK/2} |\boldsymbol{\Sigma}|^{N/2}} \exp\left[-\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \right] \tag{3.9}$$

The log of this density can be written (apart from some additive constants) as follows:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

$$= -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}}) - \frac{1}{2}\sum_{n=1}^{N}(\bar{\mathbf{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \quad (3.10)$$

The value of $\boldsymbol{\mu}$ that maximizes $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for any $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$, since that makes the third term in this loglikelihood equal to zero, so that $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ is the MLE. Furthermore, the first two terms of $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ may be rewritten as $(N/2)\log|\boldsymbol{\Sigma}^{-1}| - (1/2)\text{tr}\mathbf{A}\boldsymbol{\Sigma}^{-1}$ where $\mathbf{A} = \sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$, and from this and a little calculus, the MLE for $\boldsymbol{\Sigma}$ may be deduced. Summarizing,

Theorem 3.2

If $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the MLE for $\boldsymbol{\mu}$ is as follows:

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \quad (3.11)$$

The MLE for $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \quad (3.12)$$

Note that $\bar{\mathbf{X}}$ and $\widehat{\boldsymbol{\Sigma}}$ are sufficient statistics; they contain all of the information about $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the data matrix $\mathbf{X}$. Furthermore, note that $\widehat{\boldsymbol{\Sigma}}$ is a biased estimate of $\boldsymbol{\Sigma}$; the unbiased estimator $\mathbf{S} = (1/N-1)\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$ is often used instead.

3.3.2 Sampling Distribution of $\bar{\mathbf{X}}$ and $\mathbf{S}$

The sampling distributions of $\bar{\mathbf{X}}$ and $\mathbf{S}$ are easily generalized from the univariate case. In the univariate case, $\bar{X}$ and S are independent, $\bar{X} \sim N(\mu, \sigma^2/N)$, and $(N-1)\cdot S/\sigma^2 \sim \chi^2_{N-1}$, a χ-squared distribution with $N-1$ degrees of freedom. Recall that $\sum_{n=1}^{N}(X_n - \mu)^2/\sigma^2 = Z_1^2 + Z_2^2 + \cdots Z_N^2 \sim \chi^2_N$ by definition, because it is a sum of squares of independent standard normals; intuitively, we lose one degree of freedom for $(N-1)\cdot S/\sigma^2 = \sum_{n=1}^{N}(X_n - \bar{X})^2/\sigma^2$ because we are estimating the mean μ with $\bar{X}$ in calculating S.

In the multivariate case where we have the random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, the following three theorems apply:

Theorem 3.3

If $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are iid, then

1. $\bar{\mathbf{X}}$ and $\mathbf{S}$ are independent.

2. $\bar{\mathbf{X}}$ is distributed as a K-variate normal with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}/N$, $\bar{\mathbf{X}} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/N)$.
3. $(N-1)\cdot\mathbf{S}$ is distributed as a Wishart random variable with parameter $\boldsymbol{\Sigma}$ and $N-1$ degrees of freedom, $(N-1)\cdot\mathbf{S} \sim W_{N-1}(\boldsymbol{\Sigma})$.

By definition, $\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^T = \mathbf{Z}_1\mathbf{Z}_1^T + \mathbf{Z}_2\mathbf{Z}_2^T + \cdots + \mathbf{Z}_N\mathbf{Z}_N^T \sim W_N(\boldsymbol{\Sigma})$, since it is the sum of outer products of N independent $N(\mathbf{0}, \boldsymbol{\Sigma})$ random vectors. Once again, we lose one degree of freedom for $(N-1)\cdot\mathbf{S} = \sum_{n=1}^{N}(\mathbf{x}_n-\bar{\mathbf{x}})(\mathbf{x}_n-\bar{\mathbf{x}})^T$ since we are estimating the mean $\boldsymbol{\mu}$ with $\overline{\mathbf{X}}$. The Wishart density for a positive nonnegative definite random $K\times K$ matrix $\mathbf{A}$ with $D > K$ degrees of freedom and parameter $\boldsymbol{\Sigma}$ is as follows:

$$\omega_{D-1}(\mathbf{A}|\boldsymbol{\Sigma}) = \frac{|\mathbf{A}|^{(D-K-2)/2}e^{-tr[\mathbf{A}\boldsymbol{\Sigma}^{-1}]/2}}{2^{K(D-1)/2}\pi^{K(K-1)/4}|\boldsymbol{\Sigma}|^{(D-1)/2}\prod_{k=1}^{K}\Gamma((1/2)(D-k))} \tag{3.13}$$

3.4 Conjugate Families

Recall that if the *likelihood* for data $\mathbf{X}$ is (any function proportional to) the conditional density of $\mathbf{X}$ given (possibly multidimensional) parameter $\boldsymbol{\eta}$, $f(\mathbf{x}|\boldsymbol{\eta})$, then a *conjugate family of prior distributions* for $f(\mathbf{x}|\boldsymbol{\eta})$ is a parametric family of densities $f(\boldsymbol{\eta};\boldsymbol{\tau})$ for $\boldsymbol{\eta}$ with (possibly multidimensional) hyperparameter $\boldsymbol{\tau}$, such that for any member of the conjugate family, the *posterior distribution* $f(\boldsymbol{\eta}|\mathbf{x};\boldsymbol{\tau}) = f(\mathbf{x}|\boldsymbol{\eta})f(\boldsymbol{\eta};\boldsymbol{\tau})/\int_{\mathbf{h}} f(\mathbf{x}|\mathbf{h})f(\mathbf{h};\boldsymbol{\tau})\,d\mathbf{h}$ can be rewritten as $f(\boldsymbol{\eta}|\boldsymbol{\tau}^*)$, a member of the same parametric family as $f(\boldsymbol{\eta}|\boldsymbol{\tau})$, where $\boldsymbol{\tau}^* = \boldsymbol{\tau}^*(\mathbf{x},\boldsymbol{\tau})$ is some function of $\mathbf{x}$ and $\boldsymbol{\tau}$. Since only the form of $f(\boldsymbol{\eta}|\boldsymbol{\tau}^*)$ as a function of $\boldsymbol{\eta}$ matters, in verifying that $f(\boldsymbol{\eta}|\boldsymbol{\tau}^*)$ and $f(\boldsymbol{\eta};\boldsymbol{\tau})$ belong to the same parametric family, it is usual to ignore multiplicative constants that do not depend on $\boldsymbol{\eta}$.

For example, if $X_1,\ldots,X_N$ are iid $N(\mu,\sigma^2)$, then the likelihood is

$$f(x_1,\ldots,x_N|\mu,\sigma^2) = \prod_{n=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}e^{-(1/2\sigma^2)(x_n-\mu)^2} \propto \frac{1}{\sqrt{2\pi\sigma^2/n}}e^{-(1/(2\sigma^2/n))(\bar{x}-\mu)^2} \tag{3.14}$$

as a function of μ, as would be expected, since $\bar{x}$ is sufficient for μ.

If we assume σ^2 is known and place a normal $f(\mu;\mu_0,\tau_0^2) = (1/\sqrt{2\pi}\tau_0)e^{-(1/2\tau_0^2)(\mu-\mu_0)^2}$ prior on μ, then the posterior density for μ will be

$$\begin{aligned} f(\mu|x_1,\ldots,x_N;\mu_N,\tau_N^2) &\propto f(x_1,\ldots,x_N|\mu,\sigma^2)f(\mu|\mu_0,\tau_0^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2/n}}e^{-(1/(2\sigma^2/n))(\bar{x}-\mu)^2}\frac{1}{\sqrt{2\pi}\tau_0}e^{-(1/2\tau_0^2)(\mu-\mu_0)^2} \\ &\propto \frac{1}{\sqrt{2\pi}\tau_N}e^{-(1/2\tau_N^2)(\mu-\mu_N)^2} \end{aligned} \tag{3.15}$$

after completing the square, collecting terms, and identifying the normalizing constant, where

$$\tau_N^2 = \frac{1}{1/(\sigma^2/N) + 1/\tau_0^2} \tag{3.16}$$

$$\mu_N = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/N}\right)\bar{x} + \left(\frac{\sigma^2/N}{\tau_0^2 + \sigma^2/N}\right)\mu_0 \tag{3.17}$$

In this case, the posterior mean is $\mu_N = \rho_N\bar{x} + (1-\rho_N)\mu_0$ where $\rho_N = \tau_0^2/(\tau_0^2 + \sigma^2/n)$ is the classical reliability coefficient. Thus, if the prior distribution is $\mu \sim N(\mu_0, \tau_0)$, then the posterior distribution will be $\mu|x_1, \ldots, x_N \sim N(\mu_N, \tau_N)$; this shows that the normal distribution is the conjugate prior for a normal mean μ, when the variance σ^2 is known.

Further calculation (as shown in Gelman et al., 2004) shows that when *both* the mean μ and the variance σ^2 are unknown, and the data $X_1, \ldots, X_N$ are an iid sample from $N(\mu, \sigma^2)$, the joint distribution

$$\mu|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0) \tag{3.18}$$

$$\sigma^2 \sim \text{Inv–}\chi^2(\nu_0, \sigma_0^2) \tag{3.19}$$

is the conjugate prior, where the notation "$\sigma^2 \sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2)$" means that $\sigma_0^2/\sigma^2 \sim \chi^2_{\nu 0}$. Here, κ_0 and ν_0 are hyperparmeters that function as "prior sample sizes"—the larger κ_0 and ν_0, the greater the influence of the prior on the posterior. In this case, the joint posterior distribution for μ, σ^2 is

$$\mu|\sigma^2, x_1, x_2, \ldots, x_N \sim N(\mu_N, \sigma_N^2/\kappa_N) \tag{3.20}$$

$$\sigma^2|x_1, x_2, \ldots, x_N \sim \text{Inv} - \chi^2(\nu_N, \sigma_N^2) \tag{3.21}$$

where

$$\kappa_N = \kappa_0 + N$$

$$\nu_N = \nu_0 + N$$

$$\nu_N\sigma_N^2 = \nu_0\sigma_0^2 + (N-1)S^2 + \frac{\kappa_0 N}{\kappa_0 + N}(x - \mu_0)^2$$

$$\mu_N = \left(\frac{\sigma^2/\kappa_0}{\sigma^2/\kappa_0 + \sigma^2/N}\right)\bar{x} + \left(\frac{\sigma^2/N}{\sigma^2/\kappa_0 + \sigma^2/N}\right)\mu_0 = \left(\frac{N}{\kappa_0 + N}\right)\bar{x} + \left(\frac{\kappa_0}{\kappa_0 + N}\right)\mu_0$$

with $S^2 = (1/N-1)\sum_{n=1}^{N}(x_n - \bar{x})^2$.

Although this is the conjugate family when μ and σ^2 are both unknown, the forced dependence between μ and σ^2 in the prior is often awkward in applications. Common alternatives are as follows:

1. To replace σ^2/κ_0 with an arbitrary τ^2 in the conditional prior for $\mu|\sigma^2$, forcing independence. In this case, the conditional posterior for $\mu|\sigma^2$ is as in the "σ^2 known" case above, but the marginal posterior for σ^2 is neither conjugate nor in closed form (it is not difficult to calculate however) or

2. To make the conditional prior for $\mu|\sigma^2$ very flat/uninformative by letting $\kappa_0 \to 0$. In this case, the posterior distributions for $\mu|\sigma^2$ and for σ^2 mimic the sampling distributions of the MLEs

If one wishes for a noninformative prior for σ^2 (analogous to the flat prior choice for μ in (b) above), a choice that preserves conjugacy would be to take the degrees of freedom $\nu_0 = 0$ in the Inv $-\chi^2(\nu_0, \sigma_0^2)$ prior for σ^2. This leads to a prior $f(\sigma^2) \propto 1/\sigma^2$. Another noninformative choice is the Jeffreys prior for σ^2 (proportional to the square root of the Fisher information for the parameter), which in this case leads to the prior $f(\sigma) \propto 1/\sigma$. The Jeffreys prior for $\log \sigma^2$ (or equivalently $\log \sigma$) is simply $f(\sigma^2) \propto 1$. All of these choices are improper priors and care must be taken that the posterior turns out to be proper. One common way to avoid this issue is to force the prior to be proper, say, by taking $\sigma^2 \sim \text{Unif}(0, M)$ for some suitably large number M.

The conjugate prior distribution for a multivariate normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ has a form similar to that of the univariate case, but with multivariate normal and inverse-Wishart densities replacing the univariate normal and inverse-χ-squared densities. In particular, assuming sampling $\mathbf{x}_1, \ldots, \mathbf{x}_N$ iid from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the joint prior for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is of the following form:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0) \tag{3.22}$$

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(\nu_0, \boldsymbol{\Sigma}_0^{-1}) \tag{3.23}$$

where the notation "$\Sigma \sim$ Inv-Wishart$(\nu_0, \boldsymbol{\Sigma}_0^{-1})$" means that $\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1} \sim \omega_{\nu_0}(\boldsymbol{\Sigma}_0)$, the usual Wishart distribution with ν_0 degrees of freedom and parameter matrix $\boldsymbol{\Sigma}_0$. Again, κ_0 and ν_0 function as prior sample sizes. Then, the joint posterior distribution will be as follows:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma}, x_1, x_2, \ldots, x_N \sim N(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}/\kappa_N) \tag{3.24}$$

$$\boldsymbol{\Sigma}|x_1, x_2, \ldots, x_N \sim \text{Inv-Wishart}(\nu_N, \boldsymbol{\Sigma}_N^{-1}) \tag{3.25}$$

where

$$\begin{aligned}
\kappa_N &= \kappa_0 + N \\
\nu_N &= \nu_0 + N \\
\nu_N\boldsymbol{\Sigma}_N &= \nu_0\boldsymbol{\Sigma}_0 + (N-1)\mathbf{S} + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\mathbf{x}} - \mu_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T \\
\boldsymbol{\mu}_N &= \left(\frac{N}{\kappa_0 + N}\right)\bar{\mathbf{x}} + \left(\frac{\kappa_0}{\kappa_0 + N}\right)\boldsymbol{\mu}_0
\end{aligned}$$

with $\mathbf{S} = (1/N-1)\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$ (Gelman et al., 2004). Once again the joint conjugate prior forces some prior dependence between $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that may be awkward in practice. And once again the common fixes are as follows:

1. Replace $\boldsymbol{\Sigma}/\kappa_0$ with an arbitrary $\boldsymbol{\Lambda}_0$ in the conditional prior for $\boldsymbol{\mu}|\boldsymbol{\Sigma}$, forcing independence or
2. Make the conditional prior for $\boldsymbol{\mu}|\boldsymbol{\Sigma}$ very flat/uninformative by letting $\kappa_0 \to 0$

The details, generally analogous to the univariate normal case, are worked out in many places; in particular, see Gelman et al. (2004). Gelman et al. (2004) suggest another way to reduce the prior dependence between $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Sun and Berger (2007) provide an extensive discussion of several "default" objective/noninformative prior choices for the multivariate normal.

3.5 Generalizations of the Multivariate Normal Distribution

Assessing multivariate normality in higher dimensions is challenging. A well-known theorem (Johnson and Wichern, 1998; see also Anderson, 2003) states that $\mathbf{X}$ is a multivariate normal vector if and only if each linear combination $\mathbf{a}^T\mathbf{X}$ of its components is univariate normal, but this is seldom feasible to show in practice. Instead one often only checks the one- and two-dimensional margins of $\mathbf{X}$; for example, examining a Q—Q plot for each of the K components of $\mathbf{X}$, and in addition, examining bivariate scatterplots of each possible pair of variables to determine whether the data points yield an elliptical appearance. (See Johnson and Wichern (1998) for more information on techniques for assessing multivariate normality.)

There is no doubt that true multivariate normality is a rare property for a multivariate dataset. Although the latent ability variable in IRT is often assumed to be normally distributed, the data may not conform well to this assumption at all (Casabianca, 2011; Casabianca et al., 2010; Moran and Dresher, 2007; Woods and Lin, 2009; Woods and Thissen, 2006). In these and other cases, robust alternatives to the multivariate normal distribution can be considered.

An important class of generalizations of the multivariate normal distribution is the family of *multivariate elliptical distributions*, so named because each of its level sets defines an ellipsoid (Fang and Zhang, 1990; Branco and Dey, 2002). The K-dimensional random variable $\mathbf{X}$ has an elliptical distribution if and only if its characteristic function (Billingsley, 1995; Lukacs, 1970) is of the form

$$\Psi(\mathbf{t}) = E[\exp(i\mathbf{t}^T\mathbf{X})] = -\exp(i\mathbf{t}^T\mu)\psi(\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}) \tag{3.26}$$

where as usual $\boldsymbol{\mu}$ is a K-dimensional vector and $\boldsymbol{\Sigma}$ is a symmetric nonnegative definite $K \times K$ matrix, and $\mathbf{t} \in \Re^K$. When a density $f(\mathbf{x})$ exists for $\mathbf{X}$, it has the form

$$f_g(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} g[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] \tag{3.27}$$

where $g()$ is itself a univariate density; $g()$ is called the generator density for $f()$. The density defines a location/scale family with location parameter $\boldsymbol{\mu}$ and scale parameter $\boldsymbol{\Sigma}$. The parameter $\boldsymbol{\mu}$ is the median of $f()$ in all cases, and if $E[\mathbf{X}]$ exists, $E[\mathbf{X}] = \boldsymbol{\mu}$. If $\mathrm{Var}(\mathbf{X})$ exists, $\mathrm{Var}(\mathbf{X}) = -(\partial\psi(\mathbf{0})/\partial\mathbf{t}) \cdot \boldsymbol{\Sigma}$. A little calculation shows that if $\mathbf{W} = \mathbf{A}\mathbf{X} + \mathbf{b}$ is elliptically distributed with location $\boldsymbol{\mu}$ and scale $\boldsymbol{\Sigma}$, then $\mathbf{W} = \mathbf{A}\mathbf{X} + \mathbf{b}$ is again elliptically distributed, with location $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and scale $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$.

Elliptical distributions are used as a tool for generalizing normal-theory structural equations modeling (e.g., Shapiro and Browne, 1987; Schumacker and Cheevatanarak, 2000). The special case of the K-dimensional multivariate-t distribution on ν degrees of freedom

with density

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\Gamma((v+K)/2)}{\Gamma(v/2)v^{K/2}\pi^{K/2}}|\boldsymbol{\Sigma}|^{-1/2} \times \left(1+\frac{1}{v}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-(v+K)/2} \tag{3.28}$$

has been used as the error distribution in robust Bayesian and non-Bayesian linear regression modeling since at least Zellner (1976); more recently, robust regression with general elliptical error distributions and the particular case of scale mixtures of normals (of which the univariate-t and multivariate-t are also examples) has also been studied (e.g., Fernandez and Steel, 2000).

A further generalization is the family of *skew-elliptical distributions* (e.g., Branco and Dey, 2001, 2002). When it exists, the density of a skew-elliptical distribution is of the form

$$f_{g_1,g_2}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda}) = 2f_{g_1}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})F_{g_2}(\boldsymbol{\lambda}^T(\mathbf{x}-\boldsymbol{\mu})) \tag{3.29}$$

where $f_{g1}()$ is the density of a multivariate elliptical distribution, and $F_{g2}()$ is the cumulative distribution function of a (possibly different) univariate elliptical distribution with location parameter 0 and scale parameter 1. The vector parameter $\boldsymbol{\lambda}$ is a skewness parameter; when $\boldsymbol{\lambda} = \mathbf{0}$, the skew-elliptical density reduces to a symmetric elliptical density.

When the generator densities $g_1(x) = g_2(x) = \phi(x)$, the standard normal density, we obtain the special case of the *skew-normal distributions*. It has been observed (Moran and Dresher, 2007) that the empirical distribution of the latent proficiency variable in IRT models applied to large-scale educational surveys sometimes exhibits some nontrivial skewing, which if unmodeled can cause bias in estimating item parameters and features of the proficiency distribution. Skew-normal and related distributions have been proposed (Xu and Jia, 2011) as a way of accounting for this skewing in the modeling of such data, with as few extra parameters as possible.

Two additional classes of transformed normal distributions used in IRT are the *lognormal* and *logit-normal distributions*. A random variable X has a lognormal distribution when its logarithm is normally distributed. The density of the lognormal distribution is of the following form:

$$f(x;\mu,\sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}}e^{-(\ln x-\mu)^2/2\sigma^2} \quad x>0 \tag{3.30}$$

where μ and σ^2 are the mean and variance of the variable's natural log. This distribution is an alternative to Gamma and Weibull distributions for nonnegative continuous random variables and is used in psychometrics to model examinee response times (e.g., van der Linden, 2006) and latent processes in decision making (e.g., Rouder et al., 2014). It is also used as a prior distribution in Bayesian estimation of nonnegative parameters, for example the discrimination parameter in two- and three-parameter logistic IRT models (van der Linden, 2006).

Similarly, a random variable X has a logit-normal distribution when its logit is normally distributed. The density of the logit-normal distribution is of the following form:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(\text{logit}(x)-\mu)^2/2\sigma^2}\frac{1}{x(1-x)} \quad 0<x<1 \tag{3.31}$$

Here, μ and σ^2 are the mean and variance of the variable's logit, and x is a proportion, bounded by 0 and 1. A multivariate generalization of the logit-normal distribution (Aitchison, 1985) has been used in latent Dirichlet allocation models for text classification (Blei and

Lafferty, 2007) and in mixed membership models for strategy choice in cognitive diagnosis (Galyardt, 2012).

Acknowledgment

This work was supported by a postdoctoral fellowship at Carnegie Mellon University and Rand Corporation, through Grant #R305B1000012 from the Institute of Education Sciences, U.S. Department of Education.

References

Aitchison, J. 1985. A general class of distributions on the simplex. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47,136–146.

Anderson, T. W. 2003. *Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons.

Bartholomew, D. J. and Knott, M. 1999. *Latent Variable Models and Factor Analysis*. London: Arnold. (Kendall's Library of Statistics 7).

Billingsley, P. 1995. *Probability and Measure* (3rd ed.). New York: John Wiley & Sons.

Blei, D. and Lafferty, J. 2007. A correlated topic model of science. *Annals of Applied Statistics*, 1, 17–35.

Branco, M. and Dey, D. K. 2001. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79, 99–113.

Branco, M. and Dey, D. K. 2002. Regression model under skew-elliptical error distribution. *The Journal of Mathematical Sciences, Delhi, New Series*, 1, 151–169.

Casabianca, J. M. 2011. *Loglinear Smoothing for the Latent Trait Distribution: A Two-Tiered Evaluation*. (Doctoral dissertation), ProQuest dissertations and theses. (Accession Order No. AAT 3474125.)

Casabianca, J. M., Xu, X., Jia, Y., and Lewis, C. 2010. Estimation of item parameters when the underlying latent trait distribution of test takers is nonnormal. *Paper Presented at the Meeting of the National Council for Measurement in Education*, Denver, Colorado.

Fang, K. T. and Zhang, Y. T. 1990. *Generalized Multivariate Analysis*. New York: Springer.

Fernandez, C. and Steel, M. F. J. 2000. Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, 16, 80–101.

Fox, J. P. 2003. Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Stat Psychology*, 56, 65–81.

Fox, J. P. 2005a. Multilevel IRT model assessment. In van der Ark, L. A., Croon, M. A., and Sijtsma, K. (Eds.), *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences* (pp. 227–252). Mahwah, NJ: Lawrence Erlbaum.

Fox, J. P. 2005b. Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, 58, 145–172.

Fox, J. P. 2010. *Bayesian Item Response Modeling*. New York: Springer.

Gelman, A., Carlin, J. B., Stern, H. A., and Rubin, D. B. 2004. *Bayesian Data Analysis*. New York: John Wiley and Sons.

Galyardt, A. 2012. *Mixed Membership Distributions with Applications to Modeling Multiple Strategy Usage*. PhD dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Johnson, R. A. and Wichern, D. W. 1998. *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice-Hall.

Lukacs, E. 1970. *Characteristic Functions*. London: Griffin.

Moran, R. and Dresher, A. 2007. Results from NAEP marginal estimation research on multivariate scales. *Paper Presented at the Meeting of the National Council for Measurement in Education,* Chicago, IL.

Morrison, D. F. 2005. *Multivariate Statistical Methods*. Belmont, CA: Thomson Brooks Cole.
Rencher, A. C. 2002. *Methods of Multivariate Analysis*. New York: John Wiley and Sons.
Rouder, J. N., Province, J. M., Morey, R. D., and Heathcote, A. 2014. The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 1–23.
Schumacker, R. E. and Cheevatanarak, S. 2000. A comparison of normal and elliptical estimation methods in structural equations models. *Paper Presented at the Meeting of the American Educational Research Association*, New Orleans, LA (ERIC Document Reproduction Service No. ED441872). Retrieved April 21, 2012, from http://www.eric.ed.gov/PDFS/ED441872.pdf.
Shapiro, A. and Browne, M. W. 1987. Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association*, 82, 1092–1097.
Sun, D. and Berger, J. O. 2007. Objective Bayesian analysis for the multivariate normal model. *Bayesian Statistics*, 8, 525–562.
van der Linden, W. J. 2006. A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31, 181–204.
Woods, C. M. and Lin, N. 2009. IRT with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, 33, 102–117.
Woods, C. M. and Thissen, D. 2006. IRT with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71, 281–301.
Xu, X. and Jia, Y. 2011. *The Sensitivity of Parameter Estimates to the Latent Ability Distribution* (Research Report 11–40). Princeton, NJ: Educational Testing Service.
Zellner, A. 1976. Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. *Journal of the American Statistical Association*, 71, 400–405.