Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 2, 2020 (New Series), pp 187-205

Using Conditionally Specified Joint Distribution to Simultaneously Model Discrete and Continuous Data

Nadeesri Wijekoon[1] **and Nagaraj K. Neerchal**[1,2]
[1]*Department of Mathematics and Statistics*
University of Maryland, Baltimore County, USA
[2]*Chinmaya Vishwavidyapeeth, India*

Received: 29 May 2020; Revised: 21 July 2020; Accepted: 24 July 2020

Abstract

Often, in practice, conditionals are easier to model and interpret while the joint distribution itself is either intractable or not available in a closed form. Conditionally specified statistical models offer several advantages over joint models. Conditionally specified models are intuitively appealing and enrich our ability to build interpretable models in practice. In this paper, we derive the likelihood of a joint distribution obtained from Binomial and Bivariate Normal conditionals. Properties of maximum likelihood estimates and pseudolikelihood estimates are explored using a simulation. A conditionally specified model is obtained by assuming that closing prices are conditionally normally distributed and that the buy-sell recommendation by an Analyst follows a logistic regression model given the prices.

Key words: Conditionally specified models; Compatibility; Maximum likelihood; Pseudolikelihood; Gibb's sampling.

AMS Subject Classifications: 60G50, 05C81

1. Introduction

When solving real-world problems, the main difficulty could be selecting a suitable model to reflect the reality being observed (Ghosh and Nadarajah, 2017). Especially when the observed response consists of both continuous and discrete components, it is not very convenient to directly specify a joint distribution. Often, in practice, conditionals are easier to model and interpret while the joint distribution itself is either intractable or not available in a closed form. According to Arnold *et al.* (1999, 2001), although the joint distribution is less tractable, availability of easily handled conditionals enables us to consider computationally more efficient estimation methods such as pseudolikelihood. Arnold *et al.* (1999, 2001) provide an extensive account of conditionally specified joint distributions. They proved theorems describing the conditions for the existence of joint distributions consistent with the given conditionals and provided general forms of such joint distributions. They also consider examples of joint distributions determined by discrete-continuous pairs of conditionals. It

Correponding Author: Nagaraj K. Neerchal
Email: nagaraj@umbc.edu

turns out that the joint distributions obtained by following the theoretical developments are not always tractable. In particular, often the normalizing constant does not lend itself to a closed form. In this paper, we consider a joint distribution for data consisting of a binary random variable and bivariate continuous measurements. The model postulates a joint distribution determined by Binomial (in fact by a logistic regression model) and bivariate Normal conditionals. For this model, we can derive a closed form for the normalizing constant, however, it turns out that the parameter estimation is numerically intensive. Therefore, we consider an alternative method using a pseudolikelihood function. The two approaches are compared in a simulation study.

As an illustrative application, we consider stock price data along with the corresponding buy-sell recommendation of an analyst. Stock market analysts classify a stock as either a "buy" or a "sell" based on their own research into the history of the stock as well as their assessment of other market dynamics which have a bearing on the price of the stock. Here, the distribution of the stock price on the "buy" days is clearly different from the price distribution on "sell" days. Thus, our choice of the model parameters would differ if we knew that the stock has been classified as a buy instead of as a sell. On the other hand, the price history of the stock will influence the classification (buy or sell) decision of the analyst. Thus, even though it is cumbersome and inconvenient to think of a joint distribution of the stock prices (continuous) and the analyst recommendations (binary), it is easier to think of the conditional distribution of the recommendations given the price history and the conditional distribution of the stock price given its buy or sell status. Thus, conditionally specified distribution is a convenient way to model both, expert recommendations and closing prices simultaneously.

Another motivating example can be found in the area of Gerontology. In health care studies involving aged subjects, due to progressively deteriorating health conditions over time, subjects become unable to respond to questions. To avoid the resulting missing data situation, sometimes the study protocol would allow the investigator to collect a proxy response from another person who is familiar with the non-responding subject. Thus, for each subject, we record a pair of responses. One of the variables in the pair is a discrete random variable (0 or 1 depending upon whether the respondent is either the subject or the proxy) and the other is a continuous (a composite score from a mental health related questionnaire) measurement. Note that the ability to respond is usually related to the overall health condition of the patient. A proxy is needed when a subject is not well enough to answer the questions of the study. On the other hand, the distributional properties of the proxy responses and subject responses would be different. Thus, the conditional relationships between responses and self/proxy indication can be specified using commonly used models and a single joint distribution can be derived for analyzing both subject data and proxy data. The reader is referred to Hosseini (2017) for a detailed description of this approach along with a working code implementing the parameter estimation.

The contents of this paper are organized as follows. In Section 2, we briefly introduce conditionally specified models and present the existing theories of deriving the joint distribution and some issues like compatibility of conditionals. In Section 3, we consider the problem of compatibility and present the relevant restrictions on the original problem. In

Section 4, we propose our new joint distributions. In Section 5, we present an illustrative example using the stock data. Finally, some concluding remarks are made in Section 6.

All computations are done using the freeware R.

2. Conditionally Specified Distributions in Exponential Family

In a comprehensive review given by Arnold *et al.*(2001), in *Statistical Science,* it is stated that a bivariate density is easy to understand/visualize in the terms of its conditional densities. In practice, researchers often have better insight into the form of conditional distributions of experimental variables rather than the joint distribution (See Castillo and Galambos, 1989). For instance, instead of providing a model for (X, Y), one can propose families of conditional distributions of X given values of Y, and of Y given values of X. Castillo and Galambos (1989) identified the complete class of such bivariate distributions with given specified conditional distributions. Arnold and Strauss (1988) extended their work to arbitrary exponential family of conditionals. The key result in this area, which gives the form of the joint distribution which is consistent with the given specific pair of conditional distributions is provided in Arnold and Strauss (1988). Below is a brief statement of this key result, which provides a general form of the joint distribution starting with conditionals belonging to the exponential family of distributions. We start with the following notations.

Notations: Define an l_1-parameter family of densities $\{f_1(x;\underline{\theta}) : \underline{\theta} \in \Theta\}$ with respect to μ_1 (frequently, Lebesgue measure or counting measure) on D_1, a subset of Euclidean space of finite dimension, of the form

$$f_1(x;\underline{\theta}) = r_1(x)\beta_1(\underline{\theta})exp\left\{\sum_{i=1}^{l_1}\theta_i q_{1i}(x)\right\} \tag{1}$$

where $q_{1i}(x)$'s (sufficient statistics) are linearly independent, and $\underline{\theta} = (\theta_1, \cdots, \theta_{l_1})^T$. Similarly, we define, an l_2-parameter family of densities $\{f_2(x;\underline{\tau}) : \underline{\tau} \in\}$ with respect to μ_2 (frequently, Lebesgue measure or counting measure) on D_2, a subset of Euclidean space of finite dimension, of the form

$$f_2(y;\underline{\tau}) = r_2(y)\beta_2(\underline{\tau})exp\left\{\sum_{j=1}^{l_2}\tau_j q_{2j}(y)\right\} \tag{2}$$

where $q_{2j}(y)$'s (sufficient statistics) are linearly independent , and $\underline{\tau} = (\tau_1, \cdots, \tau_{l_2})^T$.

Our goal is to identify the class of bivariate densities $f(x, y)$ with respect to $\mu_1 \times \mu_2$ on $D_1 \times D_2$, whose conditionals belong to the above families of densities respectively. That is, we want to find a joint distribution $f(x, y)$ such that $f(x|y) = f_1(x;\underline{\theta}(y))$ and $f(y|x) = f_2(y;\underline{\tau}(x))$. Arnold *et al.*(1988) show the existence and provide a general form of the joint distribution. Their result is stated in the theorem below.

Theorem 1: Let $f(x, y)$ be a bivariate density whose conditional densities satisfy $f(x|y) = f_1(x;\underline{\theta}(y))$ and $f(y|x) = f_2(y;\underline{\tau}(x))$ for some function $\underline{\theta}(y)$ and $\underline{\tau}(x)$ where f_1 and f_2 are as defined in (3) and (4). It follows that $f(x, y)$ is of the form

$$f(x, y) = r_1(x)r_2(y)exp\left\{\underline{q}^{(1)}(x)^T M \underline{q}^{(2)}(y)\right\} \tag{3}$$

where,

$$\underline{q}^{(1)}(x) = (q_{10}(x), q_{11}(x), q_{12}(x), ..., q_{1l_1}(x))^T,$$

$$\underline{q}^{(2)}(y) = (q_{20}(y), q_{21}(y), q_{22}(y), ..., q_{2l_2}(y))^T$$

and where, $q_{10}(x) = q_{20}(y) = 1$ and M is an $(l_1+1) \times (l_2+1)$ matrix of parameters, subject to the requirement that $\int_{D_1} \int_{D_2} f(x, y) d\mu_1(x) d\mu_2(y) = 1$. The family of these joint distributions is referred to as conditional exponential family (CEF). Note that the elements of M may be denoted by m_{ij} for $i = 0, 1, \cdots, (l_1 + 1)$ and $j = 0, 1, \cdots, (l_2 + 1)$.

To illustrate the application of Theorem 1 we present an example where one of the conditionals is the Poisson distribution and the other is the Gamma distribution.

Example 1: This example was given in Arnold *et al.*(1999). Suppose we are seeking a joint distribution of a random vector (X, Y) such that, $X|Y = y \sim Poi(y)$ and assume $Y|X = x \sim \Gamma(x + \alpha, \lambda + 1)$. Since both conditionals belong to the exponential family of distributions, we can put them in the notations of Theorem 1 as follows: l_1=1 and l_2=2. The M matrix is 2×3. Further, $r_1(x) = \frac{1}{x!}$ and $r_2(y) = \frac{1}{y}$. Similarly, the sufficient statistics $q_{1i}(x)$'s and $q_{2j}(y)$'s can be identified as $q^{(1)}(x) = (1, x)^T$ and $q^{(2)}(y) = (1, -y, ln(y))^T$. Therefore, joint density belongs to CEF and its general form is given by,

$$f(x, y) = \frac{1}{x!y} exp\left(\begin{bmatrix} 1 & x \end{bmatrix} M \begin{bmatrix} 1 \\ -y \\ ln(y) \end{bmatrix} \right) \qquad x = 0, 1, ...; y > 0$$

where,

$$m_{01} > 0, m_{02} > 0, m_{11} \geq 0, m_{12} \geq 0.$$

Note that, when $m_{11} = m_{12} = 0$, X and Y are independent. And when $m_{11} = 0$ and $m_{12} = 1$, it can be shown that the marginal of X is given by

$$f(x) = \frac{\Gamma(x + \alpha)}{\Gamma(\alpha) x!} \left(\frac{\lambda}{\lambda + 1} \right)^{\alpha} \left(\frac{1}{\lambda + 1} \right)^{x} \qquad x = 0, 1, 2, ...$$

Thus, we obtain the familiar compound Poisson distribution.

An interesting point to be noted here is that, for $m_{11} > 0$ the joint distribution does not yield the compound Poisson distribution as the marginal of X. This indicates that, even though we started with the Poisson and Gamma conditionals, the resulting CEF is a much larger class than that obtained by combining $X|Y = y \sim Poi(y)$ and $Y \sim \Gamma(\alpha, \lambda + 1)$.

It turns out that the candidate functions for the conditional distributions will have to satisfy certain conditions for the existence of a corresponding proper joint distribution. This issue is also referred to as the problem of compatibility of conditionals. According to Chen (2010), the incompatibility of the conditionally specified models may lead to serious consequences on the statistical inference and interpretation in the data analysis and on the convergence of the Gibbs sampling. Thus, the compatibility issue is a widely researched area and there are several computational/theoretical approaches in literature to identify the possible compatibility of given families of conditional distributions. Besag (1974), Arnold and

Press (1989), Hobert and Casella (1998), Arnold, Castillo and Serabia (2002) and recently Ghosh and Nadarajah (2017) studied compatibility extensively. We refer the reader to theorems introduced by Arnold and Press (1989) which are used in checking the compatibility of the conditionals in this paper.

Example 2: (Logistic Regression) Suppose X takes values in the set $\{x_1, x_2, ..., x_k\}$, and is real valued. For each x we have $Y|X = x \sim N(\theta_x, \sigma_x^2)$.

And for each y we have,

$$P(X = x|Y = y) = \frac{exp[-(a_x + b_x y)]}{\sum_{x=1}^{k} exp[-(a_x + b_x y)]}.$$

We apply Theorem 4.1 in the Arnold and Press (1989) paper as follows. The Theorem is stated in the Appendix (A.2).

Proof of Compatibility: Let,

$$a(x, y) = f(Y|X = x) \sim N(\theta_x, \sigma_x^2)$$

$$b(x, y) = f(X|Y = y) \sim \frac{exp[-(a_x + b_x y)]}{\sum_{x=1}^{k} exp[-(a_x + b_x y)]}$$

In order to show compatibility, we need to prove that the ratio of $a(x, y)/b(x, y)$ factors into a product such as $U(x) \times V(y)$. Consider,

$$\begin{aligned}\frac{a(x,y)}{b(x,y)} &= \frac{\frac{1}{\sqrt{2\pi}\sigma_x} exp\left\{ -\frac{(y-\theta_x)^2}{2\sigma_x^2} \right\} \sum_{x=1}^{k} exp[-(a_x + b_x y)]}{exp[-(a_x + b_x y)]} \\ &= \frac{1}{\sqrt{2\pi}\sigma_x} exp\left\{ a_x - \frac{\theta_x^2}{2\sigma_x^2} \right\} exp\left\{ -\frac{1}{2\sigma_x^2}\left[y^2 - 2(\theta_x + \sigma_x^2 b_x) y \right] \right\} \sum_{x=1}^{k} exp[-(a_x + b_x y)]\end{aligned}$$

If $b_x = -\frac{\theta_x}{\sigma_x^2}$ and $\sigma_x^2 = \sigma^2$ where a_x's are unconstrained,

$$\begin{aligned}\frac{a(x,y)}{b(x,y)} &= \frac{1}{\sqrt{2\pi}\sigma} exp\left\{ a_x + \frac{\theta_x b_x}{2} \right\} exp\left\{ -\frac{y^2}{2\sigma^2} \right\} \sum_{x=1}^{k} exp[-(a_x + b_x y)] \\ &= \underbrace{exp\left\{ a_x + \frac{\theta_x b_x}{2} \right\}}_{U(x)} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} exp\left\{ -\frac{y^2}{2\sigma^2} \right\} \sum_{x=1}^{k} exp[-(a_x + b_x y)]}_{V(y)} \\ &= U(x)V(y)\end{aligned}$$

where, $\int_{y \in R} V(y)dy < \infty$ and $\sum_{x=1}^{k} U(x) < \infty$.

Therefore, according to Arnold and Press (1989) the two distributions are compatible provided $b_x = -\frac{\theta_x}{\sigma_x^2}$ and $\sigma_x^2 = \sigma^2$ where a_x's are unconstrained.

3. Logistic and Bivariate Conditionals

In this section, we present the derivation of a conditionally specified model starting with Logistic distribution and Bivariate Normal distribution as conditionals. We consider the stock price data as the motivating example. We propose a logistic regression model for the conditional distribution of the binary valued analyst recommendation given the stock prices from the first and last day of the trading week and the conditional distribution of the stock prices given the analyst recommendation as a Bivariate Normal. We first set up the problem in the notations of Theorem 1 and then obtain the form of the joint. It turns out that the normalizing constant can be obtained in a closed form using some results on multivariate normal integrals. We discuss some properties of the resulting joint distribution. We have created a shiny application (an interactive tool in the freeware R) which can be used to explore the structure of the joint distribution for various parameter values.

3.1. Setting up the problem

Suppose we have the distribution of observed beginning and end price vector (2×1) of a single trading week (say y) given the analyst recommendation (say r), $f_{Y|R}(y|R=r)$ and the distribution of analyst recommendation given observed data vector of a week, $f_{R|Y}(R|Y=y) \sim Ber[\pi(y,\alpha)]$, where, $\pi(y,\alpha)$ is a function of y parameterized by α. We assume a logistic link:

$$\text{logit}[\pi(y,\alpha)] = log\left(\frac{\pi(y,\alpha)}{1-\pi(y,\alpha)}\right) = \alpha_0 + \alpha_1 y_1 + \alpha_2 y_2$$

Further, R is a binary variable and Y is a continuous variable such that $f_{Y|R}(y|R=r) \sim N_2(\mu^{(r)}, \Sigma^{(r)})$. where,

$$\mu^{(r)} = \begin{pmatrix} \mu_1^{(r)} \\ \mu_2^{(r)} \end{pmatrix} \text{ and } \Sigma^{(r)} = \begin{pmatrix} \sigma_1^{(r)2} & \rho\sigma_1^{(r)}\sigma_2^{(r)} \\ \rho\sigma_1^{(r)}\sigma_2^{(r)} & \sigma_2^{(r)2} \end{pmatrix}$$

Under this model, the conditional distributions of the stock price given its buy/sell status is assumed to be normal, with different set of parameter values depending upon the classification. As we shall see in the next section, we will need to assume that the variance-covariance matrices of the two conditionals must be the same (that is, $\Sigma^{(0)} = \Sigma^{(1)}$) in order to ensure compatibility.

3.2. Deriving conditionally specified model

Confirming the existence of the joint model given the two conditionals is essential. In other words, the two conditionals should be compatible. Therefore, for this problem, we will start by checking whether $f_{R|Y}(R|Y=y)$ and $f_{Y|R}(y|R=r)$ are compatible. We will use the Arnold and Press (1989) Theorem given in the Appendix (A.2). Assume,

$$a(r,y) = (2\pi)^{-1}|\Sigma^{(r)}|^{-1/2}exp[-\frac{1}{2}(y-\mu^{(r)})^T\Sigma^{(r)-1}(y-\mu^{(r)})]$$

and

$$b(r,y) = \pi(y)^r[1-\pi(y)].^{(1-r)}$$

Let us form the ratio,

$$\frac{a(r,y)}{b(r,y)} = \frac{(2\pi)^{-1}|\Sigma^{(r)}|^{-1/2}exp[-\frac{1}{2}(y-\mu^{(r)})^T\Sigma^{(r)-1}(y-\mu^{(r)})]}{\pi(y)^r[1-\pi(y)]^{(1-r)}}$$

Let $\Sigma^{(r)} = \Sigma$, be the common variance-covariance matrix. As in Example 2, letting $-\Sigma^{-1}\mu^{(1)}$ $= (\alpha_1, \alpha_2)^T$, and leaving α_0 unconstrained we can rewrite the above ratio as follows:

$$\begin{aligned}\frac{a(r,y)}{b(r,y)} &= \frac{(2\pi)^{-1}|\Sigma|^{-1/2}exp[-\frac{1}{2}(y^T\Sigma^{-1}y - 2y^T\Sigma^{-1}\mu^{(r)} + \mu^{(r)T}\Sigma^{-1}\mu^{(r)}]}{exp(r\alpha_0).exp(\alpha_0+\alpha_1y_1+\alpha_2y_2)^{-1}} \\ &= \underbrace{\frac{(2\pi)^{-1}|\Sigma|^{-1/2}exp[-\frac{1}{2}(y^T\Sigma^{-1}y)}{exp[\alpha_0+\alpha_1y_1+\alpha_2y_2]^{-1}}}_{\text{U(y)}} \cdot \underbrace{\frac{exp[-\frac{1}{2}(\mu^{(r)T}\Sigma^{-1}\mu^{(r)})]}{exp(r\alpha_0)}}_{\text{V(r)}}\end{aligned}$$

Since $\sum_{r=0}^{1} V(r) < \infty$, compatibility of the given family of conditional densities is assured provided the integrability restriction is also satisfied. That is, the two conditionals are compatible under common variance-covariance matrix Σ and under the condition that $\exp[y^T\Sigma^{-1}y - \alpha_1y_1 - \alpha_2y_2]$ integrates to 1.

Now that the compatibility condition is satisfied, we can apply Theorem 1 to obtain the form of the joint distribution. Theorem 1 was used to obtain the joint distribution of $f(y,r)$. Recall that,

$$\ln(f(y,r)) = (1,r)\begin{bmatrix} m_{00} & m_{01} & m_{02} & m_{03} & m_{04} & m_{05} \\ m_{10} & m_{11} & m_{12} & m_{13} & m_{14} & m_{15}\end{bmatrix}\begin{bmatrix}1 \\ y_1 \\ y_2 \\ y_1^2 \\ y_1y_2 \\ y_2^2\end{bmatrix}$$

For simplicity we write,

$$f(y,r) = exp(m_{00}+rm_{10}).exp(Q(y,r)); r = 0,1 \text{ and } y \in \mathbf{R}$$

where, $Q(y,r) = (m_{01}+rm_{11})y_1 + (m_{02}+rm_{12})y_2 + (m_{03}+rm_{13})y_1^2 + (m_{04}+rm_{14})y_1y_2 + (m_{05}+rm_{15})y_2^2$

Obviously, the tiresome part of this derivation is finding the m values. We will now start by finding the solutions for all the m_{ij} values except for m_{00} which is the normalizing constant. We will present a solution for m_{00} at the end of this section. The usual way to find m values (except m_{00}) is comparing $f(Y|R=r)$ and $f(R|Y=y)$ derived from the joint distribution with the original $f(Y|R=r)$ and $f(R|Y=y)$. Thus, by comparing we obtain

$$f(R=r|Y=y) = \frac{exp(r[m_{10}+m_{11}y_1+m_{12}y_2+m_{13}y_1^2+m_{14}y_1y_2+m_{15}y_2^2])}{1+exp(m_{10}+m_{11}y_1+m_{12}y_2+m_{13}y_1^2+m_{14}y_1y_2+m_{15}y_2^2)} \quad ; \quad r=0,1$$

By comparing the true density and the density of the desired logistic regression model, we have that

$$\pi(y,\alpha) = \frac{exp(m_{10}+m_{11}y_1+m_{12}y_2+m_{13}y_1^2+m_{14}y_1y_2+m_{15}y_2^2)}{1+exp(m_{10}+m_{11}y_1+m_{12}y_2+m_{13}y_1^2+m_{14}y_1y_2+m_{15}y_2^2)}$$

Desired $\pi(y,\alpha)$ can be obtained by setting the quadratic terms into zero. However, having quadratic terms in the general setting reveals that the derived joint density represents a larger class of joint densities. The particular problem that we are interested in is a special case where the logit link is constructed with a linear function only. Thus, by comparing we get

$$m_{10} = \alpha_0 \tag{4}$$

$$m_{11} = \alpha_1 \tag{5}$$

$$m_{12} = \alpha_2 \tag{6}$$

$$m_{13} = m_{14} = m_{15} = 0 \tag{7}$$

Similarly, to obtain $f(Y|R=r)$, we first derive $f(R=r)$ using $f(Y=y,R=r)$.

$$\begin{aligned} &f(R=r) \\ &= \int_{y_2=-\infty}^{\infty}\int_{y_1=-\infty}^{\infty} f(y,r)dy_1dy_2 \\ &= \int_{y_2=-\infty}^{\infty}\int_{y_1=-\infty}^{\infty} \frac{f(y,r)}{f(y|r)} \times f(y|r)dy_1dy_2 \\ &= exp(m_{00}+rm_{10})(2\pi)|\Sigma|^{1/2}\int_{y_1,y_2=-\infty}^{\infty} exp\left\{\frac{1}{2}(y-\mu^{(r)})^T\Sigma^{-1}(y-\mu^{(r)})+Q(y,r)\right\}f(y|r)d\underline{y} \end{aligned}$$

In order to obtain a closed form expression for the normalizing constant, we will need to finish the integration. It turns out that an old result involving multivariate normal density function can be used to accomplish this. The complete result is given in the Appendix (A.1) for the reader's convenience. Once the joint distribution becomes available, the marginals of Y and R can be obtained from the joint distribution. Thus, $f(R=r)$ can be written as follows:

$$\begin{aligned} &(R=r) \\ &= exp(m_{00}+rm_{10})(2\pi)|\Sigma|^{1/2}\int_{y_1,y_2=-\infty}^{\infty} exp\left\{2b\underline{y}+y^TAy\right\}f(y|r)d\underline{y} \\ &= \frac{exp(m_{00}+rm_{10})(2\pi)|\Sigma|^{1/2}}{|I-2A\Sigma|^{1/2}} exp\left\{2b^T\Sigma(I-2A\Sigma)^{-1}b+\mu^{(r)T}(I-2A\Sigma)^{-1}(2b+A\mu^{(r)})\right\} \end{aligned}$$

where,

$$\mathrm{b}=\left(\frac{2\rho\sigma_1\sigma_2\mu_2^{(r)}-2\sigma_2^2\mu_1^{(r)}}{4(1-\rho^2)}+\frac{m_{01}+rm_{11}}{2},\quad \frac{2\rho\sigma_1\sigma_2\mu_1^{(r)}-2\mu_2^{(r)}\sigma_1^2}{4(1-\rho^2)\sigma_1^2\sigma_2^2}+\frac{m_{02}+rm_{12}}{2}\right)^T$$

$$\text{and}\quad \mathrm{A}=\begin{pmatrix}\frac{1}{2(1-\rho^2)\sigma_1^2}+m_{03}+rm_{13} & m_{04}+rm_{14}\\ \frac{-\rho}{(1-\rho^2)\sigma_1\sigma_2} & \frac{1}{2(1-\rho^2)\sigma_2^2}+m_{05}+rm_{15}\end{pmatrix}$$

Let us now obtain the conditional distribution $f(Y|R=r)$.

$$f(Y|R=r)$$
$$=\frac{f(Y=y,R=r)}{f(R=r)}$$

$$=\frac{|I-2A\Sigma_r|^{1/2}}{(2\pi)|\Sigma_r|^{1/2}}exp\left\{Q(y,r)-2b^T\Sigma_r(I-2A\Sigma_r)^{-1}b-\mu_r^T(I-2A\Sigma_r)^{-1}(2b+A\mu_r)\right\}$$

We compare the conditional distribution $f(Y|R=r)$ expressed in terms of the m_{ij} values to the originally specified form of the same distribution $(Y|R=r)$ expressed in terms of the parameters $\underset{\sim}{\mu}^{(r)}$'s and Σ's etc. to obtain the relationships between the two sets of parameters. These relationships are captured in the following equations.

$$m_{01}+rm_{11}=\frac{\mu_1^{(r)}-\rho_r(\sigma_{11}/\sigma_{22})\mu_2^{(r)}}{(1-\rho^2)\sigma_{11}^2}$$

$$m_{02}+rm_{12}=\frac{\mu_2^{(r)}(\sigma_{11}/\sigma_{22})^2-\rho(\sigma_{11}/\sigma_{22})\mu_1^{(r)}}{(1-\rho^2)\sigma_{11}^2}$$

$$m_{03}+rm_{13}=\frac{-1}{2(1-\rho^2)\sigma_{11}^2}$$

$$m_{04}+rm_{14}=\frac{\rho}{(1-\rho^2)\sigma_{11}\sigma_{22}}$$

$$m_{05}+rm_{15}=\frac{-1}{2(1-\rho^2)\sigma_{22}^2} \tag{8}$$

The above equations can be solved for m_{ij} values in terms of the $\mu_1^{(r)}$ and $\mu_2^{(r)}$ values etc.

Expressions for $m_{01}, m_{02}, m_{03}, m_{04}$ and m_{05} are as follows:

$$m_{01} = \frac{\mu_1^{(r)} - \rho(\sigma_{11}/\sigma_{22})\mu_2^{(r)}}{(1-\rho^2)\sigma_{11}^2} - r\alpha_1$$

$$m_{02} = \frac{\mu_2^{(r)}(\sigma_{11}/\sigma_{22})^2 - \rho(\sigma_{11}/\sigma_{22})\mu_1^{(r)}}{(1-\rho^2)\sigma_{11}^2} - r\alpha_2$$

$$m_{03} = \frac{-1}{2(1-\rho^2)\sigma_{11}^2} \tag{9}$$

$$m_{04} = \frac{\rho}{(1-\rho^2)\sigma_{11}\sigma_{22}}$$

$$m_{05} = \frac{-1}{2(1-\rho^2)\sigma_{22}^2}$$

From (9), one can also obtain the original parameters $\mu_1^{(r)}, \mu_2^{(r)}, \rho_r, \sigma_{11}^2$ and σ_{22}^2 in terms of m_{ij}'s. Thus, we have,

$$\rho = \frac{m_{04}}{2\sqrt{m_{03}.m_{05}}}, \quad \sigma_{22}^2 = \frac{2m_{03}}{m_{04}^2 - 4m_{03}.m_{05}} \text{ and } \quad \sigma_{11}^2 = \frac{2m_{05}}{m_{04}^2 - 4m_{03}.m_{05}}. \tag{10}$$

According to (10) it is clear that ρ,σ_{11}^2 and σ_{22}^2 do not depend on the value of R. This implies a common variance-covariance matrix for the conditional distributions $f(Y = y|R = 0)$ and $f(Y = y|R = 1)$. Further, $\mu_1^{(r)}$ and $\mu_2^{(r)}$ as follows.

$$\mu_2^{(r)} = \frac{\rho\left\{\frac{m_{01}+r\alpha_1}{m_{03}}\right\} + \left\{\frac{m_{02}+r\alpha_2}{m_{05}}\right\}}{\rho^2\sqrt{\left\{\frac{m_{05}}{m_{03}}\right\} - \left\{\frac{m_{05}}{m_{03}}\right\}}} \quad \text{and} \quad \mu_1^{(r)} = \rho\sqrt{\frac{m_{05}}{m_{03}}} \times \mu_2^{(r)} - \frac{m_{01}+r\alpha_1}{m_{03}} \tag{11}$$

The above solutions are verified using simulation studies.

3.3. Deriving normalizing constant (m_{00})

In general, obtaining closed form solution to m_{00} is known to be difficult and sometimes such a closed form may not even exist. For our problem however, we were able to derive a closed form expression for m_{00}. Since m_{00} should be such that $\int_{y=-\infty}^{\infty} f(y,r) = 1$, it follows that

$$m_{00} = -ln\left[\sum_{r=0}^{1} exp(rm_{10}) \int_{y1=-\infty}^{\infty} \int_{y2=-\infty}^{\infty} exp(Q(y,r))dy_2 dy_1\right]$$

As shown in the Appendix A.1, the double integral above is given by,

$$\int_{y_1=-\infty}^{\infty} \int_{y_2=-\infty}^{\infty} exp(Q(y,r)) = \frac{2\pi\sigma_{11}\sigma_{22}\sqrt{1-\rho^2}}{exp(C(r))}|I - 2A\Sigma|^{-1/2}$$
$$exp\left[2b^T\Sigma(I-2A\Sigma)^{-1}b + \mu^{(r)T}(I-2A\Sigma)^{-1}(2b + A\mu^{(r)})\right]$$

where,

$$C(r) = \frac{-(\sigma_{22}^2\mu_1^{(r)2} - 2\rho\sigma_{11}\sigma_{22}\mu_1^{(r)}\mu_2^{(r)} + \sigma_{11}^2\mu_2^{(r)2})}{2(1-\rho^2)\sigma_{11}^2\sigma_{22}^2}$$

$$\mathrm{b}=\left(\frac{2\rho\sigma_{11}\sigma_{22}\mu_2^{(r)}-2\sigma_{22}^2\mu_1^{(r)}}{4(1-\rho^2)} + \frac{m_{01}+rm_{11}}{2}, \quad \frac{2\rho\sigma_{11}\sigma_{22}\mu_1^{(r)}-2\mu_2^{(r)}\sigma_{11}^2}{4(1-\rho^2)\sigma_{11}^2\sigma_{22}^2} + \frac{m_{02}+rm_{12}}{2}\right)^T$$

$$\text{and A}=\begin{pmatrix} \frac{1}{2(1-\rho^2)\sigma_{11}^2} + m_{03} + rm_{13} & m_{04} + rm_{14} \\ \frac{-\rho}{(1-\rho^2)\sigma_{11}\sigma_{22}} & \frac{1}{2(1-\rho^2)\sigma_{22}^2} + m_{05} + rm_{15} \end{pmatrix}.$$

Therefore, m_{00} is given by,

$$m_{00} = -ln\Bigg[2\pi\sigma_{11}\sigma_{22}\sqrt{1-\rho^2}\sum_{r=0}^{1}|I-2A\Sigma|^{-1/2}.$$
$$exp\Bigg(rm_{10} - C(r) + 2b^T\Sigma(I-2A\Sigma)^{-1}b + \mu^{(r)T}(I-2A\Sigma)^{-1}(2b + A\mu^{(r)})\Bigg)\Bigg]$$

4. Data Generation and Estimation

In this section, we discuss how to generate data from the new conditionally specified joint distribution. We also discuss estimation of the parameters of the joint distribution using the maximum likelihood estimation (MLE) method and pseudolikelihood estimation (PLE) method. We will also provide some numerical results to investigate the computational effort involved in computing MLEs and PLEs.

4.1. Data generation from the model

Although the proposed joint model has a closed form expression, it is very complex and has a messy normalizing constant. Therefore, generating data directly from the joint model is immensely difficult and may even not be feasible. However, since the model is conditionally specified we can apply other numerical algorithms such as Gibb's Sampling.

Gibbs sampling algorithm, named by Geman and Geman (1984), is a special case of Metropolis-Hastings algorithm. It is a way to generate data from multivariate distributions when the univariate conditional densities are fully specified. For more details the reader is referred to Rizzo(2008). For instance, let us assume that we want to generate data from a bivariate density $f_{X,Y}(x, y)$ and the conditional distributions of the model are fully specified. Let $f_{X|Y}(x/y)$ and $f_{Y|X}(y/x)$ be conditional densities of x given y and y given x respectively. Suppose simulating data from the bivariate density is complicated, Gibbs sampling algorithm is an effective method to generate data from the conditional densities despite having information about the marginal distributions $f_X(x)$ and $f_Y(y)$. Presented below is the algorithm which can be used to generate data from the proposed density.

Gibb's Sampling Algorithm

1. Initial value : Set $R^{(0)} = 1$. So, $Y^{(0)} \sim f_1(y)$.
2. Set $\mu^{(0)}$,$\mu^{(1)}$,Σ and the vector α.
3. Suppose we generated $(Y^{(0)}, R^{(0)}), (Y^{(1)}, R^{(1)}), ..., (Y^{(t)}, R^{(t)})$.
   (a) if $R^{(t)} = 0$ Generate $Y^{(t+1)} \sim f(y/R = 0)$ and $R^{(t+1)} \sim f(r/Y = y^{(t+1)})$
   (b) if $R^{(t)} = 1$ Generate $Y^{(t+1)} \sim f(y/R = 1)$ and $R^{(t+1)} \sim f(r/Y = y^{(t+1)})$

Contour plots and Surface plots (Figure 1) for $f(Y, R = 0)$ and $f(Y, R = 1)$ suggest that the joint model is not a unimodal distribution. We have built an interactive tool for exploring the shape of the joint distribution in Shiny. The reader may download this tool from *https://github.com/nadeesriw/ShinyApp.git*.

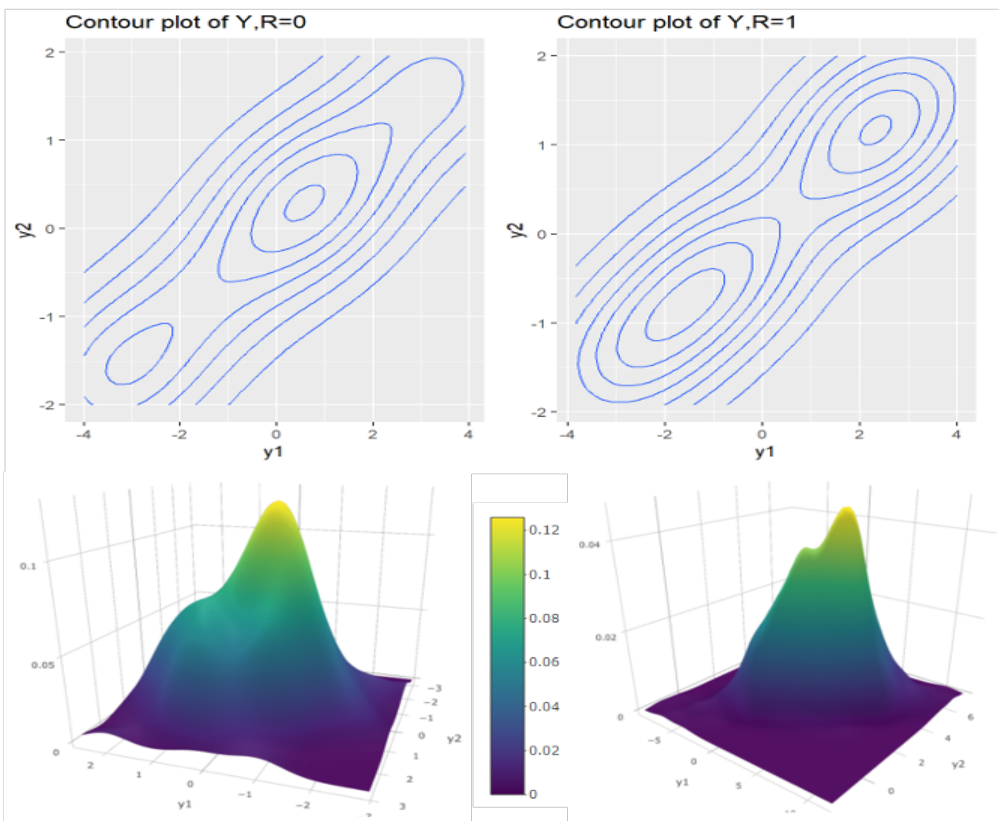

Figure 1: Contour plots and Surface plots of the joint distribution. Top and bottom left: $f(Y, R = 0)$ and top and bottom right: $f(Y, R = 1)$.

4.2. Estimating parameters

As stated in Arnold *et al.* (1991, 2001) papers, standard estimation methods are often difficult to implement when dealing with conditionally specified models. This is mainly because of the awkward normalizing constant m_{00} which is often intractable. And, if an explicit expression is available for m_{00}, it is usually complicated, as is the case here. Thus, differentiating the likelihood and deriving the maximum likelihood equations is challenging. Although in practice MLE is the preferred estimation method for parametric models, in our case it comes with a heavy computational burden. Therefore, it behooves upon us to explore other methods such as pseudolikelihood estimation (PLE). In this section, we will compare MLE and PLE in terms of relative efficiency and computational cost.

The likelihood function of the proposed joint distribution is given by,

$$L(f(Y_i, R_i; \hat{\underset{\sim}{M}})) = \prod_{i=1}^{n} \frac{exp(rm_{10} + Q(\underset{\sim}{y}, r))}{\sum_{r=0}^{1} exp(rm_{10}) \int_{y1=-\infty}^{\infty} \int_{y2=-\infty}^{\infty} exp(Q(\underset{\sim}{y}, r)) dy_2 dy_1} \tag{12}$$

where,

$$Q(\underset{\sim}{y}, r) = (m_{01} + rm_{11})y_1 + (m_{02} + rm_{12})y_2 + (m_{03} + rm_{13})y_1^2 + (m_{04} + rm_{14})y_1 y_2 + (m_{05} + rm_{15})y_2^2$$

Deriving the score function and obtaining closed form expressions for estimates is not feasible due to the complexity of the model. Therefore, numerical methods were considered to obtain the estimates of parameters.

Arnold and Strauss (1991) proposed PLE as an alternative method to maximum likelihood estimation. Further, they proposed the product of the two conditional distributions as a possible pseudolikelihood function. The advantages of this method are that the function is far more tractable compared to the original likelihood and the awkward normalizing constant m_{00} is absent. The proposed pseudolikelihood function can be written as follows:

$$PL(\mu^{(0)}, \mu^{(1)}, \Sigma, \alpha_0, \alpha_1, \alpha_2) = \prod_{i=1}^{n} f(y_i | R_i = r_i) f(R_i = r_i | y_i)$$

$$\begin{aligned}
&PL(\mu^{(r)}, \Sigma, \alpha_0, \alpha_1, \alpha_2) \\
&= \prod_{i=1}^{n} \frac{1}{2\pi|\Sigma|^{1/2}} \times \\
&exp\left\{ -\frac{1}{2}(\underset{\sim}{y_i} - \mu^{(1)} r_i - \mu^{(0)}(1 - r_i))^T [\Sigma]^{-1} (\underset{\sim}{y_i} - \mu^{(1)} r_i - \mu^{(0)}(1 - r_i)) \right\} \times \\
&[exp\{\underset{\sim}{\alpha}^T y_i^*\}]^{r_i} \frac{1}{1 + exp\{\underset{\sim}{\alpha}^T y_i^*\}}
\end{aligned}$$

where, $\alpha = (\alpha_0, \alpha_1, \alpha_2)^T$ and $\underset{\sim}{y_i^*} = (1, \underset{\sim}{y_i})^T$. While the computation of PLE is also only feasible numerically, the computations involved are significantly less complex than the computation of the original likelihood.

We carried out a simulation study for different sample sizes varying from 100 to 1000. Estimates of the model parameters were obtained using both MLE and PLE methods. Data were generated using the Gibb's algorithms. 100 such data sets were used for the study. Apart from the estimates, we also calculated the variance, bias and time. The entire analysis was performed using the R 3.5.1 software. The R package called *Rsolnp*, which performs constrained optimization, was used to obtain estimates. Results are shown in Table 1 and Table 2.

Table 1: Wall times of simulation study with respect to sample size and method used: TOL 10^{-3} and 100 data sets

Sample Size (n)	Wall time (in hours)	
	Maximum Likelihood Estimates	Pseudolikelihood Estimates
100	0.24876	0.03046
200	0.48388	0.05411
500	1.01297	0.12695
1000	18.74536	13.94225
10000	29.14374	15.74069

In Table 1, the first column shows the wall times for different sample sizes n, when the maximum likelihood method was used. The second column of Table 1 presents the wall times for pseudolikelihood method. According to results, we can see that pseudolikelihood method is superior to maximum likelihood method in terms of computation efficiency. For n=100, wall time of pseudolikelihood estimation for 100 iterations is 0.0304 hours (1.83 mins) while the wall time of maximum likelihood estimation for 100 iterations is 0.2488 hours ($\sim$ 15 mins). It is apparent that for small sample sizes such as $n = 100$, 200 there is no significant computational cost difference between the two methods. However, for a larger sample size, the computational advantage of PLE surpasses that of MLE quickly. Further, we note that even the PLE method shows large wall times when n increases. In that case, paralleling the code would be more effective. We are planning to explore this in the future.

In Table 2, we present the estimates, the bias and the variances for different sample sizes. By looking at the bias values it is clear that the MLE method has less bias (and nearly zero in some cases) than the PLE method. In both methods, the variances of the estimates decrease as the sample size increases. The MLE of the α vector has less bias compared to the PLE. Overall, it is evident that the MLE outperforms the PLE in terms of accuracy (less bias) and efficiency (lower variance). Therefore, we can conclude that choosing PLE over MLE is a trade off between efficiency and computational cost.

5. Illustrative Example

In this section we provide an illustrative example. We apply the derived joint distribution to a data set with stock prices and a weekly buy/sell recommendation for stocks. All the data has been downloaded from *www.Barchart.com*. The dataset consists of closing prices and recommendations from Monday to Friday for 40 weeks. As the starting step, we only considered the two closing prices from Monday and Friday of each week and the recommendation from Monday of the following week. The data structure is given in Table 3.

Table 2: Maximum likelihood estimates and pseudolikelihood estimates for different sample sizes (n): TOL 10^{-3}

Parameters	True Values	Maximum Likelihood		Pseudolikelihood	
		Estimate (std)	Bias	Estimate (std)	Bias
		$n = 100$			
σ_0^2	1.2000	1.1986(0.0100)	-0.0014	1.1836(0.0213)	-0.0164
σ_{01}	0.6481	0.6455(0.0100)	-0.0026	0.6510(0.0135)	0.0029
σ_1^2	1.4000	1.3981(0.0100)	-0.0018	1.4273(0.0213)	0.0273
$\mu_1^{(1)}$	2.0000	2.1828(0.0173)	0.1828	1.9879(0.0176)	-0.0120
$\mu_2^{(1)}$	3.0000	3.2157(0.0141)	0.2157	3.0179(0.0253)	0.0179
$\mu_1^{(0)}$	0.0000	-0.0104(0.0223)	-0.0104	0.0053(0.0182)	0.0053
$\mu_2^{(0)}$	0.0000	-0.0213(0.0173)	-0.0213	-0.0008(0.0218)	-0.0008
α_0	0.0010	0.0756(0.0012)	0.0746	-2.3291(0.2972)	-2.3301
α_1	0.0010	0.0009(0.0000)*	-0.0001	0.5229(0.1117)	0.5219
α_2	0.0010	0.0009(0.0000)*	-0.0001	1.2877(0.1710)	1.2867
		$n = 200$			
σ_0^2	1.2000	1.1993(0.0001)	-0.0007	1.2119(0.0085)	0.0119
σ_{01}	0.6481	0.6458(0.0007)	-0.0023	0.6498(0.0053)	0.0017
σ_1^2	1.4000	1.3990(0.0017)	-0.0009	1.4205(0.0073)	0.0205
$\mu_1^{(1)}$	2.0000	2.1824(0.0001)	0.1824	1.9785(0.0076)	-0.0215
$\mu_2^{(1)}$	3.0000	3.2145(0.0001)	0.2145	2.9683(0.0115)	-0.0317
$\mu_1^{(0)}$	0.0000	-0.0124(0.0002)	-0.0124	-0.0049(0.0073)	-0.0049
$\mu_2^{(0)}$	0.0000	-0.0219(0.0001)	-0.0219	0.0013(0.0130)	0.0013
α_0	0.0010	0.0736(0.0000)*	0.0726	-2.2607(0.1565)	-2.2617
α_1	0.0010	0.0009(0.0000)*	-0.0001	0.5220(0.0565)	0.5210
α_2	0.0010	0.0009(0.0000)*	-0.0001	1.1872(0.0994)	1.1862
		$n = 500$			
σ_0^2	1.2000	1.1996(0.0000)*	-0.0003	1.2116(0.0009)	0.0116
σ_{01}	0.6481	0.6458(0.0008)	-0.0023	0.6434(0.0014)	-0.0046
σ_1^2	1.4000	1.3995(0.0000)*	-0.0005	1.4109(0.0010)	0.0109
$\mu_1^{(1)}$	2.0000	2.1842(0.0078)	0.1842	1.9995(0.0036)	-0.0005
$\mu_2^{(1)}$	3.0000	3.2158(0.0080)	0.2158	3.0081(0.0056)	0.0081
$\mu_1^{(0)}$	0.0000	-0.0107(0.0112)	-0.0107	0.0035(0.0045)	0.0035
$\mu_2^{(0)}$	0.0000	-0.0210(0.0069)	-0.0210	0.0069(0.0047)	0.0069
α_0	0.0010	0.0755(0.0037)	0.0745	-2.3054(0.0940)	-2.3064
α_1	0.0010	0.0009(0.0000)*	-0.0001	0.4747(0.0171)	0.4737
α_2	0.0010	0.0009(0.0000)*	-0.0001	1.1533(0.0153)	1.1523
		$n = 1000$			
σ_0^2	1.2000	1.1998(0.0000)*	-0.0002	1.2039(0.0002)	0.0039
σ_{01}	0.6481	0.6457(0.0009)	-0.0024	0.6434(0.0009)	-0.0046
σ_1^2	1.4000	1.3997(0.0000)*	-0.0003	1.4085(0.0003)	0.0085
$\mu_1^{(1)}$	2.0000	2.1814(0.0088)	0.1814	1.9947(0.0019)	-0.0053
$\mu_2^{(1)}$	3.0000	3.2237(0.0107)	0.2237	2.9945(0.0032)	-0.0055
$\mu_1^{(0)}$	0.0000	-0.0250(0.0096)	-0.0250	-0.0042(0.0021)	-0.0042
$\mu_2^{(0)}$	0.0000	-0.0241(0.0064)	-0.0241	0.0009(0.0019)	0.0009
α_0	0.0010	0.0702(0.0035)	0.0692	-2.3886(0.0976)	-2.3896
α_1	0.0010	0.0009(0.0000)*	-0.0001	0.4790(0.0096)	0.4780
α_2	0.0010	0.0009(0.0000)*	-0.0001	1.1613(0.0094)	1.1603

$*$: very small non zero values

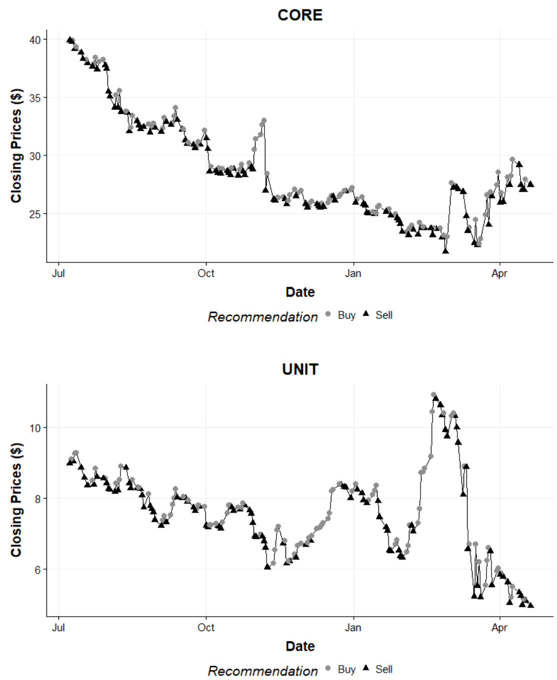

Figure 2: Closing prices, recommendations

Table 3: Data structure

		Closing Prices (Mon.price, Fri.price)	Recommendation
Stock 1	Week 1	$(y_{11}^{(1)}, y_{11}^{(2)})$	Buy
	Week 2	$(y_{12}^{(1)}, y_{12}^{(2)})$	Buy
	⋮	⋮	⋮
	Week 40	$(y_{1,32}^{(1)}, y_{1,40}^{(2)})$	Sell
Stock 2	Week 1	$(y_{21}^{(1)}, y_{21}^{(2)})$	Buy
	Week 2	$(y_{22}^{(1)}, y_{22}^{(2)})$	Sell
	⋮	⋮	⋮
	Week 40	$(y_{2,32}^{(1)}, y_{2,40}^{(2)})$	Sell
Stock$_i$	week$_j$	$(y_{ij}^{(1)}, y_{ij}^{(2)})$	r_{ij}

Conditional distributions of closing prices of Monday and Friday of each week given the end of the week recommendation, is assumed to follow a bivariate normal distribution and the end of the week recommendation given the closing prices of Monday and Friday is assumed to follow a standard logistic regression model. The normality assumption is crucial. In this example, we choose stocks where closing prices follow bivariate normal distribution. However, most of the closing prices of stock data do not follow a bivariate normal distribution and hence one might need to transform the data using methods such as Box-Cox transformation. The pseudolikelihood and the maximum likelihood methods are used to estimate the parameters. Results are given in Table 4. Note that the means, variances and correlation estimates

Table 4: Maximum likelihood and pseudolikelihood estimates for stock price data

Stock	Method	$\mu^{(0)}$	$\mu^{(1)}$	σ_1^2	σ_2^2	$\rho\sigma_1\sigma_2$	α_0	α_1	α_2
CORE	Pseudolikelihood	(28.80,28.50)	(29.14,28.59)	19.92	18.29	18.47	-0.36	0.20	-0.19
	Maximum Likelihood	(28.80,28.51)	(29.14,28.60)	19.55	17.96	18.12	-0.54	0.40	-0.39
UNIT	Pseudolikelihood	(7.51,7.33)	(7.70,7.76)	1.73	1.62	1.54	-2.37	-0.807	1.02
	Maximum Likelihood	(7.51,7.33)	(7.70,7.76)	1.72	1.62	1.54	-4.01	-1.64	2.08

obtained by the two methods are very close to each other. However, the estimates of the parameters of the logistic regression model are not so close. The other important point to note is that the α_2 parameter seems to be significantly different from zero. This implies that the product of the Monday and Friday prices is significant in the logistic regression model.

6. Remarks

A researcher can get better insight into the data using conditional specification of experimental variables rather than a joint distribution. Using conditional specification makes the visualization easier. Furthermore, it makes it easier to use Gibbs sampling for data generation. We derived Joint distribution starting from $Y|R = r$ bivariate normal and $R|Y = y$ logistic regression.

In practice, the maximum likelihood method is a preferred estimation method for parametric models. However, if the likelihood function is complicated or contains an awkward normalizing constant, the implementation becomes difficult. Among other methods, pseudolikelihood estimation is intuitively appealing and computationally less burdensome. According to the simulation study, choosing the pseudolikelihood method over the maximum likelihood method is a trade off between the efficiency and accuracy of estimates and the computational cost.

Acknowledgments

The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). The facility is partially supported by the U.S. National Science Foundation through the MRI program (OAC–1726023). See *hpcf.umbc.edu* for more information on HPCF and the projects using its resources.

References

Arnold, B. C., Castillo, E. and Sarabia, J. M. (1999). *Conditional Specification of Statistical Models.* Springer.

Arnold, B. C., Castillo, E. and Sarabia, J. M. (2001). Conditionally specified distributions: An introduction. *Statistical Science,* **16**, 249–274.

Arnold, B. C., Castillo, E. and Sarabia, J. M. (2002). Exact and near compatibility of discrete conditional distributions. *Computational Statistics and Data Analysis,* **40**, 231–252.

Arnold, B. C. and Gokhale, D. (1994). On uniform marginal representation of contingency tables. *Statistic and Probability Letters,* **21**, 311–316.

Arnold, B. C., James, S. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association,* **84**, 152–156.

Arnold, B. C. and Strauss, D. (1988). Bivariate distributions with exponential conditionals. *Journal of the American Statistical Association,* **83**, 522–527.

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: Some examples. *Sankhya: The Indian Journal of Statistics* **B(53)**, 233–243.

Besag, J. (1974). Spatial interaction and the statistical analysis of Lattice systems. *Journal of the Royal Statistical Society,* **B(36)**, 192–225.

Castillo, E. and Galambos, J. (1989). Conditional distributions and the bivariate normal distribution. *Metrika,* **36**, 209–214.

Chen, H. Y. (2010). Compatibility of conditionally specified models. *Statistics and Probability Letters,* **80**, 670–677.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **6(6)**, 721–741.

Ghosh, I. and Nadarajah, S. (2017). On the construction of a joint distribution given two discrete conditionals. *Studia Scientiarum Mathematicarum Hungarica,* **54**, 178–204.

Hobert, J. P. and Casella, G. (1998). Functional compatibility, Markov chains, and Gibbs sampling with improper posteriors. *Journal of Computational and Graphical Statistics,* **7**,42–60.

Hosseini, M. (2017). *Statistical Modeling of Self and Proxy Responses with Gerontological Applications.* PhD thesis, University of Maryland Baltimore County.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rizzo, M. L. (2008). *Statistical Computing with R.* Chapman Hall/CRC.

APPENDIX

1. Moment Generating Function Theorem

This theorem can be used to derive the moment generating function of a linear or a quadratic form in a normal random vector $\underline{x}$ or, more generally, of a second degree polynomial in $\underline{x}$.

Theorem A.1: Let $\underline{x} \sim N_n(\mu, \Sigma)$ and take $q_i = 2b_i\underline{x} + \underline{x}^T A\underline{x}$ where b_i is an $n \times 1$ non random vector and A_i is an $n \times n$ non random symmetric matrix (i=1,2,...,k). Take Γ to be any $r \times n$ matrix such that $\Sigma = \Gamma^T\Gamma$, where $r = \text{rank}\Sigma$. The joint m.g.f. of $q_1, ..., q_k$ is given by,

$$m_{q_1,...,q_k}(t_1,...,t_k) = |I - 2\sum_{i=1}^{k} t_i\Gamma A_i\Gamma^T|^{-1/2}.exp\left\{2\left[\sum_{i=1}^{k} t_i(b_i + A_i\mu)\right]^T \Gamma^T \left[I - 2\sum_{i=1}^{k} t_i\Gamma A_i\Gamma^T\right]^{-1} \Gamma \left[\sum_{i=1}^{k} t_i(b_i + A_i\mu)\right] + \mu^T \sum_{i=1}^{k} t_i(2b_i + A_i\mu)\right\}$$

$$= |I - 2\sum_{i=1}^{k} t_i\Gamma A_i\Gamma^T|^{-1/2}.exp\left\{2\left[\sum_{i=1}^{k} t_ib_i\right]^T \Sigma \left[I - 2\sum_{i=1}^{k} t_iA_i\Sigma\right]^{-1} \left[\sum_{i=1}^{k} t_ib_i\right] + \mu^T[I - 2\sum_{i=1}^{k} t_iA_i\Sigma\Big]^{-1} \sum_{i=1}^{k} t_i(2b_i + A_i\mu)\right\}$$

where $(|t_i| < h_i; i = 1, ..., k)$. for sufficiently small positive constants $h_1, ..., h_k$.

2. Compatibility Theorem: Arnold and Press (1989)

we have,

$a(x,y) = f_{X|Y}(x|y),\ \ x \in S(X),\ \ y \in S(Y),$

$b(x,y) = f_{Y|X}(y|x),\ \ x \in S(X),\ \ y \in S(Y),$

and let, $N_A = (x,y) : a(x,y) > 0$ and $N_B = (x,y) : b(x,y) > 0$

Theorem A.2: A joint density $f(x,y)$, with $a(x,y)$ and $b(x,y)$ as its conditional densities, will exist if and only if,

1. N_A=N_B=N
2. ∃ functions u and v such that ∀ x,y,∈ N

$$a(x,y)/b(x,y) = u(x)v(y)$$

where

$$\int_{S(X)} u(x) d\mu_1(x) < \infty$$