Applications and Applied Mathematics: An International Journal (AAM)

Volume 6 | Issue 1 Article 7

6-2011

A Multivariate Variable Model with Possibility of Missing Data on a Stochastic Process

Ehsan B. Samani
Shahid Beheshti University

Follow this and additional works at: https://digitalcommons.pvamu.edu/aam

Part of the Statistics and Probability Commons

Recommended Citation

Samani, Ehsan B. (2011). A Multivariate Variable Model with Possibility of Missing Data on a Stochastic Process, Applications and Applied Mathematics: An International Journal (AAM), Vol. 6, Iss. 1, Article 7.
Available at: https://digitalcommons.pvamu.edu/aam/vol6/iss1/7

This Article is brought to you for free and open access by Digital Commons @PVAMU. It has been accepted for inclusion in Applications and Applied Mathematics: An International Journal (AAM) by an authorized editor of Digital Commons @PVAMU. For more information, please contact hvkoshy@pvamu.edu.

Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466

Vol. 6, Issue 1 (June 2011) pp. 90 – 104
(Previously, Vol. 6, Issue 11, pp. 1830 – 1844)

Applications and Applied Mathematics:
An International Journal
(AAM)

A Multivariate Variable Model with Possibility of Missing Data on a Stochastic Process

Ehsan Bahrami Samani

Department of Statistics
Faculty of Mathematical Science
Shahid Beheshti University
Tehran, Iran
ehsan_bahrami_samani@yahoo.com

Received: January 14, 2011; Accepted: May 13, 2011

Abstract

A joint model for multivariate responses with potentially non-random missing values on a stochastic process is proposed. A full likelihood-based approach that allows yielding maximum likelihood estimates of the model parameters is used. Sensitivity of the results to the assumptions is also investigated. A common way to investigate whether perturbations of model components influence key results of the analysis is to compare the results derived from the original and perturbed models using a general index of sensitivity (ISNI). The approach is illustrated by analyzing a finance data set.

Keywords: Brownian motion, Sensitivity Analysis, Missing data, ISNI, Finance data

AMS (2010) No.: 62F03

1. Introduction

Financial data with possibility of missing values, are pervasive and research on analyzing them needs to be promoted. Covariates, often relevant to the study at hand are not considered for reasons of parsimony. Specific financial time series which may be useful predictors are omitted from the analysis because they are either entirely missing or only partially observed. Continuous

time models in finance, typically diffusions may (or may not) provide a reasonable model for the fluctuations in prices of a given asset. In any case these prices, even if the market were essentially perfect, are only observed at specific times when trades are made. This may be of little consequence for highly liquid equities and benchmark bonds, but is a much more significant issue with thinly traded or illiquid assets.

Although most modern models for interest rates, bond yields, equity prices, etc. are continuous time multivariate models, these are the very models most susceptible to problems associated with asynchronous trading and missing data. DeCesare (2006), for example, describes data set from consists of one minute discretized tick data from the Toronto Stock Exchange (TSE) dated on February 2, 2005 on three bank stocks: the Bank of Nova Scotia (BNS), Royal Bank (RY) and Bank of Montreal (BMO). The data consist of 392 time points with 12, 51 and 45 missing values for BNS, RY and BMO, respectively. A stock price at a particular time point is reported missing if no sales were made in the last minute of trading on that stock.

The Black- Scholes Model was the first and is the most widely used model for pricing options. The Black- Scholes (1973) used the equilibrium Capital Asset Pricing Model (CAPM) to derive an equation for the option price and had the insight that they could assume for valuation purposes the option had expected return equal to the riskless rate. The assumptions underlying the Black and Scholes option pricing model are as follows:

(a) The market is arbitrage free.
(b) Frictionless and continuous markets. There are no transaction cost such as differential taxes, trading takes place continuously, assets are infinitely divisible, unlimited borrowing and short selling are allowed and borrowing and lending rates are equal.
(c) The riskless instantaneous interest rate is constant over time.
(d) The dynamics for the price of the risky traded asset S (which pays no dividends) are given by

$$d[\log(S_t)] = \mu dt + \sigma dW_t, \tag{1}$$

where μ is the instantaneous expected rate of return on assets S, σ is its instantaneous volatility, both constants and W is a standard dimensional Brownaian motion process.
(e) Investors prefer more to less, agree on σ^2 and dynamics (1).

The model and associated call and put option formulas have revolutionized finance theory and practice and the surviving inventors Merton and Scholes received the nobel prize in economics in 1997 for their contributions. Black and Scholes (1973) and Metron (1971) introduced the key concept of dynamic hedging whereby the option pay off is replicated by a trading strategy in the underlying asset. They derive their formula under lognormal dynamics for the asset price ($\log(S_t)$). Suppose that there are various segments of the components of S_t that are missing. This happens, for example, if S_t is the price of various assets and these prices are only observed at specific discrete trading times. Imputation (or conditional simulation) of the missing pieces of the sample paths of S_t is discussed in several settings. When S_t is a Brownian motion the conditioned process is a tied down Brownian motion.

Rubin (1976), Little and Rubin (2002), Diggle and Kenward (1994) made important distinctions between the various types of missing mechanisms for each of the above mentioned patterns. They define the missing mechanism as missing completely at random (MCAR) if missingness is dependent neither on the observed responses nor on the missing responses, and missing at random (MAR) if, given the observed responses, it is not dependent on the missing responses. Missingness is defined as non-random if it depends on the unobserved responses. From a likelihood point of view MCAR and MAR are ignorable but not missing at random (NMAR) is non-ignorable.

Standard methods of analysis based on the strong and unverifiable assumption of missing at random mechanism could be highly misleading. A way out of this problem is to model both responses and the missing mechanisms jointly. One can then use a direct estimation process to fit a non-ignorable model for the data. As there is always little information about the missing process, these models lead to challenging calculations or even unidentifiability problem. Another alternative is sensitivity analysis, in which one estimates models under a range of assumptions about non-ignorability parameters to study the impact of these parameters on key inferences. Some previous authors have carried out this type of sensitivity analysis exactly for specific complete and missing data models [Bahrami Samani and Ganjali (2008) and Baker et al. (1993)], whereas others have proposed approximate analysis that assess sensitivity in the neighborhood of the ignorable model [Berridige and Dos Santos (1996), Catalano and Ryan (1992) and Cox and Wermuth (1992)].

We want to apply the approximate sensitivity analysis of Cox and Wermuth (1992) who introduced a general index of sensitivity to non-ignorability (ISNI) to measure sensitivity of key inferences in a neighborhood of MAR model needless of fitting a complicated NMAR model. They presented this index for univariate generalized linear models. Recently, Diggle and Kenward (1994) applied ISNI methodology to examine non-ignorability for univariate longitudinal non-Gaussian data. However, since in practice, statistical analysis involving responses of both discrete and continuous types are extremely common, in this paper we will extend ISNI methodology to analyze multivariate longitudinal mixed data subject to non-ignorable dropout.

There is a class of models that is particularly well-suited to the treatment of missing or incomplete data, these are the multivariate processes which are transformations of a Gaussian process. This includes multivariate Brownian motion of these under a transformation of time (sometimes referred to as a subordinated Brownian motion), as well as common stationary parameter μ and diffusion matrix Σ. Kofman and Sharpe (2003) provided a small survey of papers in financial journals which explicitly recognized the presence of missing data. Using multiple imputation in the analysis of incomplete observations in Finance. Malhotra (1987) applied market research data with incomplete information on the dependent variable DeCesare (2006) used imputation of the missing pieces of the sample paths of Brownian motion is discussed in several setting and applied in financial data, who assume an ignorable missing mechanism, for these data. In the presence of a (possibly large) number of missing observations, these estimators can not be used. In theory, of course, since the joint distribution of all of the data (observed and unobserved) is multivariate normal, we could obtain the parameters of the conditional distribution of the unobserved given the observed data.

In this paper a general multivariate model simultaneously handling response and non-response in Brownian motion with potentially non random missing values in responses is presented. This model also considers a probit model. The presented model simultaneously considers probit regression models to allow the examination of the missing mechanisms. The remainder of this paper is organized as follows: Section 2 states the model and its likelihood in the general case with and without the assumption of a non-ignorable dropout. Section 3 derives ISNI calculation in the above mentioned model where there is a vector of nonignorability parameter. Section 4 applies the methodology and simulation to the Finance data. Finally some concluding remarks are given in Section 5.

2. Model and Likelihood

2.1 Complete Data Model

The $p-$dimensional multivariate Brownian motion $X(t) = (X_1(t),..., X_p(t))'$ process with drift and diffusion parameters given by $\Theta = (\mu, \Sigma)$ and expressed with a stochastic differential equation:

$$dX(t) = \mu dt + \Sigma^{\frac{1}{2}} dW(t)$$

for a standard $p-$dimensional multivariate Brownaian motion process W, where $\Sigma^{\frac{1}{2}}$ denotes a matrix square root of Σ and $X(t)$ is multivariate normal distribution with $\mu = (\mu_1,..., \mu_p) \in R^p$ mean and Σ is $p \times p$ matrix covariance. Component of $X(t)$ at times $t_0 < t_1 < t_2 < ... < t_n$ is observed. For the moment assume that the time points t_0, ..., t_n are equally spaced, and denote the common time increment as $(t_i - t_{i-1})$. The natural approach taken by many authors is to use an Euler approximation of the process which is of the form

$$X_{t_i - t_{i-1}} \mid X(t_0): \quad N(X(t_0) + (t_i - t_{i-1})\mu_j, (t_i - t_{i-1})\Sigma),$$

where $X(t_0) = 0$. Given the data the matrix of natural sufficient statistics used in estimating $\Theta = (\mu, \Sigma)$ is easily constructed by appealing to the independent increments property. Let $X_{ij} = X_j(t_i)$, $Y_{ij} = X_{ij} - X_{i-1,j}$, $i = 1,..., n$ and $j = 1,..., p$. The model takes the form:

$$Y_{ij} = (t_i - t_{i-1})\mu_j + \varepsilon_{ij}, \quad i = 1,..., n, \; j = 1,..., p, \tag{2}$$

where random vectors $\varepsilon_i = (\varepsilon_{i1},..., \varepsilon_{ip})'$ are independent with multivariate norma with zero mean and $(t_i - t_{i-1})\Sigma$ matrix covariance.

The likelihood function of $Y = (Y_1,..., Y_n)'$ is

$$L(\Theta \mid y) = \prod_{i=1}^{n} f(y_i \mid \theta) = \frac{1}{(2\pi)^{\frac{np}{2}} |\Sigma|^{\frac{n}{2}} \prod_{i=1}^{n} (t_i - t_{i-1})^{\frac{1}{2}}}$$
$$\times \exp(-\frac{1}{2}\sum_{i=1}^{n} \frac{[Y(t_i) - \mu(t_i - t_{i-1})]' \Sigma^{-1} [Y(t_i) - \mu(t_i - t_{i-1})]}{t_i - t_{i-1}}),$$

where $X(t_i) = (X_1(t_i),..., X_p(t_i))'$, $Y_i = (Y_{i1},..., Y_{ip})'$, $Y_{ij} = Y_j(t_i) = X_j(t_i) - X_j(t_{i-1})$ and the maximum likelihood estimator of μ and Σ are easily, respectively, determined as:

$$\hat{\mu} = \frac{Y(t_i)}{t_i - t_{i-1}}$$

and

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \frac{[Y(t_i) - \hat{\mu}(t_i - t_{i-1})]'[Y(t_i) - \hat{\mu}(t_i - t_{i-1})]}{t_i - t_{i-1}}.$$

2.2. Incomplete Data Model

Suppose we partially observe the process at times $t_0,..., t_n$. In other words, we observed at last one vector component of $X(t)$ at each of the given times and for simplicity we will take $X(t_0)$ to be observed. Typically, when missing data occur in an outcome, assume $R_{y_i} = (R_{y_{i1}},..., R_{y_{ip}})'$ as the indicator vector of responding to Y_i and $R_{y_{ij}}$ is defined as

$$R_{y_{ij}} = \begin{cases} 0 & \text{if } y_{ij} \text{ is not observed} \\ 1 & \text{if } y_{ij} \text{ is observed.} \end{cases}$$

We consider Bernoulli distribution for $R_{y_{ij}}$ with success probability of begin observed dependents on the outcome Y_{ij} according to a specified link function, follows:

$$logit(\xi_{ij}) = logit[P(R_{Y_{ij}} = 1 \mid Y_{i,obs}, Y_{i,mis}, \Gamma_0, \Gamma_1)] = (t_i - t_{i-1})\mu_j + \Gamma_0' Y_{i,obs} + \Gamma_1' Y_{i,mis},$$

where

$$J_{obs}^{y} = \{ j : y_{ij} \text{ is observed}\},\ J_{mis}^{y} = (J_{obs}^{y})^C,\ \ Y_{i,obs} = \{Y_{ij}, \forall j \in J_{obs}^{y}\},\ \ Y_{i,mis} = \{Y_{ij}, \forall j \in J_{mis}^{y}\}.$$

Supposed the g_1- elements of Y_i is observed, and the g_2- elements of Y_i is missing, so

$J_{obs}^{y} = \{o_1,...,o_{g_1}\}$, $J_{mis}^{y} = \{M_1,...,M_{g_2}\}$, $\Gamma_0 = (\gamma_{0o_1},...,\gamma_{0o_{g_1}})$ and $\Gamma_1 = (\gamma_{1M_1},...,\gamma_{1M_{g_2}})$.

The joint model takes the form:

$$\begin{aligned} &Y_{ij} = (t_i - t_{i-1})\mu_j + \varepsilon_{ij}, \quad i = 1,...,n, \ j = 1,..., p \\ &logit[P(R_{Y_{ij}} = 1 \mid Y_{i,obs}, Y_{i,mis}, \Gamma_0, \Gamma_1)] = (t_i - t_{i-1})\mu_j + \Gamma_0' Y_{i,obs} + \Gamma_1' Y_{i,mis}, \end{aligned} \tag{3}$$

where Γ_0 and Γ_1 lead to observed and missing mechanism in which the parameter estimates of the Y_{ij} could be obtained ignoring the missing mechanism (provided disjoint parameter spaces for response models and the missing mechanism). The vector of μ and Σ should be estimated.

To obtain the log likelihood function as follows,

$$\begin{aligned} l(\Theta) = \sum_{i=1}^{n}[\log(f(y_i)) &+ \sum_{\{i:\, y_{ij} \text{ is observed}\}} \log(P(\bigcap_{j \in J_{obs}^{y}} \{R_{y_{ij}} = 1\} \mid y_i) \\ &+ \sum_{\{i:\, y_{ij} \text{ is not observed}\}} \log(P(\bigcap_{j \in J_{mis}^{y}} \{R_{y_{ij}} = 0\} \mid y_i). \end{aligned}$$

Suppose the g_1- elements of Y_i is observed, so $J_{obs}^{y} = \{o_1,...,o_{g_1}\}$ and the g_2- elements of Y_i is missing, so $J_{Mis}^{y} = \{M_1,...,M_{g_2}\}$. Let $I_{ij} = I_{\{R_{y_{ij}}=1|Y_i\}}$ and $J_{ij} = J_{\{R_{y_{ij}}=0|Y_i\}}$

$$\begin{aligned} l(\Theta) \square \sum_{i=1}^{n}\{\log(f(y_i)) \\ &+ \sum_{\{i:\, y_{ij} \text{ is observed}\}} \log[E(I_{io_1})\prod_{j=o_2}^{o_{g_1}}\{E(I_{ij}) + \Omega_{21}\Omega_{11}^{-1}(1 - E(I_{io_1}),...,1 - E(I_{i,j-1}))'\})] \\ &+ \sum_{\{i:\, y_{ij} \text{ is not observed}\}} \log[E(J_{iM_1})\prod_{j=M_2}^{M_{g_1}}\{E(J_{ij}) + \Delta_{21}\Delta_{11}^{-1}(1 - E(J_{iM_1}),...,1 - E(J_{i,j-1}))'\}]\}, \end{aligned}$$

where Ω_{21} and Δ_{21} are a row vector consisting of the entries

$$\begin{aligned} &cov(I_{ij}, I_{ih}) = E(I_{ij}I_{ih}) - E(I_{ij})E(I_{ih}) \text{ and} \\ &cov(J_{ij}, J_{ih}) = E(J_{ij}J_{ih}) - E(J_{ij})E(J_{ih}), \ h = 1,..., j-1 \end{aligned}$$

and Ω_{11} and Δ_{11} are a $(j-1)\times(j-1)$ matrix with (i, j) element $cov(I_h, I_k)$ and $cov(J_h, J_k)$, $1 \le h, k \le j-1$.

3. Derivation of ISNI

We considered estimation of a vector parameter Θ of the distribution of an outcome variable $Y = (Y_1,...,Y_n)$, whose independent components $Yi = (Y_{i1},...,Y_{ip})'$ have densities $MVN(\mu\delta_i, \Sigma\delta_i)$, $\delta_i = t_i - t_{i-1}$. An incomplete data assumes that the probability of begin observed dependents on the outcome Y_{ij} according to a specified link function, follows:

$$logit[P(R_{Y_{ij}} = 1 | Y_{i,obs}, Y_{i,mis}, \Gamma_0, \Gamma_1)] = (t_i - t_{i-1})\mu_j + \Gamma_0' Y_{i,obs} + \Gamma_1' Y_{i,mis}.$$

The index of sensitivity to nonignorability (ISNI) measured the the extent to which the maximum likelihood estimation (MLE) of Θ for a given vector Γ_1 of the nonignorability parameter [denoted as $\hat{\Theta}(\Gamma_1)$ depends on Γ_1] Specifically, it measures the sensitivity of $\hat{\Theta}(\Gamma_1)$ to small departures of Γ_1 from its MAR vector of zero. Troxel et al. (1998) defined ISNI as the derivative of $\hat{\Theta}$ with respect to Γ_1 at $\Gamma_1 = 0$, i.e.,

$$ISNI = \frac{\partial \hat{\Theta}(\Gamma_1)}{\partial \Gamma_1^T} |_{\Gamma_1 = 0} .$$

One obtains $\hat{\Theta}(\Gamma_1)$ from a Taylor- series expansion of the log likelihood around $\Theta = \hat{\Theta}_0$ (the MLE of Θ assuming ignorability). A large ISNI implies subtantial sensitivity.

To measure the sensitivity of the MLEs when the $dim(Y_{i,mis})$ dimensional vector of Γ_1 is perturbed around the ignorable model ($\Gamma_1 = 0$), assume $\hat{\Theta}(\Gamma_1)$ as the MLE of Θ for a fixed Γ_1 in a neighborhood of $\Gamma_1 = 0$. Hence, $\hat{\Theta}(0)$ is the MLE for Θ in the ignorable model. The difference $\hat{\Theta}(\Gamma_1) - \hat{\Theta}(0)$ is a sensible measure of the sensitivity when Γ_1 is perturbed around the ignorable model. Having a vector of non-ignorability parameter Γ_1, we need to adjust ISNI proposed by Troxel et al. (1998)

$$ISNI(\hat{\Theta}) = \frac{\partial \hat{\Theta}(\Gamma_1)}{\partial \Gamma_1^T} |_{\Gamma_1 = 0} = -[\frac{\partial^2 L}{\partial \Theta \partial \Theta^T}]^{-1} \frac{\partial^2 L}{\partial \Theta \partial \Gamma_1^T} |_{\hat{\Theta}(0)} . \tag{4}$$

These index vectors measure sensitivity of the MLEs to perturbations in the individual nonignorability parameters. Also we can approximate the MLE of a scalar smooth function $f(.)$ of Θ using the first order Taylor series expansion around $\Gamma_1 = 0$ as follows:

$$f(\hat{\Theta}(\Gamma_1)) \approx f(\hat{\Theta}(0)) + \left[\frac{\partial f}{\partial \Theta^T} |_{\hat{\Theta}(0)} \times \frac{\partial \hat{\Theta}(\Gamma_1)}{\partial \Gamma_1^T} |_{\Gamma_1 = 0} \right] \times \Gamma_1,$$

where $\frac{\partial\hat{\Theta}(\Gamma_1)}{\partial\Gamma_1^T}|_{\Gamma_1=0}$ is the sensitivity vector defined in (4). It is clear that we need to know the values of Γ_1 to approximate the effect of non-ignorability on $f(\hat{\Theta})$ such as $\Gamma_1 = 1_q$, $q = g_2$, which assumes that the effect of different responses on the MLEs for moderately large nonignorability is the same. However, when there are no preferable direction, we would want to take the direction where the sensitivity is greatest among all possible perturbations whose norm is $\sqrt{q}$:

$$ISNI(f(\hat{\Theta})) = \sqrt{q}\left\|\frac{\partial f}{\partial\Theta^T}\times\frac{\partial\hat{\Theta}(\Gamma_1)}{\partial\Gamma_1^T}\right\|^{1/2}_{\hat{\Theta}(0),\Gamma_1=0}.$$

To derive the formulas in the case of Brownian motion, then we have:

$$\begin{aligned}\frac{\partial^2 L}{\partial\Theta\partial\Theta^T} &= \sum_{i=1}^{n}\{\frac{\partial^2}{\partial\Theta\partial\Theta^T}\log(f(y_{i*}))\\ &+\sum_{\{i:\, y_{ij}\ is\ observed\}}\frac{\partial^2}{\partial\Theta\partial\Theta^T}\log[E(I_{io_1})\prod_{j=o_2}^{o_{g_1}}\{E(I_{ij})+\Omega_{21}\Omega_{11}^{-1}(1-E(I_{io_1}),...,1-E(I_{i,j-1}))'\})]\\ &+\sum_{\{i:\, y_{ij}\ is\ not\ observed\}}\frac{\partial^2}{\partial\Theta\partial\Theta^T}\log[E(J_{iM_1})\prod_{j=M_2}^{M_{g_1}}\{E(J_{ij})+\Delta_{21}\Delta_{11}^{-1}(1-E(J_{iM_1}),...,1-E(J_{i,j-1}))'\}]\}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 L}{\partial\Theta\partial\Gamma_1^T} &= \sum_{i=1}^{n}\{\sum_{\{i:\, y_{ij}\ is\ observed\}}\frac{\partial^2}{\partial\Theta\partial\Gamma_1^T}\log[E(I_{io_1})\prod_{j=o_2}^{o_{g_1}}\{E(I_{ij})+\Omega_{21}\Omega_{11}^{-1}(1-E(I_{io_1}),...,1-E(I_{i,j-1}))'\})]\\ &+\sum_{\{i:\, y_{ij}\ is\ not\ observed\}}\frac{\partial^2}{\partial\Theta\partial\Gamma_1^T}\ln[E(J_{iM_1})\prod_{j=M_2}^{M_{g_1}}\{E(J_{ij})+\Delta_{21}\Delta_{11}^{-1}(1-E(J_{iM_1}),...,1-E(J_{i,j-1}))'\}]\}.\end{aligned}$$

To obtain $\frac{\partial^2 L}{\partial\Theta\partial\Theta^T}$ one can use the Hessian matrix of Θ under the MAR model. For calculating $\frac{\partial^2 L}{\partial\Theta\partial\Gamma_1}$, the Monte Carlo methods of approximating integrals can be utilized to calculate corresponding conditional expectations.

Because ISNI depends on the units of measurement of Y_{ij}, Troxel et al. (1998) proposed a scale free measure called the sensitivity transformation c, defined as

$$c(\hat{\Theta}) = \left| \frac{var(Y_{ij})^{\frac{1}{2}} SE(\hat{\Theta})}{ISNI(\hat{\Theta})} \right|,$$

where $SE(\hat{\Theta})$ is the standard error (SE) of $\hat{\Theta}$. Large values of c suggest that sensitivity ocurrs only in cases of extreme nongnorability, whereas small values suggest that sensitivity may be a problem even when the nonignorability is modest. Troxel et al. (1998) have suggested using $c < 1$ as a cutoff value for important sensitivity.

4. Simulation

A simulation study was used to investigate coverage of 95 percent confiedence intervals for estimates obtained by the complete data model (model 2) and incomplete data model (model 3). Let us suppose that $X = (S^{(1)}, S^{(2)})'$ represents the logarithm of stock price as is postulated under the Black-Scholes options pricing model. Two different sets of simulations were considered. In the first set, data were generated from:

We generated from a bivariate Brownian motion $X = (X_1, X_2)$ with drift $\mu = (0.100, 0.050)$ and covariance matrix:

$$\Sigma = \begin{pmatrix} 0.060 & 0.700 \\ 0.700 & 0.160 \end{pmatrix}.$$

Supposed $n = 3500$ equidistant data points were simulated with a time step of $dt = 1/252$ and let $Y_{ij} = X_{ij} - X_{i-1,j}$. The data generated from compelet data were modeled using the following:

$$Y_{ij} = (t_i - t_{i-1})\mu_j + \varepsilon_{ij} \qquad i = 1,...,3500\ ,\ j = 1,2$$

with $(\varepsilon_{i1}, \varepsilon_{i2})$ generated from a multivariate normal distribution with zero mean and $(t_i - t_{i-1})\Sigma$, where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_1\rho \\ \sigma_1\sigma_1\rho & \sigma_2^2 \end{pmatrix},$$

$$\sigma_j^2 = Var(Y_{ij})$$

and

$$\rho = Corr(Y_{i1}, Y_{i2}).$$

Suppose that the first component of X, $(S^{(1)})$ is a heavily traded asset and that none of its observations are missing. We will assume that the second component of X, $(S^{(2)})$ is not traded at every time instant giving rise to some missing data. In the second set of simulations, missing mechanism was added to each responses, (Y_{ij}) and data were generated from:

Missing data are generated from the missingness mechanism

$$logit\ \ P(R_{Y_{i2}} = 1 \mid Y_{i1}, Y_{i2}, \gamma_0, \gamma_1) = (t_i - t_{i-1})\mu_2 + \gamma_0 Y_{i1} + \gamma_1 Y_{i2}$$

with true parameters $\gamma = (\gamma_0, \gamma_1) = (1,1)$, which implies that this missingness mechanism is NMAR. The response Y_{i2} was obtained by second generating a $uniform(0,1)$ random vector V of length n and then assigning Y_{i2} is missing if $V < P(R_{y_{i2}} = 1 \mid Y_{i1}, Y_{i2}, \gamma_0, \gamma_1)$ and y_{i2} is observed, otherwise. We assume the percentage of missing values of Y_{i2} is 28.000%. The data generated from incompelet data were modeled using the following:

$$Y_{ij} = (t_i - t_{i-1})\mu_j + \varepsilon_{ij} \qquad i = 1,...,3500,\ j = 1,2$$
$$logit\ \ P(R_{Y_{i2}} = 1 \mid Y_{i1}, Y_{i2}, \gamma_0, \gamma_1) = (t_i - t_{i-1})\mu_2 + \gamma_0 Y_{i1} + \gamma_1 Y_{i2}.$$

The vector of μ and Σ should be estimated. The models (1) and (2) were fitted using *nlminb* from R to assure that the same numerical algorithms were used to maximize the likelihoods.

Results of using model (Section 2.1) is given in Table I. The complete data maximum likelihood estimates based on the uncensored data can be easily calculated and are displayed in Table I. In Table I, the diffusion coefficient is estimated quite well. The finer our observation of the process the closer to the true value we can expect our estimate to be. In the limit, observating a continuous path of any time length would lead to perfect estimation of the diffusion coefficient which is a consequence of the quadratic variation of the process. The same is not true for the drift. As expected, the estimates of the drift parameters are very poor with relatively large standard errors because our estimate of drift is based used only one observed sample path of the process. In fact it depends values of this path and these can be quite variable.

Table I. Results of the simulation study for complete data

Parameter	μ_1	σ_1^2	μ_2	σ_2^2	ρ
Model (3)					
Real Value	0.100	0.060	0.050	0.160	0.700
Est.	0.048	0.062	-0.021	0.163	0.720
S.E.	0.067	0.001	0.010	0.003	0.002

Table II presents the estimates of the parameters of the incomplet data with assumption of NMAR. The results are summarized as follows. The parameter estimates by the model are close to the true values of the parameters.

Table II. Results of the simulation study for incomplete data with assumption of NMAR

Parameter	μ_1	σ_1^2	μ_2	σ_2^2	ρ	γ_0	γ_1
Model (3)							
Real Value	0.100	0.060	0.050	0.160	0.700	1.000	1.000
Est.	0.059	0.073	0.034	0.167	0.842	1.003	1.006
S.E.	0.068	0.004	0.015	0.045	0.064	0.073	0.059

5. Application

5.1 Data and Model

Toronto Stock Exchange (TSX) is the largest stock exchange in Canada, the third largest in North America and the seventh largest in the world by market capitalisation. Based in Canada's largest city, Toronto, it is owned by and operated as a subsidiary of the TMX Group for the trading of senior equities. A broad range of businesses from Canada, the United States, Europe, and other countries are represented on the exchange. In addition to conventional securities, the exchange lists various exchange-traded funds, split share corporations, income trusts and investment funds. The Toronto Stock Exchange is the leader in the mining and oil gas sector; more mining and oil gas companies are listed on Toronto Stock Exchange than any other exchange in the world. The data set consists of one minute discredited tick data the Toronto Stock Exchange(TSE) dated on February 2, 2005 on three bank stocks: the Bank of Nova Scotia (BNS), Royal Bank (RY) and Bank of Montreal (BMO). The data consist of 392 time points with 12, 51 and 45 missing values for BNS, RY and BMO, respectively. A stock price at a particular time point is reported missing if no sales were made in the last minute of trading on that stock.

We use a joint model similar to the one proposed in Section 2 along with a logistic model for the non-dropout indicator for the TSE data set which can be summarized as follows:

$$Y_{i,BNS} = (t_i - t_{i-1})\mu_{BNS} + \varepsilon_{i,BNS}$$
$$Y_{i,RY} = (t_i - t_{i-1})\mu_{RY} + \varepsilon_{i,RY}$$
$$Y_{i,BMO} = (t_i - t_{i-1})\mu_{BMO} + \varepsilon_{i,BMO}$$

and

$$logit\ \ P(R_{Y_{i,BNS}} = 1 \mid Y_{i,BNS}, Y_{i,RY}, Y_{i,BMO}\xi_0, \xi_1, \xi_2) = (t_i - t_{i-1})\mu_{BMO} + \xi_0 Y_{iBNS} + \xi_1 Y_{i,RY} + \xi_2 Y_{i,BMO}$$
$$logit\ \ P(R_{Y_{i,RY}} = 1 \mid Y_{i,BNS}, Y_{i,RY}, Y_{i,BMO}\gamma_0, \gamma_1, \gamma_2) = (t_i - t_{i-1})\mu_{BMO} + \gamma_0 Y_{iBNS} + \gamma_1 Y_{i,RY} + \gamma_2 Y_{i,BMO}$$
$$logit\ \ P(R_{Y_{i,BMO}} = 1 \mid Y_{i,BNS}, Y_{i,RY}, Y_{i,BMO}\eta_0, \eta_1, \eta_2) = (t_i - t_{i-1})\mu_{BMO} + \eta_0 Y_{iBNS} + \eta_1 Y_{i,RY} + \eta_2 Y_{i,BMO},$$

where

$$Y_{i,BNS} = BNS(t_i) - BNS(t_{i-1}),\ Y_{i,RY} = RY(t_i) - RY(t_{i-1}) \text{ and } Y_{i,BMO} = BMO(t_i) - BMO(t_{i-1}).$$

Let

$$\Gamma_1 = (\xi_0, \xi_1, \xi_2, \gamma_0, \gamma_1, \gamma_2, \eta_0, \eta_1 \eta_2)$$

lead to observed and missing mechanism in which the parameter estimates of the BNS, RY and BMO could be obtained ignoring the missing mechanism (provided disjoint parameter spaces for response models and the missing mechanism). The vector of μ_{BNS}, μ_{RY}, μ_{BMO}, σ^2_{BNS}, σ^2_{RY}, σ^2_{BMO}, $\rho_{BNS,RY}$, $\rho_{BNS,BMO}$ and $\rho_{RY,BMO}$ should be estimated. The dropout mechanism would be ignorable if $\Gamma_1 = 0$. The log likelihood function for the above non-ignorable model could be obtained from appendix in which the joint distribution of $Y_{i,BNS}, Y_{i,RY}$ and $Y_{i,BMO}$.

According to Table III, all the model parameters are not highly sensitive to even little non-ignorability which reveals the need for examining the ignorability assumption before conducting a simple MAR analysis.

Table III. Results of using model data

Parameter	E. S.	S.E.	ISNI	c
μ_{BNS}	-0.760	2.700	-4.371	0.617
μ_{RY}	1.755	3.280	4.391	0.746
μ_{BMO}	- 1.700	2.300	-4.428	0.519
σ^2_{BNS},	0.025	0.001	0.147	0.006
σ^2_{RY}	0.039	0.003	0.008	0.375
σ^2_{BMO}	0.018	0.001	-0.005	0.200
$\rho_{BNS,RY}$,	0.002	0.001	-0.703	0.001
$\rho_{BNS,BMO}$	0.001	0.001	0.167	0.005
$\rho_{RY,BMO}$	0.001	0.002	-1.620	0.001
ξ_0	0.043	0.056	0.601	0.003
ξ_1	0.005	0.006	0.176	0.010
ξ_2	0.003	0.012	1.789	0.067
γ_0,	0.054	0.031	0.654	0.021
γ_1	0.006	0.005	0.148	0.003
γ_2	0.023	0.032	-1.450	0.001
η_0	0.076	0.039	-0.444	0.005
η_1	0.002	0.001	0.132	0.005
η_2	0.009	0.004	-1.321	0.007

6. Conclusion

Increasing application of Finance data studies reveals the improvements needed for modelling such data. One common vexing problem in these kinds of studies is the individual dropout which may be non-ignorable and hence a MAR analysis might not be appropriate. Examining the non-ignorability of the dropout before drawing inferences could prevent both wrong simple MAR or unnecessary complicated NMAR analysis. In this paper we have extended the ISNI measure to assess the likely effect of non-ignorable dropout for multivariate Brownian motion.

Appendix

Likelihood for Incomplete Data Model with Assumption of NMAR

$$
\begin{aligned}
l(\Theta) &= \sum_{i=1}^{n} \log(f(y_i, R_{y_i})) \\
&= \sum_{i=1}^{n} [\log(f(y_i)) + \log(f(R_{y_i} | y_i))] \\
&= \sum_{i=1}^{n} [\log(f(y_i)) + \sum_{\{i:\, y_{ij} \text{ is observed}\}} \log(P(\bigcap_{j \in J_{obs}^{y}} \{R_{y_{ij}} = 1\} | y_i) \\
&\qquad + \sum_{\{i:\, y_{ij} \text{ is not observed}\}} \log(P(\bigcap_{j \in J_{mis}^{y}} \{R_{y_{ij}} = 0\} | y_i).
\end{aligned}
$$

Supposed the g_1- elements of Y_i is observed, so $J_{obs}^{y} = \{o_1, ..., o_{g_1}\}$ and the g_2- elements of Y_i is missing, so $J_{Mis}^{y} = \{M_1, ..., M_{g_2}\}$. Let $I_{ij} = I_{\{R_{y_{ij}} = 1 | Y_i\}}$. So

$$
P(\bigcap_{j=o_1}^{o_{g_1}} \{R_{y_{ij}} = 1\} | y_i) = E(I_{io_1}) \prod_{j=o_2}^{o_{g_1}} E[I_{ij} \,\big|\, I_{io_1} = 1, I_{io_2} = 1, ..., I_{i,j-1} = 1].
$$

Now, the expectation on the right hand can be approximated as (vide, Harry,1995). We have,

$$
E[I_{ij} \,\big|\, I_{io_1} = 1, I_{io_2} = 1, ..., I_{i,j-1} = 1] \approx E(I_{ij}) + \Omega_{21}\Omega_{11}^{-1}(1 - E(I_{io_1}), ..., 1 - E(I_{i,j-1}))'.
$$

So,

$$
P(\bigcap_{j=o_1}^{o_{g_1}} \{R_{y_{ij}} = 1\} | y_i) \approx E(I_{io_1}) \prod_{j=o_2}^{o_{g_1}} E(I_{ij}) + \Omega_{21}\Omega_{11}^{-1}(1 - E(I_{io_1}), ..., 1 - E(I_{i,j-1}))'.
$$

Let

$$J_{ij} = J_{\{R_{y_{ij}}=0|Y_i\}}. \text{ So}$$

$$P(\bigcap_{j=M_1}^{M_{g_1}}\{R_{y_{ij}}=0\}|\, y_i) \approx E(J_{iM_1})\prod_{j=M_2}^{M_{g_1}} E(J_{ij}) + \Delta_{21}\Delta_{11}^{-1}(1-E(J_{iM_1}),\ldots,1-E(J_{i,j-1}))'.$$

The log-likelihood is

$$l(\Theta) \approx \sum_{i=1}^{n}\{\ln[f(y_i)]$$

$$+ \sum_{\{i:\, y_{ij} \text{ is observed}\}} \ln[E(I_{io_1})\prod_{j=o_2}^{o_{g_1}}\{E(I_{ij}) + \Omega_{21}\Omega_{11}^{-1}(1-E(I_{io_1}),\ldots,1-E(I_{i,j-1}))'\})]$$

$$+ \sum_{\{i:\, y_{ij} \text{ is not observed}\}} \ln[E(J_{iM_1})\prod_{j=M_2}^{M_{g_1}}\{E(J_{ij}) + \Delta_{21}\Delta_{11}^{-1}(1-E(J_{iM_1}),\ldots,1-E(J_{i,j-1}))'\}]\}.$$

REFERENCES

Bahrami Samani, E. and Ganjali, M. (2008). A Latent Variable Model for Mixed Continuous and Ordinal Responses, *Journal of Statistical Theory and Applications*, **7**, 337-348.

Baker, P. C., Keck, C. K., Mott, F. L. and Quinlan, S. V. (1993). NLSY child Handbook: Guide to the 1986-1990 National Longitudinal Survey of Youth Child Data, Columbus, OH: Center fo Human Resource Research.

Berridge, D.M. and Dos Santos, D.M. (1996). Fitting a random effects model to ordinal recurrent events using existing software, *J. Statist. Comput. Simul.*, **55**, 73-86.

Black, F.and Scholes, M (1973). The pricing of options and coroprate liabilities, *Journal of Plotitical Economy*, **81**, 637-659.

Catalano, P. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcoms, *Journal of the American Statistical Association*, **50**, 3, 1078-1095.

Cox, D. R. and Wermuth, N. (1992). Response models for mixed binary and quantitative variables, *Biometrika*, **79**(3): 441-461.

DeCesare, G. (2006). Imputation, Estimation and Missing Data in Finance *the thesis requirement for the degree of Doctor of statistics*.

Diggle, P.J. and Kenward, M. G. (1994). Informative Drop-out in longitudinal data analysis, *Journal of Applied Statistics*, **43**, 49-93.

Harry, J. (1995). Approximations multivariate normal rectangle probabilites based on conditional expectation, *Journal of the Royal Stat Soc, Series C*, **90**: 957-967 .

Kofman, P. and Sharpe, I. (2003). Using Multiple Imputation in the Analysis of Incomplete Observations in Finance, *J. Financial Econometrics*, **1**(2):216-249.

Little, R. J. and Rubin, D. (2002). Statistical analysis with missing data, Second edition. New york, Wiley.

McCulloch, C. (2007). Joint modelling of mixed outcome type using latend variables, *statistical methods in Medical Research*, 13: 1-27.

Merton, R. (1971). Theory of rational option pricing, *Bell Journal of Economics and Management*, **4**, 141-183.

Rubin, D. B. (1976). Inference and missing data, *Biometrika*, 63(3): 581-592.

Troxel, A. B., Harrington, D. P., and Lipsitz, S. R. (1998). Analysis of Longitudinal Measurements with Non-ignorable Non-monotone Missing Values, *Applied Statistics*, **47**, 425 - 438.

Xie, H. and Heitjan, D.F. (2004). Sensitivity analysis of causal inference in a clinical trial subject to crossover, *Clinical Trials*, 21-30.

Yang, Y., Kang, J., Mao, K. and Zhang, J. (2007). Regression models for mixed Poisson and continuous longitudinal data, *Statistics in Medicine*; **26**: 3782-3800.