Methods for a Longitudinal Quantitative Outcome With a Multivariate Gaussian Mixture Distribution Multi-dimensionally Censored by Therapeutic Intervention

Wanjie Sun[1], Michael D. Larsen[12], John M. Lachin[123]
[1] The Biostatsitics Center, The George Washington University,
6110 Executive Blvd., Suite 750, Rockville, MD 20852
[2] Department of Statistics, The George Washington University
[3] Department of Biostatistics and Epidemiology, GWU
wsun@bsc.gwu.edu; mlarsen@bsc.gwu.edu; jml@bsc.gwu.edu

Abstract

In longitudinal clinical trials, epidemiologic, or genetic studies, a quantitative outcome may be altered by the administration of a non-randomized, non-trial intervention during follow up. The resulting effect of the non-trial intervention may seriously bias the study results, including treatment or exposure effects or associations. Current methods to address this issue including multilevel models (White *et al* 2001) or multiple imputation (MI) (Cook 1997, Cook 2006), are either restricted to a specific longitudinal data structure or are valid only under special circumstances. We propose two new methods for general longitudinal data – a modified Expectation-Maximization (EM)-type model and a modified Monte Carlo EM-MI model. These combine Monte Carlo EM (Wei and Tanner 1990) and MI, and are extensions of Censored Normal Regression (Tobin *et al* 2005) to longitudinal data. These replace the intractable calculation of a multi-dimensionally truncated MVN posterior distribution with a simplified but sufficiently accurate approximation. Simulation shows that the two proposed methods have the least biased treatment effect estimate in a majority of simulated scenarios amongst six methods applied.

Key Words: Censored Normal Regression; Monte Carlo EM; Multiple imputation; Non-ignoreable Missingness; Rescue Medication; Truncated MVN.

1 Introduction

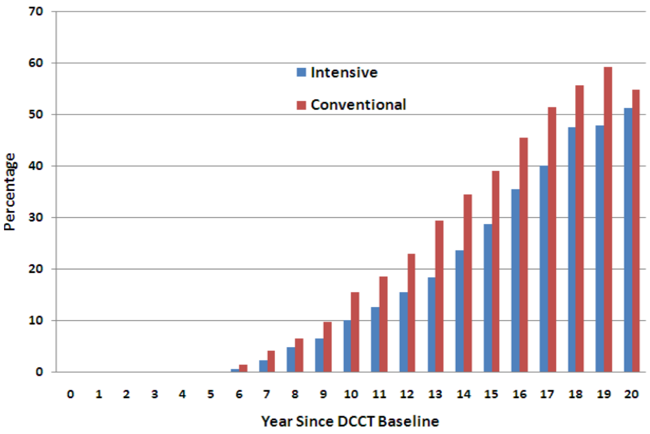
In clinical trials and in epidemiologic or genetic studies, the value of a quantitative outcome may be altered by the administration of a non-randomized, non-trial therapeutic intervention. The resulting effect of the non-randomized, non-trial intervention may seriously distort the analysis and undermine the scientific aims of the study.

1.1 DCCT/EDIC Example

The Diabetes Control and Complications Trial (DCCT, 1982-93) was a landmark type 1 diabetes clinical trial funded by the National Institute of Diabetes and Digestive and Kidney diseases (NIDDK). At enrollment, 1,441 type 1 diabetics were randomly assigned to intensive (INT) versus conventional (CON) diabetes therapy. DCCT demonstrated that INT reduced the risk of development and progression of retinopathy, nephropathy, and neuropathy compared to CON (DCCT Research Group 1993, 1995, 1998). In 1994, after completion of the DCCT, 1375 subjects agreed to participate in the follow-up study, the Epidemiology of Diabetes Interventions and Complications (EDIC, 1994-2016), which aimed to examine the longer-term effects of the original DCCT interventions on cardiovascular and more advanced stages of retinal and renal disease.

Among many others, two objectives EDIC aimed to address are, “What is the prolonged treatment effect of the former intensive therapy on blood pressure (BP) and albuminuria excretion rate (AER) twenty years after the start of the DCCT?”, and “What are the genetic determinants for eleveated BP or AER in the DCCT/EDIC cohort?”. These two questions, however, are complicated by the non-randomized administration of antihypertensive medications (ACEI or ARBs) in the EDIC cohort. During the DCCT, participants were proscribed from taking antihypertensive medications. This restriction was relaxed, however, during the EDIC follow up. Figure 1 shows the prevalence of anti-hypertensive

Figure 1: Medication use (ACEI or ARBs) in DCCT and EDIC.



medication use in DCCT and EDIC. Medication use started six years after DCCT baseline and reached 50% in both groups twenty years after the start of the DCCT. Overall, patients were intervened at 20% of the visits in CON and 15% in INT in the combined DCCT and EDIC study.

ACEI/ARBs are used primarily in the treatment of hypertension, though they can also be used to treat diabetic nephropathy. DCCT/EDIC patients took ACEI/ARBs for different reasons – hypertension (10.4%), albuminuria (0.9%), or prophylactic reasons (1.6%), Another 4.8% of patients were on medication but did not report reasons because this information was not collected in early EDIC. Subjects intervened for hypertension or who did not report intervention reasons had higher SBP that those intervened for renal or prevention reasons. The latter had about the same SBP levels after intervention as those free of intervention.

The administration of antihypertensives altered the distribution of the underlying Y. Throughout this paper, “underlying” outcome refers to the true value of Y if a patient were not intervened. This underlying outcome is not observed once a patient is intervened. The observed outcome Y_{obs}, after non-trial intervention, was lower than the underlying true value Y. The difference between the two depended upon how large the effect of medication was for each patient, which may vary according to age, gender, the level of Y prior to intervention, and genetic susceptibility to the medication. Figure 2 shows the distribution of SBP prior to and post use of ACEI/ARBs in the DCCT/EDIC. The x axis was rescaled to make Year 0 an index intervening visit. SBP dropped immediately after the administration of medication and discontinued the previous increasing trend.

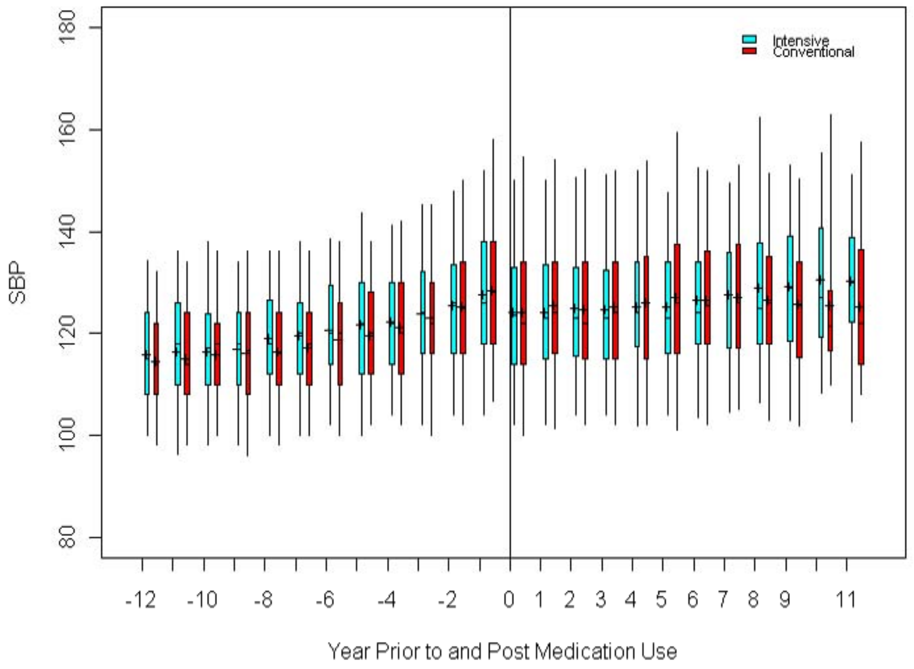
Therefore, the real questions EDIC aimed to address are, “What is the prolonged treatment effect of the former INT on BP or AER, had no one on non-randomized, non-trial medication, twenty years after DCCT baseline”, and “What are the genetic determinants for elevated BP or AER in the DCCT/EDIC cohort, had no one on non-randomized, non-trial medication?”.

1.2 Impact of therapeutic intervention on scientific interest

In clinical trials, the analysis of interest can be ”explanatory” (Schwartz and Lellouch 1967), i.e., the group difference that would have been observed in the absence of rescue medication, or ”pragmatic”, i.e., the observed difference between the two treatment groups with intervention as necessary. In this paper, we focus on the ”explanatory” evaluation. In epidemiologic or genetic studies, the scientific aim is usually to identify risk factors/exposures associated with, or etiological genetic determinants of the increased risk of future disease, which is often the more informative underlying outcome.

White *et al.* (2001) argued that in clinical trials, if there is a therapeutic difference between trial treatments, rescue medication may well be more used in the inferior treatment group, and the true difference between the treatment groups would be reduced. Tobin *et al.* and others (Tobin *et al.* 2005, Levy *et al.* 2000, Cui et. al. 2003, White *et al.* 1994) showed that in cross-sectional epidemiologic/genetic

Figure 2: SBP prior to and post medication use



studies, if without appropriate correction, analysis based on the observed outcome Y_{obs} can seriously distort the analysis and undermine the scientific aims of the study. Analysis of Y_{obs} ignoring the effect of non-trial intervention leads to a substantial shrinkage in the estimated effects of etiological determinants of scientific interest (bias), and a marked reduction in statistical power.

Section 2 reviews current statistical methods for a quantitative outcome censored by therapeutic intervention. Section 3 introduces a model of a multivariate Gaussian data with multi-dimensionally right-censored data and two proposed methods. Section 4 compares the performance of four existing methods versus the proposed two models in various simulation scenarios. Application of the proposed methods in the DCCT/EDIC example is deferred to future clincial papers. We conclude with overall recommendation and discussion in Section 5.

2 Current statistical methods for a quantitative outcomes censored by therapeutic intervention

Current methods to address a quantitative outcome altered by a non-randomized, non-trial therapeutic intervention are mainly for cross-sectional studies, including the 'Ignore' method (Schunkert *et al.* 1998; Matsubara *et al.* 2001; Iwai *et al.* 2001; Brand *et al.* 2003), the 'exclude' method (Schunkert *et al.* 1998; Rice *et al.* 2000; Matsubara *et al.* 2001; Iwai *et al.* 2001; Brand *et al.* 2003), the "median" approach (White *et al.* 1994, 1996), censored normal regression (Tobin *et al.* 2005), or adding a constant to the treated outcome (Cui *et al.* 2005). For longitudinal data, the proposed methods including multi-level models (White *et al.* 2001), multiple imputation (MI) (Cook 1997, 2006), or a two-step approach (McClelland *et al.* 2008), are either restricted to a specific longitudinal data structure or are valid only under special circumstances.

White et al (1994) were the first to address this problem in statistical literature. They proposed a median approach which assumed that only hypertensive patients were intervened and no more than

50% of the sample were undergoing intervention. White et al. suggested imputing the underlying unobserved BP by some arbitrarily large values (larger than the median). A between-group comparison using median as the measure of location would be robust to the imputed values. In a series of subsequent papers (White 1996, 1998, 1999, 2000, 2001, 2003 I, 2003 II, 2004), they discussed different variations of the median approach and applied them to clinical trials and epidemiology studies and extended to time to event data. Tobin et al. (2005) reviewed nine ad-hoc and post-hoc statistical approaches currently used to assess genetic association with BP when BP is modified by antihypertensive medication in cross-sectional data. Based on the simulation results, Tobin recommended two methods out of nine, which performed well across a range of realistic settings, namely, adding a constant to the treated BP values (Cui et al. 2005) and a censored normal regression, or a modified Tobit model (Tobin 2005). Adding a constant, however, requires a pre-specified size of treatment effect based on prior knowledge, which may not be available in practice. Masca, Sheehan and Tobin (2011) conducted a simulation study to test the influence of pharmacogenetic interactions on various methods to analyses of BP in a cross-sectional genetic association study.

For longitudinal data, White *et al.* (2001) proposed a multilevel regression model for longitudinal clinical trial data. It is basically a LMM with time-dependent use of medication as a confounder. White *et al.* concludes that adjustment for rescue medication will not radically alter the randomized treatment comparison unless rescue medication is substantially imbalanced between randomized groups and has a substantial effect on the outcome. Cook (1997) proposed an imputation method for a special longitudinal data set where post-medication measurements were all missing by design, medication was prescribed to hypertensives only, and an out-of-study measurement prior to intervention was assumed to be measured which should exceed a single known threshold. Cook proposed an EM algorithm and a nested random effect model to impute the BP values that are missing not at random (MNAR), and model the treatment effect on the augmented BP. Cook 2006 was a variation of Cook's 1997 paper where the extra out-of-study measurement prior to medication intervention was observed. This is not common in a typical study though. McClelland *et al.* (2008) proposed a two-stage imputation method for a special longitudinal epidemiology context with only two repeated measurements. It performed better than the censored normal regression when there were different intervening thresholds for two groups especially when medication has a big treatment effect on the outcome.

The aim of this paper is to develop methods for a more general longitudinal data set altered by therapeutic intervention that can be used in clinical trials and epidemiologic and genetic studies. In particular, our aim is to extend Cook's (1997) multiple imputation method for a restrictive longitudinal data structure to a more general longitudinal data set up, and to extend Tobin *et al.*'s (2005) censored normal regression model for cross-sectional data to one for longitudinal data.

3 Structural Model: A Multivariate Gaussian Mixture Distribution with Multi-dimensional Right censoring and Available Methods

Suppose in a long-term clinical trail or epidemiologic or genetic study, a quantitative outcome Y is measured at multiple time points for each subject; the time points can be equally or unequally spaced. Assume at the beginning of the study, all participants are free of non-trial medication. During the follow up, some patients are intervened by non-study medications for treatment, or intolerance of trial medication, or prevention. Here we assume that once intervened, a patient will continue to be intervened in subsequent visits. The objective is to assess the effect of treatment, or an exposure, or genetic factors on the rate of change or slope of the underlying outcome Y over time.

Let y_n be the value of a t_n-dimensional vector of underlying outcome for subject n, which can be partitioned into a p_n-dimensional vector of observed underlying value prior to intervention, $y_{o1,n}$, and a q_n-dimensional vector of missing underlying value post intervention, $y_{m,n}$. Let $y_{obs,n}$ be the value of a t_n-dimensional vector of observed outcome for subject n. Likewise, it is composed of observed value prior to intervention which is equivalent to the underlying value, $y_{o1,n}$, and observed or treated value post intervention, $y_{o2,n}$. For simplicity of notation, the subscript n denoting subject n is sometimes omitted.

3.1 Assumption

Two assumptions are made for each subject. *First*, conditional upon covariates, the underlying outcome Y is assumed to be normally distributed. For longitudinal data, this is a MVN. *Second*, the underlying outcome Y is assumed to be at least as high as the treated value y_{obs} once intervention has begun. That is, Y is right censored at $y_{obs}, Y \geq y_{obs}$. These are the same assumptions as in censored normal regression (Tobin *et al.* 1958, Tobin *et al.* 2005). In reality, most of the medication effect is to reduce the elevated level of a disease. For few cases when medication improves the outcome Y, similar approach can be easily derived for left-censoring.

In Tobin *et al.*'s (2005) censored normal regression, a third assumption is made that conditional upon covariates, the distribution for those above any specific value is the same in treated ($R = 1$) and untreated individuals ($R = 0$), or $f(y \mid X, Y \geq c, R = 1) = f(y \mid X, Y \geq c, R = 0)$. This assumption is often not true in reality. In Tobin *et al.*'s (2005) simulated cases, departure from this assumption does not affect the estimate of treatment/exposure effect on Y. McClelland (2008), however, showed that when this assumption is seriously violated, the estimate from censored normal regression is biased (McClelland 2008). In this paper, this assumption is relaxed.

3.2 Model

Following Cook (1997, 2005)'s structure, the joint distribution of the observed underlying Y prior to intervention, $Y_{o1,n}$,and the missing underlying Y post intervention, $Y_{m,n}$, which is right censored at the observed treated value $y_{o2,n}$, $Y_{m,n} \geq y_{o2,n}$, is,

$$\begin{bmatrix} Y_{o1,n} \\ Y_{m,n} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{o1,n}(\beta) \\ \mu_{m,n}(\beta) \end{bmatrix}, \begin{bmatrix} \Sigma_{o1,n}(R_n, D) & \Sigma_{(o1,m),n}(R_n, D) \\ \Sigma_{(m,o1),n}(R_n, D) & \Sigma_{m,n}(R_n, D) \end{bmatrix} \right)$$

Parameters β, R_n, and D are from a general linear mixed model (LMM) of the underlying outcome,

$$y_{nj} = u + \alpha G_n + b_{0n} + (\beta_1 + \beta_2 G_n + b_{1n})T_j + W_n^T \gamma + \varepsilon_{nj},$$

where u denotes the intercept or overall mean when all the covariates are 0, G_n is the treatment group or exposure or genetic factor for subject n ($G_n = 1$ for treatment/exposure and $G_n = 0$ for reference), α is the intercept for the treatment/exposure group when all other covariates are 0, T_j is the visit number, β_1 is the slope or rate of change in y_n per unit of time for the reference group, $\beta_1 + \beta_2$ is the slope for the treatment/exposure group, β_2 is the slope difference between the two groups, W_n is a group of other covariates adjusted for in the model, γ is a vector of parameters for W_n, ε_{nj} is the measurement error for subject n at visit j, and b_{0n} and b_{1n} are the random intercept and random slope for subject n covering individual departure from the population average intercept and slope. The distribution of the random components is assumed to be a MVN, $(b_{0n}, b_{1n})' \sim N(0, D)$,

independently for each subject $n = 1, ..., N$. Here D can be any 2× 2 variance-covariance matrix like unstructured or compound sysmmetry. The joint distribution of ε_n for subject n across t_n visits is,$(\varepsilon_{n1}, \ldots, \varepsilon_{n,t_n})' \sim N(0, R_n)$, where R_n can be any variance covariance matrix.

The objective here is to estimate the rate of change or slope of Y for each group (β_1, $\beta_1 + \beta_2$) and the slope difference between groups (β_2) to assess treatment or exposure or genetic effect on rate of change in Y. For simplicity, we assume a linear slope, though other non-linear growth curves can also be defined.

3.3 EM approach?

When no individuals are intervened by a non-trial medication, or the underlying y_{nj} can be directly observed, this is a regular LMM model which can be solved by standard methods (Laird and Ware 1982, Lindstrom and Bates 1988). When the underlying y_{nj} is missing for some visits of certain subjects, a natural approach is the EM algorithm. The expected complete data log likelihood conditioned upon the observed data y_{obs} (including the pre-medicated outcome y_{o1} and treated outcome y_{o2} after intervention) under the current estimate of the parameter $\Theta^{(t)} = (\beta^{(t)}, R_N^{(t)}, D^{(t)})$ can be derived as follows,

$$\begin{aligned} Q(\Theta \mid y_{obs}, \Theta^{(t)}) &= E(l(\Theta) \mid y_{obs}, \Theta^{(t)}) \\ &= \sum\nolimits_{n=1}^{N} \left(-\frac{1}{2} \ln |\Sigma_n(R_n, D)| - \frac{1}{2} tr \left(\Sigma_n(R_n, D)^{-1} \left[E^T E + V \right] \right) \right), \end{aligned}$$

where $E = \begin{pmatrix} y_{o1,n} - \mu_{o1,n}(\beta) \\ E(Y_{m,n} \mid y_{obs,n}, \Theta^{(t)}) - \mu_{m,n}(\beta) \end{pmatrix}$, $V = \begin{pmatrix} 0 & 0 \\ 0 & V(Y_{m,n} \mid y_{obs,n}, \Theta^{(t)}) \end{pmatrix}$ with
$E(Y_{m,n} \mid y_{obs,n}, \Theta^{(t)}) = E(Y_{m,n} \mid y_{o1,n}, Y_{m_{1,n}} \geq y_{o2_1,n}, ..., Y^n_{m_{q_n,,n}} \geq y^n_{02_{q_n,n}}, \Theta^{(t)})$
and $V(Y_{m,n} \mid y_{obs,n}, \Theta^{(t)}) = V(Y_{m,n} \mid y_{o1,n}, Y_{m_1,n} \geq y_{o2_1,n}, ..., Y_{m_{q_n,n}} \geq y_{o2_{q_n,n}}, \Theta^{(t)})$.

3.4 Difficulties in calculating and maximizing $Q(\Theta \mid y_{obs}, \Theta^{(t)})$

There are two complications involved in estimating parameters $\Theta = (\beta, b, R_n, D)$ based on the above $Q(\Theta \mid y_{obs}, \Theta^{(t)})$ if using a traditional EM approach, namely, the intractable calculation of the posterior distribution of the missing underlying outcome Y_m conditioned upon the observed data y_{o1} and a q_n-dimensional censored distribution, $f(y_m \mid y_{o1}, Y_{m_1} \geq y_{o2_1}, ..., Y_{m_{q_n}} \geq y_{o2_{q_n}}, \Theta^{(t)})$, and the use of complicated LMM with parameters $\Theta = (\beta, b, R_n, D)$ in the M step for a censored MVN mixture data.

In details, the calculation of $Q(\Theta \mid y_{obs}, \Theta^{(t)})$ entails solving the first and second moments of the posterior censored MVN distribution for the missing underlying outcome after intervention for each subject n, i. e. $E(Y_m \mid y_{01}, Y_{m_1} \geq y_{02_1}, ..., Y_{m_{q_n}} \geq y_{02_{q_n}}, \Theta^{(t)})$.and $V(Y_m \mid y_{01}, Y_{m_1} \geq y_{02_1}, ..., Y_{m_q} \geq y_{02_{q_n}}, \Theta^{(t)})$. As Cadez et al. (2002) stated, multivariate integration often entail numerical integration, which is usually subject to exponential time complexity and numerical instability as the dimension of integration increases, so called "the curse of dimensionality". Numerical integration allows moderate accurate multivariate normal probabilities to be quickly computed for problems with as many as ten dimensions (Cadez et cal 1993). However, studies may not be limited to ten visits. For a multi-dimensionally censored MVN, these moments are computationally difficult and time-consuming. Furthermore, estimates of the two moments need to be done for each subject at each post-intervention visit in each iteration. The dimension of missing data q_n is also different for each subject n. All of these adds to the complexity of the computation.

A second complexity comes from the complex regression parameters, $\Theta = (\beta, b, R_n, D)$. Most papers with censored multivariate Gaussian mixture data aim to estimate simple component parameters μ_i and Σ_i where there are K component distributions among N subjects ($N > K$) (Cadez et al 2002, Makarim et al 2006). In our case, not only is it a mixture of N censored component MVN among N subjects, but also the aim is to estimate complex regression parameters $\Theta = (\beta, b, R_n, D)$, which was non-trivial when there was complete data (Laird and Ware 1982).

3.5 Extensions to parameter estimation $\Theta = (\beta, b, R_n, D)$ in a MVN mixture distribution with multi-dimensional censored data

3.5.1 Proper imputation using Monte Carlo in EM to simplify the integration and optimization of Q

When it is hard to calculate or maximize $Q(\Theta \mid y_{obs}, \Theta^{(t)})$ analytically, a viable alternative is to use Monte Carlo as in the Monte Carlo EM (Wei and Tanner 1990, McLachlan and Krishnan 2008, Hughes 1999). In Wei and Tanner's (1990) paper, Monte Carlo was used in the E step by random drawing m samples from the posterior predictive distribution of the missing data $f(y_m \mid y_{obs}, \Theta^{(t)})$ at each iteration to avoid an otherwise intractable calculation of conditional expectation of log likelihood. The Q function can be estimated by averaging the conditional log-likelihoods of m simulated sets of "complete" data

$$Q_{t+1}(\Theta \mid \Theta^{(t)}) = \frac{1}{m} \sum\nolimits_{j=1}^{m} l(\theta \mid y_{obs}, y_m^{(j)}, \Theta^{(t)}).$$

Maximization is then simplified based on complete data loglikelihoods on the right side. When $m = 1$ and y_m is imputed by some "good" summary of the posterior distribution $f(y_m \mid y_{obs}, \theta^{(t)})$, this reduces to an "EM-type" algorithm (McLachlan and Krishnan 1997).

In the application in this paper, if we can find the posterior predictive distribution of the missing data conditional upon the observed data and a multi-dimensionally censored data, $f(Y_m \mid y_{01}, Y_{m_1} \geq y_{02_1}, ..., Y_{m_{q_n}} \geq y_{02_{q_n}}, \Theta^{(t)})$, we can then follow the logic of MCEM or EM-type algorithm to draw samples (random or deterministic) from the posterior distribution. According to Rubin (1996), this would be a proper or confidence proper imputation because the imputation is based on the posterior predictive distribution of the missing data, $P(Y_m \mid y_{obs})$, which is an approximate Bayesian method. Following the EM-type algorithm, if one good draw is taken in the E step ($m = 1$), and assuming that the imputed data is the true data, one then get an augmented "complete" data $\widetilde{y}$. The originally intractable $Q(\Theta \mid y_{obs}, \Theta^{(t)})$ is then reduced to a regular log-likelihood for LMM,

$$Q(\Theta \mid y_{obs}, \Theta^{(t)}) = \log l(\Theta \mid \widetilde{y}) = \sum\nolimits_{n=1}^{N} -\frac{1}{2} \ln |\Sigma_n(R_n, D)| - \frac{1}{2} tr \Sigma_n(R_n, D)^{-1} E^T E,$$

where $E = (\widetilde{y} - U_n(\beta))$.

The term $E(Y_{m,n} \mid y_{obs,n}, \Theta^{(t)})$ is replaced by the imputed post-intervention data $\widetilde{y_{m,n}}$, and $V(Y_{m,n} \mid y_{obs,n}, \Theta^{(t)})$ disappeared due to the replacement of the "complete" data. An originally intractale data reference is then reduced to a complete data inference for LMM. One can then follow the regular procedures of LMM to get the MLE or REML estimates of parameters using available software. This would simplify the computation significantly and provide a lot of modeling convenience like utilizing the existing model of fit, existing variance-covariance matrices, and univariate-multivariate inference.

3.5.2 A simplified calculation for a multi-dimensionally censored MVN posterior distribution $f(y_m \mid y_{o1}, Y_{m_1} \geq y_{o2_1}, ..., Y_{m_{q_n}} \geq y_{o2_{q_n}}, \Theta^{(t)})$

The next question is how to derive the posterior predictive distribution of the missing data in a N-component MVN mixture data in presence of a multi-dimensionally censored data, and the dimension of missing data q_n varies for each component MVN. Given the computational complexity and instability of numerical integration for larger dimensions, can we find a simple yet sufficiently accurate method to achieve this without numerical integration?

The posterior predictive distribution of the underlying Y_m given a multi-dimensionally truncated distribution is skewed. However, according to Schafer (1997), MVN is robust in imputation to distributions that are manifestly not normal when the amount of missing data is not large. Several authors showed that approximate methods for sampling from the posterior distribution of parameter estimates may be sufficient when the proportion of information lost of censoring is moderate or small (Rubin 1996, Dorey et al 1993). Furthermore, MVN has a nice property that any subset of a MVN is still distributed as a MVN which other distributions do not possess. Therefore, MVN is employed to get the posterior predictive distribution of Y_m. Transformations can be applied to deviations from MVN within the frame work of MVN. For MVN, the first and second moments are sufficient to define the distribution.

Next question is how to get around multiple dimension intergration without numerical integration? One thought is to take only the most informative subset of MVN to do the imputation. In order to impute the underlying value of Y_m at a post-medication visit m_i, data that gives the most information are 1) the pre-medicated data y_{o1} which gives the exact location of the prior data, and 2) the treated value at the visit of interest m_i, y_{o2_i}, which gives a lower boundary for the underlying value Y at that visit. Often times the rest treated data $y_{o2_{(-i)}}$ do not add much more information than y_{o1} and y_{o2_i} as they do not give exact location for values at the rest of the visits, nor do they give direct information for the value at the visit of interest m_i. For a MVN, a subset of the MVN is still distributed as a MVN. Therefore, the pre-medicated p_n-dimensional data y_{o1} and the treated value at a specific visit m_i after medication, y_{o2_i}, are distributed as follows,

$$\begin{bmatrix} Y_{o1} \\ Y_{m_i} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{o1}(\beta) \\ \mu_{m_i}(\beta) \end{bmatrix}, \begin{bmatrix} \Sigma_{o1}(R_n, D) & \Sigma_{o1,m_i}(R_n, D) \\ \Sigma_{m_i,o1}(R_n, D) & \Sigma_{m_i}(R_n, D) \end{bmatrix} \right)$$

The first two moments are then simplified to,

$$E(Y_{m_i} \mid y_{o1}, Y_{m_1} \geq y_{o2_1}, ..., Y_{m_{q_n}} \geq y_{o2_{q_n}}, \Theta^{(t)}) \simeq E(Y_{m_i} \mid y_{o1}, Y_{m_i} \geq y_{o2_i}, \Theta),$$

$$V(Y_{m_i} \mid y_{o1}, Y_{m_1} \geq y_{o2_1}, ..., Y_{m_{q_n}} \geq y_{o2_{q_n}}, \Theta^{(t)}) \simeq V(Y_{m_i} \mid y_{o1}, Y_{m_i} \geq y_{o2_i}, \Theta),$$

which can be analytically derived.

Likewise, the posterior pairwise covariance of the underlying outcome (Y_{m_i}, Y_{m_j}) at any two post-medication visits m_i and m_j, can be simplified as,

$$Cov(Y_{m_i}, Y_{m_j} \mid y_{o1}, Y_{m_1} \geq y_{o2_1}, ..., Y_{m_{q_n}} \geq y_{o2_{q_n}}, \Theta^{(t)}) \simeq Cov(Y_{m_i}, Y_{m_j} \mid y_{o1}, Y_{m_i} \geq y_{o2_i}, Y_{m_j} \geq y_{o2_j}, \Theta).$$

For brevity, the results of these derivation are not shown. The complete posterior variance-covariance matrix is then composed of diagonal entry $V(Y_{m_i} \mid y_{o1}, Y_{m_i} \geq y_{o2_1}, \Theta)$, and (i,j) off-diagonal entry $Cov(Y_{m_i}, Y_{m_j} \mid y_{o1}, Y_{m_i} \geq y_{o2_i}, Y_{m_j} \geq y_{o2_j}, \Theta)$. Since it is a simplified version of the true variance-covariance matrix, the resulting posterior var-covariance matrices, $V(Y_m \mid y_{obs}, \Theta)$ for some patients can be negative definite. Similar to ridge regression (Hoerl 1962, 1970), diagonal inflation can be used to increase the diagonal values of the matrix and force the resulting var-covariance matrix to be positive definite. Diagonal inflation is performed by adding the largest eigen-value in absolute value to the diagonal of the negative definite matrix. Simulation later proves the robustness of this approximation.

3.6 A Modified EM-type algorithm; single deterministic imputation

Following the EM-type algorithm, we propose a modified EM-type algorithm and a single deterministic imputation. That is, only one draw is taken in the E step ($m = 1$) and a good summary of the posterior distribution of the missing data, the posterior expectation, is used for imputation. Since the posterior predictive distribution is approximated by a simplified calculation, one can call this a modified EM-type algorithm.

3.6.1 Starting data

There are different ways of choosing a starting data. A good starting value is usually vital to the convergence or a faster convergence of any algorithm. In this paper, the observed data y_{obs} (observed pre-medicated data $y_{o1}$and treated post-medicated data y_{o2}) turned out to be the best starting data. Another option is to use the complete case data, or the observed y_{obs} for those never on intervention. However, since those without intervention tend to have a much lower underlying Y, this will introduce more bias than the observed data. For subject n, the starting data is,$\widetilde{y_n}^{(0)} = [y_{o1,n}, y_{o2,n}]'$.

3.6.2 Starting model

Among the current methods for a longitudinal quantitative data modified by a non-trial medication intervention, the Multi-level model (White *et al.* 2001), a LMM model after adjustment for use of medication as a time-dependent covariates, was selected as the starting model in this paper because it is the only available method that can handle a general longitudinal data structure. The others are restricted to some special longitudinal data set up (Cook 1997 and 2006, McClelland et al 2008).

The Multi-level model is specified as:

$$y_{nj} = u + \beta_0 G_n + b_{on} + (\beta_1 + \beta_2 G_n + b_{1n}) T_j + M_{nj} G_n + W_n \gamma + \varepsilon_{nj},$$

where M_{nj} is the time-dependent use of medication. To make the model robust to a heterogeneous medication effect on Y between groups, an interaction term between the medication use and treatment/exposure group, $M_{nj} G_n$, is employed. Otherwise, a main medication effect M_{nj} would suffice. Starting data $\widetilde{y_n}^{(0)}$ is fit into the Multi-level model.

3.6.3 E-step or Imputation Model

In each iteration, the missing underlying value of Y_{m_i} at a specific post-medication visit m_i for a subject is replaced by a good summary of its posterior distribution, a simplified approximation of the posterior expectation conditional upon the observed p_n-dimensional y_{o1}, and a 1-dimensional truncated distribution based on the treated value at the visit of interest m_i, $Y_{m_i} \geq y_{o2_i}$, under the current estimate of the parameters Θ, as previously discussed, $E(Y_{m_i} \mid y_{o1}, Y_{m_i} \geq y_{o2_i}, \Theta)$.The augmented data $\widetilde{y_n}$ is composed of the observed value $y_{o1,n}$ prior to medication, and the imputed underlying value after medication, $\widetilde{y_{m,n}} = E(Y_{m_i} \mid y_{o1}, Y_{m_i} \geq y_{o2_i}, \Theta)$, $\widetilde{y_n} = \left[y_{o1,n}, \widetilde{y_m, n}\right]'$.

3.6.4 M step or Analysis Model: LMM

Assuming the augmented $\widetilde{y_{nj}}$ is the true underlying Y for the nth subject $(1 \leq n \leq N)$ at the jth visit $(1 \leq j \leq t_n)$. The LMM can be modeled as previously defined,

$$\widetilde{y_{nj}} = u + \beta_0 G_n + b_{on} + (\beta_1 + \beta_2 G_n + b_{1n}) T_j + W_n^T \gamma + \varepsilon_{nj}.$$

Iterate between the E and M step, or between the imputation and analysis model, until a pre-specified stopping rule is satisfied, for example, $\left|\beta^{(t+1)} - \beta^{(t)}\right| < \varepsilon$, or the log-likelihood is stablized, where ε is a prescribed extremely small number. The MLE estimate is the final estimate of Θ.

3.7 MCEM-MI algorithm - Combination of Monte Carlo EM and MI

Single imputation is often subject to the criticism of under-estimating the true variability of the parameters. To overcome this, a multiple imputation approach, the Modified Monte Carlo EM-MI algorithm is proposed which combines the Monte Carlo EM algorithm and multiple imputation. In Wei and Tanner's (1990) MCEM, m random draws were taken from the posterior predictive distribution of the missing data in the E step to simplify the computation of $Q(\Theta \mid y_{obs}, \Theta^{(t)})$.For the modified MCEM-MI algorithm, we propose to take one random draw from the posterior predictive distribution in the E step at each iteration, rather than taking one deterministic draw as in the modified EM-type algorithm. Follow the rest modified EM-type algorithm, and iterate to get one set of parameter estimates. Repeat this whole process m times to get m set of parameters. Rubin's combination rule (1987) is then employed to get the final parameter estimates.

3.8 Convergence complications

For the EM-type algorithm, Chauveau (1995) proved that it converges linearly to the strongly consistent MLE of θ with suitable starting values. For a regular MI, Rubin (1987, 1996) concluded that when repeated imputations are proper or confidence proper for the complete data inference $(\widehat{Q}, U)$ and the complete data inference is randomization-valid for the estimand Q, then the large-m repeated-imputation inference given by $(Q - \overline{Q}_\infty) \sim N(0, T_\infty)$ is randomization or confidence valid for the scientific estimand Q under the posited response mechanism. Rubin suggested that imputation with a Bayesian or approximate Bayesian imputation. The proposed modified MCEM-MI algorithm used an approximate Bayesian method, i.e. taking random draws from a simplified approximation of the posterior predictive distribution of the missing data given the observed data.

For the proposed modified EM-type and MCEM-MI algorithms, however, the convergence is more complicated due to the following reasons. First, a simplified approximation of the posterior predictive distribution of the missing data was employed. This may introduce a bias when the amount of imputation is large or when the data is noisy, e.g. when between-visit correlation is low. Secondly, in the proposed methods, the posterior moments are based on a truncated distribution which sets a lower boundary for the underlying outcome Y_{m_i}, i.e., the treated value, y_{o2_i}, but not an upper boundary. This may over-estimate the parameter of interest sometimes. When there is a larger percentage of missing data or when the data is noisy, a modest departure in the beginning may lead to a larger deviation by the end of iteration.

These considerations lead to the proposal of three sub-algorithms within the class of modified EM-type algorithms. In order to incorporate the above convergence complications, we propose to implement the modified EM-type and MCEM-MI algorithms in three ways, yielding full-iteration, one-step and two-step sub-algorithm. When the amount of imputation is modest and data is not noisy, a full iteration algorithm can be employed. In case of a large amount of imputation or a noisy data, a one- or two-step iteration would be sufficient.

4 Simulation

A numerical study is conducted to compare the bias and precision of current methods vs. proposed methods in estimating the slope or rate of change in the underlying Y when Y is altered by the administration of a non-randomized, non-trial therapeutic intervention.

4.1 Simulated data sets

For the nth subject $(1 \leq n \leq N)$, the underlying outcome Y at the jth visit is generated from a LMM as previously introduced,

$$y_{nj} = u + b_{on} + (\beta_1 + \beta_2 G_n + b_{1n})T_j + \varepsilon_{nj}$$

Baseline u is simulated to be 65. Slopes between the two groups can be 2 vs. 1, or 5 vs. 1, or 2 vs. 2 unit per year. The variance structure of the random components are simulated based on a coefficient of variability of 0.2 in Y. The residual term can have a first-order auto-regressive or compound symmetry variance-covariance structure with different correlation coefficient $\rho = 0.6 - 0.9$. Group G_n is generated from a binomial distribution with probability of 0.5. Time T_j can have up to 3 or 5 or 10 visits. It can also be annual, every half-year or quarterly visit. Patients can be intervened for treatment (100% or 80% on med once $y_{nj} > 80$ or 90), or for prevention (20% on med if $y_{nj} < 80$ or 90). Once a patient is intervened, a random variate $N(-20, -5^2)$ or $N(-10, -2^2)$ is deducted from the underlying y which generates an observed outcome y_obs. Each simulated data is generated by a combination of different factors listed above. For each simulated scenario, 100 data sets are generated. Every data contains 1000 subjects, each with a number of visits.

4.2 Applied methods

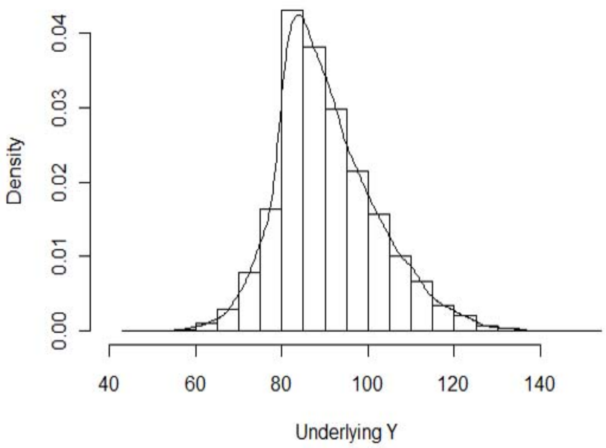
Methods applied are 0) the true model or no medication model where no one is intervened; 1) the "ignore" model where the observed Y is treated as the underlying Y; 2) the "exclude" model where only complete case data is used for analysis; 3) the Multilevel Model (White. *et al.* 2001) which is a LMM after adjustment for use of medication and an interaction term between medication use and group; 4) a class of Modified EM-type algorithms implemented in three ways, 4.1) a full iteration, 4.2) a one-step iteration. and 4.3) a two-step iteration algorithm; 5) a class of Modified MCEM-MI algorithms including 5.1) a full iteration, 5.2) a one-step iteration, and 5.3) a two-step iteration algorithm; 6) adding a constant to the treated value y_{obs} (Tobin *et al.* 2005, Cui *et al.* 2003), including 6.1) adding 5, 6.2) adding 10, 6.3) adding 15, and 6.4) adding 20 to the treated value y_{obs}.

4.3 One Scenario

We simulated over 60 scenarios. One scenario is given here as an example. This data was generated from a LMM allowing for random intercept and slope, and a first-order auto-regressive serial correlation matrix with $\rho = 0.9$. The progression rates are 2 vs. 1 unit per year for the treatment/risk group and the control group. Once the underlying outcome exceeds 80, 80% of the patients are put on medication, resulting in 33.4% and 29.3% of medication use in the two groups. Medication effect is random and heterogeneous across Y ($N(-20, -5^2)$ if $Y > 90$, and $N(-10, -2^2)$ if $80 < Y \leq 90$).

The ignore and exclude methods severly underestimate the slope in both groups and none of the 95% confidence intervals cover the true values. The medication adjustment method produces estimates that are less biased, but they are still negatively biased and confidence interval coverage is poor (less

Figure 3: Histogram of the true underlying Y variable.



than 20%). The EM algorithm produces estimates with little bias, but standard errors are a bit small to give adequate coverage (less than 70% for the two slopes). Doing just two steps of the EM algorithm produces nearly the same estimates of slopes, but larger standard errors and better coverage (nearly 90% fo the two slopes). Mutiple imputation (MI) does the best: little bias in estiamtes and slightly above 95% coverage. MI with two iterations does about as well as the full MI. Add a constant (10, 15, or 20) to the observations censored by rescue medication produce better estimates than ignore or exclude, but only work well for point estimates if the amount adjustment is correct on average (15 in this case) and do not do well in terms of confidence interval coverage. A two-step or full iteration MCEM-MI or EM type algorithm give the best estimates of the slopes for both individual groups and group difference.

To examine the performance of the simulated $\widetilde{Y}$ against the true underlying Y, Figure 3 shows the distribution of the true underlying Y for those on intervention. Notably, although the data is heavily skewed to the right with the peak at 80, there is a small tail on the left side below 80. This is because due to the within-subject variation, or regression to the mean. Some patients may still experience a measurement below 80 after commencement of intervention, especially in the control group where the progression is flat over time.

The histogram of the imputed underlying Y from a modified EM-type algorithm with full-iteration is shown in Figure 4. It basically resembles the skewed distribution of the true underlying Y as in Figure 3 except for minor differences.

A scatter plot is drawn between the true and imputed underlying Y to check the reliability of the two (Figure 5). Data are about evenly distributed above and below the symmetry line (green).

4.4 Simulation Results

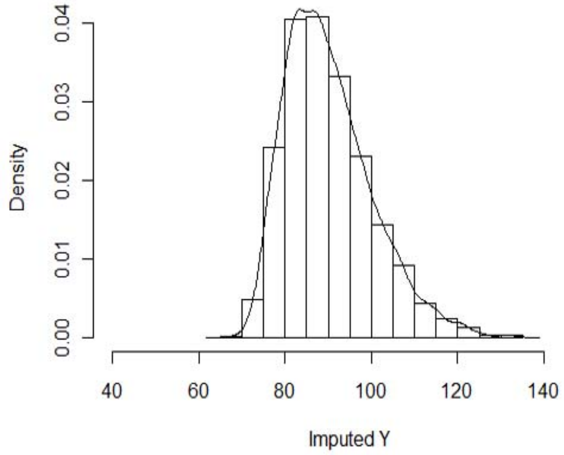
Simulation results is summarized as following based on over 60 simulated scenarios.

4.4.1 Two proposed methods vs. other four methods applied

Overall speaking, similar to what was observed in the cross-sectional studies as shown in Tobin et al. 2005, in longitudinal studies, when the quantitative outcome is altered by a non-trial, non-randomized intervention, if without appropriate correction, analysis based on the observed outcome Y_{obs} (the "ignore" or "exclude" method) can lead to a substantial bias in the estimated treatment or exposure or association effects, seriously distort the analysis and undermine the scientific aims of the study.

Simulation shows that at least one of the sub-algorithms within the proposed two classes of methods have the least biased treatment effect estimates in a majority of simulated scenarios amongst six meth-

Figure 4: Histogram of the Imputed Underlying Y Variable



ods applied. These scenarios include when there is homogeneous or heterogeneous medication effect across different Y levels, when patients are intervened for single reason (treatment) or multiple reasons (treatment or prevention), when there are fixed effects and/or random effects, when there are more or fewer visits, or when the proportion of medication use is large or small.

The other four simple methods give biased estimates for group-specific slopes in almost all of the simulated scenarios, and give biased estimates for slope difference between groups in most of the scenarios except for when medication assignment is balanced between the two arms. In the latter case, the medication adjustment model and "ignore" and "exclude" model give unbiased estimate for slope difference between groups, although these methods still under-estimate slopes for individual groups.

Notably, there are two scenarios where other simple models apply better.

First, if therapeutic intervention is randomly assigned, i.e., the assignment is unrelated to the outcome of interest Y, then the Multi-level mode and the 'exclude' model are the best models which give unbiased estimates of both group-specific slopes and slope differences between groups. In reality, however, most likely the administration of non-trial medication is not randomized, but related to the underlying Y.

Second, if a prior knowledge of the true average size of medication effect is known, adding an exact constant to the treated Y will give unbiased estimates of the slopes for both individual groups and group differences, no matter how large the amount of missing data is. However, it is often hard to get a prior knowledge of the true medication effect, and medication effect is usually heterogeneous across different brands, different dosage levels, different levels of underlying Y, and different subjects. According to a meta-analysis by Low *et al.* (2003), higher dosage has more reduction in BP, and anti-hypertensives have more reduction at a higher BP level. Medication effect for each subject also varies according to age, gender, and genetic susceptibility. Therefore, it is often hard to get the true average size of medication effect. The method of adding a constant, however, is very sensitive to the added constant. The bias can be large if the added constant departs from the true average size of medication effect.

4.4.2 Between two proposed methods - Modified EM-type vs. Modified MCEM-MI Model

A modified EM-type algorithm and modified MCEM-MI algorithm perform similarly in point estimates of slopes for individual groups and group differences. As expected, a modified MCEM-MI algorithm has a better coverage of the true parameters due to the inclusion of between-imputation variation, whereas a modified EM-type algorithm converges very fast, usually within 7 iterations, contrary to the slow convergence of a regular EM algorithm.

4.4.3 Comparing three proposed sub-algorithms within each class of methods - Full Iteration vs. 1 step vs. 2 steps

Depending upon different scenarios, one of the three sub-algorithms would fit the data the best. A rule of thumb is that the larger the amount of imputation, the more noisy the data is, less iteration gives more accurate estimates than a full-iteration. A full iteration gives the best slope estimates when the amount of missing data is small to moderate (<20%), and when the data is not noisy (between-visit correlation $\geq 90\%$. A one step iteration gives the least biased estimate when the percentage of missing data is $> 30\%$, or the data is noisy (between-visit correlation < 0.7). A two-step iteration fits in between. Although approximate methods works the best when the proportion of missing data is moderate or small (Rubin 1996, Dorey et al 1993), the extension to one step iteration relaxed this condition.

5 Discussion and Conclusions

Methods were successfully applied to data from the DCCT/EDIC study. Results were presented in Sun (2012). Due to space limitations, results are not presented in this proceedings paper. Rather, extensive applied results will be deferred to a clinical paper.

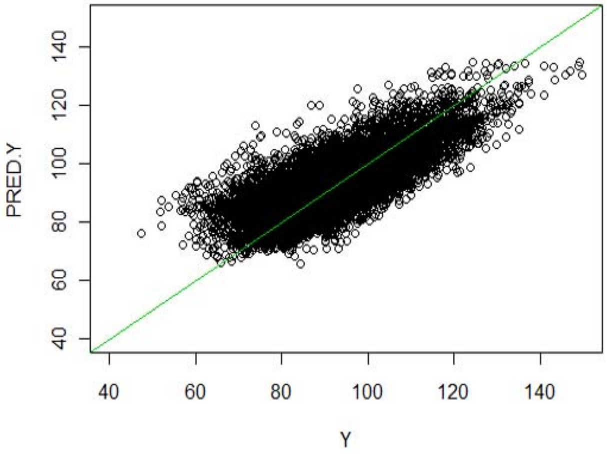
The proposed methods are an improvement over the current available methods for a quantitative outcome censored by a non-trial, non-randomized intervention, by applying to more general scenarios and more general longitudinal data structure while reducing their restrictive assumptions. These include when there is homogeneous or heterogeneous medication effect across different Y levels, when patients are intervened for single reason (treatment) or multiple reasons (treatment or prevention), when there are fixed effects and/or random effects, when there are more or fewer visits, or when the proportion of medication use is large or small. The proposed methods extend Tobin *et al.*'s (2005) censored normal regression for cross-section data to one for longitudinal data, extend Cook's 1997 MI method from a restrictive longitudinal data to a more general longitudinal data. These methods revise Wei and Tanner's Monte Carlo EM (1990) for a different application, and can be viewed as a variation of Gibbs Sampling in the frequentist framework. Different from methods using numerical integration, the proposed methods replace the intractable calculation of a multi-dimensionally censored MVN posterior distribution with a simplified approximation yet maintain sufficient accuracy. It gets around "the curse of the dimensionality" by avoiding complicated numerical integration of MVN probability. The proposed methods enjoy straightforward implementation using existing software in the M step. This simplifies the computation significantly, at the same time, provide a lot of modeling convenience (a variety of existing complex variance-covariance structures, existing model of fit and multivariate inference within the LMM and MI framework. The proposed methods also converge fast, especially the modified EM-type algorithm, usually within 7 iterations.

The limitations of the proposed methods are summarized as follows. *First*, even though a general guideline is given regarding which of the three sub-algorithms should be selected for different scenarios, it is not as convenient as having one single algorithm for all the scenarios. A sensitivity analysis needs be done by comparing different approaches to locate a best model. *Second*, as previously mentioned, there are two scenarios where other simple models apply better, although these two conditions are not realistic in practice.

Our suggestion is that when medication assignment is randomized or when the medication reduction effect is prior known, simple ad hoc methods are the best approach to go, i.e. the "exclude" method, medication adjustment model, or adding a constant to the treated outcome. However, when medication assignment is not randomized but related to the outcome Y, and when there is no or insufficient prior knowledge about medication reduction effect, the proposed methods are viable alternatives. In both cases, however, sensitivity analysis should be done to verify the results from the main analysis.

Several generalizations were discussed in Sun (2012) and will be considered in future work.

Figure 5: Scatter plot of the imputed versus true underlying Y variable



6 References

Brand, E., Wang, J. G., Herrmann, S. M., & Staessen, J. A. (2003). An epidemiological study of blood pressure and metabolic phenotypes in relation to the Gbeta3 C825T polymorphism. *Journal of Hypertension*, 21(4), 729-737.

Brown, C. H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics*, 46(1), 143-155.

Cadez, I. V., McLachlan, G. J., & McLaren, C. E. (2002). Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning*, 47(1), 7-34.

Chauveau, D. (1995). A stochastic EM algorithm for mixtures with censored data. *Journal of Statistical Planning and Inference*, 46(1), 1-25.

Cook, N. R. (1997). An imputation method for non-ignorable missing data in studies of blood pressure. *Statistics in Medicine*, 16(23), 2713-2728.

Cook, N. R. (2006). Imputation strategies fro blood pressure data nonignorably missing due to medication data. *Clinical Trials*, 3(5), 411-420.

Cui, J. S., Hopper, J. L., & Harrap, S. B. (2002). Genes and family environment explain correlations between blood pressure and body mass index. *Hypertension*, 40(1), 7-12.

Cui, J. S., Hopper, J. L., & Harrap, S. B. (2003). Antihypertensive treatments obscure familial contributions to blood pressure variation. *Hypertension*, 41(2), 207-210.

The DCCT Research Group. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. textit New England Journal of Medicine, 329(14), 977-986. doi:10.1056/NEJM199309303291401

The DCCT Research Group. (1995). The effect of intensive diabetes therapy on the development and progression of nephropathy in the diabetes control and complications trial (DCCT). *Kidney International*, 47(1), 1703-1720.

The DCCT Research Group. (1998). The effect of intensive diabetes therapy on measures of autonomic nervous system function in the diabetes control and complications trial (DCCT). *Diabetologia*, 41(1), 416-423.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963-974.

Lindstrom, M. J., & Bates, D. M. (1988). Newton-raphson and EM algorithms for linear mixed effects models for repeated-measures data. *American Statistical Association*, (83), 1014-1022.

Makarim, A. E., Aboueissa, A., Stoline, M. R. (2006) Maximum likelihood estimators of population parameters from doubly left-censored samples. Environmetrics, 17: 811-826.

Masca, N., Sheehan, N., & Tobin, M. (2011). Pharmacogenetic interactions and their potential effects on genetic analyses of blood pressure. *Statistics in Medicine*, 30(7), 769-783.

Matsubara, M., Kikuya, M., Ohkubo, T., Metoki, H., Omori, F., Fujiwara, T., et al. (2001). Aldosterone synthase gene (CYP11B2) C-334T polymorphism, ambulatory blood pressure and nocturnal decline in blood pressure in the general japanese population: The ohasama study. *Journal of Hypertension*, 19(12), 2179-2184.

McClelland, R. L., Kronmal, R. A., Haessler, J., Blumenthal, R. S., & Goff, D. C., Jr. (2008). Estimation of risk factor associations when the response is influenced by medication use: An imputation approach. *Statistics in Medicine*, 27(24), 5039-5053.

McLachlan, G. J. & Jones, P. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, 44(2), 571-578.

McLachlan, G. J. & Krishnan, T. (1997). *The EM algorithm and extensions.* New Jersey: Wiley.

McLachlan, G. J. & Krishnan, T. (2008). *The EM algorithm and extensions* (2 ed.). New Jersey: Wiley.

Rice, T., Rankinen, T., Province, M. A., Chagnon, Y. C., Perusse, L., Borecki, I. B., et al. (2000). Genome-wide linkage analysis of systolic and diastolic blood pressure: The quebec family study. *Circulation*, 102(16), 1956-1963.

Rubin, D. B. (1987). textit Imputation for non-response in surveys. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* New York: Chapman and Hall.

Schunkert, H., Hense, H. W., Doring, A., Riegger, G. A., & Siffert, W. (1998). Association between a polymorphism in the G protein beta3 subunit gene and lower renin and elevated diastolic blood pressure levels. *Hypertension*, 32(3), 510-513.

Schwartz, D., & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases*, 20, 637-648.

Sun, W. (2012). Methods for a Longitudinal Quantitative Outcome With a Multivariate Gaussian Mixture Distribution Multi-dimensionally Censored by Therapeutic Intervention. Ph.D. Dissertation, The George Washington University, Department of Biostatistics.

Tobin, M. D., Sheehan, N. A., Scurrah, K. J., & Burton, P. R. (2005). Adjusting for treatment effects in studies of quantitative traits: Antihypertensive therapy and systolic blood pressure. *Statistics in Medicine*, 24(19), 2911-2935.

Wei, G. C. G., & Tanner, M. A. (1990). A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411), 699-704.

White, I. R., Babiker, A. G., Walker, S., & Darbyshire, J. H. (1999). Randomization-based methods for correcting for treatment changes: Examples from the concorde trial. *Statistics in Medicine*, 18(19), 2617-2634.

White, I. R., Carpenter, J., Pocock, S. J., & Henderson, R. A. (2003). Adjusting treatment comparisons to account for non-randomized interventions: An example from an angina trial. *Statistics in Medicine*, 22(5), 781-793.

White, I. R., Bamias, C., Hardy, P., Pocock, S., & Warner, J. (2001). Randomized clinical trials with added rescue medication: Some approaches to their analysis and interpretation. *Statistics in Medicine*, 20(20), 2995-3008.

White, I. R., Chaturvedi, N., & McKeigue, P. M. (1994). Median analysis of blood pressure for a sample including treated hypertensives. *Statistics in Medicine*, 13(16), 1635-1641.

White, I. R., Koupilova, I., & Carpenter, J. (2003). The use of regression models for medians when observed outcomes may be modified by interventions. *Statistics in Medicine*, 22(7), 1083-1096.

White, I. R., & Pocock, S. J. (1996). Statistical reporting of clinical trials with individual changes from allocated treatment. *Statistics in Medicine*, 15(3), 249-262.

White., N. H., Sun, W., Cleary, P. A., Danis, R. P., Davis, M. D., Hainsworth, D. P., et al. (2008). Prolonged effect of intensive therapy on the risk of retinopathy complications in patients with type 1 diabetes mellitus: 10 years after the diabetes control and complications trial. *Archives of Ophthalmology*, 126(12), 1707-1715.