Supplementary Material

(Schomaker,M., Hogger, S., Johnson, L., Hoffmann, C., Bärnighausen, T., Heumann, C., *Simultaneous Treatment of Missing Data and Measurement Error in HIV Research using Multiple Overimputation,* 2015, Epidemiology)

eTable 1: Mortality in South African patients after starting antiretroviral treatment. Cox regression estimates, reported as hazard ratios, for a complete case analysis, multiple imputation and multiple overimputation. 95% confidence intervals are reported in brackets. All results relate to the data from the illustrative example and should not be interpreted causally.

		Complete Cases	**Multiple Imputation**	**Multiple Overimputation**
Baseline CD4				
	<25	1	1	1
	25-50	0.73 [0.62;0.87]	0.74 [0.67;0.82]	0.63 [0.58;0.70]
	50-100	0.47 [0.40;0.56]	0.49 [0.45;0.54]	0.47 [0.43;0.52]
	100-200	0.33 [0.28;0.38]	0.33 [0.30;0.36]	0.34 [0.30;0.38]
	>200	0.38 [0.29;0.48]	0.29 [0.25;0.34]	0.21 [0.18;0.24]
Baseline $\log_{10}$ viral load				
	<4	1	1	1
	4 to 5	1.29 [1.07;1.57]	1.18 [1.07;1.31]	1.20 [1.08;1.33]
	5 to 6	1.54 [1.26;1.88]	1.41 [1.26;1.57]	1.42 [1.27;1.59]
	>6	1.71 [1.27;2.29]	1.60 [1.34;1.91]	1.66 [1.37;2.01]
Sex				
	Female	1	1	1
	Male	1.34 [1.19;1.51]	1.31 [1.22;1.40]	1.31 [1.22;1.41]
Age				
	<25	1	1	1
	25-35	1.01 [0.81;1.26]	1.00 [0.87;1.16]	1.04 [0.90;1.20]
	35-45	1.05 [0.83;1.32]	1.09 [0.94;1.26]	1.14 [0.99;1.32]
	>45	1.22 [0.95;1.57]	1.36 [1.16;1.59]	1.37 [1.17;1.60]
Year				
	before 2004	1	1	1
	2004-2006	0.94 [0.75;1.18]	1.27 [1.06;1.50]	1.27 [1.06;1.50]
	2007 and after	1.01 [0.76;1.34]	1.52 [1.25;1.84]	1.53 [1.26;1.85]
Cohort				
	A	1	1	1
	B	0.81 [0.67;0.97]	0.93 [0.83;1.04]	0.93 [0.83;1.04]
	C	0.70 [0.44;1.13]	0.88 [0.76;1.00]	0.89 [0.78;1.02]
	D	0.85 [0.71;1.01]	0.96 [0.88;1.06]	0.95 [0.86;1.04]

eTable 2: Mortality in South African patients after starting antiretroviral treatment. Estimates from a Cox regression model, reported as hazard ratios, if baseline viral load was not included in the analysis. Results are reported for a complete case analysis, multiple imputation and multiple overimputation. 95% confidence intervals are reported in brackets. All results relate to the data from the illustrative example and should not be interpreted causally.

	Complete Cases	**Multiple Imputation**	**Multiple Overimputation**
Baseline CD4			
<25	1	1	1
25-50	0.75 [0.63;0.89]	0.73 [0.66;0.81]	0.62 [0.57;0.68]
50-100	0.47 [0.40;0.56]	0.48 [0.44;0.53]	0.46 [0.42;0.50]
100-200	0.32 [0.27;0.37]	0.32 [0.29;0.35]	0.33 [0.29;0.36]
>200	0.36 [0.28;0.46]	0.28 [0.24;0.33]	0.20 [0.17;0.23]
Sex			
Female	1	1	1
Male	1.36 [1.21;1.53]	1.33 [1.24;1.42]	1.33 [1.24;1.43]
Age			
<25	1	1	1
25-35	1.00 [0.80;1.25]	1.00 [0.87;1.15]	1.04 [0.90;1.21]
35-45	1.04 [0.83;1.31]	1.09 [0.94;1.26]	1.15 [1.00;1.33]
>45	1.22 [0.95;1.57]	1.36 [1.16;1.59]	1.39 [1.19;1.62]
Year			
before 2004	1	1	1
2004-2006	0.96 [0.76;1.20]	1.28 [1.07;1.52]	1.27 [1.07;1.52]
2007 and after	0.99 [0.74;1.31]	1.51 [1.24;1.83]	1.52 [1.25;1.85]
Cohort			
A	1	1	1
B	0.86 [0.72;1.03]	0.98 [0.88;1.10]	0.99 [0.88;1.11]
C	0.73 [0.45;1.16]	0.89 [0.77;1.02]	0.90 [0.78;1.03]
D	0.81 [0.69;0.95]	0.91 [0.83;1.00]	0.90 [0.82;0.99]

eTable 3: Mortality in South African patients after starting ART. Estimates from a Cox regression model, reported as hazard ratios, if baseline TB, WHO stage, and haemoglobin are added to the analysis. Results are reported for a complete case analysis, multiple imputation and multiple overimputation. 95% CI's are reported in brackets. All results relate to the data from the illustrative example and should not be interpreted causally.

	Complete Cases	**Multiple Imputation**	**Multiple Overimputation**
Baseline CD4			
<25	1	1	1
25-50	0.74 [0.58;0.94]	0.80 [0.72;0.88]	0.68 [0.62;0.74]
50-100	0.5 [0.39;0.63]	0.56 [0.51;0.62]	0.50 [0.45;0.55]
100-200	0.48 [0.39;0.6]	0.41 [0.37;0.45]	0.37 [0.33;0.42]
>200	0.49 [0.35;0.7]	0.37 [0.32;0.44]	0.24 [0.21;0.28]
Baseline log$_{10}$ viral load			
<4	1	1	1
4 to 5	1.43 [1.07;1.91]	1.11 [1.01;1.23]	1.14 [1.03;1.27]
5 to 6	1.53 [1.14;2.06]	1.24 [1.11;1.38]	1.26 [1.13;1.42]
>6	1.28 [0.85;1.94]	1.32 [1.11;1.58]	1.43 [1.18;1.73]
Prevalent TB			
no	1	1	1
yes	1.98 [1.38;2.83]	1.11 [0.96;1.28]	1.07 [0.93;1.24]
Baseline WHO stage			
I & II	1	1	1
III	2.45 [1.68;3.56]	1.4 [1.26;1.55]	1.34 [1.21;1.48]
IV	3.40 [2.27;5.1]	1.85 [1.67;2.05]	1.77 [1.60;1.96]
Hemoglobin			
per gm/dL	0.85 [0.82;0.88]	0.9 [0.89;0.91]	0.89 [0.88;0.9]
Sex			
Female	1	1	1
Male	1.34 [1.13;1.59]	1.44 [1.34;1.54]	1.44 [1.34;1.54]
Age			
<25	1	1	1
25-35	0.93 [0.7;1.24]	1 [0.87;1.16]	1.04 [0.91;1.21]
35-45	0.91 [0.68;1.24]	1.1 [0.95;1.27]	1.17 [1.01;1.36]
>45	1.14 [0.81;1.58]	1.39 [1.19;1.62]	1.44 [1.23;1.68]
Year			
before 2004	1	1	1
2004-2006	0.77 [0.56;1.08]	1.32 [1.11;1.57]	1.33 [1.12;1.58]
2007 and after	0.77 [0.51;1.16]	1.56 [1.29;1.89]	1.61 [1.32;1.95]
Cohort			
A	1	1	1
B	0.59 [0.45;0.76]	0.86 [0.77;0.97]	0.87 [0.77;0.98]
C	---[1]	0.87 [0.76;1]	0.88 [0.77;1.01]
D	---[1]	0.97 [0.88;1.07]	0.97 [0.88;1.07]

[1] Cohorts C and D are excluded in the complete case analysis because of missing WHO stage data

eTable 4: Mortality in South African patients after starting antiretroviral treatment. Cox regression estimates, reported as hazard ratios, based on the longitudinal data, stratified by cohort. Results are reported for a complete case analysis, multiple imputation and multiple overimputation. 95% confidence intervals are reported in brackets. All results relate to the data from the illustrative example and should not be interpreted causally.

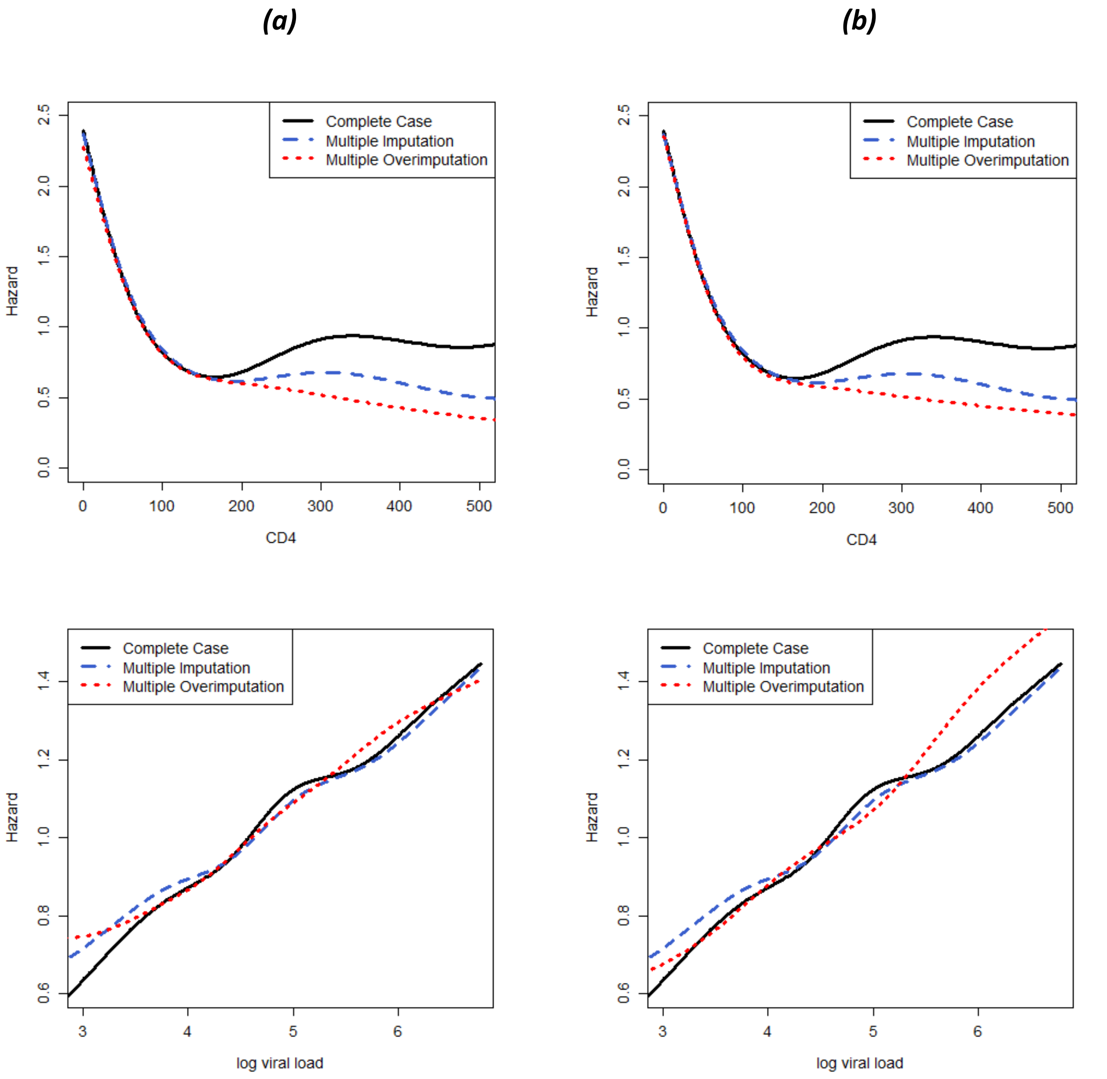
	Complete Cases	**Multiple Imputation**	**Multiple Overimputation**
Time-updated CD4			
<25	1	1	1
25-50	0.76 [0.56;1.03]	0.72 [0.63;0.82]	0.45 [0.40;0.51]
50-100	0.39 [0.29;0.52]	0.42 [0.37;0.48]	0.25 [0.22;0.29]
100-200	0.25 [0.19;0.33]	0.22 [0.19;0.25]	0.14 [0.12;0.16]
>200	0.10 [0.07;0.16]	0.13 [0.11;0.15]	0.06 [0.05;0.08]
Time-updated virological sup.			
unsuppressed	1	1	1
suppressed	0.28 [0.16;0.47]	0.67 [0.59;0.76]	0.60 [0.55;0.66]
Sex			
Female	1	1	1
Male	1.39 [1.13;1.70]	1.21 [1.11;1.33]	1.19 [1.08;1.30]
Age			
<25	1	1	1
25-35	0.92 [0.61;1.36]	0.99 [0.83;1.19]	1.03 [0.85;1.23]
35-45	1.07 [0.72;1.61]	1.19 [0.91;1.32]	1.15 [0.95;1.39]
>45	1.31 [0.85;2.03]	1.45 [1.19;1.76]	1.43 [1.17;1.75]
Year			
before 2004	1	1	1
2004-2006	0.83 [0.58;1.19]	1.20 [0.99;1.45]	1.11 [0.91;1.36]
2007 and after	0.60 [0.38;0.96]	1.08 [0.86;1.35]	0.96 [0.76;1.20]

eTable 5: Mortality in South African patients after starting antiretroviral treatment. Estimates from a Cox regression model, reported as hazard ratios, if the variables and categorizations from the predictive model of May et al.[2] (developed in 3 sub-Saharan countries) are used. The analysis is restricted to one year on antiretroviral treatment. 95% confidence intervals are reported in brackets. All results relate to the data from the illustrative example and should not be interpreted causally.

	Complete Cases	**Multiple Imputation**	**Multiple Overimputation**	**May et al.[2]**
Baseline CD4				
<25	1	1	1	1
25-50	0.67 [0.51;0.88]	0.76 [0.68;0.85]	0.61 [0.55;0.67]	0.76 [0.62; 0.94]
50-100	0.39 [0.29;0.52]	0.51 [0.46;0.57]	0.43 [0.39;0.48]	0.46 [0.38; 0.57]
100-200	0.34 [0.26;0.43]	0.34 [0.30;0.38]	0.31 [0.27;0.35]	0.35 [0.28; 0.42]
>200	0.34 [0.22;0.52]	0.32 [0.26;0.38]	0.20 [0.17;0.24]	0.29 [0.22; 0.38]
Sex				
Male	1	1	1	1
Female	0.8 [0.66;0.98]	0.71 [0.66;0.77]	0.71 [0.66;0.77]	0.68 [0.58; 0.79]
Weight (in kg)				
<45	1	1	1	1
45-50	0.65 [0.49;0.87]	0.69 [0.61;0.78]	0.68 [0.61;0.77]	0.59 [0.48; 0.72]
50-60	0.33 [0.26;0.43]	0.41 [0.37;0.46]	0.39 [0.35;0.44]	0.40 [0.33; 0.48]
>60	0.24 [0.18;0.32]	0.30 [0.27;0.34]	0.29 [0.26;0.33]	0.24 [0.19; 0.30]
WHO stage				
I and II	1	1	1	1
III and IV	2.49 [1.58;3.92]	1.61 [1.44;1.81]	1.69 [1.51;1.9]	2.72 [1.87; 3.95]
Age (in years)				
<40	1	1	1	
>40	1.07 [0.86;1.33]	1.23 [1.13;1.34]	1.21 [1.11;1.32]	1.43 [1.23; 1.66]

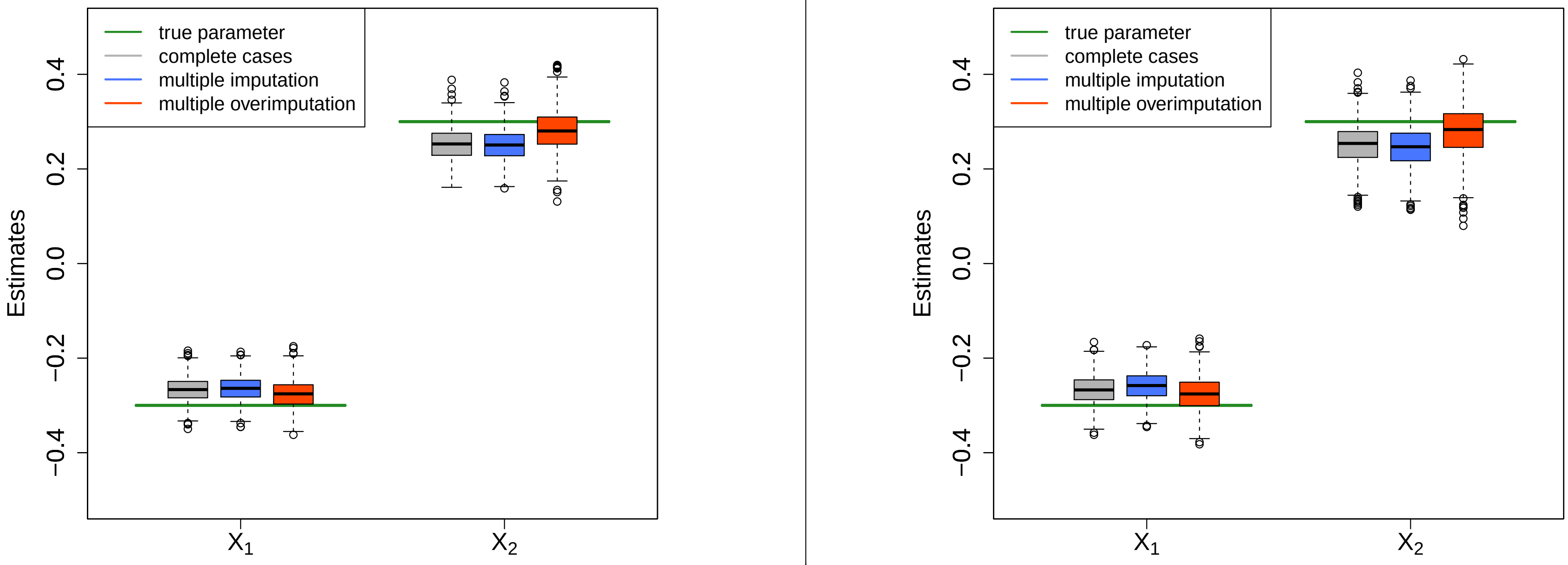
[2] May M, Boulle A, Phiri S, et al. Prognosis of patients with HIV-1 infection starting therapy in sub-Saharan Africa: a collaborative analysis of scale-up programmes. Lancet. 2010;376:449-457.

eFigure 1: Results of the Cox regression analysis when using different assumptions about the measurement error variance: (a) $\sigma^2_{U_{ij}}$*=0.20*^{2} *for CD4 count and* $\sigma^2_{U_{ij}}$*=0.15*^{2}*for log viral load (b)* $\sigma^2_{U_{ij}}$*=0.30*^{2} *for CD4 count and* $\sigma^2_{U_{ij}}$*=0.31*^{2} *for log viral load.*



eFigure 2: Results of variations of the simulation study. The settings specified in the captions refer to changes compared to the main setting in the manuscript.

Variation of missingness assumption



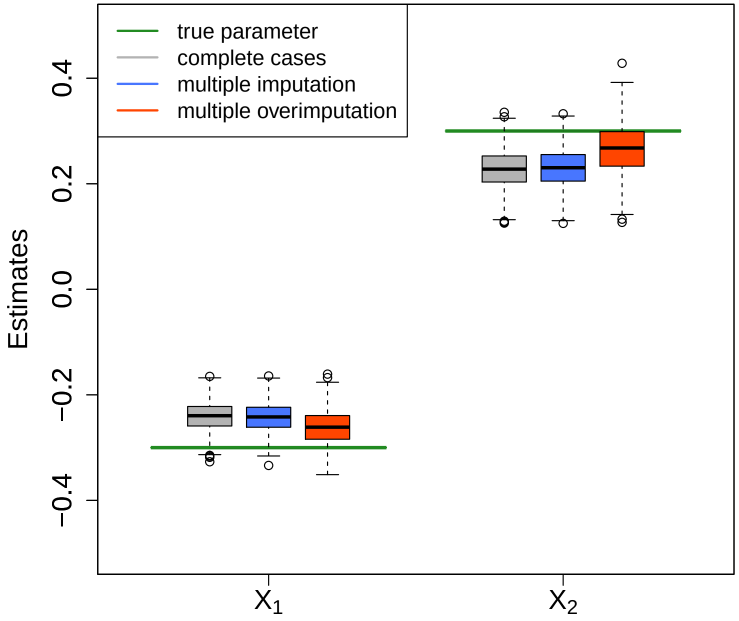
(a) Data missing completely at random, with 10% missingness for both X_1 and X_2

	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.033	-0.048	0.0018	0.0034
Multiple imputation	0.035	-0.049	0.0019	0.0036
Multiple overimputation	0.024	-0.019	0.0014	0.0023

(b) Data missing completely at random, with 40% missingness for both X_1 and X_2

	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.033	-0.048	0.0020	0.0040
Multiple imputation	0.042	-0.054	0.0028	0.0047
Multiple overimputation	0.024	-0.018	0.0019	0.0033

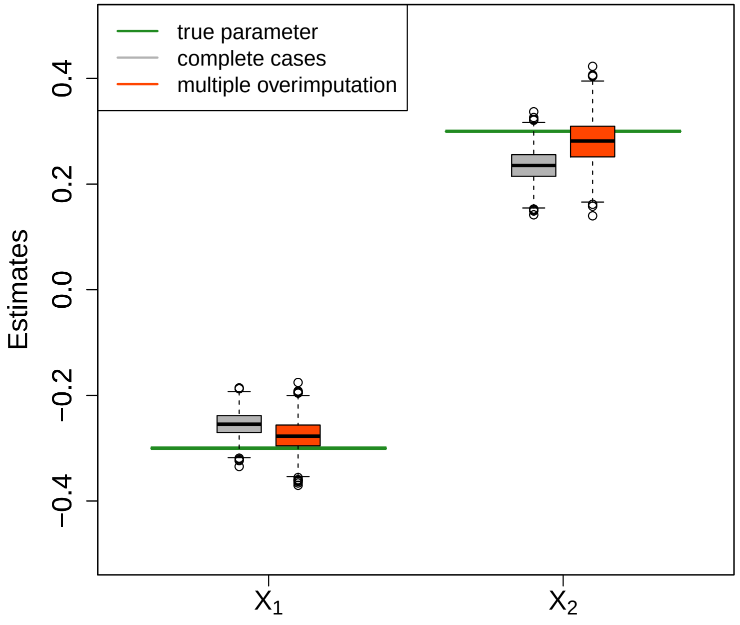
(c) Data missing at random, but with higher missingness probability, ca. 20% for both X1 and X2, defined via $\pi_X(T) = 1 - \{0.02T^2 + 1\}^{-1}$



	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.060	-0.073	0.0043	0.0065
Multiple imputation	0.058	-0.069	0.0041	0.0062
Multiple overimputation	0.038	-0.032	0.0025	0.0033

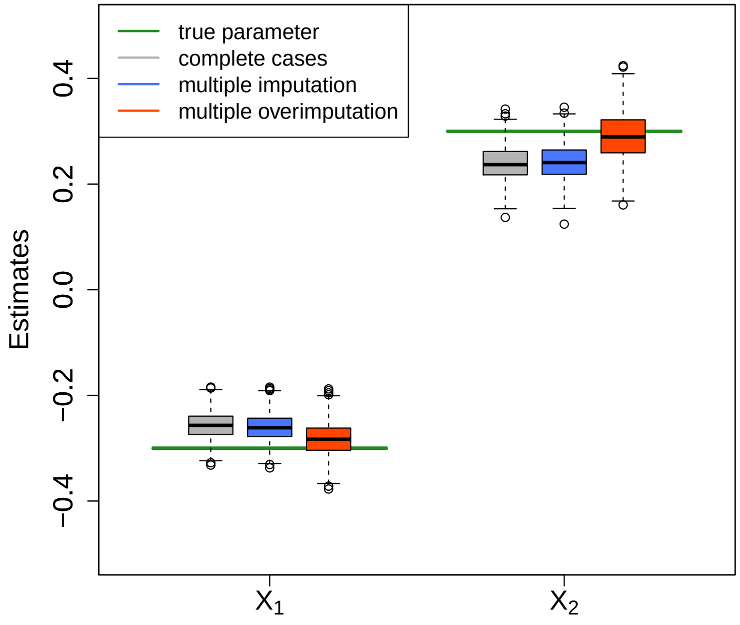
Variation of the measurement error variance

(d) Larger measurement error variance, 0.3^2 and 0.31^2 for both X_1 and X_2 respectively, no missing data



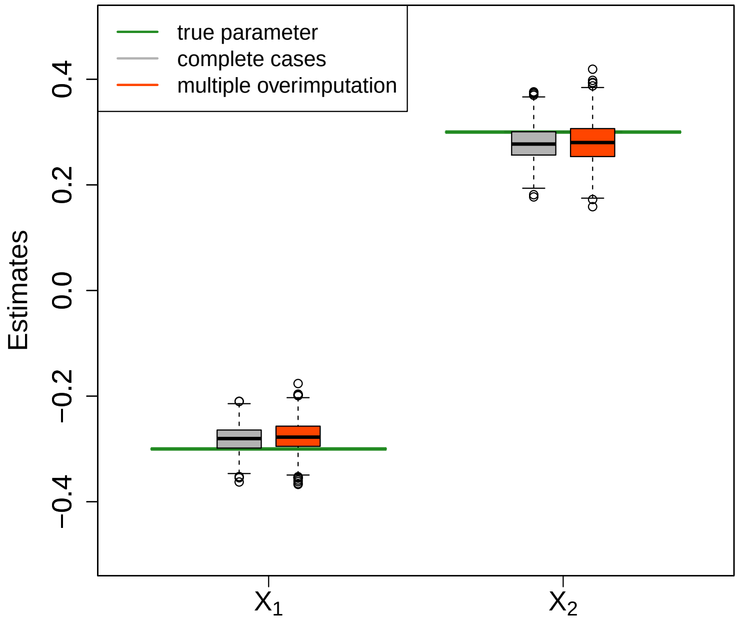
	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.045	-0.065	0.0026	0.0052
Multiple imputation	–	–	–	–
Multiple overimputation	0.023	-0.019	0.0015	0.0022

(e) Larger measurement error variance, 0.3^2 and 0.31^2 for both X_1 and X_2 respectively, with missing data



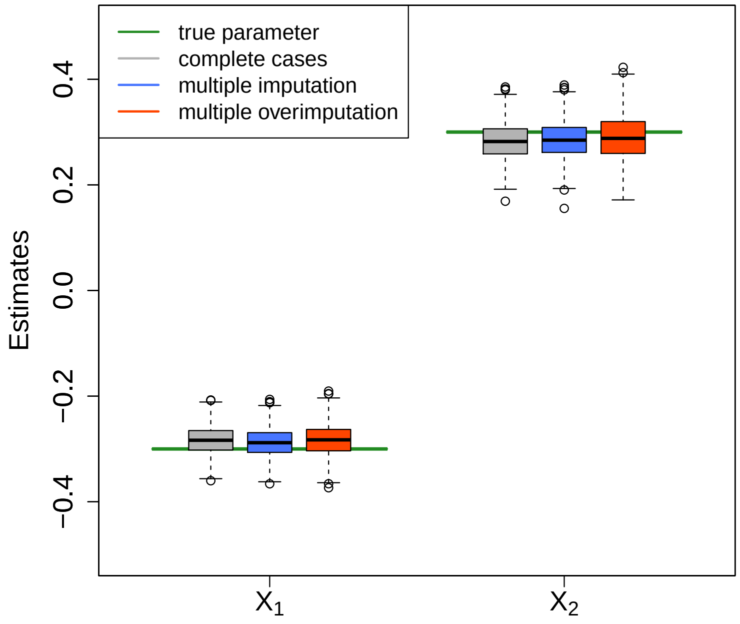
	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.043	-0.062	0.0025	0.0048
Multiple imputation	0.040	-0.059	0.0022	0.0045
Multiple overimputation	0.017	-0.009	0.0013	0.0022

(f) Smaller measurement error variance, 0.2^2 and 0.15^2 for both X_1 and X_2 respectively, no missing data



	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.019	-0.022	0.0010	0.0016
Multiple imputation	–	–	–	–
Multiple overimputation	0.023	-0.019	0.0014	0.0019

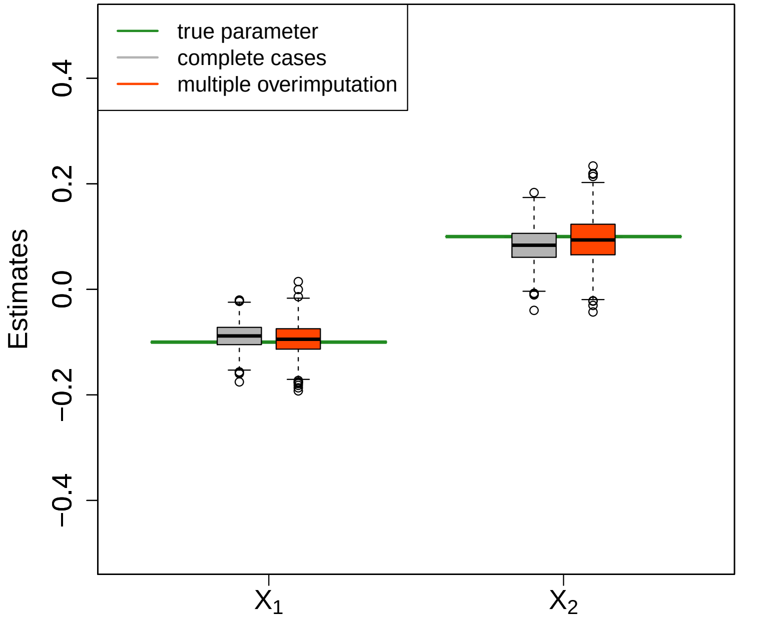
(g) Smaller measurement error variance, 0.2^2 and 0.15^2 for both X_1 and X_2 respectively, with missing data



	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.017	-0.018	0.0009	0.0015
Multiple imputation	0.012	-0.015	0.0009	0.0015
Multiple overimputation	0.017	-0.009	0.0012	0.0019

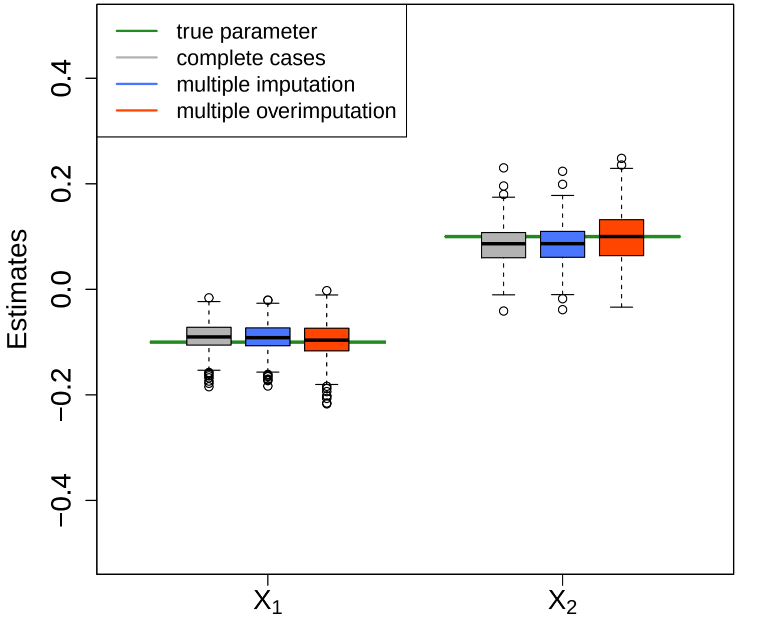
Variation of the linear predictor

(h) The linear predictor $X\beta$ is defined as $-0.1 \ln X_1 + 0.1 \log_{10} X_2$, no missing data



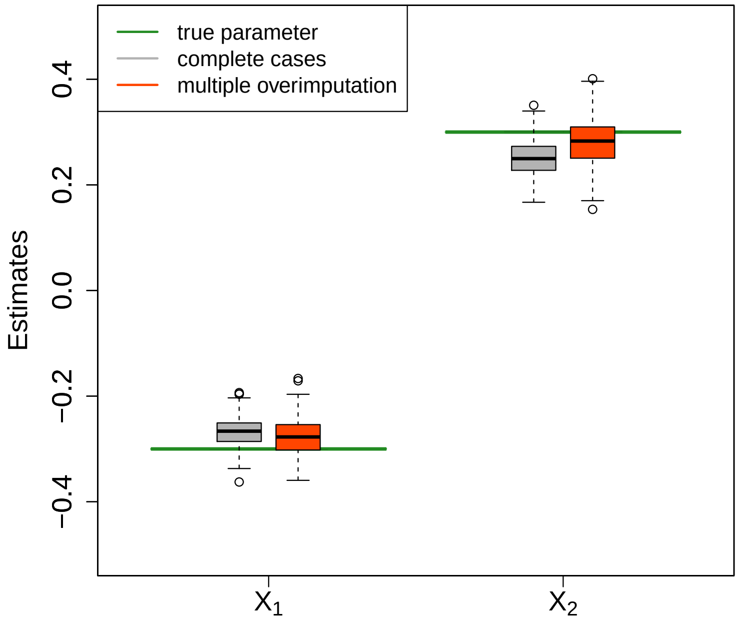
	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.012	-0.017	0.0007	0.0014
Multiple imputation	–	–	–	–
Multiple overimputation	0.005	-0.005	0.0009	0.0018

(i) The linear predictor $X\beta$ is defined as $-0.1 \ln X_1 + 0.1 \log_{10} X_2$, with missing data



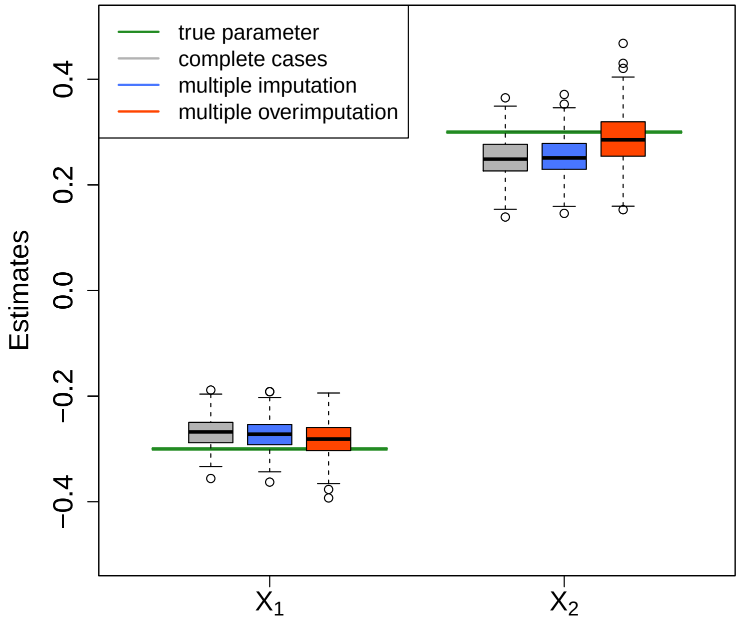
	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.011	-0.016	0.0008	0.0015
Multiple imputation	0.010	-0.015	0.0008	0.0015
Multiple overimputation	0.004	-0.001	0.0011	0.0023

(j) Additional 4 covariates: $X_3 \sim Binom(0.65)$, $X_4 \sim Weibull(1.75, 1.9)$, $X_5 \sim Exp(1)$, $X_6 \sim Gamma(0.25, 2)$, $\beta = (-0.3, 0.3, 0, 0, 0, 0)$, no missing data



	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.033	-0.050	0.0018	0.0035
Multiple imputation	–	–	–	–
Multiple overimputation	0.023	-0.018	0.0015	0.0023

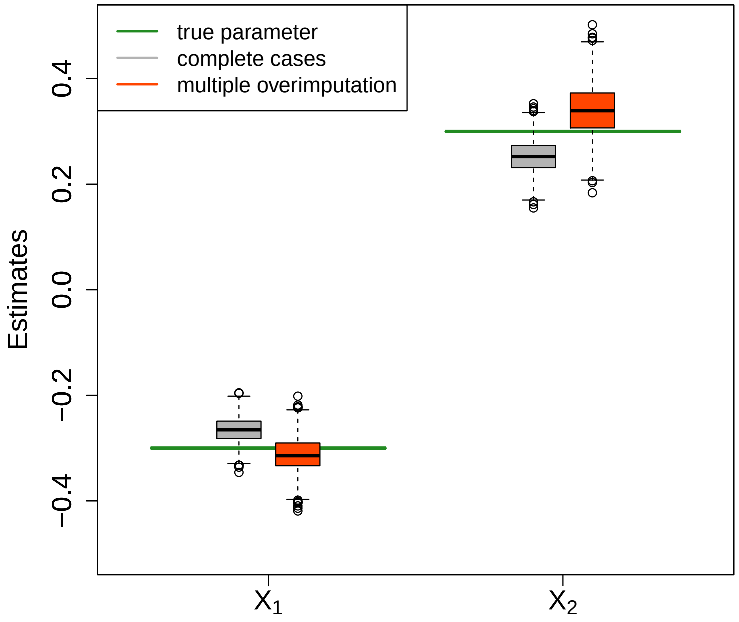
(k) Additional 4 covariates: $X_3 \sim Binom(0.65)$, $X_4 \sim Weibull(1.75, 1.9)$, $X_5 \sim Exp(1)$, $X_6 \sim Gamma(0.25, 2)$, $\beta = (-0.3, 0.3, 0, 0, 0, 0)$, with missing data



	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.031	-0.049	0.0018	0.0037
Multiple imputation	0.027	-0.047	0.0016	0.0035
Multiple overimputation	0.018	-0.014	0.0015	0.0027

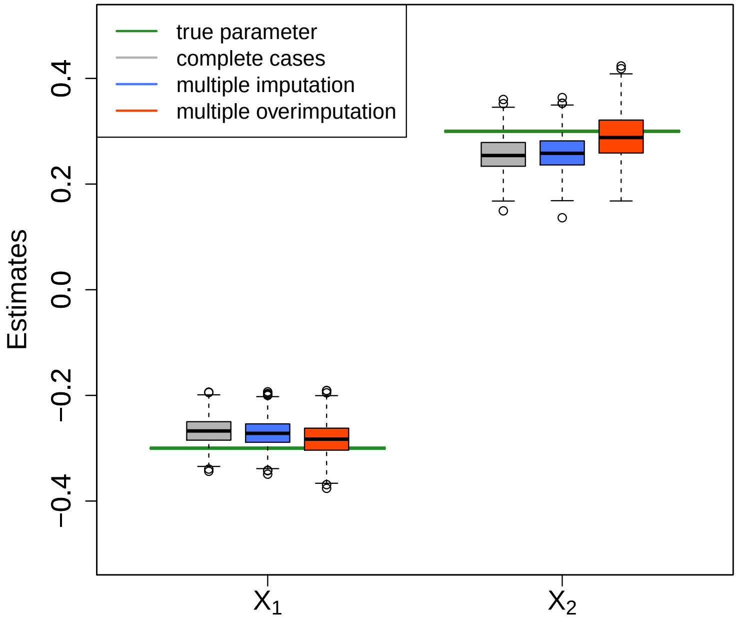
Wrong assumption for measurement error variance used

(l) Wrong assumption used for measurement error variance: 0.36^2 and 0.355^2 for X_1 and X_2 respectively, no missing data



	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.034	-0.048	0.0018	0.0033
Multiple imputation	–	–	–	–
Multiple overimputation	-0.013	0.040	0.0012	0.0039

(m) Wrong assumption used for measurement error variance: 0.36^2 and 0.355^2 for X_1 and X_2 respectively, with missing data



	Bias β_1	Bias β_2	MSE β_1	MSE β_2
Complete cases	0.033	-0.045	0.0017	0.0030
Multiple imputation	0.029	-0.042	0.0015	0.0028
Multiple overimputation	0.017	-0.009	0.0012	0.0021

eText 1: Outline of the technical background of multiple overimputation. More details can be found in both the main body and the appendices of Blackwell et al. (2014) and Honaker and King (2010). For a better understanding of the technical details the reader may also wish to consult Rubin (1996) for more insight on multiple imputation, King et al. (2001) for useful technicalities of an algorithm similar to EMB, Dempster et al. (1977) for the EM algorithm, and Goodnight (1979) for the details of the sweep operator.

1. *Data and notation:* Consider a data set $\mathbf{X}$ consisting of observations $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$. Let e_{ij} be an indicator whether x_{ij} was measured with error and m_{ij} be an indicator if x_{ij} is missing. The data may consist of perfectly measured values $\mathbf{x}_i^{obs} = \{x_{ij}; e_{ij} = m_{ij} = 0\}$, values which are missing, $\mathbf{x}_i^{mis} = \{x_{ij}; m_{ij} = 1\}$, and values measured with error (w_{ij}) as a proxy to the latent 'true' unobserved values x_{ij}^{err}, $\mathbf{x}_i^{err} = \{x_{ij}; e_{ij} = 1\}$, $\mathbf{w}_i = \{w_{ij}; e_{ij} = 1\}$. Thus, the observed data for any observation is $\mathbf{d}_i = (\mathbf{x}_i^{obs}, \mathbf{w}_i)$ while the true underlying data is $\mathbf{x}_i = (\mathbf{x}_i^{obs}, \mathbf{x}_i^{err}, \mathbf{x}_i^{mis})$.

2. *Observed data probability density function:* The probability density function for the *observed* data equates to

$$p(\mathbf{d}_i, \mathbf{m}_i, \mathbf{e}_i|\theta, \gamma, \phi) \quad = \quad \int\!\!\int p(\mathbf{x}_i|\theta)p(\mathbf{w}_i|\mathbf{x}_i, \gamma)p(\mathbf{m}_i, \mathbf{e}_i|\mathbf{d}_i, \mathbf{x}_i, \phi)\, d\mathbf{x}_i^{err} d\mathbf{x}_i^{mis} \tag{1}$$

whereby θ refers to the parameterization of the true underlying data, γ to the error distribution, and ϕ to the joint distribution of $\mathbf{m}_i$ and $\mathbf{e}_i$. Using the mismeasured at random (MMAR) assumption, which is $p(\mathbf{m}_i, \mathbf{e}_i|\mathbf{d}_i, \mathbf{x}_i, \phi) = p(\mathbf{m}_i, \mathbf{e}_i|\mathbf{d}_i, \phi)$, (1) can be written as

$$p(\mathbf{d}_i, \mathbf{m}_i, \mathbf{e}_i|\theta, \gamma, \phi) \quad = \quad p(\mathbf{m}_i, \mathbf{e}_i|\mathbf{d}_i, \phi)p(\mathbf{d}_i|\theta, \gamma) \tag{2}$$

which is proportional to

$$p(\mathbf{d}_i|\theta, \gamma) \quad = \quad \int\!\!\int p(\mathbf{x}_i|\theta)p(\mathbf{w}_i|\mathbf{x}_i, \gamma)\, d\mathbf{x}_i^{err} d\mathbf{x}_i^{mis}\,. \tag{3}$$

Note that from a Bayesian perspective, for a given prior on (θ, γ), this gives us a posterior distribution for $p(\theta, \gamma|\mathbf{d}_i)$ which makes use of only observed quantities.

3. *Posterior predictive distribution of the unobserved data:* To obtain valid inference with multiple imputation (MI), one needs to draw from the posterior predictive distribution of the unobserved data. If one were to omit mismeasured data and thus define $x_{ij}^{err} = x_{ij}^{mis}$ MI would already yield valid inference but omit important information. Given that both missing *and* latent values are unobserved, draws from the predictive posterior distribution of this unobserved data relate to:

$$p(\mathbf{x}_i^{err}, \mathbf{x}_i^{mis}) = \int p(\mathbf{x}_i^{err}, \mathbf{x}_i^{mis}|\mathbf{d}_i, \theta, \gamma)p(\theta, \gamma|\mathbf{d}_i)d\theta d\gamma\,. \tag{4}$$

4. *Multiple imputation with EMB:* To draw values from (4) one needs to (i) draw $(\theta_{(i)}, \gamma_{(i)})$ from its posterior distribution $p(\theta, \gamma|\mathbf{d}_i)$ and then (ii) draw $(\mathbf{x}_i^{err}, \mathbf{x}_i^{mis})$ from $p(\mathbf{x}_i^{err}, \mathbf{x}_i^{mis}|\mathbf{d}_i, \theta, \gamma)$. The EMB algorithm utilizes this (i) by means of the EM algorithm to obtain an unbiased estimate $\hat{\theta}$ in the presence of unobserved data, and (ii) by repeating this for different bootstrap samples of $\mathbf{d}$. Specifically, under the assumption of a multivariate normal distribution for the data, $\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$, and under the assumption of no measurement error, EMB does the following:

   i) The Expectation-Maximization (EM) algorithm estimates $\theta = (\mu, \mathbf{\Sigma})$ in the presence of unobserved data. In the E(xpectation)-Step the algorithm fills in estimates for the missing values using conditional expectations; in the M(aximazation)-Step the complete data parameters are estimated (from the sufficient statistics) using the available and filled-in data. These two steps are repeated until the parameter estimates converge and one obtains $(\hat{\mu}, \hat{\mathbf{\Sigma}})$, see Dempster et al. (1977) for the technical details. Thus, an estimate of θ can be drawn from $N(\hat{\mu}, \hat{\mathbf{\Sigma}})$. This step simulates draws from $p(\theta, \gamma|\mathbf{d}_i)$ related to (3).

   ii) The draws from $N(\hat{\mu}, \hat{\mathbf{\Sigma}})$ are used to obtain $\tilde{\beta}$ (an estimate of β) via the sweep operator and impute missing values via $x_{ij} = x_{i,-j}^{obs}\tilde{\beta} + \tilde{\epsilon}$. We refer the reader to Goodnight (1979) and the appendix of Honaker and King (2010) for the details on how $\hat{\mu}$ and $\hat{\mathbf{\Sigma}}$ relate to $\tilde{\beta}$.

iiii) Repeating this procedure for M bootstrap samples of $\mathbf{d}$ yields M different imputations adequately reflecting estimation uncertainty. They can be seen as draws from (4) for $\mathbf{x}_i^{err} = \emptyset$.

5) *Incorporating measurement error into EMB via prior distributions:* To simulate proper draws from (4) one needs to first simulate proper draws from the posterior relating to (3). Blackwell et al. (2014) show that under the setting of (3) the EM-algorithm needs to estimate

$$E(T(\mathbf{x}_i)|\mathbf{d}_i,\theta^{(t)},\gamma) \quad = \quad \int\int T(\mathbf{x}_i)\,\underbrace{p(\mathbf{x}_i^{err},\mathbf{x}_i^{mis}|\mathbf{x}_i^{obs},\theta^{(t)})}_{\text{imputation}}\ \underbrace{p(\mathbf{w}_i|\mathbf{x}_i,\gamma)}_{\text{mismeasurement}}\ d\mathbf{x}_i^{err}d\mathbf{x}_i^{mis} \tag{5}$$

in the E-Step. Note that $\theta^{(t)}$ refers to the t^{th} updated estimate of θ and $T(\mathbf{x}_i)$ to the complete data sufficient statistic (from which θ can be derived; under multivariate normality $T(\mathbf{x}) = \mathbf{X}'\mathbf{X}$). Now, if we assume a *classical measurement error* model we implicitly specify

$$w_{ij} \quad \sim \quad N(x_{ij},\lambda_{ij}^2) \qquad \forall w_{ij} \in \mathbf{w}_i\,. \tag{6}$$

Putting (6) into (5) and using the normality assumptions $\mathbf{x}_i^{err}|\mathbf{x}_i^{obs},\theta \sim N(\mu_{e|o},\mathbf{\Sigma}_{e|o})$, $\mathbf{w}_i|\mathbf{x}_i^{err},\lambda_i^2 \sim N(\mathbf{x}_i^{err},\mathbf{\Lambda}_i)$ yields the following distribution

$$(\mathbf{x}_i^{err}|\mathbf{d}_i,\theta^{(t)},\lambda_i^2) \sim N(\mu^*,\mathbf{\Sigma}^*) \quad \text{with} \quad \mathbf{\Sigma}^* = (\mathbf{\Lambda}_i^{-1}+\mathbf{\Sigma}_{e|o}^{-1})^{-1},\ \mu^* = \mathbf{\Sigma}^*(\mathbf{\Lambda}_i^{-1}\mathbf{w}_i + \mathbf{\Sigma}_{e|o}^{-1}\mu_{e|o})\,, \tag{7}$$

as demonstrated by Blackwell et al. (2014). Thus, to calculate the expectation on the left hand side of (5) for each cell with error, the E-Step needs simply make use of (7). This will result in overall proper multiple overimputations drawn from (4).

6) *Implications for the Implementation with `Amelia II`:* The standard EMB algorithm is implemented in the *R*-package `Amelia II` (Honaker et al., 2011). It allows the incorporation of prior distribution for single cells, i.e $x_{ij} \sim N(\mu_{ij,0},\kappa_{ij,0}^2)$. If using $\mu_{ij,0} = w_{ij}$ and $\kappa_{ij,0}^2 = \lambda_{ij}^2$ one obtains the same results as in (7), see the appendix of Honaker and King (2010); and thus, using priors for mismeasured cells which equal $x_{ij} \sim N(w_{ij},\lambda_{ij}^2)$ yields draws from the modified EMB algorithm described in step 5, and therefore *proper* multiple overimputations. To specify the mismeasured cells one needs the *overimp* option of the function `amelia`, and to specify the priors for the respective cells one needs the *priors* option.

7) *Combining estimates after multiple overimputation:* After generating M overimputed datasets by means of multiple overimputation, the analysis model (e.g. any regression model) can be fitted on each overimputed dataset and the M results will be combined as follows: the point estimate of θ (here implicitly referring to the parameters in the analysis model) is

$$\hat{\theta}_{\text{MI}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^{(m)} \tag{8}$$

where $\hat{\theta}^{(m)}$ refers to the estimate of θ in the mth overimputed set of data $\mathcal{D}^{(m)}$, $m = 1,\ldots,M$. Based on the average within imputation covariance $\widehat{\mathbf{W}} = M^{-1}\sum_m \widehat{\text{Cov}}(\hat{\theta}^{(m)})$ and the between imputation covariance $\hat{\mathbf{B}} = (M-1)^{-1}\sum_m(\hat{\theta}^{(m)} - \hat{\theta}_{\text{MI}})(\hat{\theta}^{(m)} - \hat{\theta}_{\text{MI}})'$ one obtains variance estimates via

$$\widehat{\text{Cov}}(\hat{\theta}_{\text{MI}}) = \widehat{\mathbf{W}} + \frac{M+1}{M}\hat{\mathbf{B}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\text{Cov}}(\hat{\theta}^{(m)}) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\theta}^{(m)} - \hat{\theta}_{\text{MI}})(\hat{\theta}^{(m)} - \hat{\theta}_{\text{MI}})' \tag{9}$$

To construct confidence intervals for $\hat{\theta}_{\text{MI}}$ in the scalar case, it may be assumed that $\widehat{\text{Var}}(\hat{\theta}_{\text{MI}})^{-\frac{1}{2}}(\hat{\theta}_{\text{MI}} - \theta)$ follows a t_R-distribution with approximately $R = (M-1)[1 + \{M\hat{W}/(M+1)\hat{B}\}]^2$ degrees of freedom. Instead of computing these quantities by hand the function `mi.inference` in the *R*-package `norm` can be used; or, alternatively, the functionalities in the *R*-package `Zelig` or the Stata commands `mi estimate` or `mim` can be used.

eText 2: Simulation of viral loads and the misclassification proportion related to measurement error.

R-code:

```
# Generating true data
n=30000
VL            <- rlnorm(n, meanlog=10.760, sdlog=1.808607)

# Generating mismeasured data
VL_measured   <- 10^(log10(VL)+rnorm(n,0,0.255))

# Virological suppression if VL<1000
VL_supp            <- as.numeric(VL<1000)
VL_measured_supp   <- as.numeric(VL_measured<1000)

# Evaluating misclassification
VL_total       <- cbind(VL_supp,VL_measured_supp)
misclass_FN    <- as.numeric(VL_total[,1]==1 &  VL_total[,2]==0)
misclass_FP    <- as.numeric(VL_total[,1]==0 &  VL_total[,2]==1)
mean(misclass_FN)+mean(misclass_FP)  # overall misclassification
```

With $z_{0.985} = 2.17$ and a standard deviation of 0.23 one obtains about 1.5% misclassification by evaluating the confidence intervals related to the prior distributions used during multiple overimputation for the mismeasured values. For example, if a patient had a virological failure ($\mathrm{VL}_{\mathrm{supp}} = 0$) we impose a prior normal distribution, $N(0, 0.23)$, on the mismeasured value which implies that the upper limit of a 98.5% confidence interval corresponds to $0+2.17\cdot0.23 = 0.499$ and thus 1.5% of values from this normal distribution are > 0.499 and therefore get rounded off to 1 which relates to virological failure and thus misclassification. The motivation and more details on how to use prior normal distributions for categorical variables can be found in Section 3.2 of Blackwell et al. (2014).

References

Blackwell, M., Honaker, J., King, G., 2014. Multiple overimputation: A unified approach to measurement error and missing data. Sociological Methods and Research, in press, copy at http://j.mp/jqdj72.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39, 1–38.

Goodnight, J. H., 1979. Tutorial on the sweep operator. American Statistician 33 (3), 149–158.

Honaker, J., King, G., 2010. What to do about missing values in time-series cross-section data? American Journal of Political Science 54, 561–581.

Honaker, J., King, G., Blackwell, M., 2011. Amelia II: A program for missing data. Journal of Statistical Software 45 (7), 1–47.

King, G., Honaker, J., Joseph, A., Scheve, K., 2001. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. American Political Science Review 95, 49–69.

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