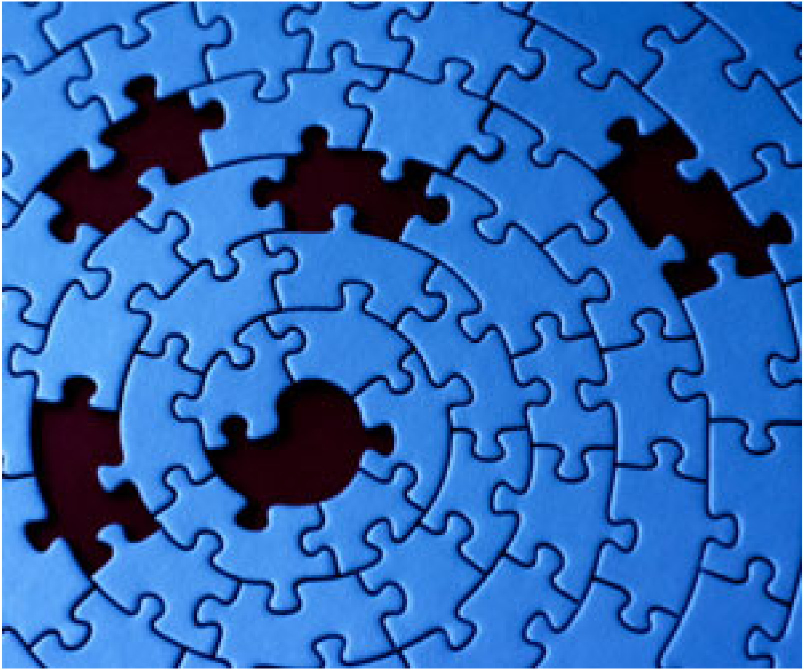
Multiple imputation as a type of stochastic EM approximation to maximum likelihood

Firouzeh Noghrehchi
Prof. David Warton, Dr. Jakub Stoklosa
School of Mathematics and Statistics,
The University of New South Wales,
Sydney, Australia.



Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

Motivation

Two popular missing data analysis methods, treated as distinct in the literature:

- Maximum likelihood estimation (MLE)
  - ▷ via expectation-maximisation (EM) algorithm, proposed by Dempster et al. in late 1970's
  - ▷ via stochastic versions of EM, developed in mid 1980's and early 1990's
- Multiple imputation (MI)
  - ▷ proposed by Rubin in late 1970's

However, close relationship between MLE and MI

$\Longrightarrow$ **A type of MI is exactly MLE!**

Motivation (II)

Aim is to explore ideas from ML literature that can be applied to MI in order to, for example,

- choose variables to be included in the imputation model
- gain insight into consequences of misspecification
- ...

Outline

- Motivation
- Missing data analysis methods
    - Multiple imputation (MI)
    - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
    - Methods for imputation model selection
    - Simulation study
    - Efficiency
- Conclusion

Missing data problem

- Notation
  - observed data y
  - missing data z
  - parameters of model θ
- Observed likelihood

$$p(y \mid \theta) = \int p(y, z \mid \theta) dz$$

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

Multiple imputation

What does MI do?

- impute missing data $M \geq 2$ times, creating M completed datasets
- analyse each M completed datasets separately
- combine the results together over M completed datasets (Rubin, 1987)

Multiple imputation (II)

How does MI impute missing data?

1. assumes a complete-data model $p(y, z \mid \theta)$
2. imputes missing data from imputation model $p(z \mid y, \theta)$ to complete dataset
3. estimates θ from the completed dataset
4. repeats steps 2-3 M times
5. combines the results

Multiple imputation (III)

Most commonly in a Bayesian manner (Tanner and Wong, 1987):

- approximation to the observed posterior

$$p(\theta \mid y) = \int p(\theta \mid y, z) p(z \mid y) dz$$
$$\simeq \frac{1}{M} \sum_{j=1}^{M} p(\theta \mid y, z^{(j)})$$

- in an iterative manner in two steps
  - I-step: imputation of missing data by randomly drawing from imputation model
  - P-step: re-estimation of parameters by randomly drawing from their postrior distribution given the completed data

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

MLE via EM algorithm

Finds MLE of parameters of observed likelihood in the presence of missing data by making use of an associated complete-data likelihood; EM iteration $\theta^{(t)} \rightarrow \theta^{(t+1)}$ consists of two steps:

- E-step

$$Q(\theta \mid \theta^{(t)}) = \int \log\left[p(y, z \mid \theta)\right] p(z \mid y, \theta^{(t)}) dz$$

- M-step

$$\theta^{(t+1)} = argmax \;\; Q(\theta \mid \theta^{(t)})$$

Stochastic versions of EM

Approximate $Q(\theta \mid \theta^{(t)})$ using Monte Carlo integration:

- E-step $\rightarrow$ I-step:
  impute missing data from imputation model

  $$z^{(j)} \sim p(z \mid y, \theta^{(t)}), \quad j = 1, \ldots, M$$

  to approximate $Q(\theta \mid \theta^{(t)})$ as

  $$Q(\theta \mid \theta^{(t)}) \simeq \frac{1}{M} \sum_{j=1}^{M} \log p(y, z^{(j)} \mid \theta^{(t)})$$

Stochastic versions of EM (II)

A popular stochastic version of EM

- stochastic EM (StEM; Celeux and Diebolt 1985): set $M = 1$ and iterate until convergence to stationary distribution $\Psi(\hat{\theta})$ at $t = T$

$$\begin{aligned}\hat{\theta} &= E(\Psi(\hat{\theta})) \\ &= \frac{1}{m}\sum_{j=1}^{m} argmax \left(\log p(y, z^{(j)} \mid \theta^{(T+j)})\right)\end{aligned}$$

- $\{\theta^{(t)}\}$ by StEM algorithm does not converge pointwise to $\hat{\theta}$ but in distribution (Biscarat et al., 1992)

- StEM estimator unbiased and consistent estimator of MLE of θ (Diebolt and Ip, 1995)

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

StEM vs MI

- Artificial distinction between MI and StEM
  $\implies$ A type of MI is equivalent to StEM

StEM:

0. Fix $\theta^{(0)}$ in Θ
1. $z^{(t+1)} \sim p(z \mid y, \theta^{(t)})$
2. $\theta^{(t+1)} = argmax\ p(y, z^{(t+1)} \mid \theta^{(t)})$
3. Repeat 1-2 until convergence
4. Combine results of next M iterations

MI ("*proper*"):

0. Fix $\theta^{(0)}$ in Θ
1. $z^{(t+1)} \sim p(z \mid y, \theta^{(t)})$
2. $\theta^{(t+1)} \sim p(\theta \mid y, z^{(t+1)})$
3. Repeat 1-2 until convergence
4. Combine results of next M iterations

StEM vs MI (II)

StEM or MI (*"improper"*):

0. Fix $\theta^{(0)}$ in Θ
1. $z^{(t+1)} \sim p(z \mid y, \theta^{(t)})$
2. $\theta^{(t+1)} = argmax\ p(y, z^{(t+1)} \mid \theta^{(t)})$
3. Repeat 1-2 until convergence
4. Combine results of next M iterations

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

Methods for imputation model selection

- In MI literature, no standard tool to choose which auxiliary variables to be included in the imputation model
- Available model selection criteria in the ML literature
  - Akaike information criterion (AIC)

    $$AIC = -2 \log p(y \mid \hat{\theta}) + 2d$$

    with d denoting number of parameters
  - Bayesian information criterion (BIC)

    $$BIC = -2 \log p(y \mid \hat{\theta}) + \log(n) \times d$$

  - Other methods specifically developed for missing data problems such as Complete AIC (AICcd) and Mixed AIC (AICmix)

Methods for imputation model selection (II)

Why likelihood can distinguish between imputation models?
$\rightarrow$ incorrect imputation model can be understood as a variational approximation to observed log-likelihood:

- Let $q(z)$ be the specified imputation model, then

$$\begin{aligned}\log p(y \mid \theta) &= \log p(y \mid \theta) \int q(z) dz = \int q(z) \log p(y \mid \theta) dz \\ &= \int q(z) \log \left(\frac{p(y, z \mid \theta)}{p(z \mid y, \theta)} \right) dz \\ &= Q(\theta \mid \theta^{(T)}) - \int q(z) \log p(z \mid y, \theta) dz\end{aligned}$$

- When $q(z) \neq p(z \mid y, \theta)$

$$\log p_q(y \mid \theta) = \log p(y \mid \theta) - KL(q||p)$$

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

Simulation study

Interested in a response variable Y, which is a function of a predictor $X1$

- $X1$ partially observed
- Two imputation models (linear regressions):
  - True model: an auxiliary variable $X2$, together with Y, to impute missing values, where $X1$ and $X2$ are correlated
  - Wrong model: an auxiliary variable $X3$, together with Y, to impute missing values, where $X3$ is independent from $X1$ and Y
- Y and $X2$ conditionally independent given $X1$
- Complete data $(Y, X1, X2, X3) \sim N_4(\mu, \Sigma)$

Simulation result

- Medium correlation (0.5) between $X1$ and $X2$
- $X1$ 60% missing below a limit of detection
- Sample size n varies between $n = 50, 100, 1000$
- Results averaged over 200 simulated datasets

True/Wrong	(n=50)	(n=100)	(n=1000)
AIC	0.80	0.83	1.00
BIC	0.80	0.83	1.00

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

Efficiency gain?

By imputation of missing data more than once in the I-step

- Monte Carlo EM (MCEM; Wei and Tanner 1990):
  set $M \geq 2$ and iterate until convergence to $\hat{\theta}$ at $t = T$

$$\hat{\theta} = argmax \left(Q(\theta \mid \theta^{(T)}) \mid z^{(1)}, ..., z^{(M)} \right)$$
$$= argmax \frac{1}{M} \sum_{j=1}^{M} \left(\log p(y, z^{(j)} \mid \theta^{(T)}) \right)$$

- MCEM more efficient than StEM
  - for finite sample size
  - for finite number of imputations (Nielsen, 2000)
    - StEM loses efficiecy due to maximise-then-average

Outline

- Motivation
- Missing data analysis methods
  - Multiple imputation (MI)
  - MLE via stochastic EM
- MI as stochastic EM
- Gains of equivalence
  - Methods for imputation model selection
  - Simulation study
  - Efficiency
- Conclusion

Conclusion

- A type of MI can be understood as a stochastic version of EM which is an approximation to MLE
- Access to standard likelihood machinery can improve MI's performance:
  - standard ICs for imputation model selection
  - methods developed for assessment of imputation model misspecification
  - efficieny gain

Reference

Celeux,G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comp.Statist.Quart.2, 73-82.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1-38.

Diebolt, J., and Ip, E. H. S. (1995). A stochastic EM algorithm for approximating the maximum likelihood estimate: Sandia National Labs., Livermore, CA (United States).

Nielsen, S. F. (2000). The stochastic EM algorithm: estimation and asymptotic results. Bernoulli, 457-489.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys (Vol. 81): John Wiley and Sons.

Wei, G. C. G., and 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.

Asymptotic variance

Let $W(\hat{\theta})$ and $B(\hat{\theta})$ denote within- and between-imputation variance of $\hat{\theta}$, respectively, and I the identity matrix:

- stochastic EM (Louis 1982, Diebolt and Ip 1995, Wang and Robins 1998, von Hippel 2012; *"Louis method"*)

$$
\begin{aligned}
\hat{var}(\hat{\theta}_{StEM}) &= E_\theta\left[\frac{\partial^2 p(y,z \mid \theta)}{\partial\theta\partial\theta'} \mid y\right] - cov_\theta\left[\frac{\partial p(y,z \mid \theta)}{\partial\theta} \mid y\right] \\
&= W(\hat{\theta})\left[I - W(\hat{\theta})^{-1}B(\hat{\theta})\right]^{-1}
\end{aligned}
$$

- MI (Rubin 1987; *"Rubin's rules"*)

$$
\begin{aligned}
\hat{var}(\hat{\theta}_{MI}) &= W(\hat{\theta}) + B(\hat{\theta}) \\
&= W(\hat{\theta})\left[I - \left(W(\hat{\theta}) + B(\hat{\theta})\right)^{-1} B(\hat{\theta})\right]^{-1}
\end{aligned}
$$