Möglichkeiten der Verbesserung des EM-Algorithmus für normalverteilte Daten

Dipl. Ing. Luis Huergo

Universität Tübingen
Lehrstuhl für Statistik, Ökonometrie und Unternehmensforschung
Prof. Dr. Dr. h.c. mult. Eberhard Schaich
(luis.huergo@uni-tuebingen.de)

08. Juli 2008

Motivation

- Neuartige Indikatoren für die wissensbasierte Wirtschaft (knowledge-based economy) sollten zu einem einzigen zusammengesetzten Indikator aggregiert und seine Eigenschaften untersucht werden.
- Es wurden 116 Indikatoren ausgewählt.
- Kein einziges der 25 teilnehmenden europäischen Länder war in der Lage, alle 116 Indikatoren zu liefern.
- Der fertiggestellte Datensatz hatte 42% fehlende Daten.
- Diese fehlenden Daten mussten imputiert werden.

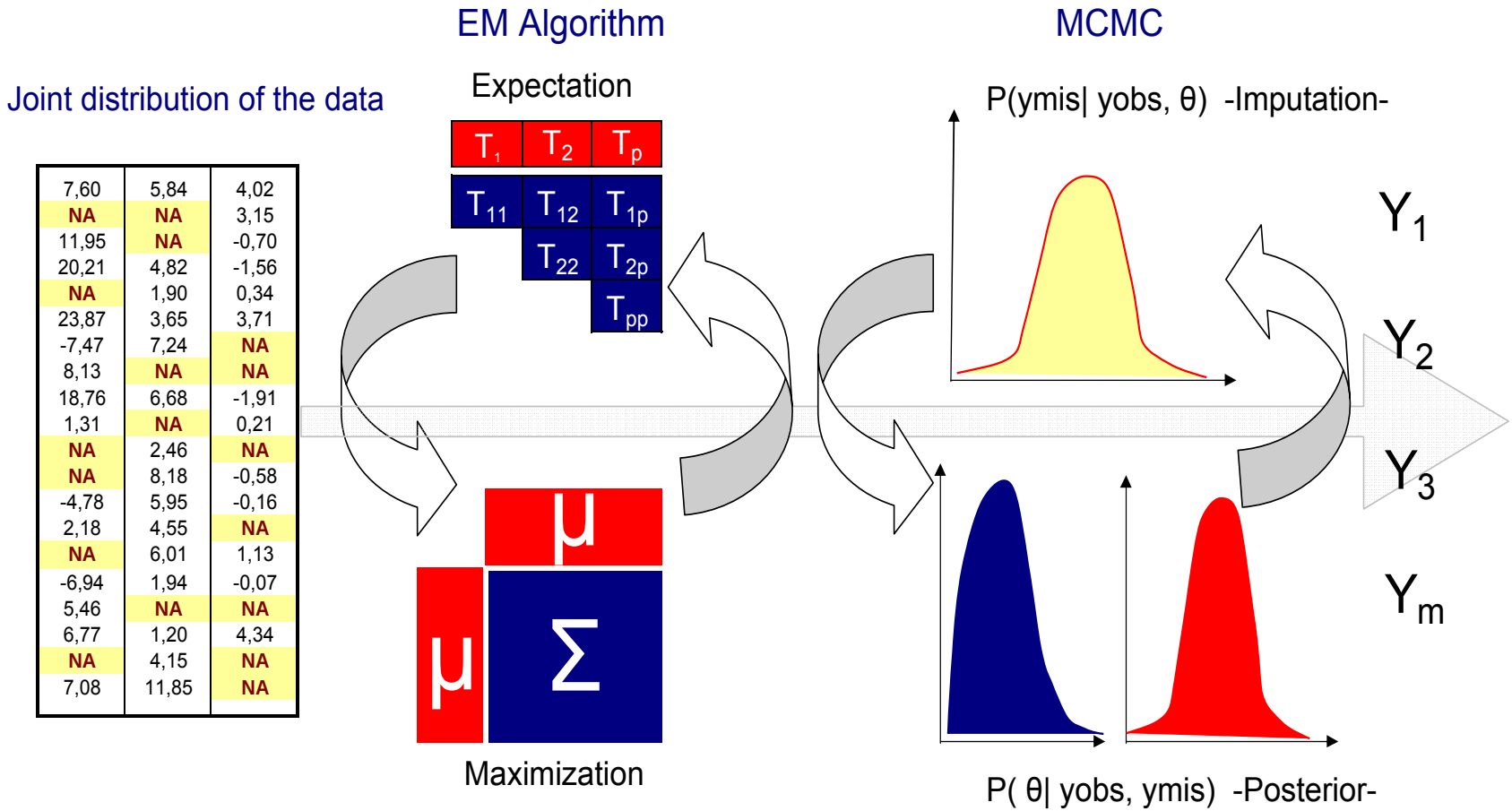
Eberhard Karls
Universität
Tübingen

Imputation

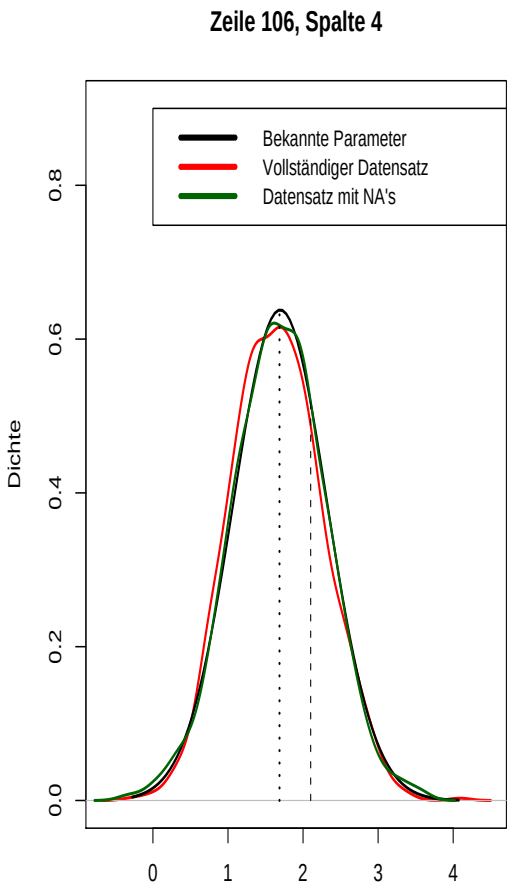
Die (moderne) Imputation eines mit fehlenden Werten behafteten Datensatzes kann mit der Rekonstruktion eines beschädigten Bildes verglichen werden.

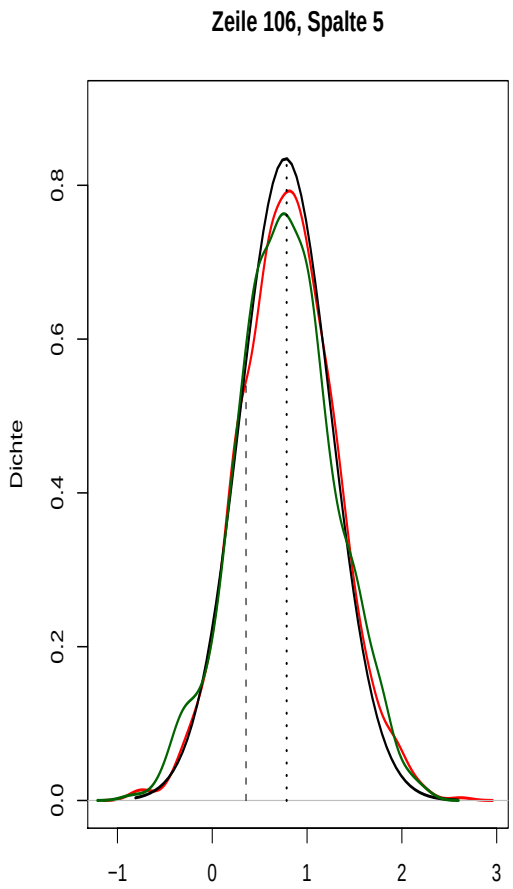
Eberhard Karls
Universität
Tübingen

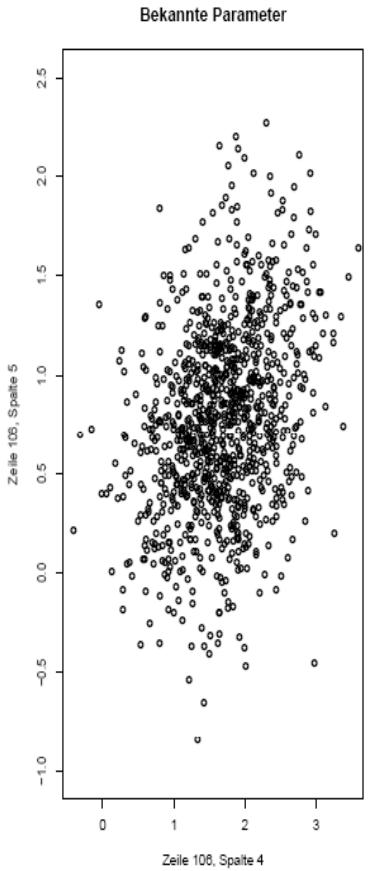
Moderne Imputationsverfahren: EM-Algorithmus und MCMC-Methoden

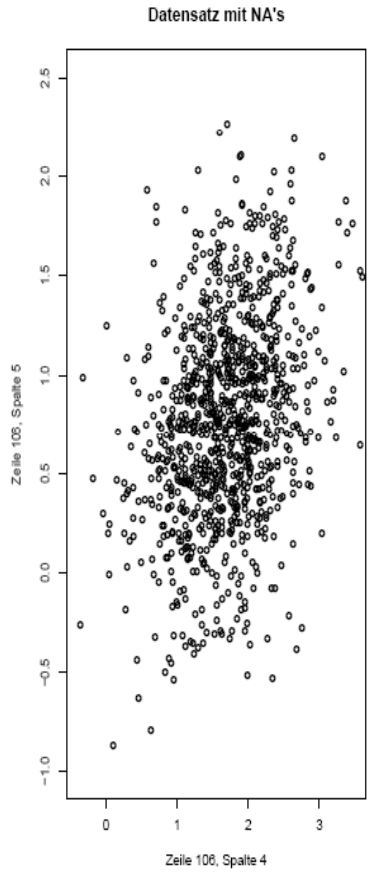


Ergebnisse der Simulation



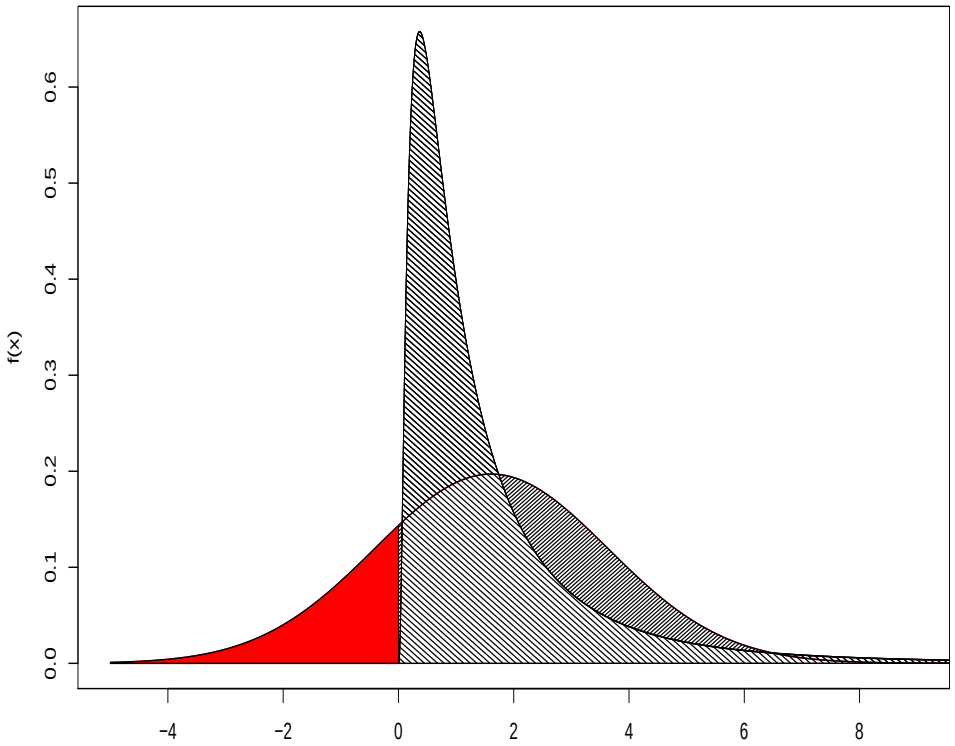






Eberhard Karls
Universität
Tübingen

Probleme bei nicht normalverteilten Daten

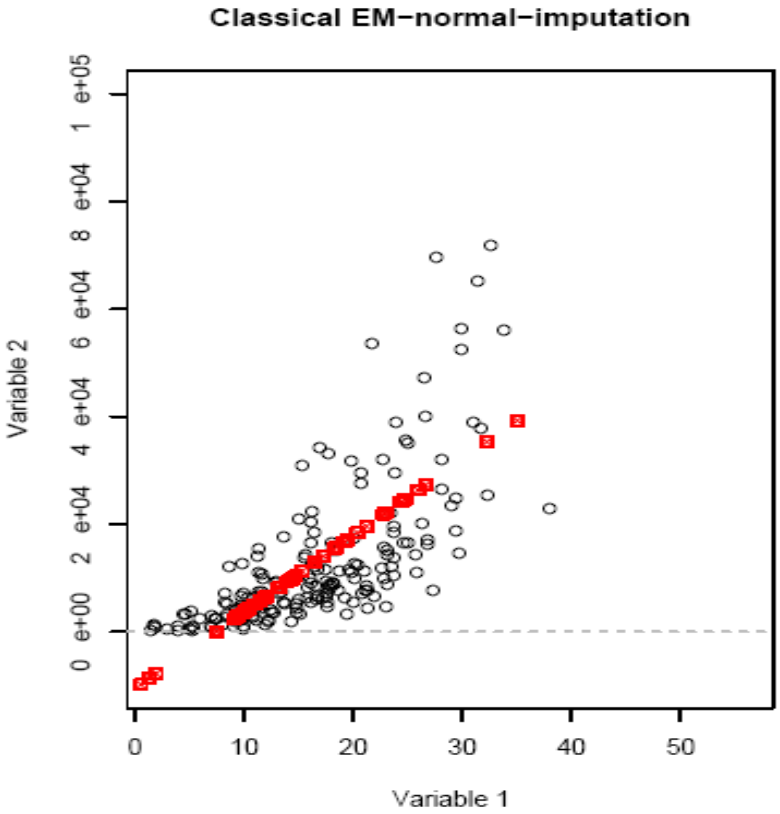


- Beide Verteilungen haben den gleichen Erwartungswert und die gleiche Varianz
- Die Werte der rot markierten Fläche sind unter der rechtsschiefen Verteilung unzulässig.

Der Algorithmus kann jedoch, anhand der ihm verpassten Parameter, ausschließlich diese Normalverteilung „sehen“.

Stark vereinfachtes Beispiel

4,02	11135,84
3,15	9011,95
0,70	20,21
1,56	554,82
0,34	1,90
3,71	8113,65
NA	0,24
NA	4,55
1,91	6,68
0,21	6,01
NA	2,46
0,58	8,18
0,16	5,95
NA	954,55
1,13	6,01
0,07	1,94
NA	0,84
4,34	20201,20
NA	4,15
NA	11,85



- Aufgrund der Schiefe der Verteilungen weist die Punktwolke einen nichtlinearen Verlauf auf.
- Die roten Punkte unterhalb der gestrichelten grauen Linie sind unter der ersten Verteilung unzulässig.

Eberhard Karls
Universität
Tübingen

Momentenbedingungen im Falle einer multivariaten Normalverteilung

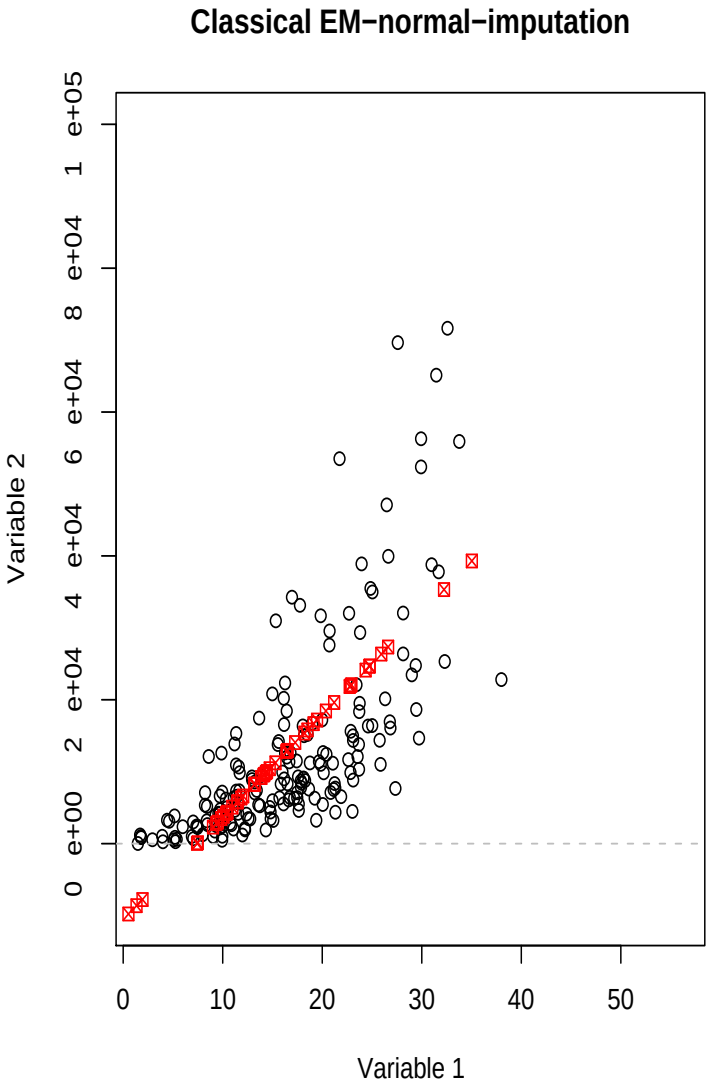
Gegeben sei $Y := \mathcal{Y}^\theta$ normalverteilt.

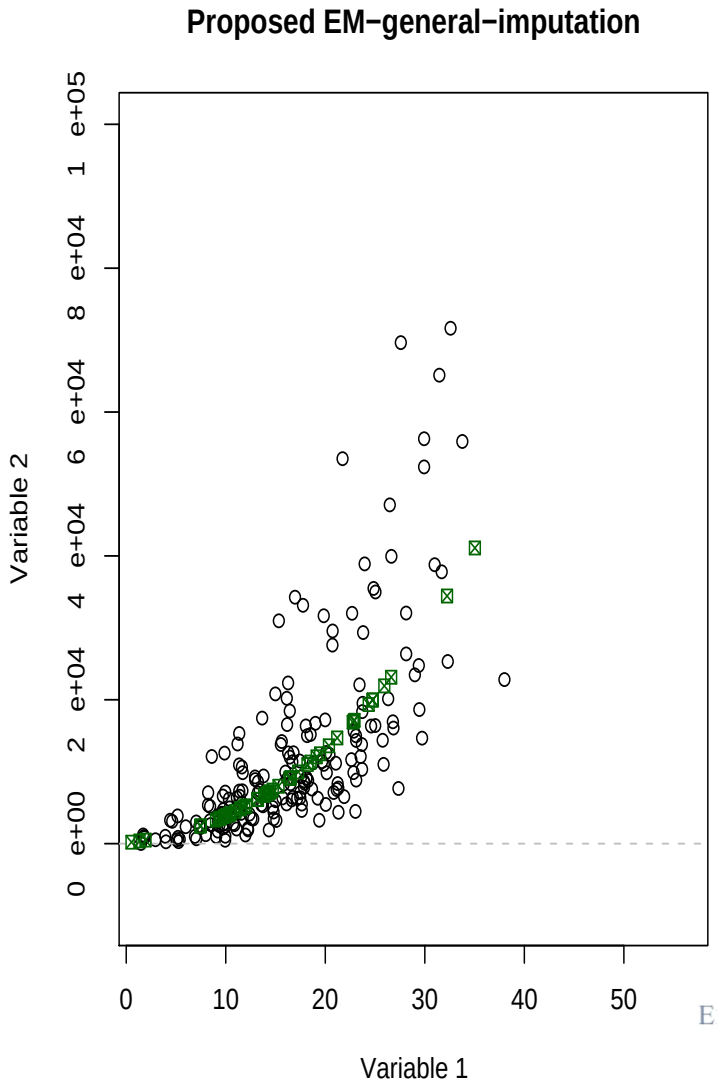
$$\left.\begin{array}{ll} \left.\begin{array}{ll} \mathrm{E}[(Y-\mu)^{3+2k}] = 0 & \text{für } k = 0, 1 \ldots \\ \mathrm{E}[(Y-\mu)^4] - 3\sigma^4 = 0 & \end{array}\right\} \text{Randverteilungen} \\ \left.\begin{array}{ll} \mathrm{E}[(\varepsilon)^{3+2k}] = 0 & \text{für } k = 0, 1 \ldots \\ \mathrm{E}[(\varepsilon)^4] - 3\sigma_\varepsilon^4 = 0 & \\ \mathrm{E}[X\beta(\varepsilon^2 - \sigma_\varepsilon^2)] = 0 & \end{array}\right\} \text{Residuen} \end{array}\right\} =: g$$

Diese Bedingungen sind im Falle einer multivariaten Normalverteilung stets erfüllt.
Die Schätzstrategie besteht nun darin, dass eine geeignet gewichtete quadratische Form der empirischen Momentenbedingungen bezüglich des Potenzparameters minimiert wird:

$$\min_\theta \hat{g}(\theta)' W \hat{g}(\theta)$$

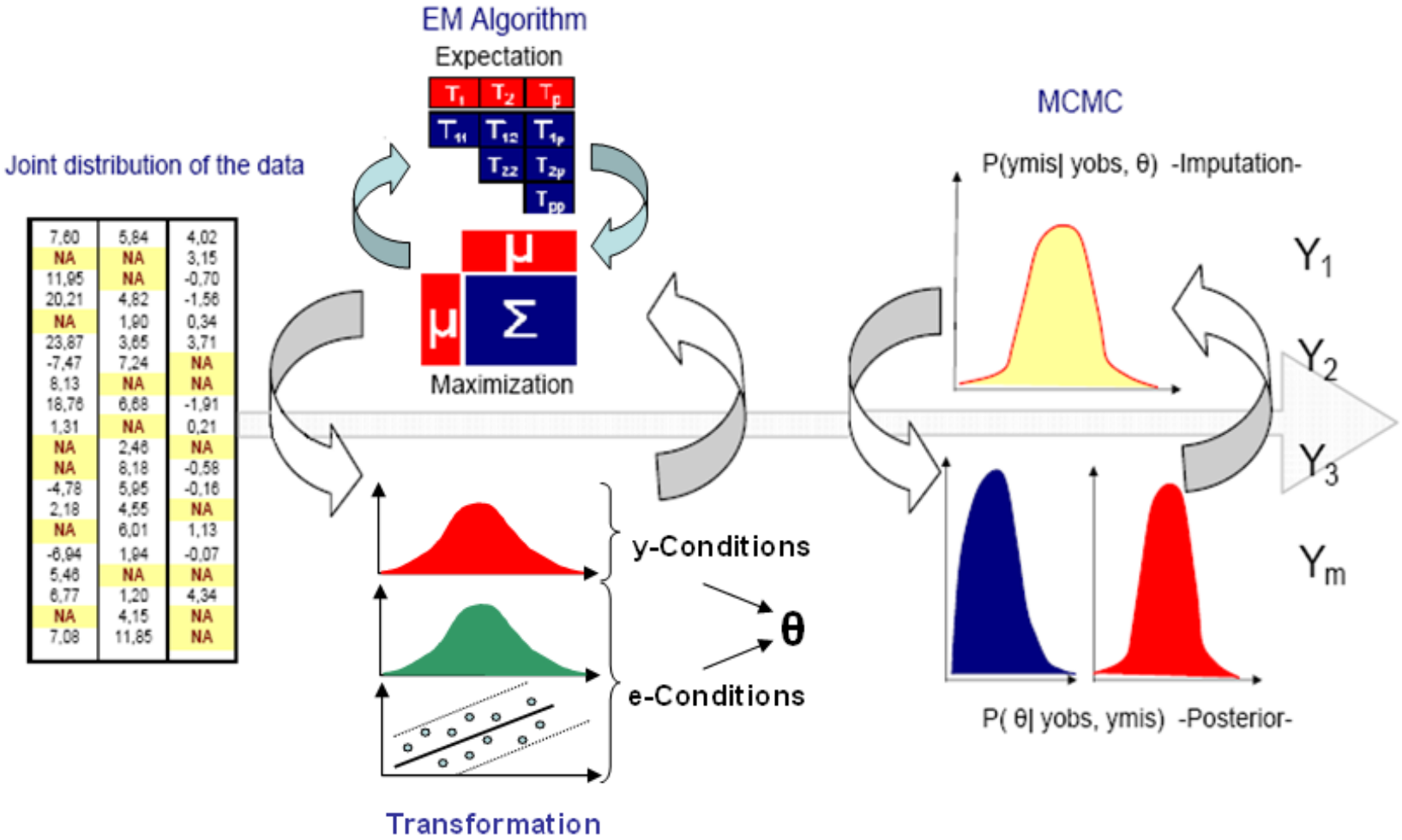
Vergleich beider Methoden





Eberhard Karls
Universität
Tübingen

EM-Algorithmus/Potenztransformation + MCMC



7,60	5,84	4,02
NA	NA	3,15
11,95	NA	-0,70
20,21	4,82	-1,56
NA	1,90	0,34
23,87	3,65	3,71
-7,47	7,24	NA
8,13	NA	NA
18,76	6,68	-1,91
1,31	NA	0,21
NA	2,46	NA
NA	8,18	-0,58
-4,78	5,95	-0,16
2,18	4,55	NA
NA	6,01	1,13
-6,94	1,94	-0,07
5,46	NA	NA
6,77	1,20	4,34
NA	4,15	NA
7,08	11,85	NA

Eberhard Karls
Universität
Tübingen

Vielen Dank für Ihre Aufmerksamkeit.