Deep Multiple Imputation:

Using Mixture Density Networks to Impute Missing Values

Marcel Neunhoeffer, University of Mannheim

Motivation

- Multiple Imputation is a popular approach to deal with **missing data**.
- Current Multiple Imputation techniques rely on restrictive assumptions – either about a joint distribution of data (e.g. **AMELIA II**) or carefully specified conditional probability distributions (e.g. **MICE**) for each data column with missing values.
- **Is it possible to come up with an imputation algorithm that is less restrictive in its assumptions?**

Mixture Density Network

- A **Mixture Density Network (MDN)** is a **combination** of a **deep neural network** and **a mixture model** first described by Bishop (1994).
- The setup of a MDN is like a standard neural network, where **the output layer is mapped to a mixture of normal distributions** with $K > 1$ kernels.
- For a sufficient number of kernels, a **MDN can model arbitrary conditional probability distributions**.

$$p(t|x) = \sum_{i=1}^{K} \alpha_i(z)\phi_i(y|z) \qquad (1)$$

- The output of the neural network is the **parameter vector** $\mathbf{z}$, which contains $K \times \alpha$ (where $\sum_{i=1}^{K} \alpha_i = 1$), $K \times \mu$ and $K \times \sigma$ (where all σ_i are constrained to be > 0).

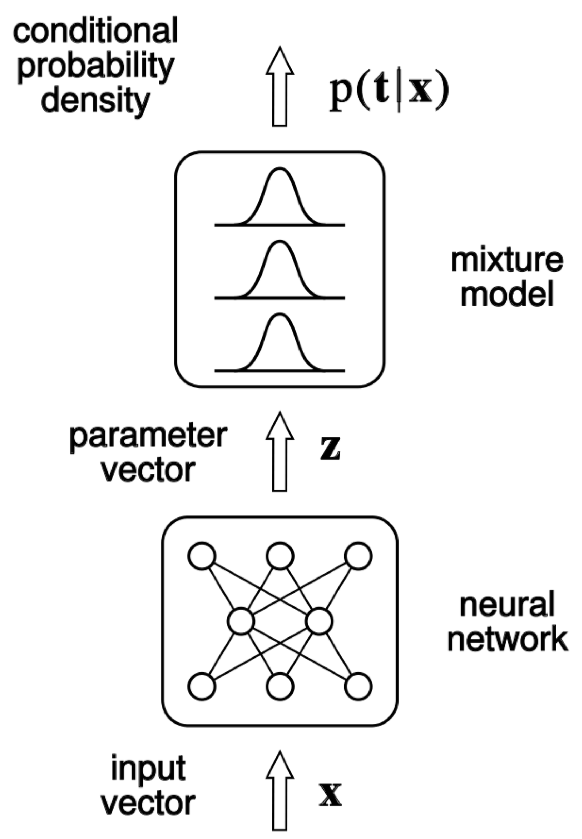

Figure 1: Overview of a MDN by Bishop (1994)

Deep Multiple Imputation

- I draw on a **conditional multiple imputation algorithm** (see Kropko et al 2014, van Buuren 2012) and modify it to make it work with MDNs.
- The setup of an MDN allows me to **draw m times from the conditional probability distribution**. One completed run of the algorithm generates m multiply imputed data sets.

Experimental Setup

- **Experiment 1 - Multivariate Normal Data:** The full data set (with the columns Y, X_1, X_2, X_3 and X_4) is drawn from a multivariate normal distribution. The quantities of interest to recover are β_1 and β_2 in the regression $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.
- **Experiment 2 - Heteroscedastic Data:** The DGP is given by $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, with $\mu_i = \beta_0 + \beta_1 X_{1i}$ and $\sigma_i^2 = exp(\gamma_0 + \gamma_1 X_{1i})$. A second variable X_2 is also included in the data set. Both X_1 and X_2 are drawn from normal distributions. The quantities of interest to recover are β_0, β_1, γ_0 and γ_1. They are calculated using maximum likelihood heteroscedastic linear regression.

Experimental Results

- **Experiment 1** shows that **Deep Multiple Imputation performs as well as current Multiple Imputation techniques** on problems that **current Multiple Imputation techniques can solve well**.
- **Experiment 2** shows that **Deep Multiple Imputation performs better than current Multiple Imputation techniques** on problems that go **beyond what current Multiple Imputation techniques are capable of**.

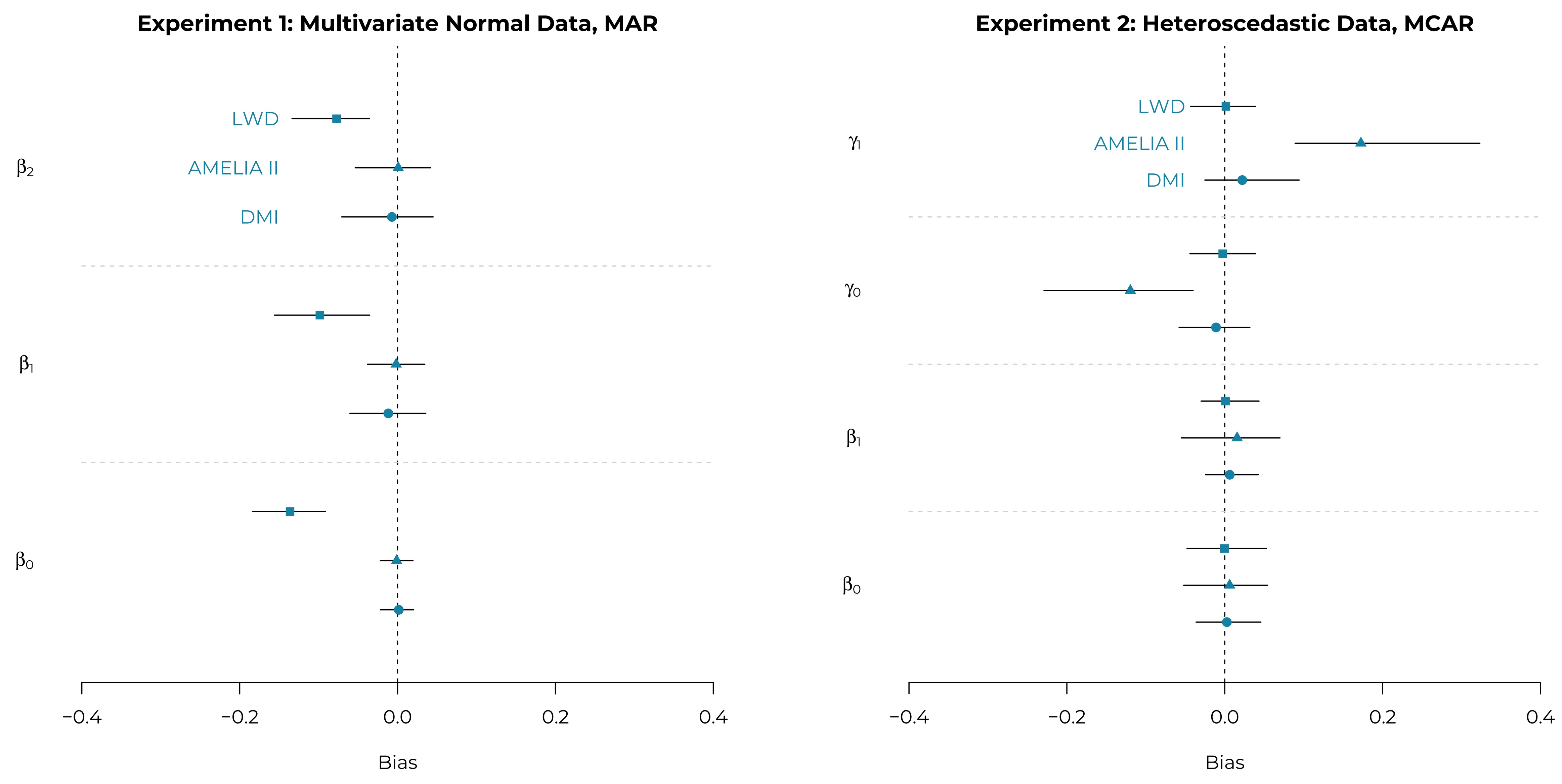

Figure 2: Results of the Monte Carlo Experiments

Contributions

- **First application** of Mixture Density Networks for Multiple Imputation.
- Deep Multiple Imputation is **less restrictive in its assumptions** than current Multiple Imputation approaches.
- Deep Multiple Imputation **decreases *Researcher Degrees of Freedom***, e.g. interactions do not need to be specified ahead of the imputation procedure.

Where to Go From Here?

- Show that it makes a difference on **real world problems**.
- Extend the algorithm to **take into account different data types**.
  - **Other distributions for the components** (e.g. Bernoulli).
  - **Time-series data**.
- Implement it as an **easy to use R-package**.

References

[1] Stef Van Buuren.
Flexible imputation of missing data.
Chapman and Hall/CRC, 2012.

[2] James Honaker, Gary King, and Matthew Blackwell.
AMELIA II : A Program for Missing Data.
Journal Of Statistical Software, 45(7):1–54, 2011.

[3] Christopher M. Bishop.
Mixture Density Networks.
NCRG, 94(004), 1994.

[4] Jonathan Kropko, Ben Goodrich, Andrew Gelman, and Jennifer Hill.
Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches.
Political Analysis, 22(4):497–519, 2014.

Contact Information

- Web: marcel-neunhoeffer.com
- Email: mneunhoe@mail.uni-mannheim.de