SONDERFORSCHUNGSBEREICH 504

Rationalitätskonzepte,
Entscheidungsverhalten und
ökonomische Modellierung

No. 07-06

A Markov Chain Monte Carlo Multiple Imputation Procedure for Dealing with Item Nonresponse in the German SAVE Survey

Daniel Schunk*

February 2007

Financial support from the Deutsche Forschungsgemeinschaft, SFB 504, at the University of Mannheim, is gratefully acknowledged.

*Sonderforschungsbereich 504, email: dschunk@uni-mannheim.de

Universität Mannheim
L 13,15
68131 Mannheim

A Markov Chain Monte Carlo Multiple Imputation Procedure for Dealing with Item Nonresponse in the German SAVE Survey[1]

Daniel Schunk*

Mannheim Research Institute for the Economics of Aging (MEA)

Department of Economics, University of Mannheim

This version: 12 February 2007

Abstract: Important empirical information on household behavior is obtained from surveys. However, various interdependent factors that can only be controlled to a limited extent lead to unit and item nonresponse, and missing data on certain items is a frequent source of difficulties in statistical practice. This paper presents the theoretical underpinnings of a Markov Chain Monte Carlo multiple imputation procedure and applies this procedure to a socio-economic survey of German households, the SAVE survey. I discuss convergence properties and results of the iterative multiple imputation method and I compare them briefly with other imputation approaches. Concerning missing data in the SAVE survey, the results suggest that item nonresponse is not occurring randomly but is related to the included covariates. The analysis further indicates that there might be differences in the character of nonresponse across asset types. Concerning the methodology of imputation, the paper underlines that it would be of particular interest to apply different imputation methods to the same dataset and to compare the findings.

---

[1] I gratefully acknowledge many astute comments by Axel Börsch-Supan, Dimitrios Christelis, Joachim Frick, Arthur Kennickell, Susanne Rässler, Arthur van Soest, Guglielmo Weber, and seminar participants at the University of Mannheim, and the conference on statistical imputation methods in Bronnbach. I am particularly grateful to Arthur Kennickell for his support during this research project. Gunhild Berg, Armin Rick, Frank Schilbach, Bjarne Steffen and Michael Ziegelmeyer provided excellent research assistance. Financial support was provided by the Deutsche Forschungsgemeinschaft (via Sonderforschungsbereich 504 at the University of Mannheim).

*Daniel Schunk
Mannheim Research Institute for the Economics of Aging (MEA)
University of Mannheim
D-68131 Mannheim
Tel.: +49-621-181-3448
E-mail: dschunk@uni-mannheim.de

1 Introduction

Important empirical information on household behavior is obtained from surveys. However, various interdependent factors that can only be controlled to a limited extent, such as privacy concerns, respondent uncertainty, cognitive burden of the questions, and survey context, lead to unit nonresponse and item nonresponse. Unit nonresponse is the lack of any information for a contacted survey participant and as such is the strongest type of refusal. The phenomenon that only a subset of the information is missing, e.g. only the response to the question on household income, is referred to as item nonresponse.

The general phenomenon of item nonresponse to questions in household surveys as well as problems of statistical analysis with missing data have been analyzed by various authors, beginning with the work by Ferber (1966) and Hartley and Hocking (1971); see Beatty and Herrmann (2002) as well as Rässler and Riphahn (2006) for reviews. Recent examples for Germany, focusing on income, saving, and asset choice, are Biewen (2001), Frick and Grabka (2005), Riphahn and Serfling (2005), and Schräpler (2003) who work with data from the German Socio-Economic Panel (GSOEP). Finally, Essig and Winter (2003) describe and analyze nonresponse patterns to financial questions in the first wave of the German SAVE survey. They exploit that this first wave has included a controlled experiment specifically designed to analyze the effects of interview mode and question format on answering behavior.

The German SAVE study (Schunk, 2006) focuses on details of households' finances, as well as households' sociological and psychological characteristics. For the large majority of variables in SAVE, item nonresponse is not a problem. For example, there is hardly any nonresponse to detailed questions about socio-demographic conditions of the household, to questions about households’ expectations and about indicators of household economic behavior. Mainly due to privacy concerns and cognitive burden, though, there are significantly higher item nonresponse rates for detailed questions about household financial circumstances than for other less private and less sensitive questions. Tables 1 and 2 show that these questions can have a missing rate of over 40%. Similar missing rates for questions about financial circumstances have been documented in various socio-economic survey contexts (e.g., Bover, 2004; Hoynes et al., 1998; Juster and Smith, 1997; Kalwij and van Soest, 2005).

Table 1: Response rates for monthly net income and for the question about total annual saving.

	Value	Bracket	Unknown
Net income	69%	25%	6%
Annual saving	88%		12%

Note: Calculations are unweighted and based on the 2003/2004 wave of the SAVE data.

Table 2: Response rates for financial and real wealth items.

	Have item			Value reported for
	Yes	No	Unknown	those having the item
Savings/term accounts	56%	36%	8%	74%
Building society savings agreements	26%	66%	8%	67%
Whole life insurance policies	28%	64%	8%	57%
Bonds	8%	84%	8%	57%
Shares & real-estate funds	18%	74%	8%	61%
Owner occupied housing	47%	49%	4%	96%

Note: Calculations are unweighted and based on the 2003/2004 wave of the SAVE data.

For studies that use the detailed financial information in the SAVE study, missing information on one of those variables is a problem. It is tempting and still very common to simply delete all observations with missing values. But deleting observations with item nonresponse, i.e. relying on a complete-case analysis, might lead to an efficiency loss due to a smaller sample size and to biased inference when item nonresponse is related to the variable of interest.[2] Particularly for multivariate analyses that involve a large number of covariates, case deletion procedures can discard a high proportion of subjects, even if the per-item rate of missingness is rather low.

The purpose of this paper is to present and discuss the theoretical underpinnings and the practical application of an iterative multiple imputation method that has been developed for the German SAVE dataset. Missing item values are imputed controlling for observed characteristics of nonrespondents and respondents in order to preserve the correlation structure of the dataset as much as possible. The method yields a multiply imputed and complete data set that can be analyzed by the public using standard software packages without discarding any observed cases. In contrast to single imputation, multiple imputation allows the uncertainty due to imputation to be reflected in subsequent analyses of the data (see, e.g., Rubin, 1987; Rubin, 1996; Rubin and Schenker, 1986).

[2] See, e.g., Rubin (1987) and Little and Rubin (2002) for discussions about efficiency and bias in a missing data context.

Iterative multiple imputation methods have recently been applied to other large-scale socio-economic survey data (Barceló, 2006; Bover, 2004; Kennickell, 1998). The imputation method for the U.S. Survey of Consumer Finances, developed by Arthur Kennickell, has recently been applied to the Spanish Survey of Household Finances (Barceló, 2006; Bover, 2004), and it has also ultimately inspired the development of the imputation method that is presented in this paper. The contribution of this paper is to investigate the convergence properties of an iterative imputation method that is applied to a large socio-economic survey, the German SAVE survey, and to analyze the resulting distributions of various imputed financial survey items.[3] The latter gives insights about item nonresponse behavior of the survey participants and about the bias that would result from a complete-case analysis. Furthermore, this paper documents in detail the imputation method that has been developed for the German SAVE study.

The paper is organized as follows: Section 2 gives an overview of the SAVE survey, section 3 describes the theoretical underpinnings of the iterative imputation algorithm, develops and documents the application of this algorithm to the SAVE survey, and describes its relationship to existing work on imputation in large surveys. Section 4 investigates the convergence properties of the algorithm and compares imputed and observed data. Section 5 discusses the presented algorithm and concludes the paper.

2 The SAVE Survey – An Overview

In Germany, there has been no dataset available that records detailed data on both financial variables such as income, savings, and asset holdings and on sociological and psychological characteristics of households. The German Socio-Economic Panel (*German SOEP*) has rich data on household behavior and records indicators of saving and asset choices; in 1988 and 2002, the quantitative composition of households' assets was covered in much more detail. Another representative survey, *Soll und Haben,* records detailed data on the composition of various financial assets, but it only has qualitative indicators and does not quantify asset holdings. Finally, the official budget and expenditure survey (*Einkommens- und Verbrauchsstichprobe, EVS*), conducted every five years by the Federal Statistical Office, has very detailed information on the amount and composition of income, expenditure, and wealth, but information on other household characteristics is very limited, in particular in the most recent waves in 1998 and in 2003. Taking as a basis the Dutch CentER Panel and the U.S. Health and Retirement Study (HRS), researchers of

[3] A short companion paper (Schunk, 2007), focuses on the theoretical aspects of MCMC imputation and summarizes the imputation method that is presented in detail here.

the University of Mannheim have cooperated with the Mannheim Center for Surveys, Methods and Analyses (ZUMA), NFO Infratest (Munich), Psychonomics (Cologne) and Sinus (Heidelberg) to produce a questionnaire on households' saving and asset choice; see Börsch-Supan and Essig (2005). The questionnaire has been designed in such a way that the interview should not exceed 45 minutes and was first fielded in 2001 using a quota sampling design. The first random sample was drawn in 2003.[4] The questionnaire consists of six parts (see table 3).

Table 3: Structure of the questionnaire of the SAVE Survey.

Part 1:	Introduction, determining which person will be surveyed in the household
Part 2:	Basic socio-economic data of the household
Part 3:	Qualitative questions concerning saving behavior, income and wealth
Part 4:	Quantitative questions concerning income and wealth
Part 5:	Psychological and social determinants of saving behavior
Part 6:	Conclusion: Interview-situation

The first, relatively short part explains the purpose of the study and describes the precautions that have been taken with respect to confidentiality and data protection. Part 2 lasts about 15 minutes and contains questions on the socio-economic structure of the household, including age, education and labor-force participation of the respondent and his or her spouse. Part 3 of the questionnaire contains qualitative and simple quantitative questions on saving behavior and on how households deal with income and assets, including hypothetical choice tasks and questions on saving motives; questions are also asked on financial decision processes, rules of thumb, and attitudes towards consumption and money. Part 4 is the critical part of the questionnaire. It contains a comprehensive financial review of the household and therefore the most sensitive questions in financial items such as income from various sources and holdings of various assets. Apart from financial assets, the questions also cover private and company pensions, ownership of property and business assets. Questions are also asked about debt. Part 5 contains questions about psychological and social variables. It includes the social environment, expectations about income, the economic situation, health, life expectancy and general attitudes to life. The interview ends with open-ended questions about the interview

[4] A description of SAVE and further details on methodological aspects of the SAVE survey are found in Schunk (2006).

situation, and a question that asks whether the respondent would be willing to participate in a similar survey in the future (part 6).

3 A Multiple Imputation Method for SAVE

3.1 Motivation and Theoretical Underpinnings

To deal with item nonresponse, one can resort to a complete-case analysis, to model-based approaches that incorporate the structure of the missing data, or one can use imputation procedures.[5] A complete-case analysis may produce biased inference, if the dataset with only complete observations differs systematically from the target population; weighting of the complete cases reduces the bias but generally leads to inappropriate standard errors. Additionally, a complete-case analysis leads to less efficient estimates, since the number of individuals with complete data is often considerably smaller than the total sample size.[6] Formal modeling that incorporates the structure of the missing data involves basing inference on the likelihood or posterior distribution under a structural model for the missing-data mechanism and the incomplete survey variables, where parameters are estimated by methods such as maximum likelihood. Multiple imputation essentially is a way to solve the modeling problem by simulating the distribution of the missing data (Rubin, 1996). Ideally, the imputation procedures control for all relevant observed differences between nonrespondents and respondents, such that the results obtained from the analysis of the complete dataset are less biased overall and estimates are more efficient than in an analysis based on complete cases only.

The goal of imputation is not to create any artificial information but to use the existing information in such a way that public users can analyze the resulting complete dataset with standard statistical methods for complete data. It is often seen as the responsibility of the data provider to provide the imputations: First, because imputation is a very resources-consuming process that is not at the disposal of many users. Second, because some pieces of information which are very useful for the imputation, such as information on interviewer characteristics, are not available to the public. Users are free to ignore the imputations, all imputed values are flagged.

Assumptions

[5] An overview of approaches to deal with item nonresponse is presented in Rässler and Riphahn (2006).

[6] Rubin (1987) and Little and Rubin (2002) illustrate and discuss biased inference and efficiency losses based on complete-case analyses and weighted complete-case analyses.

Many different statistical imputation methods exist and are applied in a variety of data contexts. Examples are mean or median imputation, hotdeck imputation and regression-based imputation. Hotdeck is a very frequently used nonparametric method (e.g., in the RAND-HRS). For hotdeck, only very few conditioning variables can be used, even when the dataset is very large. Regression-based imputations need parametric assumptions. Since regression-based methods allow for conditioning on many more variables than hotdeck methods, they are better than hotdeck methods in preserving a rich correlation structure of the data, provided that an appropriate parametric assumption is made.

Ideally, to impute the missing values, a statistical model should be explicitly formulated for each incomplete survey variable and for the missing-data mechanism. The parameters should then be estimated from the existing data (and from potentially available further information, such as information about the interview process) by methods such as maximum likelihood. Identifying the probability distributions of the variables under study is often very hard and requires weakly motivated assumptions, since the mechanisms of nonresponse are often very complex (Manski, 2005).

Clearly, imputation methods have to make some statistical assumption about the nonresponse mechanism and about the distribution of the data values in the survey.[7] For the imputation method presented in this paper, the underlying assumption about the way in which missing data were lost is that missing values are *ignorable*. To define the ignorability assumption, let us first define *missing at random* (MAR):[8]

Suppose that Y is a variable with missing data and X is a vector of always observed variables in the dataset. Then, formally:

$$Y \text{ is MAR} \Longrightarrow P(Y \text{ is observed} \mid X, Y) = P(Y \text{ is observed} \mid X)$$

That is, after controlling for information in X, the probability of missingness of Y is unrelated to Y.[9] MAR implies that the imputation method should condition on all variables that are predictive of the missingness of Y, since MAR may no longer be

[7] The Bayesian nature of the presented imputation algorithm also requires specification of a prior distribution for the parameters of the imputation model. In practice, unless the data are very sparse or the sample is very small, a noninformative prior is used (see Schafer (1997) for details). Based on Schafer (1997), it can be concluded that the data in the SAVE survey are neither sparse, nor is the sample small. Consequently, I do not make any assumption about the prior distribution of parameters.

[8] Note that the MAR assumption cannot be tested from available data (Cameron and Trivedi, 2005).

[9] MAR does not imply that the missing values are a random subsample of the complete dataset. This latter condition is much more restrictive and is called 'missing completely at random' (MCAR). See Little and Rubin (2002) for further discussions.

satisfied if variables that determine the nonresponse are not included as conditioning variables (Schafer, 1997).

The missing data mechanism is said to be *ignorable*, if, (a), the data are MAR and, (b), the parameters for the missing data generating process are unrelated to the parameters that the researcher wants to estimate from the data.[10] Ignorability is the formal assumption that allows one to, first, estimate relationships among variables between observed data and, then, use these relationships to obtain predictions of the missing values from the observed values.

Of course, for these relationships to yield unbiased predictions, one would need the correct model for the observed and missing values. The imputation method presented in this paper relies on simple parametric assumptions for all core variables with high rates of missingness[11] and the method uses nonparametric hotdeck methods for discrete variables with only few categories and with very low rates of missingness. The fact that data have been multiply imputed increases robustness to departures from the true imputation model considerably compared to single imputation approaches that are based on the same imputation model. This has been demonstrated in simulation studies (Ezzati-Rice et al., 1995; Graham and Schafer, 1999; Schafer, 1997). Furthermore, using simulated and real datasets from different scientific fields and with varying rates of item nonresponse, existing research emphasizes the robustness of multiple imputation to the specifically chosen imputation model, given that appropriate conditioning variables are available in the dataset (e.g., Schafer, 1997; Bernaards et al., 2003).

The imputation method used for SAVE aims at capturing all relevant relationships between variables in order to preserve the correlation structure between the variables. The method therefore conditions on as many relevant and available variables as possible in the imputation of each single variable. All possible determinants of the variable to be imputed are included as predictors of that variable. Additionally, as has been argued above, including all variables that are potential predictors of missingness makes the MAR-assumption more plausible, because this assumption depends on the availability of

[10] In the literature, MAR and "*ignorability*" are often treated as equivalent under the assumption that condition (b) for ignorability is almost always satisfied (Cameron and Trivedi, 2005).

[11] In line with other iterative or non-iterative and regression-based imputation methods for survey data, e.g. Bover (2004), Frick and Grabka (2005), and Kennickell (1998), I generally assume a linear model for the imputation of continuous variables with high missingness.

variables that can explain missingness and that are correlated with the variable to be imputed.[12]

Multiple Imputation

Single imputation does not reflect the true distributional relationship between observed and missing values and it does not allow the uncertainty about the missing data to be reflected in the subsequent analyses. Estimated standard errors are generally too small (see also appendix, section 6.2), and even if an appropriate imputation model is chosen, single imputation is more prone to generate biased estimates than multiple imputation. These defects – documented and discussed in, e.g., Li et al. (1991) and Rubin and Schenker (1986) – can seriously affect the subsequent interpretation of the analyses.

In multiple imputation, M>1 plausible data sets are generated with all missing values replaced by imputed values. All M complete datasets are then used separately for the analysis and the results of all M analyses are combined such that the uncertainty due to imputation is reflected in the results (see appendix, section 6.2). Briefly, multiple imputation simulates the distribution of missing data and the resulting overall estimates then incorporate the uncertainty about which values to impute. This involves two types of uncertainty: Sampling variation assuming the mechanisms of nonresponse are known and variation due to uncertainty about the mechanisms of nonresponse (Rubin, 1987).

Unless the fraction of missing data is extremely large, it is sufficient to obtain a relatively small number M of imputed datasets, usually not more than five, which is the choice for M in the SAVE imputation method.[13] The relative gains in efficiency from larger numbers are minor under the rates of missing data that are observed in surveys such as the SAVE survey.[14]

Markov Chain Monte Carlo Simulation

Tanner and Wong (1987) present an iterative simulation framework for imputation based on an argument that involves the estimation of a set of parameters from conditioning information that is potentially unobserved. I review briefly their arguments to motivate the iterative imputation method that is used for the SAVE study:

---

[12] Details about the inclusion of conditioning variables in the SAVE imputation method are discussed in section 3.2.4.

[13] Both, the Spanish Survey of Household Finances (Barceló, 2006; Bover, 2004) and the U.S. Survey of Consumer Finances (Kennickell, 1998) also provide 5 imputations.

[14] Rubin (1987) and Schafer (1997) define efficiency in the context of multiply imputed datasets and discuss the choice of M and its impact on efficiency in detail.

Let x_u be unobserved values of a larger set x and let $x_o = x \setminus x_u$. X_u is the sample space of the unobserved data, θ is a set of parameter values to be estimated for which the parameter space is denoted by Θ. The desired posterior distribution of the parameter values, given the observed data, can be written as:

$$f(\theta \mid x_o) = \int_{X_u} f(\theta \mid x_o, x_u) f(x_u \mid x_o) dx_u \tag{1}$$

Here, $f(\theta \mid x_o, x_u)$ is the conditional density of θ given the complete data X, and $f(x_u \mid x_o)$ is the predictive density of the unobserved data given the observed data. The predictive density of the unobserved data given the observed data can be related to the posterior distribution that is shown above as follows:

$$f(x_u \mid x_o) = \int_{\Theta} f(x_u \mid \phi, x_o) f(\phi \mid x_o) d\phi \tag{2}$$

The basic idea of Tanner and Wong is that the desired posterior is intractable based on only the observed data, but it is tractable after the data are augmented by unobserved data x_u in an iterative framework. The suggested iterative method for the calculation of the posterior starts with an initial approximation of the posterior. Then, a new draw of x_u is made from $f(x_u \mid x_o)$ given the current draw from the posterior $f(\theta \mid x_o)$, and this draw is then used for the next draw of $f(\theta \mid x_o)$. Tanner and Wong show that under mild regularity conditions, this iterative procedure converges to the desired posterior.

In an imputation framework, the target distribution is the joint conditional distribution of x_u and θ, given x_o. Based on the ideas of Tanner and Wong, the iterative simulation method is summarized as follows: First, replace all missing data by plausible starting values. Given certain parametric assumptions, θ can then be estimated from the resulting complete data posterior distribution $f(\theta \mid x_o, x_u)$. Let now θ^t be the current value of θ. The next iterative sample of x_u can then be drawn from the predictive distribution of x_u given x_o and θ^t:

$$x_u^{t+1} \sim f(x_u \mid x_o, \theta^t) \qquad \text{[Imputation step (I-step)]} \tag{3}$$

The next step is again to simulate the next iteration of θ from the complete data posterior distribution:

$$\theta^{t+1} \sim f(\theta \mid x_o, x_u^{t+1}) \qquad \text{[Prediction step (P-step)]} \tag{4}$$

Repeating steps (3) and (4), i.e. sequential sampling from the two distributions, generates an iterative Markovian procedure $\{(\theta^t, x_u^t) : t = 1,2,...,N\}$. For the purpose of imputation, this procedure yields a successive simulation of the distribution of missing

values, conditioned on both, observed data and distributions of missing data previously simulated. The set of conditioning variables in this algorithm is not necessarily the entire set of all possible values (Tanner and Wong, 1987). Geman and Geman (1984) apply a similar procedure in the field of image processing and show that the stochastic sequence is a Markov chain that has the correct stationary distribution under certain regularity conditions. Li (1988) presents an additional formal argument that the process moves closer to the true latent distribution with each iteration and finally converges. The method is called Markov Chain Monte Carlo (MCMC) because it involves simulation and the sequence is a Markov Chain. Formally, the method is also related to Gibbs sampling (Hastings, 1970), and in the missing data literature, it is often referred to as data augmentation. This method has been used in many statistical applications (e.g., Bover 2004; Kennickell, 1998; Schafer 1997). Sequential simulation algorithms of the MCMC-type can be modified and implemented in different ways, I briefly come back to this issue in section 5.

3.2 The MIMS-Model

3.2.1 Variable Definitions

The multiple imputation method for SAVE (MIMS) distinguishes between core variables and non-core variables. The core variables have been chosen such that they cover the financial modules of the SAVE survey that involve all questions related to income, saving(s), and wealth of the household. The non-core variables include socio-demographic and psychometric variables, as well as indicator variables for household economic behavior. Except for the participation questions of the core variables (e.g., “Did you or your partner own asset X?”) and the question about the value of owner-occupied housing, all core variables have missing rates of at least 6%. The non-core variables have considerably lower missing rates, in almost all cases much less than 2%. The following variables (grouped into three categories) are defined as core-variables:

- *Income variables (E):* 40 binary variables indicating income components, 1 continuous variable for monthly net income, and 1 ordinal variable indicating net income in follow-up brackets.
- *Savings variables (S):* 1 binary variable indicating whether the household has a certain savings goal, 1 continuous variable indicating the amount of this savings goal, and 1 continuous variable indicating the amount of total annual saving.
- *Asset variables (A):* 48 binary variables indicating asset ownership and credit, 44 continuous variables indicating the particular amounts.

All other variables in the dataset are non-core variables.

3.2.2 Algorithmic Overview

MIMS is a multiple imputation procedure that is based on the idea of a Markovian process that I have described in the previous subsection. The general algorithmic structure of MIMS is similar to the FRITZ imputation method that is used for the multiple imputation of the Survey of Consumer Finances and for the Spanish Survey of Household Finances (Kennickell, 1998; Bover, 2004). To set the stage for a more detailed discussion of MIMS in the next section, this section gives a brief algorithmic overview of MIMS.

For this purpose, all variables are categorized as follows:

- All variables that are not core variables are called other variables, **O**.
- **P** is a subset of **O**, the subset of all variables that is used as conditioning variables or predictors for the current imputation step.
- The union of all variables from **P** and all core variables that are used as conditioning variables for the current imputation step is referred to as the set **C** (= conditioning variables). In the following algorithmic description, **C** always contains the updated information based on the most recent iteration step. It contains, in particular, the imputed core variables that have been obtained in the last iteration step.

The complete imputation algorithm for the SAVE data works as follows:

- Impute all variables using logical imputation, whenever possible.

Outer Loop *– REPEAT 5 times, j = 1,..., 5 (= Generate 5 datasets)*

- Impute variables from O using (sequential) hotdeck imputation, obtain complete data ***O****.*

- Impute the income variables E using ***P****, obtain complete data* ***E****.*

- Impute the savings variables S using ***P**** *and* ***E****, obtain complete data* ***S****.*

- Impute the asset variables A using ***P****,* ***E****, and* ***S****, obtain complete data* ***A****.*

Inner Loop *– REPEAT N times (= Iterate N times)*

- Impute the income variables E using ***C****.*

- Impute the savings variables S using ***C****.*

- Impute the asset variables A using ***C****.*

Inner Loop *– END*

Outer Loop *– END*

The five repetitions in the outer loop generate one imputed dataset each. After the complete algorithm, five complete datasets are obtained, which I henceforth refer to as implicates. The algorithm generates an additional flag-dataset which contains binary indicators that identify for each value whether it has been imputed or observed.

3.2.3 Description of MIMS

As the algorithmic description shows, MIMS follows a fixed path through the dataset. The first step of the procedure consists of logical imputation. In many cases, the complex tree structure of the SAVE survey or cross-variable relationships allow for the possibility to logically impute missing values. The following path through the dataset is guided by the knowledge of the missing item rates and by cross-variable relationships. The path starts with variables with low missing rates, such that those variables can subsequently be used as conditioning variables for variables with higher missing rates. For example, among the core variables, the net income variable is imputed first, since its missing rate is generally lower than the missing rates of other core variables.[15] The algorithmic description shows that as soon as the iteration loop starts, all variables are already imputed, i.e. starting values for the iteration process have been obtained, and all variables can be used as conditioning variables during the iteration.

Each variable is imputed based on one of the following three general methods:[16]

(1) For all *categorical or ordinal variables* with only few categories and with a low missing rate, a hotdeck procedure with several conditioning variables is used.

(2) For all *binary, categorical, or ordinal core variables,* binomial or ordered Probit models are used.

(3) For all c*ontinuous or quasi-continuous variables,* randomized linear regressions with normally distributed errors are used. This regression procedure, in particular the handling of constraints and restrictions, follows Barceló 2006 and Kennickell (1998). First, the conditional expected value is estimated and an error term, drawn from a symmetrically censored normal distribution, is added. This normal distribution has mean zero and its variance is the residual variance of the estimation. The error term is always restricted to the central three standard deviations of the distribution in order to avoid imputing extreme

[15] The lower missing rate for the net income variable is – at least partly – due to the survey design. The net income question was presented using an open-ended format with follow-up brackets for those who did not answer the open-ended question. The imputation of the bracket answers is described later in this paper.

[16] These methods and their application to binary, categorical, ordinal and (quasi-)continuous variables with high and low missing rates are illustrated and discussed in more detail in Little and Rubin (2002).

values. In few cases, logical or other constraints require that the error term has to be further restricted; examples are non-negativity constraints. The imputed value is also restricted to lie in the observed range of values for the corresponding variable. That is, in particular, imputed values will not be higher than observed values for a certain variable.

Due to the skip patterns in the questionnaire, the SAVE data have a very complex tree structure that imposes a logical structure and that has to be accounted for in the imputation process. Further constraints stem from these logical conditions of the data, from the ranges provided (e.g., bracket respondents), from cross-relationships with other variables, or from any prior knowledge about feasible outcomes. For several variables, the specification of all relevant constraints is the most complex part of the imputation software. If necessary, the procedure draws from the estimated conditional distribution limited to the central three standard deviations, until an outcome is found that satisfies all possible constraints that apply in the particular case.

Two remarks are important at this point to gain an understanding of key procedures of the algorithm.

(1) Ownership and amount imputations

For certain quantities, e.g. the amount of assets held by a household, the SAVE survey uses a two-step question mode: In step one, households are asked about ownership of assets from a certain asset category and a binary variable records the answer. In step two, those households that have reported that they own assets from the particular category are asked about the exact value of the corresponding assets. From a modeling point of view, this is a corner solution application. Following Bover (2004) and Kennickell (1998), a hurdle model is used in MIMS to impute the missing values in these two steps: First, a Probit model is estimated for the binary ownership variable, and missing information is predicted. Then, as described above, randomized linear regressions with normally distributed errors are used for imputing continuous amounts. These regressions are estimated based on all observations that own the asset. Alternatively, Tobit models or sample-selection models might be appropriate. Tobit models are less attractive for the given problem, since they include the implicit assumption that the model governing selection and the model governing the estimation of the amounts are the same. Heckman selection models are theoretically attractive, but cause estimation problems in practice: First, the necessary exclusion restrictions differ substantially across asset categories, but there is no theoretical reason why they should differ. Second, in most cases, strong

exclusion restrictions are needed to ensure identification and convergence of the Heckman procedure in each iteration step of MIMS. This means that in practice only a very small set of conditioning variables can be used for the estimation of the second step of the Heckman model. Under these circumstances and given that the goal of the multiple imputation method is to simulate the distribution of amounts conditional on ownership and conditional on a maximally large set of potentially correlated variables, MIMS uses hurdle models for ownership and amount imputations.

(2) Net income variables

To alleviate the problem of item nonresponse to income questions (see, e.g., Juster and Smith, 1997), the survey question on monthly net income was presented using an open-ended format with follow-up brackets for those who did not answer the open-ended question. That is, there are two types of income information available: Exact (in the sense of point data) income information for households that answered the open-ended question, and interval information on household income for those who only answered the bracket question. To make best possible use of all the available income information, the imputation procedure uses a maximum-likelihood estimation procedure. The likelihood is a mixture of discrete terms (for the interval information) and continuous terms (for the point data information). After prediction of the missing income values and the addition of the randomized error term, a nearest neighbor approach is used to determine the imputed amount for household net income.[17] The procedure works as follows: First, an income bracket is predicted for all complete nonrespondents to both (i.e., open-ended *and* bracket) income questions. Now, all observations have either exact income information (if they have reported this information) or bracket information (either they have reported this information, or it has been imputed in the preceding step). Then, each observation *i* for whom an exact net income value has to be imputed and whose net income lies in bracket *j* is matched with the continuous reporter *r* from bracket *j* whose predicted net income value is closest to the predicted value of respondent *i*. The net income value assigned to observation *i* is then the reported continuous income value of the respondent *r*.[18]

---

[17] Nearest neighbor methods have been motivated in a statistical missing data context by Little et al. (1988) and they have subsequently used in the context of bracketed follow-up questions by, e.g., Hoynes et al. (1998) in the AHEAD.

[18] In contrast to this procedure, Hoynes et al. (1998) impute the brackets for the full nonrespondents using an ordered Probit model that is estimated using *only* those respondents that have provided bracket answers. The chosen procedure in MIMS has the advantage of making better use of the available information (since it uses the information from bracket respondents *and* from contiuous, i.e. open-ended, respondents) and it

3.2.4 Selection of Conditioning Variables

As is clear from the descriptions above, each regression or hotdeck method is tailored specifically to the variable to be imputed.[19] Of particular importance are the conditioning variables which have been selected individually for every single variable with missing information according to the following guidelines:

(A) Hotdeck imputations: *Hotdeck imputations,* which have been used for discrete variables with very low missing rates, allow for only few and discrete conditioning variables due to the quickly increasing number of the corresponding conditioning cells. The conditioning variables have first been selected based on theoretical relationships if available and, second, based on the strength of a correlation with the variable to be imputed; those correlations have been systematically explored. As an example for the latter, consider the question which asks respondents to rate their expectation concerning the future development of their own health situation on a scale from 0 (negative) to 10 (positive), which has a missing rate of 0.6%. As conditioning variables, the respondents' age (subdivided into five age classes), self-assessed information on the respondents' current health status (rated on a scale from 0 to 10 and subdivided into three classes), and self-assessed information on how optimistic the respondent generally is (rated on a scale from 0 to 10 and subdivided into three classes) are used.[20] All these conditioning variables are significantly correlated with the variable to be imputed, both individually, as well as jointly in a multiple regression. In some cases, it would be desirable to include core variables as additional conditioning variables in the hotdeck imputations. For example, net income is clearly expected to be correlated with educational status. Generally, the pattern of nonresponse makes this impossible, since the set of nonrespondents to the qualitative questions is in almost all cases a subset of the set of nonrespondents to the relevant core questions.

---

circumvents the practical problem in SAVE that the subsample of bracket respondents is too small to be able to include much conditioning information into the estimation of an ordered Probit model. Hoynes et al. (1998) motivate their procedure by arguing that full nonrespondents are more similar to bracket respondents than to continuous reporters. Note, however, that the evidence on the similarity between nonrespondents, bracket respondents and continuous respondents is mixed (Kennickell, 1997).

[19] A spreadsheet with information on the specific imputation methods for each imputed variable in SAVE (e.g., hotdeck, various regression techniques), as well as information on the used conditioning variables can be obtained from the author upon request.

[20] Note that these three conditioning variables already correspond to $5 \cdot 3 \cdot 3 = 45$ different cells.

(B) Regression-based imputations: In theory, every *regression-based imputation* should use all relevant variables in the dataset, as well as higher powers and interactions of those terms as conditioning variables (see section 3.1 and Little and Raghunathan, 1997). The imputation procedure should, in particular, attempt to preserve the relationships between all variables that might be jointly analyzed in future studies based on the imputed data (Schafer, 1997). In practice, a limit to the number of included conditioning variables is imposed by the degrees of freedom of the regressions. Additionally, there must not be collinearity between conditioning variables, which can easily arise in some cases due to the tree structure of the questions. Due to these constraints concerning the inclusion of conditioning variables, it is of particular importance to select these variables following certain guidelines such that best possible use is made of the available information. For that purpose, the variables used in the regression-based imputations of the core variables have been classified into three non-disjoint categories:

(B-1) *Determinants of the nonresponse.*

Research in psychology, economics, and survey methodology has investigated the relationship between observed respondent and household characteristics and item nonresponse behavior in various survey contexts (for an overview, see Groves et al., 2002). Findings from empirical studies that focus particularly on financial survey items suggest that certain variables might be useful predictors of nonresponse to wealth and income questions (Hoynes et al., 1998; Riphahn and Serfling, 2005). Following these findings, MIMS considers the following variables as determinants of nonresponse to the core variables: Age (as well as squared and cubic age), gender, dummy variables for educational achievement and employment status, as well as household size. Riphahn and Serfling (2005) and Schräpler and Wagner (2001) provide evidence that it is not only the individual respondent's characteristics that may be associated with item nonresponse to financial variables, but also the combination of interviewer and respondent characteristics. In this spirit, the following variables that capture the relationship between interviewer and interviewee characteristics are also considered as determinants of nonresponse to the core financial variables in SAVE: Dummies for whether the interviewer is older than the interviewee, for her/his educational status relative to the interviewee, for the interviewer's gender, and for the gender combination of interviewer and interviewee.

(B-2) *Variables that are related to the variable to be imputed based on different economic models.*

This category contains essentially all core variables, since financial characteristics of households, e.g. saving(s), income and asset categories, are all interrelated. Certain qualitative variables on household socio-economic and financial characteristics that are not already part of the variables in *(B-1)* are also included, for example an indicator for marital status. Variables that measure individual preferences, such as measures for risk attitude, are further included into this category.

(B-3) *Other variables that might be related to the variables to be imputed.*

This category includes variables that are correlated with the variables to be imputed but this relationship is not captured in any formal established economic theory that the author knows of. An example is the smoking habit of the respondent: While there is no formal theory that *directly* relates smoking habits to economic characteristics of a household, there is abundant evidence for a statistically strong association between smoking habits and economic characteristics (e.g., Hersch, 2000; Hersch and Viscusi, 1990; Levine et al., 1997).

The selection of the conditioning variables for the regression is based on the following procedure: First, since the goal is to include as many conditioning variables as possible, all variables from categories *(B-1)*, *(B-2)*, and *(B-3)* are included for each imputation regression. If necessary – because of multicollinearity or insufficient degrees of freedom – variables are removed in the following order: First, variables from *(B-3)* are removed. Then, variables from *(B-2)* are aggregated if possible: E.g., instead of including information on the value of owner-occupied housing and on other real estate as two separate conditioning variables, these two variables can be combined to form a variable for total real estate wealth. In a few cases, notably variables with very low variability, such as the measure of wealth in "other contractually agreed private pension schemes", further conditioning variables from category *(B-2)* have to be removed. In this case, the decision is based on the significance of the variables in the regression. Generally, psychometric variables are removed first and credit variables are removed subsequently, since those variables have the lowest variability and the highest missing rate among the core variables.

4 Results

MIMS has been applied to the 2003/2004 wave of the SAVE survey which contains 3154 observed households and all statistics presented in this section are based on this wave. This section discusses the convergence properties of the algorithm and presents descriptive analyses of the imputed and the observed data. The presented analyses serve to illustrate the differences between the five implicates, the impact of imputation on the distribution of values in the complete dataset, and they are informative concerning the differences in the character of nonresponse across various financial survey items.

4.1 Convergence of MIMS

Assessing convergence of the sequence of draws to the target distribution is more difficult than assessing convergence of, e.g., EM-type algorithms, since there is no single target quantity to monitor, like the maximum value of the likelihood. In this subsection, I first develop a convergence criterion that is based on a measure for the average change in the values of a certain variable vector between two consecutive iteration steps. I then use a standard convergence criterion that is also mentioned in Barceló (2006) and which is defined with respect to measures of position and dispersion of the distribution of the variable to be imputed. Both convergence criteria are used for assessing convergence of three core variables of the SAVE survey.

Let us assume first that there is missing information on only one variable Y in the dataset. That is, all conditioning variables are complete data vectors without missing values. Let $Y_{i,t}$ be the imputed value of the variable of interest for household i in iteration step t, and let I be the total number of imputed observations for variable Y in the dataset. Then, the squared change in the value of variable Y between iteration step t and $t-1$ is:

$$s(t) = \frac{1}{I}\sum_{i=1}^{I}(Y_{i,t} - Y_{i,t-1})^2 \qquad (5)$$

If the procedure has converged, the parameters θ that characterize the distribution of the imputed variable have stabilized.[21] That is, after convergence has been achieved, there is no systematic component in the change of Y over iterations steps any more; only a non-

[21] Note: This suggests a further way to assess convergence: One can investigate the degree of serial dependence of a certain parameter value over iteration steps by analyzing the autocorrelation function. Ideally, this has to be done for *all* parameters of the particular imputation model, and it is preferred for datasets with only few variables and a correspondingly small set of conditioning variables and parameters (Schafer, 1997).

systematic component remains. $Y_{i,t}$ and $Y_{i,t-1}$ can then be assumed to be draws from the same distribution. This implies that – as soon as convergence has been achieved – we have:

$$s(t) = \frac{1}{I}\sum_{i=1}^{I}(Y_{i,t} - Y_{i,t-1})^2 = Var(Y_{i,t} - Y_{i,t-1}) = Var(Y_{i,t}) + Var(Y_{i,t-1}) = 2Var(Y_{i,t}) \quad (6)$$

Indeed, if the procedure has converged, the distribution of the remaining non-systematic component is well known, since it is characterized by the distribution of the simulated error term that is added to the particular predicted value of in each iteration step. I.e., $Var(Y_{i,t})$ can be calculated as the variance of the simulated error term: This error term, ε, is drawn from a normal distribution, the variance of which is – by construction – the residual variance of the particular estimation (see section 3.2.3). This normal distribution is then double censored to the central three standard deviations. I derive the variance of a double censored variable ε in the appendix (see section 6.1).

From these deliberations follows: If the process has converged, $s(t)$, calculated based on the imputed values of the variable $Y_{i,t}$ and $Y_{i,t-1}$, should be equal to $e(t) = 2Var(\varepsilon_t)$, i.e. it should be equal to two times the variance of the simulated error term in iteration step *t*. Furthermore, if convergence has been achieved, $s(t)$ and $e(t)$ are stationary, i.e. they should not have any trend over iterations steps and the sample autocorrelation function for $s(t)$ and $e(t)$ should not indicate autocorrelations at any lag.

In real world data-sets, such as in the SAVE data, it is rarely the case that all conditioning variables are non-missing, as I have assumed for the derivation above. In particular, this condition will not be satisfied in MIMS, since – for reasons given above (see section 3.2.4) – MIMS conditions on as many core variables as possible which have rather high missing rates themselves. But even if the conditioning variables themselves have been imputed, the parameters θ that characterize the distribution of imputed variables should, of course, have stabilized if the process has converged. That is, if the process converges, $s(t)$ and $e(t)$ are stationary, i.e. they should not have any trend over iterations steps, and the corresponding autocorrelation functions should not indicate any autocorrelations. Therefore, displaying $s(t)$ and $e(t)$ over time provides an intuitive graphical way to

investigate convergence of the process.[22] Note, however, that the fact that the conditioning variables are also imputed has the effect that $s(t)$ should be in fact larger than $e(t)$ even if the process has converged, since the imputed conditioning variables themselves are drawn from the corresponding posterior distribution in the particular iteration step.

Figure 1 shows $s(t)$ and $e(t)$. Five different iteration runs are shown for $t = 1,..., 30$ and one additional run is shown for $t = 1,..., 100$ in the last row of the figure. The runs are displayed for three variables that are used to assess convergence, one from each category of the core variables.[23] In all simulation runs, $e(t)$ quickly resembles a horizontal line. As expected due to the sample size, $s(t)$ is very volatile. It lies above the value $e(t)$, and after few iterations, it does not exhibit any trend over the following iteration steps.[24] The results indicate quick convergence in the first few iteration steps for *net income* and for *annual saving*. For the *net income* variable, $s(t)$ is lower than $e(t)$[25]; this is due to the nearest neighbor algorithm and the available bracket information for many nonrespondents which reduces variability of a certain imputed value over iteration steps.

A further investigation of the sample autocorrelation functions of $s(t)$ and $e(t)$ does not reveal any correlations. The corresponding autocorrelation data and figures can be obtained from the author upon request.

---

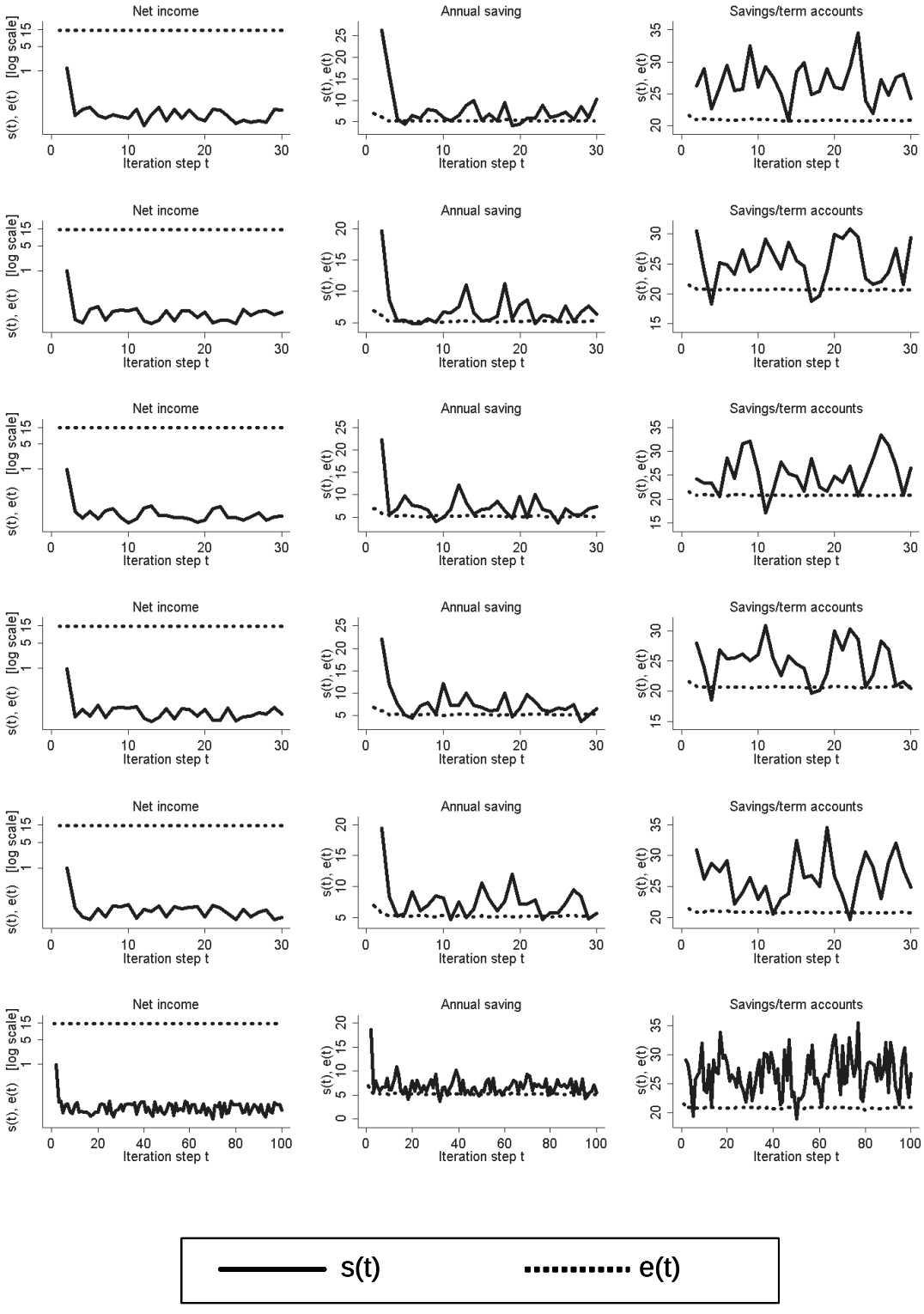
[22] The purpose of these derivations is to suggest a simple graphical convergence diagnostic for an MCMC-method that is applied to a large dataset and that uses a very large set of conditioning variables. I do *not* claim an equivalence result: While convergence of the algorithm would imply that *s(t)* and *e(t)* do not exhibit any downward or upward trend, the converse is not true; i.e. stationarity of *s(t)* and *e(t)* does *not* imply convergence of the algorithm.

[23] Note that only those values for whom no further constraints apply in all iteration steps (e.g., neither non-negativity constraints nor maximum-value constraints), are used for the calculation of *s(t)* and *e(t)*.

[24] If the calculation of *s(t)* is restricted to those observations for which the conditioning variables are almost complete, i.e. non-missing, then the plot reveals that *s(t)* fluctuates around *e(t)*, as predicted. However, the number of observations is even smaller in this case.

[25] Note, that *s(t)* and *e(t)* are plotted on a logarithmic scale for the net income variable in order to be able to plot both variables in one graph.

Figure 1: Convergence diagnostics: *s(t)* and *e(t)* displayed for three key variables.



Note: For net income, s(t) and e(t) are divided by 1,000,000, for annual saving and savings/term accounts, s(t) and e(t) are divided by 10,000,000.

A common criterion for assessing the convergence of a distribution, also suggested in Bover (2004), is to compare (functions of) quantiles, e.g., the median and the interquartile range, resulting from successive iterations of the variable *Y*:

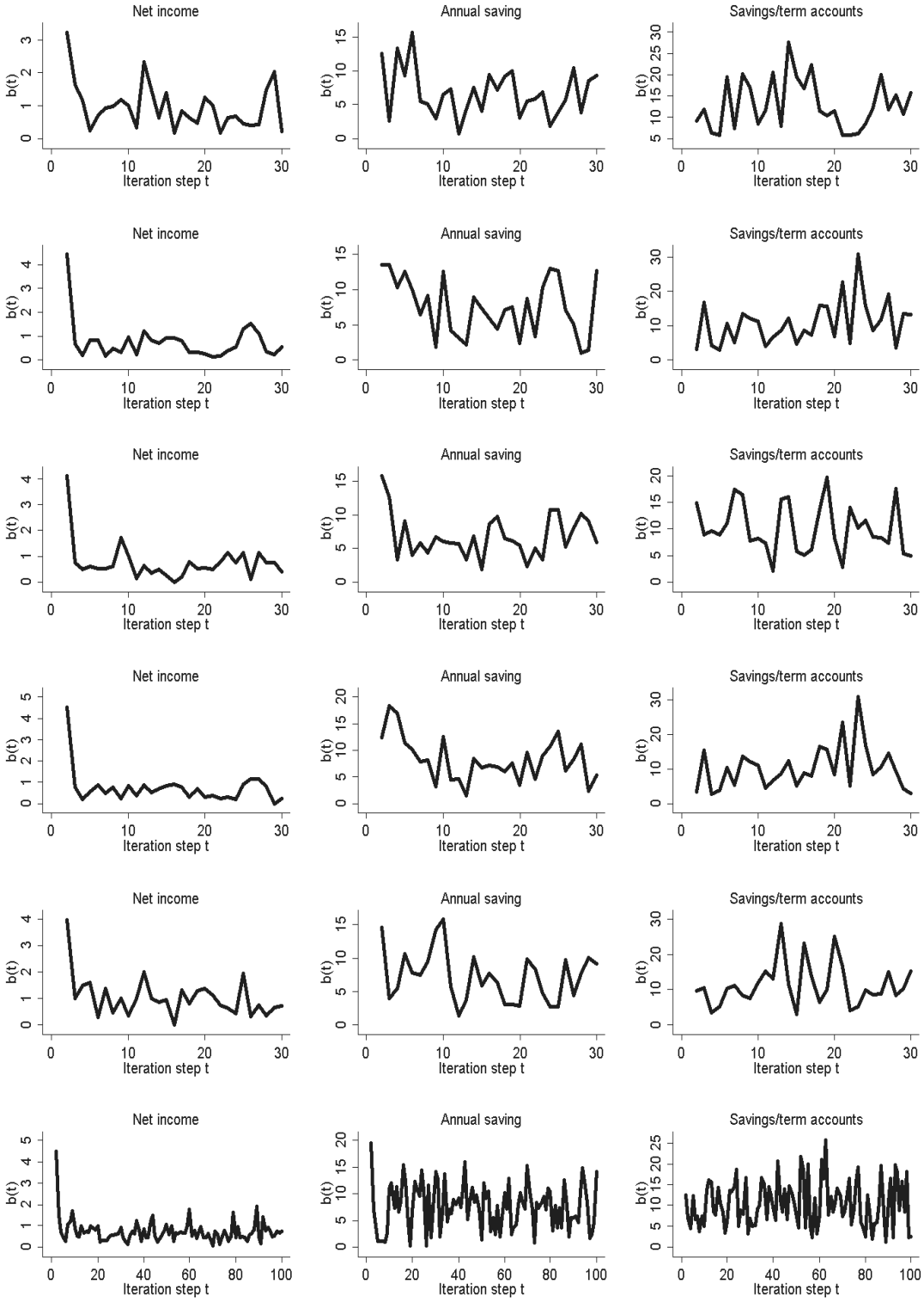
$$b(t) = \sqrt{\left(\begin{bmatrix} Q50_t^Y \\ (Q75-Q25)_t^Y \end{bmatrix} - \begin{bmatrix} Q50_{t-1}^Y \\ (Q75-Q25)_{t-1}^Y \end{bmatrix}\right)' \left(\begin{bmatrix} Q50_t^Y \\ (Q75-Q25)_t^Y \end{bmatrix} - \begin{bmatrix} Q50_{t-1}^Y \\ (Q75-Q25)_{t-1}^Y \end{bmatrix}\right)} \qquad (7)$$

Here, Q25, Q50, and Q75 denote the 25th, 50th, and 75th quantile, respectively, of the particular distribution of imputed values. As long as the process converges, b_t has a downward trend. As soon as the process has converged, $b(t)$ should not exhibit a trend any more. Figure 2 shows b_t for the three variables that are used for convergence diagnosis. As before, five iteration runs are shown for *t = 1,..., 30* and one run is shown for *t = 1,..., 100*. The figures reveal convergence for the *net income* variable, and some indication for convergence of the *annual saving* variable, which is, however, not really convincing.

Overall, the findings from the two convergence diagnostics presented above suggest relatively quick convergence of the algorithm on the *net income* variable, and mixed evidence for the *annual saving* variable. The convergence properties of the algorithm have been investigated on all other core variables. No indication for divergent behavior or long-term drift has been found, in all cases, $s(t)$, $e(t)$, and $b(t)$ are stationary after few iteration steps and no autocorrelation is present in $s(t)$, $e(t)$, and $b(t)$. However, $s(t)$ and $b(t)$ do *not* exhibit a clear downward trend for many variables in the early iteration steps; that is, they are stationary from the first iteration step on. The variable s*avings and term accounts* which is displayed in the presented figures, is an example of such a variable. This result, which is also mentioned by Kennickell (1998), suggests that those variables have essentially converged in the first iteration step; i.e. convergence has already been achieved in the first prediction step which has served to generate the starting values for the iteration.

Note, that iteration runs with t = 1000, which are not displayed graphically in this paper, have also been analyzed for both suggested convergence criteria; as well, the corresponding autocorrelation functions have been investigated. The findings show that even longer iteration procedures do not achieve better convergence results based on the presented diagnostics; in particular, no autocorrelation at longer lags is found.

Figure 2: Convergence diagnostics: *b(t)* displayed for three key variables.



Note: Values b(t) are divided by 100.

Overall, the results are in line with findings based on the iterative algorithm implemented for the imputation of the Survey of Consumer Finances (Kennickell, 1998). Kennickell reports quick convergence on key variables, the algorithm is run for 6 iteration steps overall. Given the findings about convergence in this section, MIMS is run for 20 iteration steps.

4.2 Observed, Imputed, and Complete Data

This subsection has two main purposes: First, the reader should get an impression of the differences across the five imputed and across the five complete data implicates. For this reason, the following tables report descriptive statistics of key financial variables for all five implicates. Second, the section presents and briefly discusses differences between the distributions of observed and imputed data. The section ends with a graphical comparison between observed and imputed data.

The following table 4 reports descriptive statistics for the observed data, for the five imputed implicates, and for the five complete data implicates (complete data implicates consist of observed *and* imputed data, i.e. the full rectangular data matrix). The means of all variables vary across complete data implicates and across imputed data implicates. Medians of all variables vary only across imputed data implicates, not across complete data implicates.[26] I first turn to the financial wealth variables. The table shows a consistent pattern for all financial wealth variables and for the saving variable: The mean of the imputed data is considerably higher than the mean of the observed data. This finding deserves further investigation.

[26] The fact that summary distributional characteristics, such as mean values, are similar across implicates is in line with our finding that the imputation for all 5 implicates – which have all started with different initial values for the imputed variables – have indeed converged, and not diverged. Again, longer simulations lead to similar results.

Table 4: Descriptive statistics for the observed data, for the 5 imputed implicates, and for the 5 complete data implicates.

	Observed data	Imputed data Implicate No.					Complete data Implicate No.				
		1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
						Net income [€]					
Mean	2,554	2,382	2,390	2,400	2,386	2,388	2,501	2,504	2,507	2,502	2,503
Median	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000
Min.	25	25	25	25	25	25	25	25	25	25	25
Max.	120,000	20,000	20,000	23,333	20,000	20,000	120,000	120,000	120,000	120,000	120,000
						Annual saving [€]					
Mean	2,624	5,453	5,336	5,624	5,553	5,784	2,948	2,940	2,971	2,970	2,994
Median	1,000	3,929	3,738	3,895	3,946	3,772	1,000	1,000	1,000	1,000	1,000
Min.	0	0	0	0	0	0	0	0	0	0	0
Max.	150,000	78,206	56,586	98,161	55,435	112,878	150,000	150,000	150,000	150,000	150,000
						Savings/term accounts [€]					
Mean	8,174	12,155	12,272	11,755	12,274	12,129	9,068	9,094	8,978	9,094	9,062
Median	500	10,784	11,176	10,360	11,628	10,436	2,000	2,000	2,000	2,000	2,000
Min.	0	0	0	0	0	0	0	0	0	0	0
Max.	1,000,000	88,897	95,767	116,545	81,713	143,290	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000
						Building society savings agreements [€]					
Mean	1,775	3,917	3,873	3,755	3,726	3,907	2,124	2,117	2,098	2,093	2,122
Median	0	1,844	1,805	1,528	1,671	1,972	0	0	0	0	0
Min.	0	0	0	0	0	0	0	0	0	0	0
Max.	100,000	54,442	64,057	58,061	66,725	68,408	100,000	100,000	100,000	100,000	100,000

Note: All calculations are unweighted.

Table 4 (continued)

	Observed data	Imputed data Implicate No.					Complete data Implicate No.				
		1	2	3	4	5	1	2	3	4	5
Whole life insurance policies [€]											
Mean	5,042	13,881	14,333	14,393	13,981	13,821	6,813	6,904	6,916	6,833	6,801
Median	0	9,970	10,793	10,380	9,840	9,922	0	0	0	0	0
Min.	0	0	0	0	0	0	0	0	0	0	0
Max.	500,000	196,235	189,196	224,734	198,699	203,240	500,000	500,000	500,000	500,000	500,000
Bonds [€]											
Mean	1,644	10,237	10,459	11,364	11,291	10,915	2,625	2,650	2,754	2,745	2,702
Median	0	0	0	0	0	0	0	0	0	0	0
Min.	0	0	0	0	0	0	0	0	0	0	0
Max.	1,000,000	316,511	349,122	345,260	403,173	380,301	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000
Shares & real-estate funds [€]											
Mean	3,857	8,511	8,350	8,618	8,291	8,460	4,555	4,531	4,571	4,522	4,547
Median	0	526	558	962	925	845	0	0	0	0	0
Min.	0	0	0	0	0	0	0	0	0	0	0
Max.	18,000,000	250,392	264,187	249,577	270,813	277,270	1,800,000	1,800,000	1,800,000	1,800,000	1,800,000
Owner occupied housing [€]											
Mean	123,280	44,800	43,388	40,108	38,672	43,639	111,710	111,501	111,018	110,806	111,538
Median	0	0	0	0	0	0	0	0	0	0	0
Min.	0	0	0	0	0	0	0	0	0	0	0
Max.	5,000,000	934,811	1,159,067	1,243,168	876,550	1,248,193	5,000,000	5,000,000	5,000,000	5,000,000	5,000,000

Note: All calculations are unweighted.

For this purpose, table 5 gives information on the imputation of asset variables by showing the results of the ownership imputation. The first column of the table shows the asset ownership rates for those who answer the ownership question, the following columns show imputed ownership rates for all implicates. It is found that except for the item *savings and term accounts*, ownership rates among nonrespondents are in fact lower than ownership rates among respondents. Both findings, namely that the imputation overall leads to higher means for financial asset variables (table 4) but at the same time generates lower ownership rates for financial assets (table 5) is in line with findings by Hoynes et al. (1998) who use a *non*-iterative regression-based single imputation method.[27]

Table 5: Percentage of households owning assets: Observed values and 5 imputed implicates

	Observed data	**Imputed data Implicate No.**				
		1	**2**	**3**	**4**	**5**
Savings/term accounts	60.8	70.5	70.5	70.5	70.5	70.5
Building society savings agreements	27.8	15.5	16.7	15.1	14.7	15.9
Whole life insurance policies	30.4	17.1	16.7	16.7	17.5	17.5
Bonds	8.8	2.0	2.4	1.6	2.4	2.0
Shares & real-estate funds	19.8	9.2	9.2	10.0	9.6	9.2
Owner occupied housing	48.6	35.9	37.6	36.8	35.0	36.8

Note: All calculations are unweighted.

It can be concluded that, for most financial asset items, the included conditioning variables shift the distribution to higher values for financial wealth on average, compared to the original distribution of observed values, which would simply be replicated if no conditioning variables were used. The findings by Smith (1995), who reports that the

[27] Hoynes et al. (1998) find higher mean values for all complete nonrespondents on all comparable financial asset variables. They also find lower imputed ownership rates than observed ownership rates on all financial asset variables, except from the item "bonds" and the item "checking and savings accounts". For these items, they find imputed ownership rates that are similar to the observed rates. A more detailed comparison with results from other imputation procedures would be of high interest at this point. To the author's knowledge, however, a systematic evaluation of the effect of the imputation on the distribution of different wealth components is only presented in the paper by Hoynes et al. (1998). Further methodological insights about the impact and relevance of an iterative procedure could be obtained from comparing an application of the Hoynes et al. (1998)-procedure and the MIMS-procedure to the same dataset.

effect of follow-up brackets to open-ended financial wealth questions in the HRS is a substantial increase in mean wealth, go into the same direction.

In contrast to the findings concerning the financial wealth variables, table 4 shows that the mean of imputed values of owner occupied housing are lower than observed values. How are home ownership and owner-occupied housing values distributed across observed and imputed values? Table 5 has already shown that – according to the imputation – the fraction of homeowners, i.e. households with a positive value for owner-occupied housing[28], is considerably lower among nonrespondents than among respondents. Table 6 serves to further investigate the difference between the observed and the imputed distribution of the value of owner occupied housing. Each column of the table gives the percentage distribution of home values for homeowners across four categories. The table shows that households that did not answer the corresponding question are more likely to occupy real estate with a low value. Interestingly, the results on home-ownership and owner-occupied housing values are again in line with findings by Hoynes et al. (1998), who report that those with incomplete responses on the housing questions have characteristics that make them more likely to be renters, and – given that they are homeowners – it makes them more likely to have low values for real estate.[29]

Table 6: Distribution of owner-occupied housing values for homeowners (percent).

	Observed data	**Imputed data**				
		Implicate No.				
Range (1,000 €)		**1**	**2**	**3**	**4**	**5**
0 - 49.9	9.0	12.9	13.7	14.9	14.1	14.9
50 - 99.9	8.3	12.9	16.8	17.0	18.5	8.5
100 - 199.9	29.3	26.9	27.4	31.9	26.1	34.0
> 200	53.4	47.3	42.1	36.2	41.3	42.6

Note: All calculations are unweighted.

[28] Of course, one can argue that the fraction of homeowners is not equal to the fraction of households with a positive value for owner-occupied housing, since it can also be the case that respondents own real estate and answer that its value is zero. In fact, about 5% of the respondents that report owning real estate give a value of zero in the follow-up question. In all tables above, these respondents are counted as homeowners.

[29] While the purpose of this paper is not to investigate the relationship between item nonresponse to certain questions and socio-economic characteristics, the above findings are interesting in this respect: They suggest that nonrespondents to questions about housing might have other socio-economic characteristics than nonrespondents to the financial wealth questions. A multivariate analysis indeed finds some evidence for this hypothesis.

Finally, I turn to the findings for the *net income* variable. Though medians are identical for imputed and observed values, the mean of monthly net income is lower for the imputed than for the observed values (table 4). For further investigation, table 7 compares the distribution of net income values between imputed and observed data. No substantial difference in the net income distributions of both groups is observable. The reason for the finding that the mean of monthly net income is lower for the imputed than for the observed values are a few extreme values in the observed distribution of monthly net income: If the observed distribution of monthly net income values is trimmed such that the top 0.5-percentile is left out (corresponding to 10 observations that reported having a net income between 26,000 € and 120,000 € per month), a mean monthly net income value of 2,306 € is found. This value is lower than the mean monthly net income of the imputed observations of all five implicates (see table 4); on average by about 83 €.

Table 7: Distribution of monthly net income (percent)

	Observed data	**Imputed data**				
		Implicate No.				
Range (1,000 €)		**1**	**2**	**3**	**4**	**5**
0 - 0.9	13.3	14.1	13.9	14.2	13.8	14.0
1 - 1.99	34.4	33.7	33.3	33.1	33.8	33.1
2 - 2.99	28.3	29.4	29.8	29.3	29.5	30.0
3 - 3.99	13.9	11.8	12.3	12.2	12.1	11.9
4 - 4.99	4.8	6.2	5.9	6.4	6.2	6.2
5 - 6.99	2.8	2.4	2.3	2.2	2.1	2.4
> 7	2.5	2.4	2.5	2.6	2.5	2.4

Note: All calculations are unweighted.

Overall, it is found that MIMS does not have a strong effect on the distribution of income values in SAVE. In contrast, findings from a regression-based single imputation procedure of annual income variables for the SOEP suggest that item nonresponse on income appears to be selective with respect to both tails of the income distribution (Frick and Grabka, 2005); the overall effect of their imputation is an increase in the mean of after-tax income by 1.7%.

To further illustrate the effects of imputation, figure 3 presents kernel density estimates of observed and imputed values for the above mentioned financial variables. The kernel density is estimated for positive values of the variables that have been analyzed above, an Epanechnikov kernel and Silverman's rule of thumb (Silverman, 1986) for bandwidth

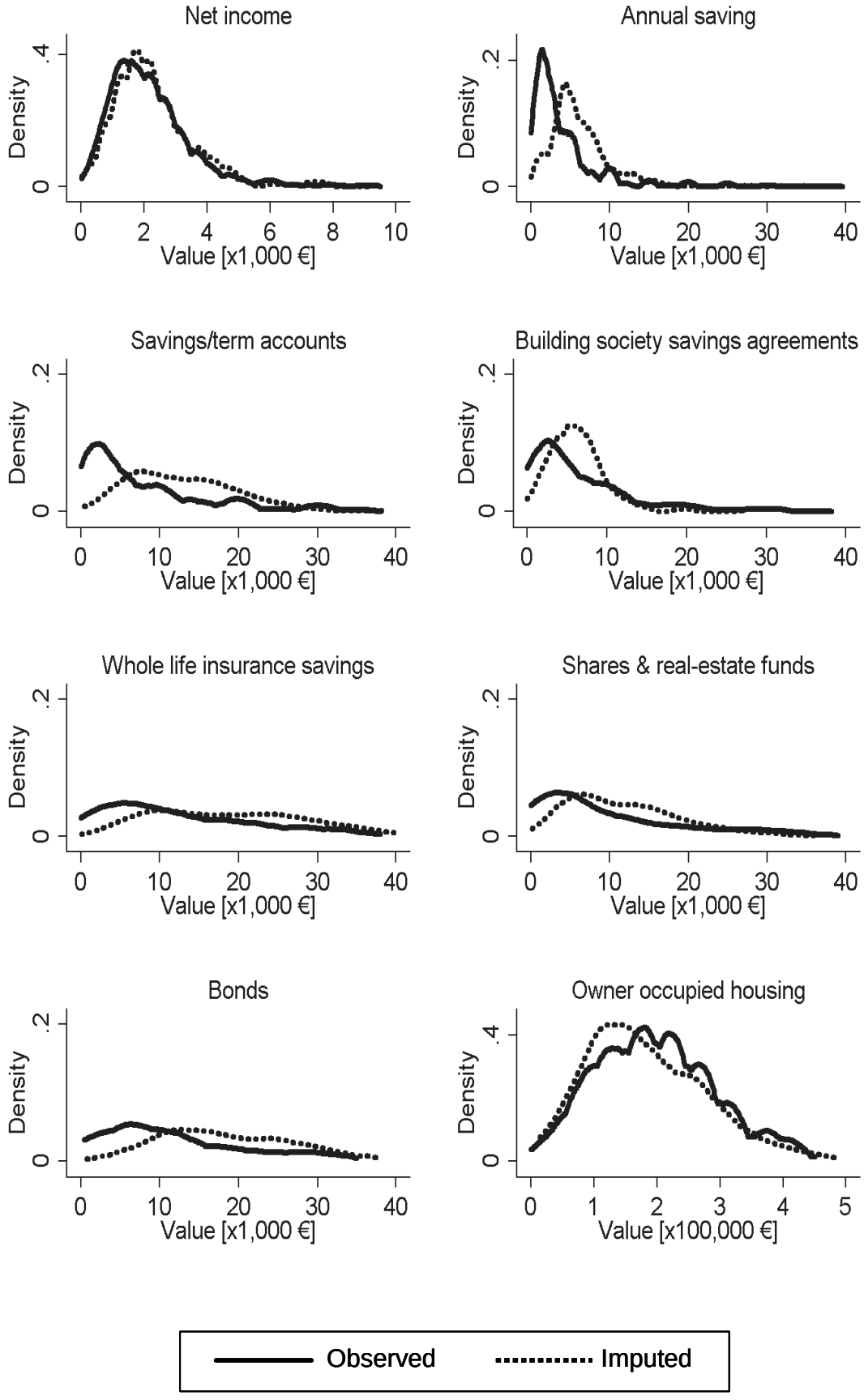
selection have been used. Kernel density estimates for the imputed data are usually obtained using Rubin's (1987) method to combine the data from the five implicates before the density estimation. According to Rubin (1987)[30], the overall imputed value $\overline{Y_i}$ of variable Y for a certain observation i is simply the average over the individual five imputed values, $m = 1,..., 5$, that is:

$$\overline{Y_i} = \frac{1}{5}\sum_{m=1}^{5} Y_{i,m}. \qquad (8)$$

In addition to the discussed findings concerning mean financial wealth differences between imputed and observed values, the figures illustrate nicely that the inclusion of covariates has a substantial effect on the distribution of asset holdings, a conclusion that is also emphasized by Hoynes et al. (1998). For the variables *annual saving* and *owner occupied housing*, the effect of focal point answers on the density is clearly visible: For example, the leftmost spike in the distribution of *annual saving* is due to the large amount of households reporting a total amount of annual saving of exactly 1,000 €. The second "spike" (or better: "plateau") stems from all households reporting 5,000 €. This multimodality is not replicated by the distribution of the imputed data, and it is debatable whether it should be replicated. One way of replicating multimodality would be to additionally use a nearest neighbor procedure after the regression-based imputation. For reasons given above, MIMS uses a nearest neighbor procedure only for variables that have follow-up brackets.

[30] Rubin (1987) derives general methods for combining the information from multiply imputed datasets. A brief summary of these methods, given in the appendix of this paper, section 6.2, informs the reader about how to work with the multiply imputed SAVE data.

Figure 3: Density functions of observed and imputed values.



Note: All calculations are unweighted.

5 Discussion and Conclusion

Except for controlled experimental settings, survey studies about human past and intended behavior rarely generate complete information. For several reasons that have been discussed in this paper, it is however desirable to provide users with a complete dataset in which all missing values have been imputed.

Missing values are rarely known with certainty. To be able to reflect the uncertainty of missing data in subsequent analyses, multiple imputation is used for the SAVE survey. This goal of this paper is to present the key theoretical underpinnings of a Markov Chain Monte Carlo multiple imputation algorithm, to describe and document the practical application of such a multiple imputation algorithm to the SAVE data, and to present and discuss properties of the algorithm as well as the resulting imputed datasets.

The Markov Chain Monte Carlo technique that is used for the algorithm presented in this paper is similar to the method presented in Schafer (1997) who uses smaller datasets with few conditioning variables, and it is similar to the method presented in Barceló, (2006) and in Kennickell (1998), who apply an iterative method to data from two large scale socio-economic surveys. It is important to note that modifications of this implementation are conceivable and should be explored: For example, the sequential simulation algorithm can be modified such that each draw from a certain conditional distribution depends not only on the conditional distribution estimated in the preceding iteration step, but also on conditional distributions estimated in earlier iteration steps (Cameron and Trivedi, 2005). Alternatively, in each iteration step the distribution of unobserved values can be simulated a certain number of times p, and the parameter values for the next iteration step can then be estimated from all p simulated distributions; this means that multiple versions of the unobserved data are generated from the predictive distribution in one iteration step. A comparison of convergence properties between these different ways of implementing the data augmentation algorithm would certainly be helpful. Considering the fact that the method proposed in this paper is based on the assumption of ignorable missing data, future research efforts should also be directed towards modeling the missing data mechanism explicitly and eventually a model should be formulated for each incomplete survey variable and for the corresponding mechanism of missingness. Particularly given the complexity of the nonresponse patterns in SAVE, this constitutes a substantial effort. A comparison with the results obtained from MIMS would be of highest scientific interest.

So far, convergence properties of MCMC methods have only been systematically analyzed on simulated datasets and datasets with fewer variables compared to the large household survey that is analyzed in this paper (see, e.g., Schafer, 1997). The findings of the present study suggest that the algorithm converges in only few iteration steps. For most variables, the process is stationary after not more than about 5-10 iterations steps. For all other variables, it is stationary from the first iteration step on, suggesting that the algorithm has already converged in the first iteration step – a phenomenon that is also reported by Kennickell (1998). It is certainly worth investigating the convergence properties of MCMC algorithms in the context of large surveys or large simulated datasets in a collaborative effort and with standardized methods. This will further contribute to a more comprehensive evaluation of the relevance of MCMC methods for survey research.

Finally, a comparison between imputed and observed values has revealed that the use of covariates in the imputation process has a substantial effect on the distributions of individual asset holdings. In general, these effects are similar to the effects reported based on other techniques. This finding suggests that item nonresponse is not occurring randomly but is related to the included covariates. The analyses also suggest that there might be differences in the character of nonresponse across asset types, and they indicate specific directions for future research on the relationship between socio-economic characteristics and nonresponse to specific items. Furthermore, from the point of view of survey methodology and data quality management which is of ultimate interest for every researcher and policy maker, the findings underline the need for an ongoing scientific discussion about imputation. In particular, this discussion will have to do with the effects of different imputation strategies on the distribution of data obtained in large-scale socio-economic surveys as well as with a systematic exploration of the feasibility of different imputation methods.

6 Appendix

6.1 Derivation of the Variance of a Normally Distributed Random Variable that is Symmetrically Censored

Consider a normally distributed random variable y^* with mean zero and standard deviation σ:

$$y^* \sim N(0,\sigma) \tag{A1}$$

Alternatively, with $\varphi(\cdot)$ being the density function of the standard normal distribution, we can write:

$$y^* \sim \frac{1}{\sigma}\varphi\left(y^*/\sigma\right) \tag{A2}$$

We now define a new random variable y, which is obtained from the original one, y^*, by symmetrically censoring the variable y:

$$y = \begin{cases} -a & if \quad y^* < -a \\ y^* & otherwise \\ a & if \quad y^* > a \end{cases} \tag{A3}$$

This variable has the following density function:

$$f(y) = \begin{cases} 0 & if \quad |y^*| > a \\ \Phi\left(-a/\sigma\right) = 1 - \Phi\left(a/\sigma\right) & if \quad |y^*| = a \\ \frac{1}{\sigma}\varphi\left(y^*/\sigma\right) & if \quad |y^*| < a \end{cases} \tag{A4}$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of $\varphi(\cdot)$.

This distribution is a mixture of discrete and continuous parts. It is the variance of the random variable y that we want to calculate as a function of the censoring value *a*.

In order to do so, I use the variance decomposition formula:

$$Var(y) = E(Var[y \mid a]) + Var(E[y \mid a]) \tag{A5}$$

I compute the first term on the right-hand side, then the second term on the right-hand side, and then combine the two results.

(a) Computation of $E(Var[y \mid a])$*:*

The expected value of the conditional variance of *y*, given the censoring value *a*, can be decomposed as follows:

$$E(Var[y \mid a]) = 2\Phi\left(-a/\sigma\right) \cdot Var[y \mid y = a] + \left[1 - 2\Phi\left(-a/\sigma\right)\right] \cdot Var[y \mid |y| < a] \tag{A6}$$

It is obvious that $Var[y \mid y = a] = 0$.

That is, $Var[y \mid |y| < a]$ remains to be computed, and it is known that $Var[y \mid |y| < a] = Var[\, y^* \mid |y^*| < a]$.

$Var[\, y^* \mid |y^*| < a]$ can be decomposed as follows:

$$\begin{aligned} Var[y^* \mid |y^*| < a] &= \int_{-a}^{a} y^{*2} \varphi(y^*) dy^* \\ &= \int_{-\infty}^{\infty} y^{*2} \varphi(y^*) dy^* - \int_{-\infty}^{-a} y^{*2} \varphi(y^*) dy^* - \int_{a}^{\infty} y^{*2} \varphi(y^*) dy^* \\ &= \sigma^2 - 2\int_{a}^{\infty} y^{*2} \varphi(y^*) dy^* \\ &= \sigma^2 - 2Var[y^* \mid y^* > a] \end{aligned} \tag{A7}$$

$Var[y^* \mid y^* > a]$ is the variance of a truncated normally distributed variable. This variance is computed as follows (see Johnson and Kotz, 1970):

$$Var[y^* \mid y^* > a] = \sigma^2(1 - \delta(a)), \tag{A8}$$

where

$\delta = \lambda(a)\left(\lambda(a) - a/\sigma\right)$, and $\lambda(a) = \dfrac{\varphi\left(a/\sigma\right)}{1 - \Phi\left(a/\sigma\right)}$.

It follows:

$$\begin{aligned} Var[y^* \mid |y^*| < a] &= \sigma^2 - 2\sigma^2(1 - \delta(a)) \\ &= \sigma^2(1 - 2 + 2\delta(a)) \\ &= \sigma^2(2\delta(a) - 1) \end{aligned} \tag{A9}$$

And therefore:

$$E(Var[y \mid a]) = \left[1 - 2\Phi\left(-a/\sigma\right)\right] \cdot \sigma^2(2\delta(a) - 1) \tag{A10}$$

(b) Computation of $Var(E[y \mid a])$*:*

We find:

$$\begin{aligned} Var(E[y \mid a]) &= 2\Phi\left(-a/\sigma\right)\cdot\{a - E[y]\}^2 + \left[1 - 2\Phi\left(-a/\sigma\right)\right]\cdot\{E[y \mid |y| < a] - E(y)\}^2 \\ &= 2\Phi\left(-a/\sigma\right)\cdot a^2, \end{aligned} \tag{A11}$$

since $E[y] = a$, $E[y \mid |y| < a] = 0$, and $E(y) = 0$ by symmetry arguments.

Combining the results of *(a)* and *(b)* finally yields the expression for the variance of a symmetrically censored normally distributed variable, with mean zero, standard deviation σ and censoring value *a*:

$$Var(y) = 2\Phi\left(-a/\sigma\right)\cdot a^2 + \left[1 - 2\Phi\left(-a/\sigma\right)\right]\cdot\sigma^2(2\delta(a) - 1) \tag{A12}$$

6.2 Rules for Inference Based on Multiply Imputed Datasets

The 5 implicates of the SAVE data can be analyzed using standard complete data methods. Every model has to be estimated 5 times, once for each complete and imputed dataset. The results across these estimations vary, this reflects the missing-data uncertainty. Rubin (1987) has derived a method for combining the results from a data analysis performed *M* times, once for each of *M* imputed data sets, to obtain a single set of results: Suppose that $\hat{Q}_m$ is the scalar point estimate of interest, obtained from data set *m*. Suppose further that $\hat{U}_m$ is the variance estimate associated with $\hat{Q}_m$.The overall estimate is then the average of the individual estimates,

$$\overline{Q} = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m. \tag{A13}$$

For the overall standard error, one must first calculate the within-imputation variance,

$$\overline{U} = \frac{1}{M}\sum_{m=1}^{M}\hat{U}_m \tag{A14}$$

and the between-imputation variance,

$$\overline{B} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}_m - \overline{Q})^2. \tag{A15}$$

The total estimated variance of the multiple-imputation point estimate is then

$$T = \overline{U} + (1 + \frac{1}{M})\overline{B}. \tag{A16}$$

Single imputation underestimates the standard errors of the estimates because it has zero between imputation variance.

Additional methods for combining the results from multiply imputed data that hold under certain special assumptions about the data are presented in Schafer (1997).

7 References

Barceló, C. (2006): Imputation of the 2002 wave of the Spanish Survey of Household Finances (EFF). Occasional Paper No. 0603, Bank of Spain.

Beatty, P. and D. Herrmann (2002): To answer or not to answer: Decision processes related to survey item nonresponse. In: R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (Eds.), *Survey Nonresponse*, 71-85. New York: Wiley.

Bernaards, C. A., M. M. Farmer, K. Qi, G. S. Dulai, P. A. Ganz, and K. L. Kahn (2003): Comparison of Two Multiple Imputation Procedures in a Cancer Screening Survey. *Journal of Data Science,* 1 (3), 293-312.

Biewen, M. (2001): Item non-response and inequality measurement: Evidence from the German earnings distribution. *Allgemeines Statistisches Archiv*, 85(4), 409-425.

Börsch-Supan, A. and L. Essig (2005): Household saving in Germany: Results of the first SAVE study. In: D. A. Wise (Ed.), *Analyses in the Economics of Aging*, 317-352. Chicago: The University of Chicago Press.

Bover, O. (2004): The Spanish Survey of Household Finances (EFF): Description and Methods of the 2002 Wave. *Documentos Ocasionales N. 0409*. Banco de Espana.

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics. Methods and Applications*. New York: Cambridge University Press.

Essig, L. and J. Winter (2003): Item Nonresponse to Financial Questions in Household Surveys: An Experimental Study of Interviewer and Mode Effects. *MEA-Discussion Paper 39-03*, MEA – Mannheim Research Institute for the Economics of Aging. University of Mannheim.

Ezzati-Rice, T. M., W. Johnson, M. Khare, R. J. A. Little, D. B. Rubin, and J. L. Schafer (1995): Multiple imputation of missing data in NHANES III. *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, 459-487.

Ferber, R. (1966): Item nonresponse in a consumer survey. *Public Opinion Quarterly*, 30 (3), 399-415.

Frick, J. R. and M. M. Grabka (2005): Item nonresponse on income questions in panel surveys: Incidence, imputation and the impact on inequality and mobility. *Allgemeines Statistisches Archiv,* 90 (1), 49-62.

Geman, S. and D. Geman (1984): Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6 (6), 721-741.

Graham, J. W. and J. L. Schafer (1999): On the performance of multiple imputation for multivariate data with small sample size. In: R. Hoyle (Ed.), *Statistical Strategies for Small Sample Research*, 1-29, Thousand Oaks, CA: Sage.

Groves, R. M., D. A. Dillman, J. L. Eltinge, and R. J. A. Little (2002): Survey nonresponse. New York: Wiley.

Hartley, H. O. and R. R. Hocking (1971): The analysis of incomplete data. *Biometrics*, 27, 783-808.

Hastings, W. K. (1970): Monte Carlo Sampling Methods Using Markov Chain and Their Applications. *Biometrika*, 57, 97–109.

Hersch, J. (2000): Gender, Income Levels, and the Demand for Cigarettes. *Journal of Risk and Uncertainty*, 21, 263-282.

Hersch, J. and W. K. Viscusi (1990): Cigarette Smoking, Seatbelt Use, and Differences in Wage-Risk Tradeoffs. *The Journal of Human Resources*, 25(2), 202-227.

Hoynes, H., M. Hurd, and H. Chand (1998): Household Wealth of the Elderly under Alternative Imputation Procedures. In: D. A. Wise (Ed.), *Inquiries in the Economics of Aging*, 229-257. Chicago: The University of Chicago Press.

Johnson, N. and S. Kotz (1970): Distributions in Statistics – Continuous Univariate Distributions. Vol. 2. New York: Wiley.

Juster, F. T. and J. P. Smith (1997): Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association*, 92 (440), 1268-1278.

Kalwij, A. and A. van Soest (2005): Item Non-Response and Alternative Imputation Procedures. In: A. Börsch-Supan and H. Jürges (Eds.), *The Survey of Health, Ageing and Retirement in Europe – Methodology,* 128-150. Mannheim: Mannheim Research Institute for the Economics of Aging.

Kennickell, A. B. (1997): Using range techniques with CAPI in the 1995 Survey of Consumer Finances. Board of Governors of the Federal Reserve System, Washington, D.C.

Kennickell, A. B. (1998): Multiple Imputation in the Survey of Consumer Finances. *Proceedings of the 1998 Joint Statistical Meetings*, Dallas TX.

Levine, P. B., T. A. Gustafson, and A. D. Velenchik (1997): More Bad News for Smokers? The Effects of Cigarette Smoking on Wages. *Industrial and Labor Relations Review*, 51, 493-509.

Li, K. (1988): Imputation Using Markov Chains. *Journal of Statistical Computing and Simulation*, 30, 57-79.

Li, K., T. Raghunathan, and D. Rubin (1991): Large sample significance levels from multiply-imputed data using moment-based statistics and an *F* reference distribution. *Journal of the American Statistical Association,* 86, 1065–1073.

Little, R. J. A. and D. B. Rubin (2002): Statistical Analysis with Missing Data. New York: Wiley.

Little, R. J.A. and T. Raghunathan (1997): Should Imputation of Missing Data Condition on All Observed Variables? *Proceedings of the Section on Survey Research Methods*, Joint Statistical Meetings, Anaheim, California.

Little, R. J. A., I. G. Sande, and F. Scheuren (1988): Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics,* 6 (3), 117-131.

Manski, C. (2005): Partial Identification with Missing Data: Concepts and Findings. *International Journal of Approximate Reasoning,* 39 (2-3), 151-165.

Rässler, S. and R. Riphahn (2006): Survey item nonresponse and its treatment. *Allgemeines Statistisches Archiv*, 90, 217 – 232.

Riphahn, R. and O. Serfling (2004): Item Non-response on Income and Wealth Questions. *Empirical Economics,* 30 (2), 521-538.

Rubin, D. B. (1987): Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D. B. (1996): Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91 (434), 473-489.

Rubin, D. B. and N. Schenker (1986): Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, 81 (394), 366-374.

Schafer, J. L. (1997): Analysis of incomplete multivariate data. London: Chapman & Hall.

Schunk, Daniel (2006): The German SAVE-Survey: Documentation and Methodology. *MEA-Discussion-Paper 109-2006*. Universität Mannheim.

Schunk, Daniel (2007): A Markov Chain Monte Carlo Algorithm for Multiple Imputation in Large Surveys**.** *Advances in Statistical Analysis.* Forthcoming in 2007.

Schräpler, J.-P. (2003): Gross income non-response in the German Socio-Economic Panel: Refusal or don't know? *Schmollers Jahrbuch*, 123, 109-124.

Schräpler, J.-P. and G. G. Wagner (2001): Das Verhalten von Interviewern - Darstellung und ausgewählte Analysen am Beispiel des "Interviewerpanels" des Sozio-Ökonomischen Panels. *Allgemeines Statistisches Archiv,* 85, 45-66.

Silverman, B. W. (1986): Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

Smith, J. P. (1995): Racial and Ethnic Differences in Wealth. *Journal of Human Resources,* 30, 158-183.

Tanner, M. A. and W. H. Wong (1987): The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82 (398), 528-550.

Nr.	Author	Title
07-08	Daniel Schunk	The German SAVE survey: documentation and methodology
07-07	Hans-Martin von Gaudecker Carsten Weber	Mandatory unisex policies and annuity pricing: quasi-experimental evidence from Germany
07-06	Daniel Schunk	A Markov Chain Monte Carlo Multiple Imputation Procedure for Dealing with Item Nonresponse in the German SAVE Survey
07-05	Hans-Martin von Gaudecker Rembrandt Scholz	Lifetime Earnings and Life Expectancy
07-04	Christopher Koch Daniel Schunk	The Case for Limited Auditor Liability - The Effects of Liability Size on Risk Aversion and Ambiguity Aversion
07-03	Siegfried K. Berninghaus Werner Gueth M. Vittoria Levati Jianying Qiu	Satisficing in sales competition: experimental evidence
07-02	Jannis Bischof Michael Ebert	Inconsistent measurement and disclosure of non-contingent financial derivatives under IFRS: A behavioral perspective
07-01	Jörg Oechssler Carsten Schmidt Wendelin Schnedler	Asset Bubbles without Dividends - An Experiment
06-16	Siegfried K. Berninghaus Hans Haller	Pairwise Interaction on Random Graphs
06-15	Markus Glaser Philipp Schmitz	Privatanleger am Optionsscheinmarkt
06-14	Daniel Houser Daniel Schunk Joachim Winter	Trust Games Measure Trust
06-13	Markus Glaser Sebastian Müller	Der Diversification Discount in Deutschland: Existiert ein Bewertungsabschlag für diversifizierte Unternehmen?