Psychology in econometric models: conceptual and methodological foundations

Thum, Anna-Elisabeth

20 December 2013

Psychology in Econometric Models: Conceptual and Methodological Foundations

Anna-Elisabeth Thum

December 20, 2013

Abstract

Personality, ability, trust, motivation and beliefs determine outcomes in life and in particular those of economic nature such as finding a job or earnings. A problem with this type of determinants is that they are not immanently objectively quantifiable and that there is no intrinsic scale - such as in the case of age, years of education or wages. Often we think of these concepts as complex and several items are needed to capture them. In the measurement sense, we dispose of a more or less noisy set of measures, which indirectly express and measure a concept of interest. This way of conceptualizing is used in latent variables modelling. I examine in this article in how far economic and econometric literature can contribute to specifying a framework of how to use latent variables in economic models. As a semiparametric identification strategy for models with endogeneous latent factors I propose to use existing work on identification in the presence of endogeneous variables and examine which additional assumptions are necessary to apply this strategy for models with latent variables. I discuss several estimation strategies and implement a Bayesian Markov Chain Monte Carlo (MCMC) algorithm.

1 Introduction

On an intuitive level it is clear that personality traits matter in life and for economic outcomes. A more self-confident candidate might outperform a candidate with a higher graduation grade or a candidate with higher level of motivation can acquire the job. But on a theoretical and empirical level the relation between personality and outcomes is not as clear-cut. Personality psychology studies personality and economics studies economic outcomes but research in studying the effects of personality on economic outcomes is still full of controversies and no consensual model has yet been found. There is a body literature in both economic theory and in econometric applications taking the relation between personality and economic outcomes into account and aiming at conceptualizing it and giving it an empirical back-up. Borghans et al (2008) give a detailed account on the potential of integrating insights and methodologies from personality psychology in economic and econometric models.

An example of a field of research in economics in which an integration of methods from both economics and psychology is immigration. The integration of immigrants cannot really be reduced to a single dimension, as for instance the economic one. Immigrants face a new labour market which they need to understand and to which they need to adapt. This process needs to be taken into account for their economic integration. In particular, this article shall examine the link between the psychological dimension and the economic one (in form of labour market outcomes) in the integration of immigrants.

In this article I will examine the conceptual and methodological issues in studying the relationship between personality and economic outcomes. I will justify my choice of methodology with a view to the existing work and assess identification possibilities of a generalized form of the model.

I first outline the set up for a generalized form of the model to study the effect of psychological concepts on economic outcomes. After discussing the interpretation of latent variables in econometric models and their added value, I present possible estimation methods of models involving latent variables. A special section is devoted to the Markov Chain Monte Carlo method to estimate parametric models including latent variable models. I will then discuss identification possibilities and assess an existing semiparametric identification strategy for a model with endogeneity in its capacity to identify models with endogenous latent variables.

2 The Setting

To examine the effect of psychological concepts on directly measurable outcomes we propose the following model:

$$\begin{aligned}
D_i &= \{0,1\} \\
D_i^* &= \beta^D X_i + \alpha^D \theta_i + \varepsilon_i^D \\
M_i &= \{1,2,3\} \\
M_i^* &= \alpha^M \theta_i + \varepsilon_i^M \\
\theta_i &= \gamma W_i + \varepsilon_i^\theta
\end{aligned}$$

D_i is observable and signifies a discrete outcome such as the probability to be employed. It is modelled as a probit model with a latent underlying variable D_i^*. The outcome could just as well be continuous such as earnings. There is also a set of (for example tricategorical) observable categorical dependent variables M_i which signify the set of measures we use to measure the psychological concept. M_i is again modelled as an ordered probit model with a latent underlying variable M_i^*. Both D_i and M_i are affected by the psychological concept θ_i. α^D and α^M express the effect of the psychological concept on the outcome and on its measurements respectively. In addition to θ_i there are observable explanatory variables X_i determining the dependent variable. Taking account of the the fact that psychological concepts are not exogenously given but are affected by social background and by experiences in life, there are observable variables W_i determining the latent variable θ_i. W_i and X_i can contain the same elements but cannot be exactly equal. The model needs to satisfy normality of the errors $\varepsilon_i^M, \varepsilon_i^D, \varepsilon_i^\theta$ and conditional independence conditions between X, θ and ε, given W. I will discuss the assumptions of the model below.

3 Interpretation of Latent Variables in Models of Economic Outcomes

The interpretation and correct use of latent variables in econometrics is not a clear issue. In econometrics, the notion of "latent variables" or "latent factors" does not yet have a clear position, even though they are already found in applications. In macroeconomics and in the financial literature latent factors are often used to capture unobservable factors influencing financial markets[1]. In micro-econometrics, there are several articles using them to capture unobservable skills. The work by Carneiro, Hansen and Heckman (2003) is a prominent example. Latent variables have a more established role in psychometrics where they were initially used to measure intelligence. The initial model was supposed to extract a measure of intelligence from a set of questions, which were usually verbal and arithmetic tasks. Later their use was extended to personality psychology. A main difference between psychologists and economists in this context is that the economist is interested in outcomes and the role that a personality trait can play for its determination. Borghans et al (2008)[2] show the problems of using the psychometric latent variable approach in econometrics. They also give credit to the work of Heckman et al for addressing some of the problems and somewhat adopting the latent variable approach to econometrics.

A common problem in econometric analysis is the fact that the econometrician can only observe a part of the factors relevant for an economic problem of interest - the problem of endogenous covariates. The famous example is the "ability bias" in the returns to schooling literature : if we cannot observe separately the effect of an individual's ability on his earnings, it will be captured by the measured effect of the education variable (for example years of education) and education will be endogenous in an earnings equation if one cannot control for ability. Bowles, Gintis, Osborne (2001) extend this argumentation from ability as a cognitive skill to non-cognitive skills such as self-esteem or motivation. Bowles et al make it clear that even when controlling for ability in addition to conventional observable explanatory variables in an earnings regression there is still a large amount of relevant unobservable variation. They do this by calculating the size of variance unexplained by conventional observable factors and cognitive ability.

To resolve this problem we might seek a different variable,which can replace education, but is sufficiently correlated with it and not correlated with anything

[1] An example for the use of latent factors in macroeconomics is the work by Marco Lippi.

[2] Section III B of their paper gives account of the limits of the psychometric approach in economics.

unobservable and relevant for the dependent variable. This would be a valid instrument. If we know several instruments for one endogenous variable, we can use a linear projection of the instruments on the endogenous variable to be replaced. The more abstract the perturbing unobservable concept is, such as self esteem, the more difficult it can become to argue that there is an instrument not correlated to it but correlated to education. It could be easier to just control for the perturbing concept even if it is unobservable. In the following I am interested in assessing the potential of "latent factors" - unobservable but measurable concepts that enter the economic model - to address the problem of endogeneity.

3.1 Latent Variables in Psychometrics

An overview from a psychometric point of view is given in Rabe-Hesketh and Skondral (2004). Generally, two strands of modelling settings with the presence of latent variables are taken : factor models and item response theory. DeLeeuw and Takane (1987) show that the models are equivalent in a one-dimensional parametric setting, assuming normality in a two parameter logistic item response theory model[3].

3.1.1 Factor Models

Structure Factor models assume the following structure underlying a matrix of items:

$$M_{ij} = \Lambda_j f_i + \varepsilon_{ij}$$

That is, the observable variables M are linear in latent factors f.The effect of f on M is captured by the factor loadings Λ_j.i is the observational unit and can signify for example individuals in micro-econometrics or points in time in stock market models. j is the indicator for the number of observable items.

Additionally it is assumed,that

- factors f and error terms ε are orthogonal to each other
- $\Lambda_j f_i \perp \varepsilon_{ij}$
- f_i and ε_{ij} are typically assumed to be standard normally distributed, so M_{ij} is normally distributed
- M_{ij} is implicitly assumed to be continuous but can be discrete. In that case a latent variable M^*_{ij} (not to be confused with f_i) is assumed. Suitable cutoff points need to be specified.

To determine, whether it is possible to fit the data according to the model, the correlation matrix of items needs to be analyzed.

[3] The two parameter logistic model in item response theory is explained in section 1.3.3.2.

Interpretation To understand how to interpret the model we look at the covariance matrix of M dropping the subscripts:

$$cov(M) = \Lambda' cov(f)\Lambda + \Sigma_\varepsilon$$

In the special case of two observable variables and one underlying factor we can write:

$$\begin{aligned} m_1 &= \alpha_1 f + \varepsilon_1 \\ m_2 &= \alpha_2 f + \varepsilon_2 \end{aligned}$$

and for the variance-covariance matrix

$$\begin{bmatrix} \sigma_{11}^m & \sigma_{12}^m \\ \sigma_{21}^m & \sigma_{22}^m \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \sigma^f \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} + \begin{bmatrix} \sigma_1^\varepsilon & 0 \\ 0 & \sigma_2^\varepsilon \end{bmatrix}$$

This follows from the assumptions of linearity and independence of factors and errors of and between each other.

We can then write

$$\begin{aligned} \sigma_{11}^m &= \alpha_1^2 \sigma^f + \sigma_1^\varepsilon \\ & \quad etc \end{aligned}$$

Due to the standard normality of f and ε,we know that

$$\sigma_{11}^m = \alpha_1^2 + 1 \Longleftrightarrow \alpha_1^2 = \sigma_{11}^m - 1$$

So squared factor loadings show how much variance in the observable variable is explained by the latent factor.

To see another interpretation of the factor loadings, we can write

$$\begin{aligned} E(m_1\theta) &= E(\alpha_1 f + \varepsilon_1, f) \\ &= E(\alpha_1 f^2 + \varepsilon_1 f) \\ &= E(\alpha_1 f^2) \\ &= \alpha_1 E(f^2) \\ &= \alpha_1 \\ &= cov(m_1 f) \end{aligned}$$

That is, factor loadings can be interpreted as the covariance between the observable variables m and the latent variable f.

Furthermore, if we standardize m, we can write

$$r_{m_1,f} = cov(m_1 f)/\sqrt{\sigma_{11}^m \sigma_1^\varepsilon} = cov(m_1 f) = \alpha_1$$

That is, factor loadings can be interpreted as the correlation between the observable variables m and the latent variable f.

Exploratory vs Confirmatory Factor Analysis Depending on how much structure the researcher is able to impose, factor analysis can be confirmatory or exploratory. In exploratory analysis no assumption is made on the number of factors explaining a set of items. The aim of exploratory factor analysis is to explain the total variance (unique and common variance) in the set of items by the smallest possible number of factors. In confirmatory factor analysis an assumption following from theory about the number of factors is made. The aim of confirmatory factor analysis is to explain the common variance among the set of items by a supposed number of factors.

3.1.2 Item Response Theory

Item Response Theory originates in educational psychology and can be seen as the first version of factor analysis with discrete dependent variables. The most common item response models are the Rasch Model,the 2pl and the 3pl model. They all have a similar specification and assume additivity and a logistically distributed error term. In the 2pl model the probability to answer "1" to an educational test is given by

$$\Pr(Y_{ij} = 1) = \frac{\exp^{\alpha_j - \beta_j \theta_i}}{1 + \exp^{\alpha_j - \beta_j \theta_i}}$$

where θ_i is the score an individual has on a latent ability scale - it is considered as a continuous latent variable. α_j can be interpreted as an item difficulty parameter and β_j as the discrimination parameter. A probit link in this model is also possible assuming normally distributed errors.

Usually item response models are estimated by conditional likelihood, marginal likelihood or conditional likelihood. These methods resemble maximum likelihood estimation, but involve an additional step of integrating out the unknown parameters θ_i. The method suffers from problems of joint consistency when letting the number of persons and of items become infinitely large (see Douglas (1997)).

Exploratory vs confirmatory Analysis As in factor analysis, the researcher can choose between assuming a number of factors or testing for an adequate number of scales. The latter is called Mokken scaling and is based on a concept of a total score of ordered items (see Mokken (1971)).

3.2 Latent Variables in Econometrics

There is an acknowledged concept in econometrics, which is close to the concept of latent variables. Tom Wansbeek (2000) shows a relation between the latent variable concept and the concept of measurement error, a concept that has already been examined in econometric theory. Matzkin (2007) and others further develop this relation between the concept of latent variables and that of measurement error. The virtue of this relation is that econometrically relevant results for the concept of measurement error can be used for the analysis of latent variables.

Measurement error can cause an estimation bias because the independent variables might be endogenous if the measurement error is correlated with the error term of the model.

Assumptions in economics are usually motivated either by theory or by (previous) empirical observations in a similar context as the model of interest. Assumptions on the nature of the latent variable cannot be based on the latter criterion since these variables can obviously not be observed and at least in economics there is not much experience with latent variables yet. So there are currently two views on how to make assumptions in the field of latent variable modelling. One is to argue, that the variable is latent and therefore we are relatively free in our assumptions. For instance, the support or the variance of the latent variable can be argued to be assumed freely. The second view is to require theoretical backing of the assumptions. This backing can come from other sciences, such as psychology, genetics or neuroscience. For example, we may argue that latent ability is genetic and therefore exogenous. To unify these two points of view one could argue that assumptions on the latent variable itself may be in a sense arbitrary[4], whereas assumptions on the relation to other variables should be based on theory.

Why? Suppose for a moment we see the latent variable simply as one specific part of the variation in the data[5]. What can we say about this variation? It is the variation in the data of interest the econometrician does not take account of by observable variables. Now we are interested in extracting the part of this variation which bears some informativeness in the sense either that we are interested in it or that it contains a relation to a variable in the model and we need to control for it to get unbiased estimates. The characteristic of a "variation " being informative comes from its being relevant for explaining a different "variation" - this relevance can be interpreted as relations (correlations) with other variables. So we are interested in extracting a part of the unobservable variation, which is correlated or uncorrelated with other variables in such a way, that we can interpret this part of the variation based on its correlations. In other words, assumptions on the relations of elements of the unobservable noise lead to its interpretability and should therefore be guided by theory. Note that the *nature* of the relation is again unknown and poorly theoretically founded. Therefore, even if the *existence* of the relation is arguably theoretically founded, the *nature* of the relation should be inferred from statistical relationships. A suitable approach used in econometrics is to assume a nonparametric relation. We thereby do not impose a possibly ad hoc parametric functional form on the relation between the latent variable and the observable variables. Assumptions on the nature of this element of the latent unobservable variation, which satisfies certain independence assumptions, are less relevant for its interpretation.

[4] Albert and Chib (1993) argue in a similar way to motivate setting the variance of a latent underlying variable in an ordered response model to 1.

[5] In a sense,we follow Matzkin (2003), who argues that "exogenous variation" can be used as an unobservable instrument in the presence of an endogenous covariate.

So, the interpretation of the latent variable, on one hand, is based on the independence assumptions. The other element for interpretation, as in psychometrics, are the items or psychometric questions and the strength of correlation between these and the latent concept. These are conventionally chosen on the basis of the criteria reliability and validity, in other words, whether the set of items reflects a latent concept and whether it reflects the concept of interest.

3.3 Problems

One problem is asymptotics : if we increase the number of individuals,we increase the number of parameters to be estimated (see Douglas 1997).

Another is that, as shown by Douglas (1997), the distribution of the estimated latent variable will never converge to its true distribution. This fact violates an assumption for most further analytical analysis, using the estimated latent variable as a fixed (in a way observed) element in a different model. This could be the case for instance if we aim to estimate the effect of latent ability on wages, having estimated latent ability in a separate model based on test scores.

4 Estimation in the Presence of Latent Variables

The presence of latent variables increases the amount of parameters to be estimated and the complexity of the likelihood function. Since the latent variable is unknown it is integrated out in the likelihood function. For a set of items (of which outcome can be seen as one item) $y = (y_1, ..., y_M)$ the likelihood function takes the form

$$\begin{aligned} p(y|\theta) &= \prod_{j=1}^{M} p(y_j|\theta_j) \\ p(y) &= \prod_{j=1}^{M} \int_{\theta} p(y_j|\theta_j) p(\theta_j) d\theta_j \end{aligned}$$

This integral needs to be solved numerically. There are several ways to do this.

4.1 Likelihood Approach

In the following section I will briefly mention two alternatives to MCMC to estimate the posterior density : the EM algorithm, which has been used much in the past to solve Bayesian models, before MCMC became popular, and quadrature, which is a deterministic technique to solve analytically intractable integrals. It does not rely on sampling techniques.

Sampling methods seem to be more powerful than quadrature, if the integral of concern is of higher dimensions, since sampling is independent of the number of function evaluations. This can be the case if a likelihood function conditional on more than one latent variable is of concern. A problem with Monte Carlo integration used for multidimensional integrals is that is biased and needs a large sample for the bias to decrease sufficiently.

4.1.1 EM algorithm

The EM algorithm specifies a rule, which implies alternating between computing the expectation of a likelihood function including latent variables as if they were observed (E-step) and maximizing the expected likelihood from the E-step (M-step). The parameters from the M-step are then used in the next E-step. The algorithm is able to incorporate missing data and unobserved variables. There is no guarantee that the estimator converges to a maximum likelihood estimator. For multimodal likelihood functions the algorithm will converge to a local maximum. EM is partially Bayesian since it produces a point estimate of a latent variable together with a distribution of the latent variable.

Within sampling algorithms, MCMC seems superior to EM if the underlying model is more complex: the algorithm is likely to converge merely to local maxima if the likelihood function is multimodal.

4.1.2 Numerical Integration: Quadrature and Cubature Rules (Deterministic)

To approximate a complex function, the numerical value of definite integrals across the function can be calculated by an algorithm (combining evaluations of the integrand by a weighted sum). The collection of rules of this type are called quadrature for one-dimensional integrals and cubature for higher dimensional integrals. For now, let us write an approximation of $f(x)$ by numerical integration as

$$Q(f(x)) = \sum_{i=1}^{N} w(i) \int_{a(i)}^{b(i)} f(c)dc$$

where $w(i)$ are the weights assigned to each interval of integration and N denotes the number of intervals. So the numerical integration rule is characterized by the spacing of the subintervals and the number and weights of subintervals.

This procedure comes in hand if the integrand $f(x)$ is known only at certain points or if a formula for the integrand may be known. A small number of evaluation combined with a small error are desired for the numerical integration method. Gaussian quadrature is suitable if the function is smooth and the limits are well defined. In the following I will discuss different types of numerical integration.

4.1.3 Quadrature Rules based on interpolating functions

A function - typically a polynomial - is used to interpolate an integrand between point a and b. For a polynomial of order 0 an interpolating function passing through the point $((a+b)/2, f((a+b)/2)$ can look like this:

$$\int_a^b f(c)dc = (b-a)f((a+b)/2)$$

The polynomial can be of higher order. For more accuracy the interval can be divided in subintervals, which are approximated separately and added up (composited, iterated rule). Whether the subintervals are equally spaced on $b-a$, yields different sets of rules. (Gaussian quadrature is not equally spaced.)

4.2 MCMC

In this section I focus on the Markov Chain Monte Carlo methodology. With this method, the likelihood function is approximated by constructing a sample from it. It is a method used in Bayesian statistics. The Bayesian paradigm is a suitable environment to estimate models with latent variables since in Bayesian statistics latent variables are treated as random parameters to be estimated. Below I will first give a brief overview of Bayesian statistics. Then I will explain the MCMC algorithm and discuss its advantages and disadvantages.

4.2.1 Bayesian Statistics

This section gives a brief outline of Bayesian statistics. For further reading introductory textbooks on Bayesian statistics include Berry (1996) and Lee (2004)[6]. A comprehensive treatment of Bayesian statistics specifically in the latent variable context (and in psychometrics) is given in Rupp, Dey, Zumbo (2004).

In classic frequentist statistics the unknown parameters of a model are considered as unknown but fixed quantities. In the Bayesian paradigm however, the unknown parameters are considered as random variables, which follow a probability distribution. The aim of estimation in a Bayesian framework is therefore to estimate the probability distribution of the parameters, given the data.

Consider a set of parameters θ and a set of data y, then the probability distribution of the parameters given the data, and the main element of interest of the Bayesian statistician, is

$$p(\theta|y)$$

which is called the posterior distribution function. The posterior distribution function of a model is rewritten by applying Bayes' theorem

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \quad (1)$$

[6] See Raach(2005).

where the denominator is constant since it does not depend on θ.Therefore we can write

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

where $p(\theta)$ denotes the prior beliefs on the values of the parameters before the data is taken into account. A prior is an assumption on the probability distribution of a parameter formed before observing the data. It can be interpreted as a flexible assumption - since it expresses a belief the researcher has but can be revised if the data gives a stronger information about the parameter.

The posterior distribution is therefore proportional to the likelihood function $p(y|\theta)$ and the prior beliefs $p(\theta)$. The likelihood derives from the model specification. The prior beliefs need to be set by the researcher in such a way that the joint posterior is proper, which means that it should integrate to a finite value.

There are several possibilities of setting priors. If the researcher is uncertain, the prior can be set in such a way that it does not contain any information. This is called a flat or uninformative prior and could be for example the uniform distribution. In this case the posterior is proportional to the likelihood function and the analysis can be interpreted as a classical frequentist analysis. Priors can also be "subjective", which means that they derive from a theory. They can be "empirical" if they derive from data. If a prior is said to be conjugate, it is from the same class as the posterior. For example the normal distribution has this property.

When setting the priors the researcher needs to make a tread-off between strongly identifying the model by tightly set priors and to leave enough freedom for the data to give evidence on the model by leaving the priors loose enough. Due to the existence of priors for any parameter of the model and the possibility to tighten these priors, it has been indicated [7], that Bayesian methods can always be used to analyze a non-identified model. Nevertheless, the informativeness of the statistical analysis can become questionable if the priors are too tight and leave no possibility to infer information from the data.

Hypothesis testing in Bayesian statistics differs from frequentist analysis. Bayesian statisticians believe that their paradigm allows a more accurate way of testing for significance of the parameters. Once the parameter distributions are estimated, they can be characterized by their means, modes or variances. In addition, we can calculate the area of the distribution with 95 % of the probability mass - the central posterior density. This allows to make statements about the probability of a parameter lying within a certain region. Bayesians claim that these statements are more helpful than the frequentist confidence intervals[8].

I now turn to standard algorithms to calculate the posterior density described above.

[7] See Poirier (1998).

[8] See Raach (2005).

4.2.2 Markov Chain Monte Carlo methods

Two events brought forward the use of Bayesian methods in statistics in the early 1990s, which had not been widely implemented due to the intractability of the posterior density and especially of the integral in the denominator in equation (1). It needed to be estimated by cumbersome techniques such as Gauss-Hermite quadrature or the Newton Raphson method. The increase in computer power together with a publication by Gefland and Smith (1990) on a computer intense but implementable Markov Chain Monte Carlo algorithm to calculate the posterior density made Bayesian methods, and especially MCMC, more and more popular in statistics.

MCMC, in contrast to the classical numerical optimization, is a simulation based technique and relies on random number generation since it solves the integral by sampling. Robert and Casella (2004) provide a thorough account of MCMC methods and Gilks, Richardson and Spiegelhalter (1996) show different possible applications of MCMC.

MCMC combines the two elements "Markov Chains" and "Monte Carlo integration", which I will outline below before I turn to explain one of the most prominent MCMC algorithms, the Gibbs sampler.

Monte Carlo Integration Monte Carlo integration is a simulation-based method to solve an integral of the form

$$E_f[h(X)] = \int_\chi h(x)f(x)dx \tag{2}$$

where X is a random variable with probability distribution $f(x), \chi$ is the probability space and $h(x)$ is an arbitrary function of x. In the classic frequentist we are interested in point estimates of parameters but, as outlined above, in Bayesian statistics we are interested in estimating the posterior mean of a parameter θ_p.The formula for the posterior mean of a parameter θ_p takes a similar form as equation (2):

$$E(\theta_p) = \int \theta_p p(\theta|y)d\theta \tag{3}$$

where $h(X)$ is equivalent to θ_p and $f(x)$ is equivalent to $p(\theta|y)$.

To solve the integral in equation (2), Monte Carlo integration will provide an approximation of the integral by generating a sample $x^{(1)}...x^{(M)}$ from the distribution $f(x)$,evaluating the function h at each sampled value and calculating the average

$$\bar{h}_M = \frac{1}{M}\sum_{m=1}^{M} h(x^{(m)})$$

$\bar{h}_M$ converges almost surely to $E_f[h(X)]$ see Breiman (1992) in Raach (2005).

Equivalently, in our context, for calculating the posterior mean of θ_p by Monte Carlo integration, we need to compute the average

$$\bar{\theta}_p^M = \frac{1}{M} \sum_{m=1}^{M} \theta_p^{(m)} \tag{4}$$

where $\theta_p^{(m)}$ are randomly sampled from $p(\theta|y)$. This is not straight forward and we make use of the properties of Markov chains, which I will outline in the next section.

Markov Chains For a thorough account of the use of Markov chains in MCMC, see Robert, Casella (2004).

Markov chains represent random processes evolving over time. Consider a state space Ω and a random variable X_i holding different states in the state space. The Markov chain is a chain of realizations x_i of the random variables X_i. The change from one point in the state space x_i to the next x_{i+1} occurs with a certain probability. It can be seen that the chain represents probabilistic jumps through the state space from one state to the next. An important property of the Markov chain is that it has no memory of where it has been in the past. So it can be characterized fully by the transition kernel, which represents the probability to jump from one state x_i to the next x_{i+1} :

$$P(x_{i+1}) = P(X_{i+1}|x_i) \tag{5}$$

Equation (5) shows that the probability of jumping to a state x_{i+1} depends only on the previous state x_i.

As i goes to infinity and each jump of the chain occurs with the probability specified by the transition kernel, the Markov chain will reach a stationary distribution and the random variables X_i will follow the stationary distribution of the Markov chain. Usually once we know the transition kernel of a Markov chain, the stationary distribution of the Markov chain follows from this. If the Markov chain fulfills essentially the irreducibility condition[9], this distribution π is stationary or invariant and satisfies

$$\pi(dx_{i+1}) = \int_{\Omega} p(x_i, dx_{i+1})\pi(x_i)dx_i$$

which states that if x_i is distributed according to the invariant distribution $\pi(x_i)$, x_{i+1} is also distributed like π.

In our setting, the state space is the probability space of the posterior density, which we aim to explore by creating a sample from this probability space. The idea underlying MCMC is to use Markov chains in the opposite way - not to first specify a transition kernel and derive a stationary distribution of the chain, but to first specify a stationary distribution and specify the transition kernel in such a way that this stationary distribution is obtained. The aim is to construct a transition kernel for which the stationary distribution of the Markov chain is equal to the posterior density of interest. The resulting Markov chain can then be interpreted as a sample from the posterior, as i goes to infinity.

[9]See Robert, Casella (2004) for more details.

Constructing a transition kernel of a Markov chain in this way produces a Markov chain that has a stationary distribution equal to the posterior density; this provides a sample $\theta_p^{(1)}...\theta_p^{(M)}$ from $p(\theta|y)$ which allows us to construct the average $\bar{\theta}_p^M = \frac{1}{M}\sum_{m=1}^{M}\theta_p^{(m)}$ in equation (4). Markov chain properties allow us to construct a sample and Monte Carlo integration is employed to take an average over this sample in order to approximate the joint posterior.

There are two prominent algorithms among MCMC methods, to sample from the posterior. The main challenge in constructing the algorithm is to specify the correct transition kernel such that the stationary distribution is equal to the posterior distribution of interest. One is the Metropolis-Hastings algorithm and the second is the Gibbs sampler, which is easier to implement[10] and is a special form of the Metropolis-Hastings algorithm. In the next session I will explain the implementation of the Gibbs sampler.

MCMC algorithm: the Gibbs sampler The Gibbs sampler (a special case of the Metropolis-Hastings (MH) algorithm) is an algorithm to generate random samples from a multivariate distribution. When the algorithm is used in the Bayesian context, this distribution is the posterior distribution. The MH algorithm draws random values from proposal densities and accepts or rejects these according to the MH acceptance probability such that the detailed balance condition holds. If the acceptance probabilities are constructed correctly, the resulting sample is a Markov chain which has a stationary distribution equal to the target density.

For the Gibbs sampler the proposal densities are the full conditionals

$$p_{p|-p}(\theta_p|\theta_{-p})$$

for $\theta_1...\theta_p...\theta_P$ parameters.

Consider the target (posterior) density of a vector of parameters $p(\theta)$ and the parameters $\theta_1...\theta_p...\theta_P$ of interest. We begin with starting values $\theta_1^{(0)}..\theta_P^{(0)}$ and construct a Markov chain $\theta^{(1)}..\theta^{(M)}$ of length M. When the Markov chain has converged to its stationary distribution, the chain can be considered being distributed according to $p(\theta)$. This is the case because the transition kernels, so the proposal densities to draw the random values from, have been specified such that the invariant distribution of the resulting Markov chain is equal to the target distribution, the posterior density.

The Gibbs sampler algorithm is constructed in the following way:

1. choose starting values $\theta^{(0)} = (\theta_1^{(0)}..\theta_P^{(0)})$
2. repeat for $0, 1, ..., M$:
   draw $\theta_1^{(m+1)} = p_{1|-1}(\theta_1|\theta_2^{(m)}..., \theta_P^{(m)})$

[10] The advantage is that there is no need to adjust acceptance ratios for the drawn values before implementing the algortihm (see Raach(2005)).

draw $\theta_2^{(m+1)} = p_{2|-2}(\theta_2|\theta_1^{(m)}, \theta_3^{(m)}..., \theta_P^{(m)})$

:

draw $\theta_p^{(m+1)} = p_{p|-p}(\theta_p|\theta_1^{(m)}, ...\theta_{p-1}^{(m)}, \theta_{p+1}^{(m)}..., \theta_P^{(m)})$

:

draw $\theta_P^{(m+1)} = p_{P|-P}(\theta_P|\theta_1^{(m)}, ..., \theta_{P-1}^{(m)})$

3. return $\{\bar{\theta}_1^{(M)}..\bar{\theta}_P^{(M)}\} = \frac{1}{M}\sum_{m=1}^{M}\{\theta_1^{(m)}..\theta_P^{(m)}\}$

The main challenge of the Gibbs sampler is to specify the transition kernels, or the full conditionals $p_{p|-p}(\theta_p|\theta_1^{(m)}, ...\theta_{p-1}^{(m)}, \theta_{p+1}^{(m)}..., \theta_P^{(m)})$ correctly. Below, I will show the implementation of a Gibbs sampler using an algorithm by Albert and Chib (1993), where the full conditionals are normal distributions.

But first I will discuss how to determine whether the Markov chain has converged and briefly some alternatives to the Gibbs sampler.

Convergence Diagnostics After the Markov chain has converged the random sample is considered to be drawn from the posterior distribution. To determine, whether the generated Markov chain has converged, there are several diagnostics. It is possible to diagnose non-convergence but convergence can never be proven. Any ergodic chain converges. An ergodic chain satisfies the property that any state can be reached from any other state in a certain number of steps. The speed of convergence depends on the form of the posterior, the smoother it is the faster the convergence. There is no rule for the number of iterations, sample size, number of parameters to guarantee convergence.

First of all one needs to look at the traceplots, which show the development of the draws for each parameter. If there is no trend in them and the draws reverse around the mode of the distribution, this is a first indicator of convergence. There are also more formal tools to assess the autocorrelation of the chain. Low or medium correlations are not a problem, but high autocorrelation can be an indicator that the chain has not converged. Cowles and Carlin (1996) give an overview over convergence diagnostics[11].

Convergence can be sped up by standardizing the variables[12], using a latent concept to summarize variables or using multivariate normal priors and by picking initial values close to the posterior modes. 100 000 iterations should be used for a model with a large number of parameters. Storage problems can be overcome by thinning the chain, that is storing only a fraction of the iterations.

4.3 Why MCMC?

Before turning to the implementation of the Gibbs algorithm in the next section I will briefly discuss why MCMC methods can be suitable in the latent variable context. First of all, a Bayesian treatment of latent variables is suitable, since

[11] See Raach (2005).

[12] The correlations between parameters are then easier to calculate.

latent variables can be considered in this framework as random parameters. They are random in the sense that they vary across individuals.

Secondly, as mentioned above (in section 3.4), asymptotic analysis is a problem in the presence of latent variables because of this increase in parameters to be estimated when the sample increases. Bayesian analysis does not rely as heavily on asymptotic results as classical frequentist analysis, since

Thirdly, if one is willing to make parametric assumptions, the Gibbs sampler is an easy to implement tool and requires less computation than numerical integration methods even though it also requires a high amount of computing time due to slow convergence, relatively to numerical integration.

A possible drawback is mentioned by Imbens (2009). The choice of the prior can be arbitrary. If there is much uncertainty about the parameters prior to considering the data, to address this problem, the priors should just be chosen to be flat or uninformative enough, in order not to give arbitrariness too much weight. Priors can also be seen as less restrictive than assumptions in the classic frequentist framework since they are flexible assumptions. If the data is more informative and gives other indications than the prior, it will overpower the prior in the posterior distribution.

Another problem mentioned by Imbens (2009) is that MCMC methods are need high computer power, which is less and less a problem due to the fast advances in computer power.

To give more reason to see the advantages of MCMC in the latent variable context, I will show an implementation of the Gibbs sampler.

4.4 An Implementation of the Gibbs sampler: Estimating an Endogenous Latent Variable Model

In the following section I will show some simulation results of a Gibbs sampler to solve a parametric model including latent variables. This implication is strongly related to work by Albert and Chib (1993), Carneiro, Hansen and Heckman (2003), Heckman, Stixrud and Urzua (2006), Fahrmeir and Raach (2006) and Raach (2005).

The joint posterior distribution takes the following form:

$$\begin{aligned} & \prod_{i=1}^{N} f(\beta, \alpha, \gamma, \theta_i, M_i^*, D_i^*, c | M_i, D_i, X_i, W_i) \\ \propto \quad & f(\beta) f(\alpha) f(\gamma) f(c) \prod_{i=1}^{N} f(M_i, D_{i,} M_i^*, D_i^*, \theta_i | X_i, W_i, \alpha, \beta, \gamma, c) \end{aligned}$$

where $f(\beta)f(\alpha)f(\delta)f(\gamma)f(c)$ are the priors and the factor loadings and coefficients are written as $\alpha = (\alpha^M, \alpha^D)$ and $\beta = \beta^D$. M_i is a vector containing the polytomous psychometric items of the model, D_i is a scalar containing a binary economic outcome variable. The likelihood function can be simplified as

$$
\begin{aligned}
& \prod_{i=1}^{N} f(M_i, D_i, M_i^*, D_i^*, \theta_i | X_i, W_i, \alpha, \beta, \gamma, c) \\
= & \prod_{i=1}^{N} f(M_i^*, D_i^*, \theta_i | X_i, W_i, \alpha, \beta, \gamma, c) \prod_{i=1}^{N} f(M_i, D_i | \theta_i, M_i^*, D_i^*, X_i, W_i, \alpha, \beta, \gamma, c) \\
= & \prod_{i=1}^{N} f(M_i^*, D_i^*, \theta_i | X_i, W_i, \alpha, \beta, \gamma, c) \prod_{i=1}^{N} f(M_i, D_i | c)
\end{aligned}
$$

The first simplification follows from the application of the product rule. The second step follows from the fact that the ordinal responses M_i and D_i are determined solely by the underlying variables $M_i^*$and D_i^* and by the cutpoints c. The likelihood function can factored out into $f(M_i^*, \theta_i|.)f(D_i^*, \theta_i|.)$ since we made the conditional independence assumptions above. The factors of the likelihood function can be written as

$$
\begin{aligned}
& \prod_{i=1}^{N} [f(M_i^*, \theta_i | \alpha, \gamma, c, M_i, W_i) \{ \sum_{k_M=1}^{K_M} 1(M_i = k_M) 1(c_{k_M-1} < M_i^* < c_{k_M}) \}] \\
& \prod_{i=1}^{N} [f(D_i^*, \theta_i | \alpha, \beta, \gamma, D_i, X_i, W_i) \{ \sum_{k_D=1}^{K_D} 1(D_i = k_D) 1(c_{k_D-1} < D_i^* < c_{k_D}) \}]
\end{aligned}
$$

Each of the factors $f(M_i^*, \theta_i|.)$ and $f(D_i^*, \theta_i|.)$ needs to be multiplied by two indicators - and indicator which equals one if the observation $M_i(D_i)$ falls in category $k_M(k_D)$ and an operator indicating that $M_i^*(D_i^*)$ must fall between two cutpoints $c_{k_M-1}(c_{k_D-1})$ and $c_{k_M}(c_{k_D})$ according to its category.

θ is unobservable and will be estimated. To make the mechanism by which θ_i determines M_i^* and D_i^* perspicuous we integrate out θ_i and obtain the conditional distributions of M_i^* and D_i^* conditional on the parameters of the model and the data.

$$
\begin{aligned}
f(M_i^* | \alpha, \gamma, c, M_i, W_i) &= \int_{\theta} f(M_i^* | \alpha, \gamma, c, \theta_i, M_i, X_i) f(\theta_i | W_i) d(\theta_i) \\
f(D_i^* | \alpha, \beta, \gamma, D_i, X_i, W_i) &= \int_{\theta} f(D_i^* | \alpha, \gamma, \beta, \theta_i, D_i, X_i) f(\theta_i | W_i) d(\theta_i)
\end{aligned}
$$

As described above the Gibbs sampler is an algorithm which samples from the joint posterior distribution in a sequential way. The idea of the Gibbs sampler is to sample one of the elements $M_i^*, D_i^*, \theta_i, \alpha, \beta, \gamma, c$ at a time, conditioning on the last sampled values for the remaining elements. This procedure is equivalent to sampling from a set of conditional distributions separately. Each conditional distribution is a posterior conditional distribution of a parameter given the last sampled parameter values and the data. These conditionals - each of them constitutes one step of the Gibbs sampling algorithm - are called "full conditionals". In the following, I will derive the full conditionals constituting the Gibbs sampler for the model of this paper.

4.4.1 The Posterior Conditional Distribution of the Latent Underlying Variables

Albert and Chib (1993) propose a data augmentation procedure to sample latent underlying variables in a threshold model. It follows from their work, that the full conditional for the latent underlying variable of the binary response is

$$\begin{aligned} f(D^*|\alpha^D, \beta^D, \theta, D, X) &\propto \\ \prod_{i=1}^{N} f(D_i^*|\beta^D X_i^D + \alpha^D \theta_i, 1)\{\sum_{k_D=1}^{1} 1(D_i^* &= k_D)1(c_{k_D-1} < D_i^* < c_{k_D})\} \end{aligned}$$

$\alpha^D, \beta^D, \theta$ signify the last sampled values (or the initial values for the first iteration of the algorithm). It follows from the normality assumptions on θ and ε that $f(D^*|\alpha^D, \beta^D, \theta, D, X)$ is normally distributed with mean $\beta^D X_i^D + \alpha^D \theta_i$ and $V(D_i^*)$ normalized to unity. The latent underlying variable is distributed as the following truncated normal distributions

$$\begin{aligned} D_i^*|\alpha, \beta, \theta_i, D_i, X_i \tilde{} TN_{(-\infty,0)}(\beta^D X_i^D + \alpha^D \theta_i, 1) \text{ if } D_i &= 0 \\ D_i^*|\alpha, \beta, \theta_i, D_i, X_i \tilde{} TN_{(0,\infty)}(\beta^D X_i^D + \alpha^D \theta_i, 1) \text{ if } D_i &= 1 \end{aligned}$$

Similarly, the full conditionals for each the polytomous variables are

$$\begin{aligned} f(M^*|\alpha, \beta, \theta, c, M, X) &\propto \\ \prod_{i=1}^{N} f(M_i^*|\alpha^M \theta_i, 1)\{\sum_{k_M=1}^{K_M} 1(M_i^* &= k_M)1(c_{k_M-1} < M_i^* < c_{k_M})\} \end{aligned}$$

The latent underlying variables of the polytomous indicators are distributed as the following truncated normal distribution:

$$M_i^*|\alpha, \theta, c, M, X \sim TN_{(c_{k_M-1}, c_{k_M})}(\alpha^M \theta_i + \beta^M X_i, 1)$$

4.4.2 The Posterior Conditional Distribution of the Factor Loadings

The full conditional for the factor loadings for D can be written as

$$f(\alpha^D|\beta, \theta, D, X, D^*) \propto f(\alpha^D) \prod_{i=1}^{N} f(D_i^*|\beta^D X_i^D + \alpha^D \theta_i, 1)$$

where we choose normal priors $f(\alpha^D) = N(0,1)$ and $f(\alpha^D) = N(0,1)$.If we rewrite the equation for D_i^* and M_i^*as

$$\begin{aligned} D_i^* - \beta^D X_i^D &= \alpha^D \theta_i + \varepsilon_i^D \\ M_i^* &= \alpha^M \theta_i + \varepsilon_i^M \end{aligned}$$

we can treat it as a normal regression model and derive for M_i and D_i

$$\begin{aligned} \alpha^M | \theta_i, M_i, X_i, M_i^* &\sim N\left[(\theta_i' \theta_i + 1)^{-1} \theta_i' (M_i^* - \beta^M X_i^M), (\theta_i' \theta_i + 1)^{-1} \right] \\ \alpha^D | \beta^D, \theta_i, D_i, X_i, D_i^* &\sim N\left[(\theta_i' \theta_i + 1)^{-1} \theta_i' (D_i^* - \beta^D X_i^D), (\theta_i' \theta_i + 1)^{-1} \right] \end{aligned}$$

4.4.3 The Posterior Conditional Distribution of the Direct Coefficients

Similarly to the procedure for the factor loadings, we can write the model as

$$D_i^* - \alpha^D \theta_i = \beta^D X_i^D + \varepsilon_i^D$$

For the coefficients, we choose to set diffuse priors as well. The full conditionals for the intercepts are, according to Albert and Chib (1993,p.671)

$$\beta^D | \alpha^D, \theta_i, D_i, X_i, D_i^* \sim N\left[(X_i' X_i)^{-1} X_i' (D_i^* - \alpha^D \theta_i^D), (X_i' X_i)^{-1} \right]$$

4.4.4 The Posterior Conditional Distribution of the Cutpoints

We assume a uniform prior for the cutpoints and can write for the full conditionals for the polytomous responses

$$c^M | \alpha^M, \theta, M, X, M^* \sim unif \left[\begin{array}{c} \max\{\max\{M_i^* : M_i = k_M\}, c_{M-1}\}, \\ \min\{\min\{M_i^* : M_i = k_{M+1}\}, c_{M+1}\} \end{array} \right]$$

4.4.5 The Posterior Conditional Distribution of the Latent Factors

Similarly as for the procedure for coefficients and factor loadings, we can rewrite the model as

$$\begin{aligned} D_i^* - \beta^D X_i^D &= \alpha^D \theta_i + \varepsilon_i^D \\ M_i^* &= \alpha^M \theta_i + \varepsilon_i^M \end{aligned}$$

and treat it as a normal regression model,where θ_i is the parameter to be estimated. Carneiro, Hansen and Heckman (2003) specify a mixture of normals as prior for the latent factors. We treat the latent factors as endogenous depending on γW_i. We treat θ_i in the same way as M_i^* and D_i^* for which the priors are implicitly determined by the prior distributions of the other parameters of the model and by the assumptions on the distribution of ε_i^D and ε_i^M. The prior of θ_i is therefore implicitly determined by the priors on the other parameters of the model and by the assumptions on the distributions of ε_i^D, ε_i^M and ε_i^θ.

We can then derive the full conditional for the latent factor as:

$$\begin{aligned} & f(\theta|\beta, \alpha, c, \gamma, X, W, D^*, M^*) \\ \propto \; & \prod_{i=1}^{N} f(M_i^*|\alpha^M\theta_i, 1) f(D_i^*|\beta^D X_i^D + \alpha^D\theta_i, 1) \end{aligned}$$

We do not need to condition on M_i and D_i since they are implicitly known through M_i^* and D_i^* and c

$$\begin{aligned} & \theta|\beta, \alpha, \gamma, \delta, c, M, D, X, W, D^*, M^* \\ \sim \; & N\left[\begin{array}{c} \gamma W_i + (\alpha^{D\prime}(\alpha^D + \alpha^{M\prime}\alpha^M + 1)^{-1} \\ (\alpha^{M\prime}(M_i^* - \beta^M X_i^M - \alpha^{M\prime}\gamma W_i) + \alpha^D(D_i^* - \beta^D X_i^D - \alpha^D\gamma W_i)), \\ I - \alpha^{D\prime}(\alpha^{D\prime}\alpha^D + \alpha^{M\prime}\alpha^M + 1)^{-1}\alpha^D \\ -\alpha^{M\prime}(\alpha^{D\prime}\alpha^D + \alpha^{M\prime}\alpha^M + 1)^{-1}\alpha^M \end{array}\right] \end{aligned}$$

4.4.6 The Posterior Conditional Distribution of the Indirect Coefficients

The posterior we sample from can be written as

$$\begin{aligned} & f(\gamma|\theta, W) \\ \propto \; & f(\gamma) f(\theta|\gamma, W) \end{aligned}$$

The model for the latent variable is

$$\theta = \gamma W + \varepsilon_\theta$$

We assume a diffuse prior for the coefficient γ. Similar to the procedures above we get:

$$f(\gamma|\theta, W) \sim N((W'W)^{-1}W'\theta), (W'W)^{-1})$$

I simulated the data for $N = 1000$ and ran the algorithm for 100000 iterations. The table below in appendix B shows the results. The algorithm has converged since the traceplots of all estimated parameters do not show any trends. There are no evident identification problems since the posteriors are not flat, they all have a single mode and they do are not equal to the prior. The estimated values are always close to the true value and the standard errors show that the estimated values fall into a confidence interval around the true value.

5 Identification in the Presence of Latent Variables

Even if latent variables can be seen as an alternative to instrumental variable techniques, most approaches to identify models in the presence of latent variables rely nevertheless on the existence of an instrument (see Matzkin (2003,2007)). Carneiro,Hansen & Heckman (2003) provide a semi-parametric identification strategy of a simultaneous equation model with the presence of latent variables, which is not based on the existence of such variables.

There is a literature on nonparametric identification of models with endogenous regressors - models with measurement error. Latent terms might be considered in this literature but it is not the main interest to identify and estimate these terms and their effect on the observable terms.

5.1 Parametric Approaches

Identification in conventional parametric factor analysis uses the terms of the variance-covariance matrix of the observable variables and fits a linear and additive model to express this covariance matrix with a latent factor and a random error term. Additionally either the scale of the latent variables needs to be set or alternatively one of the factor loadings is set to a fixed term. Distributional assumptions are made on the distribution of the random error term and the latent variable. Rosenbaum (1984) establishes a condition for identification of parametric models involving latent variables, which says, that the number of parameters to be estimated needs to be equal to the number of covariances in the model.

In item response theory the assumption of conditional independence - independence between the items conditioning on the latent variable - yields identification. A parametric ordered response is assumed to underlie the observed response pattern. The distributional assumption on the error term in this model then yields the functional form of the probability of answering a specific category to a psychometric question.

Heckman, Stixrud, Urzua (2006) implement a version of the semi-parametrically identified model of Carneiro, Hansen, Heckman (2003). These authors embed a classic factor model into a linear model for economic outcomes. They use independence conditions, exclusion restrictions and distributional assumptions for the unobservable terms.

5.2 Nonparametric Approaches

In Psychometrics there is a nonparametric literature on identification and estimation of latent variable models. Pioneering work can be found in Rosenbaum (1984) and Holland, Rosenbaum (1986), Ramsay (1991) and Samejima (1979, 1981, 1984, 1988, 1990)[13] More recent work based on a total score of items is

[13] References to the last two authors are given in Douglas (1997).

found in Molenaar & Sijtsma (2002).

In economics, Spady (2007) developed a strategy to infer a latent underlying scale, which is based on the notion of stochastic dominance, from a set of psychometric items concerning political attitudes. His method relies on a minimal set of assumptions. Matzkin (2003) develops nonparametric methods to identify functions for continuous and discrete dependent variables in the presence of endogenous observable explanatory variables and unobservable instruments. Endogeneity can result from omitted unobservable variables,measurement error or simultaneity. She claims it is a non-parametric version of the work of Heckman et al cited above[14].

5.3 Identification of the model in its generalized form

Latent variable modelling can express different conceptions. The latent variable in economics is most commonly a latent underlying variable governing an ordinal response. An interest in the effect of a latent variable on observable variables is fairly recent in economics. In other fields studying latent variables has so far been subject of mainly a parametric analysis. As mentioned above there is a nonparametric literature in item response theory, based on the total score of the items. In economics a total score is of less interest since economists are usually not interested in ordering the dependent variables by their degree of discrimination.

We are interested here explicitly in combining the existing literature in several fields to establish well-formulated conditions to identify semiparametrically the effect a latent variable has on observable variables. We are additionally interested in the interpretation of the latent variable, which we base upon the choice of dependent variables and on conditional independence assumptions.

In the following I explore, how the identification of the model with endogenous regressors in section two in Matzkin (2003) and section 4.1 in Matzkin (2007) changes when the endogenous regressor is considered as unobservable. We find that we can apply Matzkin's identification proof, but we need to add assumptions on the model for the unobservable regressor.

The model in its generalized form takes the form

$$\begin{aligned} Y &= g_1(\theta, \varepsilon_1) \\ \theta &= g_2(X^\theta, \varepsilon_2) \end{aligned}$$

θ is not independent of ε_1.Y is an observable continuous dependent variable, X^θ are continuous or discrete independent variables and θ is a continuous endogenous latent factor. $\varepsilon_1, \varepsilon_2$ are random error terms.

In the following we aim to identify the function g_1.

[14] Matzkin (2004) mentions this on page 3.

5.4 Assumptions

In the following exposition the symbol $\perp$ stands for "independent of".

Condition 1 $\theta \perp \varepsilon_1 | X^\theta$ *(for first line in the proof below)*

In other words $F(\theta, \varepsilon_1 | X^\theta) = F(\theta | X^\theta) F(\varepsilon_1 | X^\theta)$.

Condition 2 *The function $g_1(.,.)$ is increasing in its second argument ε_1.(for third line in the proof below)*

Condition 3 *The conditional distribution $F(Y|\theta = \tilde{\theta}, X^\theta = x^\theta)$ is strictly increasing. (for invertibility of $F(Y|\theta = \tilde{\theta}, X^\theta = x^\theta)$)*

Condition 4 $F_{\varepsilon_1|X^\theta}(e_1) = U(0,1)$ *(normalization)*

Condition 5 $g_2(X^\theta, \varepsilon_2) = \bar{g}_2(X^\theta) + \varepsilon_2$ *(for identification of $F(\theta|X^\theta = x^\theta)$)*

Condition 6 $F(\varepsilon_2) = N(0,1)$ *(for identification of $F(\theta|X^\theta = x^\theta)$)*

From conditions 5 and 6 it follows that

$$F(\theta|X^\theta = x^\theta) = N(\bar{g}_2(x^\theta), 1)$$

5.5 Identification

In the following we aim to identify the function g_1.

Theorem 7 *If conditions 1-3 are satisfied, then for all X^θ, ε_1*

$$g_1(\theta, \varepsilon_1) = F^{-1}_{Y|\theta, X^\theta}(F_{\varepsilon_1|X^\theta})$$

Proof.

$$\begin{aligned}
F_{\varepsilon_1|X^\theta} &= \Pr(\varepsilon_1 \leq e_1 | X^\theta = x^\theta) \\
&= \Pr(\varepsilon_1 \leq e_1 | X^\theta = x^\theta, \theta = \tilde{\theta}) \\
&= \Pr(g_1(\theta, \varepsilon_1) \leq g_1(\tilde{\theta}, e_1) | X^\theta = x^\theta, \theta = \tilde{\theta}) \\
&= \Pr(Y \leq g_1(\tilde{\theta}, e_1) | X^\theta = x^\theta, \theta = \tilde{\theta}) \\
&= F_{Y|X^\theta = x^\theta, \theta = \tilde{\theta}}(g_1(\tilde{\theta}, e_1))
\end{aligned}$$

The second line follows from condition 1, the third line follows from condition 2. The fourth line follows from substituting $g_1(\theta, \varepsilon_1)$ by Y.

Given condition 3 we can take the inverse of the last line and get

$$g_1(\tilde{\theta}, e_1) = F^{-1}_{Y|X^\theta = x^\theta, \theta = \tilde{\theta}}(F_{\varepsilon_1|X^\theta}(e_1))$$

Given the normalization $F_{\varepsilon_1|X^\theta}(e_1) = U(0,1)$ we get

$$g_1(\tilde{\theta}, e_1) = F^{-1}_{Y|X^\theta = x^\theta, \theta = \tilde{\theta}}(e_1)$$

■

This result is still incomplete since θ is unobservable and we cannot condition on it.

We can then apply Bayes rule to eliminate the conditioning

$$F_{Y|X^\theta=x^\theta,\theta=\tilde{\theta}} = \frac{F_{\theta|Y,X^\theta}F_{Y|X^\theta}}{F_{\theta|X^\theta}}$$

From the assumptions above, we have that

$$F(\theta|X^\theta = x^\theta) = N(\bar{g}_2(x^\theta), 1)$$

$F_{Y|X^\theta}$ can be estimated by any nonparametric estimator for conditional densities, such as kernels. It follows from the model that $F(\theta|X, Y)$ is normal.

$E(\theta|X, Y)$ can be shown to be linear in X and Y and $V(\theta|X, Y)$ is constant. [check and add more...] With this we have identified the function $g_1(\tilde{\theta}, e_1)$ for all $\tilde{\theta}$.

5.6 Discrete Outcome Variables

In the previous section we have shown how one could use existing literature on nonparametric identification to identify the effect of an endogenous latent variable on a continuous outcome variable using cross-sectional data, such as earnings. Especially micro-econometric and typically psychometric outcome variables are often discrete, such as employment or answers to any qualitative question. It goes beyond the scope of this article to provide a new methodology to a nonparametric or semiparametric identification of the effects of latent variables on discrete outcomes, but I would like to point the reader towards existing literature in this field and show up several possibilities of approaching the problem. I will focus on

Carneiro, Heckman and Hansen (2003) have developed a semiparametric identification strategy of factor models with discrete choices and continuous outcomes. They estimate the model parametrically, using an MCMC method. They assume that the latent factors are generated from a mixture of normal distributions. Error terms are assumed to be normal but they are theoretically nonparametrically identified.

Spady (2006, 2007) proposes yet another way of semiparametrically identifying and estimating a discrete choice model with latent factors. He uses discrete data on voting behavior and attitudes in the US and is able to estimate semiparametrically the effects of a cultural and an economic factor on US voting behavior. Spady specifies an item response theory model and imposes minimum assumptions on the distributions of responses as a function of the latent factor. His first assumption is responses of individuals with a higher position on the scale of the latent factor stochastically dominate the responses of those with a lower position on the scale of the latent factor. His second assumption is a monotonic scale representation for the scale of the latent factor. His model can then be estimated by sieve maximum likelihood estimation.

In the field of psychology Douglas (1997) has also contributed to the theory of nonparametric identification and estimation of nonparametric item response models. He develops a methodology to simultaneously and nonparametrically estimate the latent factors and their effects on the responses. Douglas applies a kernel smoothing methodology to estimate the unknown quantities. This methodology has mainly been developed for ability testing framework in psychometrics.

We can also turn to the literature on non- or semiparametric identification of discrete choice models in the presence of endogenous regressors and see how one can possible reinterpret these models in order to make them fit into the framework of meaningful latent variables. Chesher (2007) has studied the issue of endogeneity and discrete outcomes. He shows, in a nonparametric framework, how to partially identify important structural effects with minimal assumptions. Lewbel (2000) proposes estimators for binary, ordered and multinomial response models, which can deal with endogeneity problems. His methodology builds upon one special regressor with a coefficient normalized to one and the existence of instruments. This methodology, however, relies heavily on the right choice of the special regressor.

The most suitable paper among the literature on nonparametric identification and estimation of discrete choice models with endogenous regressors seems as in the continuous case again Matzkin (2003). Matzkin (2003) develops in section 5.1 the discrete case of the continuous model, which we extended above. She proposes to use the same methodology as she uses for the continuous model to identify and estimate discrete choice models with an unobservable variable correlated with an observable variable. Matzkin refers to Blundell, Powell (2003) in this section. In the following I will summarize her approach, which builds upon that of Blundell, Powell (2003). I will rewrite and reinterpret the model such that it fits into the framework used in this article.

$$\begin{aligned} M &= 1 \text{ if } g_1(X, Y, \theta) > \varepsilon \\ & \quad\ 0 \text{ otherwise} \\ M^* &= g_1(X, Y, \theta) \\ Y &= g_3(W, \eta) \end{aligned}$$

where
M - binary psychometric item
X, Y - observable explanatory variables (exogenous and endogenous)
θ - latent characteristic (endogenous)
$\varepsilon, \delta, \eta$ - error terms

5.6.1 Assumptions

Condition 8 $\theta \perp (X, \varepsilon)$

Condition 9 $\varepsilon \perp (X, Y, M)$

Condition 10 *there exists a function* $g_2(.,.)$ *is strictly increasing in its last argument and a random error term* δ *such that* $\theta = g_2(W, \delta)$

Condition 11 δ *is distributed independently of* (X, Y, W)

Condition 12 $\delta \sim U(0, 1)$

Condition 13 $F(\theta|W = w) = U(0, 1)$

Condition 14 *one of the coefficients of* X *equals one*

It follows clearly from this setup that the latent characteristic θ is correlated with the regressor Y, and θ is therefore endogenous. Matzkin shows in section 5.1 that under the first two conditions, one can identify $g_1(X, Y, \theta)$ as well as the distribution of (θ, Y), under conditional independence assumptions between X and θ.

Matzkin then discusses a more parameterized form of the model above

$$\begin{aligned} M &= 1 \text{ if } \beta X + \gamma Y + \theta > \varepsilon \\ &\quad 0 \text{ otherwise} \end{aligned}$$

and introduces the next five conditions above. There are no restrictions that the function g_2 or on the distribution of ε or δ need to belong to parametric families. Matzkin then shows that she can identify the model using a modified version of Blundell, Powell (2003). Given the assumptions, Matzkin can rewrite the expression for $E(M = 1|X, Y, W) = G$ in such a form that it corresponds to an expression identified and used by Blundell, Powell (2003). Once the estimators for G, β, γ are obtained, Matzkin shows that one can estimate the distribution of ε and the function $g_2(W, \delta)$.

I have shown, by reinterpreting and rewriting the model in section 5.1 in Matzkin (2003), that the latter can be used for the framework of latent endogenous variables in a discrete choice model. So far the model outlined above does not include a parameter α, which signifies the effect of the latent variable θ on the discrete outcome M. This would involve specifying some additional assumptions.

6 Conclusion

Latent variables are being used in some economic models, but there is not yet an established framework in economics of how to use them - how to interpret and identify them and which estimation strategy should be used. The econometric paradigm does not propose explicitly how to treat unobservable concepts, which are not straightforward to quantify. The unobservable is usually treated as an error term and there is no special interest in the information included in this term. In this paper I explored, how the identification of the model with endogenous regressors in section 2 in Matzkin (2003) and section 4.1 in Matzkin (2007) changes when the endogenous regressor is considered as unobservable. I find that we can apply Matzkin's identification proof, but we need to add assumptions on the model for the unobservable regressor. I additionally proposed and implemented a Bayesian Markov Chain Monte Carlo estimator for an endogenous latent variable model and find satisfying results for estimated parameters of simulated data.

7 Appendix A: Tables

	loadings	true values	strd errors
m1	0.27	0.40	0.08
m2	0.34	0.20	0.27
m3	0.17	0.20	0.06
d	0.22	0.20	0.07

Table 1: Simulated Model: Loadings

	coefficients	true values	strd errors
m11	-6.07	-7.00	0.65
m12	0.24	0.30	0.04
m13	0.41	0.40	0.05
m14	0.15	0.20	0.04
m15	0.63	0.60	0.05
m16	-0.05	0.00	0.04
m21	-8.31	-7.00	1.40
m22	0.00	0.00	0.04
m23	0.27	0.20	0.06
m24	0.46	0.50	0.08
m25	0.15	0.10	0.05
m26	0.38	0.30	0.07
m31	-7.01	-8.00	0.60
m32	0.13	0.20	0.04
m33	0.08	0.10	0.04
m34	0.36	0.40	0.04
m35	0.03	0.10	0.04
m36	0.47	0.50	0.04

Table 2: Simulated Model: Direct Coefficients Tri-categorical Items

	coefficients	true values	strd errors
d1	-10.34	-10.00	0.81
d2	0.26	0.30	0.05
d3	0.44	0.40	0.05
d4	0.43	0.50	0.05
d5	0.22	0.20	0.05
d6	0.34	0.30	0.05

Table 3: Simulated Model: Direct Coefficients Binary Item

	coefficients	true values	strd errors
w1	0.07	0.00	0.09
w2	0.74	0.70	0.21

Table 4: Simulated Model: Indirect Coefficients

References

[1] Albert, J.H. & Chib, S. (1993) : Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association, 88 (422), 669-679.*

[2] Berry, D. A. (1996) : Statistics: A Bayesian Perspective. Duxbury, London, UK.

[3] Blundell, R. & Powell, C. (2003) : Endogeneity in Semiparametric Binary Response Models, *The Review of Economic Studies, 71 (3), 655-679.*

[4] Breiman, L. (1992) : Probability. SIAM, Philadelphia, USA.

[5] Borghans, L.; Duckworth, A.L.; Heckman, J. & terWeel, B. (2008) : The Economics of Psychology and Personality Traits, *Journal of Human Resources.*

[6] Bowles, S.; Gintis, H. & Osborne, M. (2001) : The determinants of earnings: A Behavioral Approach, *Journal of Economic Literature 39 (4), 1137-1176.*

[7] Canerio, P.; Hansen, K & Heckman J. (2003) : Estimating Distributions of Treatment Effects with an Application to the returns of Schooling and Measurement of the Effects of Uncertainty of College Choice, *International Economic Review 44(2), 361-442*

[8] Cowles, M. K. & Carlin, B.P. (1996) : Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association 91, 883-904.*

[9] Chesher, A. (2007) : Endogeneity and discrete outcomes. *cemmap Working Papers (CWP05/07).* Institute for Fiscal Studies, London, UK.

[10] DeLeeuw, J. & Takane,Y. (1987) : On the Relationship between Item Response Theory and Factor Analysis of Discretized Variables, *Psychometrika 52 (3).*

[11] Douglas, J. (1997) : Joint Consistency of Nonparametric Item Characteristic Curve and Ability Estimation, *Psychometrika 62 (1).*

[12] Fahrmeir, L. & Raach, A. (2006) : A Bayesian semiparametric latent variable model for mixed responses, *Psychometrika.*

[13] Gelfand, A. E. & Smith, A. F. M. (1990) : Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85, 398 - 409.*

[14] Gilks, W. R.; Richardson, S. & Spiegelhalter, D. J. (Eds.) : Markov Chain Monte Carlo in practice. Chapman and Hall.

[15] Heckman J.; Stixrud, J. & Urzua, S. (2006) : The Effects of Cognitive and Non-cognitive Abilities on Labor Market Outcomes and Social Behavior, *Journal of Labor Economics.*

[16] Holland, P.W. & Rosenbaum, P.R. (1986) : Conditional Association and Unidimensionality in Monotone Latent Variable Models, *Annals of Statistics 14 (4).*

[17] Imbens, G. (2009) : New Developments in Econometrics, Lecture 13, Bayesian Inference, *cemmap lectures, University College London.*

[18] Lee, P. M. (2004) : Bayesian statistics: an introduction (3rd edition). Arnold, London.

[19] Lewbel, A. (2000) : Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables, *Journal of Econometrics 97 (1), 145-177.*

[20] Matzkin, R. (2003) : Unobservable Instruments, *mimeo*, Northwestern University.

[21] Matzkin, R. (2007) : Nonparametric Identification, *Handbook of Econometrics Vol 6B.*

[22] Mokken, R. J. (1971) : A Theory and Procedure of Scale Analysis, Berlin, Germany: De Gruyter.

[23] Poirier, D. J. (1998) : Revising Beliefs in Non-identified Models, *Econometric Theory, 14, 483-509.*

[24] Molenaar, I.W. & Sijtsma, K. (2002) : Introduction to nonparametric item response theory, Sage Publications, Thousand Oaks, USA.

[25] Rabe-Hesketh, S. & Skondral, A. (2004) : Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Boca Raton, FL: Chapman & Hall/CRC.

[26] Raach, A. (2005) : A Bayesian semiparametric latent variable model for binary, ordinal and continuous response, Thesis Ludwig-Maximilians Universitaet Muenchen, Germany.

[27] Robert, C.P. & Casella, G. (2004) : Monte Carlo statistical methods (2nd edition). Springer, New York.

[28] Rosenbaum,P.R. (1984) : Testing Conditional Independence and Monotonicity Assumptions in Item Response Theory, *Psychometrika 49 (3).*

[29] Rupp, A. ; Dey, D. K. & Zumbo, B. D. (2004) : To Bayes or not to Bayes, from whether to when: applications of Bayesian methodology to modeling, *Structural Equation Modeling 11, 424-451.*

[30] Spady, R. (2006) : Identification and estimation of latent attitudes and their behavioral implications, *cemmap Working Papers (CWP12/06).* Institute for Fiscal Studies, London, UK.

[31] Spady, R. (2007) : Semiparametric Methods for the Measurement of Latent Attitudes and the Estimation of their Behavioral Consequences, *cemmap Working Papers (CWP26/07).* Institute for Fiscal Studies, London, UK.

[32] Wansbeek, T. (2000) : Measurement Error and Latent Variables in Econometrics, North Holland.