Exploring family life circumstances and their relationship to a child's school achievement – an econometric analysis in large data contexts

Inaugural-Dissertation
zur Erlangung des akademischen Grades eines Doktors
der Wirtschafts- und Sozialwissenschaftlichen Fakultät der
Christian-Albrechts-Universität zu Kiel

vorgelegt von
Malte Hoffmann, Master of Science Quantitative Economics
aus Helmstedt

Hamburg, 2017

Gedruckt mit Genehmigung der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

Dekan:	Prof. Dr. Till Requate
Erstberichterstattender:	Prof. Dr. Uwe Jensen
Zweitberichterstattender:	Prof. Dr. Katrin Rehdanz
Tag der Abgabe der Arbeit:	21.11.2016
Tag der mündlichen Prüfung:	29.03.2017

Acknowledgements

First of all, a 'thank you' to everyone who supported me and my research.

I am sincerely grateful to my supervisor, Prof. Dr. Uwe Jensen, who not only supervised my master's thesis a long time ago but also encouraged me to start working on a PhD-thesis. He supported and guided me through this dissertation and always had time for my questions and interests. Moreover, he helped me to gather ideas and overcome obstacles, but also to focus, which is the reason why I was able to learn an incredible amount during the process. He skillfully created a collaborative and relaxed atmosphere in which to provide this support.

Many thanks also to my former colleagues at the HWWI. Christina Boll from whom I could learn a lot during my time there and who gave me the possibility to experience various facets of academic work. Also Prof. Dr. Thomas Straubhaar, with whom an inspiring discussion was never far away. Moreover, I enjoyed working with Jonas, Julian and Malte; I benefited from and still appreciate the vivid discussions we had.

Above all, I am indebted to my family who paved the way for so many of my choices in life and who have supported and encouraged me through all times. My friends, in particular Andre and Benjamin, for always being available and supportive. Last but by no means least, Laura who encouraged and supported me through my research.

Contents

Table of Abbreviations

BSSR	Backward Stepwise Selection regression
DGP	Data generating process
FA	Factor Analysis
FAR	Factor Regression
FSSR	Forward Stepwise Selection regression
Lasso	Least Absolute Shrinkage and Selection Operator
LTSS	Lower track secondary school
MSE	Mean squared error
MTSS	Middle track secondary school
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PCovR	Principal Covariates Regression
PCR	Principal Component Regression
PLS	Partial Least Squares
SD	Spectral decomposition
SES	Socio-economic status
SOEP	German Socio-Economic Panel Study
UTSS	Upper track secondary school

List of Tables

List of Figures

Table of Symbols

Roman letters	**Meaning**
E	Vector of error terms of dim $N \times q$
f_j	The j-th factor or component, where $f_j \in \mathbb{R}^N$ and $j = 1, \ldots, k$.
F	Matrix of factors or components or scores of dim $N \times k$.
I_m	Identity matrix of dim $m \times m$.
L	Matrix of (pattern) loadings of dim $q \times k$.
N	Total number of observations, $n = 1, \ldots, N$.
n_1	Number of observations in training sample.
n_2	Number of observations in test sample.
R^2	Multiple coefficient of determination.
u	Residual term of dim $N \times 1$.
x_i	The i-th independent variable, where $x_i \in \mathbb{R}^N$ and $i = 1, \ldots, q$.
X	Matrix of independent variables of dim $N \times q$.
$w_{i,j}$	Weight of the i-th variable on f_j.
W	Matrix of loading weights of dim $q \times k$.
y	Dependent (endogenous) variable of dim $N \times 1$.
Greek letters	**Meaning**.
β_i	Coefficient parameter of variable i.
β	$q \times 1$ parameter vector of coefficient parameters.
ϵ_i	Error term of variable i.
γ_j	Coefficient parameter of factor/component j.
γ	$1 \times k$ parameter vector of coefficient parameters.
λ	Eigenvalue / Tuning parameter.
Φ	Correlation matrix of factors of dim $k \times k$.
Ψ	Covariance matrix of error terms E of dim $q \times q$.
ψ_i	Variance of error term of variable i.
σ^2	Variance.
θ	Threshold parameter.

Functions	**Meaning**
$Corr\,[a, b]$	Returns the correlation between a and b.
$Cov\,[a, b]$	Returns the covariance between a and b.
$E\,[a]$	Returns the expected value of a.
$Sign\,[a]$	Returns the sign of element a.
$tr\,[A]$	Returns the trace of matrix A.
$Var\,[a]$	Returns the variance of a.

Distributions	**Meaning**
$Beta(\alpha, \beta)$	Beta distribution with parameters α and β.
$N(\mu, \sigma)$	Normal distribution with mean μ and variance σ.
$U(a, b)$	Uniform distribution with minimum value a and maximum value b.

Chapter 1

Introduction

Ever since the 1960's, labor demand in Germany has shown a clear trend: job requirements have become increasingly related to higher human capital (Bremer, 2007, p. 26). Human capital, broadly defined as a set of skills relating to knowledge or individual characteristics which increase productivity, has thus become a crucial asset for workers when competing for jobs. This development especially intensified as the service sector and the information economy expanded and aiding computer technology found its way into working life. Working in this knowledge-based society requires the capability to handle information and communication techniques as well as meta-knowledge on how to gather, process and create information (Allmendinger, 2009, p. 3).

The economic literature has suggested several mechanisms on how human capital increases a worker's productivity and ultimately their income. One established view is put forward by Becker (1964). His approach regards collected qualifications as an input factor which directly raises productivity.[1] Schultz (1963) and Nelson and Phelps (1966), on the other hand, regard human capital primarily as leading to a higher adaptive capability – a beneficial asset in a world marked by increasingly faster technological change. Since such changes often affect occupational tasks and structures, the authors emphasize that enhanced ability to adapt to new situations and thus offering an increased flexibility helps to deal with these changes more quickly and thereby increase productivity.

These views leave open how human capital materializes and how it is acquired.[2] The level of human capital could be determined, for instance, through training, inherent ability, even the level of motivation or a mix of these. Often, however, a

[1] Examples for evidence in favor of this theory can be found in Kroch and Sjoblom (1994) and Chevalier et al. (2004).

[2] Since Becker (1964) distinguishes between general and firm-specific human capital, it has to be mentioned that in this thesis only the former is of interest.

fundamental requirement is the successful completion of school education, for the reason alone that it imparts essential basic knowledge and builds the foundation for acquiring further human capital (see e.g. Currie et al., 2009). This idea has been seized in numerous empirical studies. Assuming that higher productivity is rewarded monetarily, it is natural to estimate the rate of return to education, as for instance in the work of Mincer (1974).[3] Mincer uses years of schooling, others use the highest formal education degree, but the two measures have an inherent relationship since a higher degree typically requires more years of schooling.

Using the highest formal education degree as a measure of human capital is not obvious from a strictly Beckerian perspective since it is unlikely to carry any influence on productivity as such. This influence rather stems from the process to attain the degree because it is where knowledge is acquired. Practically, however, the completion generates a certificate which proves the existence of pertinent knowledge to uninformed others and hence signals certain abilities and skills. This argument corresponds to the human capital view of Spence (1973): Since employers cannot fully validate an applicant's skills, a sensible guess can be based on certificates which are a sign of ability and can be translated into higher productivity.[4] Individuals with higher education will be offered the better rewarded jobs. Another argument why a formal degree is relevant is due to institutional norms. Education systems are often highly hierarchical – also in Germany: Norms usually require the formal completion of lower education stages before advancing to the higher, more specialized ones. For example, first successfully finishing school with an adequate certificate will open up certain vocational opportunities.

Using school-related characteristics as measures of human capital can be criticized, however. Years of schooling, in particular, are prone to error due to class repetition. Class repetition adds an additional year of schooling, but the reason for this extra year is qualitatively different to a normal year. But there are also disadvantages to the highest formal education degree. The disadvantage which school achievement measures share is to not take the quality of schooling into account. A year of schooling at a bad school is not the same as at a good school which is the reason differences in quality are likely to have a substantial impact on human capital. The larger the differences between a good and a bad school are, the more imprecise a school achievement indicator becomes. Hanushek and Woessmann (2015) contend an alternative measure. The authors propose using indicators of cognitive skills, which can be approximated by test scores as, for instance, surveyed in the PISA studies. Cognitive skills are a compound of many factors of which school quality

[3] Minding that the schooling effect is likely to be overstated due to the correlation between earnings and investment in human capital through inherent ability.

[4] Supportive evidence for the signaling theory is, for instance, due to Lang and Kropp (1986).

can be one, but they are also affected by families, peers, neighborhood and health status (Hanushek, 2009, p. 40). This measure hence exhibits variation where the mentioned school-related measures would remain constant. It is, in particular, the international perspective on the aggregate economy in which the strengths of the use of cognitive skills become clear. By referring to numerous empirical studies, a robust relationship between cognitive skills and the growth trajectory of both industrialized and developing countries is demonstrated – a relation found to be more fragile for school achievement indicators (Ibid., p. 41). For the most part, this can be ascribed to the observation that school quality is more heterogeneous across nations than within nations, which would make cognitive skills a more precise measure of human capital.

The concept of cognitive skills as an alternative measure can refer to a large extent of abilities – some inborn, some learned. The empirical literature typically differentiates between two types of cognitive skills which are sometimes referred to as intelligence. Difference in origin is what demarcates the two types from each other. One is fluid intelligence which relates to inherent abilities; the other is crystallized intelligence which refers to acquired knowledge or behavior (Dahmann, 2015, p. 13). As examples Dahmann names for the first concept the ability to reason and the capability to process information. Examples for crystallized intelligence are the abilities to read or to calculate, capabilities that can be acquired through practice. While fluid intelligence can be viewed as given, crystallized intelligence is considered influenced by environmental factors, education amongst others. A clear attribution, however, appears rather difficult as interaction effects between the two types of intelligence are likely to play a role. For crystallized intelligence is thought to be more malleable, it is of higher interest in this thesis. When speaking of cognitive skills in the following, crystallized intelligence alone is referred to.

There being an influence of schooling on cognitive skills, these measures typically feature a close relation; in fact, the latter are also termed as "a key dimension of schooling outcomes" (Hanushek, 2009, p. 42). Emphasizing the importance of school quality, Hanushek points out that the effect of schooling on key economic measures can be confounded not only through ability but also through a selection effect that is caused by differences in school quality. Hanushek et al. (2008) observed that dropout rates are higher in low-quality schools than in high-quality schools. The individuals who have enjoyed longer schooling have therefore also enjoyed better school quality on average. Together with a positive correlation of school quality and key economic outcomes, there is a quality bias in addition to the well-known ability bias.

Based on these arguments, deciding to measure human capital by school achievement indicators or test scores depends predominantly on four factors: which theory

is followed, in particular, how important signaling is considered, the aim and context of the analysis, the heterogeneity of the data sample with respect to school quality and finally data availability.

Regardless which measure is chosen, there are certain individual benefits that go along with the investment into human capital or, more specifically, education. Not only does income rise on average, the personal risk of unemployment also decreases through more (occupational) opportunities and the mentioned improved capability to adapt. Not solely economic benefits complement the many desirable job-related returns: As a modern society offers a lot of possibilities, options arise. Where options arise, decisions have to be made and schooling can improve them. Making better decisions can, for instance, refer to more distant areas like consumption or marriage (Oreopoulos and Salvanes, 2011). In this sense, more education corresponds to a larger set of accessible information, which the individual can, moreover, assess better. There are also indications that higher education enhances personal health outcomes, like longevity (Lampert et al., 2013, p. 262f.). However, due to long-term effects there does not exist any causal evidence to my knowledge. And although an indirect link could be established via unemployment, which negatively correlates to both education and health indicators, a causal effect of unemployment on health could not yet be verified (Schmitz, 2011).

The relationships do not only hold on an individual, but also on an aggregate level. In the context of endogenous growth models, it has been noted that the accumulation of relevant human capital affects overall productivity and the growth rate of the economy (Mankiw et al., 1992). But, in particular, cognitive skills significantly explain variation in economic growth in developed and developing countries. Causality concerns in this context, for instance through (unobserved) third-variable effects or reverse causality, are addressed by Hanushek and Woessmann (2012) and Kimko and Hanushek (2000) whose evidence points towards the negligibility of such concerns. Their results portend positive long-term effects of human capital on the economy.
The relationship between cognitive skills and growth holds internationally. Due to its demographic development, there are benefits which are particularly important to Germany. Current calculations predict a rising share of retirees in the total population, leading to a situation where fewer people in the labor force will have to sustain more people outside the labor force. This especially concerns the pay-as-you-go pension scheme, which is the standard in Germany. To sustain financial feasibility, it is hence all the more important to have sufficient numbers of people inside the workforce and avoid unemployment for reasons of scarce qualifications. From a civic viewpoint, education has been found to benefit democratically or-

ganized societies by increasing vote participation (Milligan et al., 2004) and the quality of civic knowledge (Dee, 2004). Moreover, studies have found a causal effect of schooling on the propensity to commit crimes. Exploiting law changes in the duration of compulsory school, researchers found that longer schooling/higher education reduces the risk of breaking the law (Lochner, 2004; Machin et al., 2011).[5]

A different way to look at the effects of education is to consider the costs that emerge by insufficient education. For Germany, the costs of insufficient education were estimated by Wößmann and Piopiunik (2009), who define insufficient education as having PISA test scores lower than a particular threshold which is associated with low competencies. According to their calculations (for data from 2000 and 2003) around 24 % of German pupils' attained PISA scores which were lower than this threshold. Defining insufficient education differently, namely as not having a school leaving certificate or having one less than A-levels but no vocational qualification, also Allmendinger et al. (2011) estimated the costs of insufficient education. Their definition concerns about 15 % of a yearly cohort at that time. Both author groups estimate the economic benefits if a certain share of insufficiently educated individuals had been sufficiently educated. The basis for the calculations of Wößmann and Piopiunik stems from a hypothetical educational reform which would reduce the share of the insufficiently educated by 90 % in 2010. The corresponding economic gains to the year 2090 would add up to 2,8 trillions in 2010–Euros. In the calculations of Allmendinger et al., the number of insufficiently educated individuals would be halved which would accumulate in net economic profits of around 615–1539 million 2011–Euros over a time span of 35 years.

From an individual's and the society's perspective, it appears therefore sensible to foster educational outcomes so that every individual attains the optimal level conditional upon their inherent potential.

Having demonstrated the importance of education in various contexts, the question arises what determines educational outcomes and whether it is possible to improve them. As the phrase "conditional upon their inherent potential" suggests, inherent abilities and other uncontrollable environmental influences may determine an upper bound. Remaining variation can be due to several sources, the most important of which typically being the environmental context in which a child grows up. When influences on school achievement are discussed, consensus is that a major part of this context is family background - as different as it may be (Funcke and Menne, 2010). Its relevance can be observed in sibling correlation studies or by looking at intergenerational changes in achieved education or wealth.

[5]Or at the very least: the probability of being caught.

Sibling correlation studies are based on the idea that if family or community factors have an influence on a child's outcome, the correlation between the siblings' outcomes should be higher than for two randomly selected children (Schnitzlein, 2014). The sibling correlation approach has the advantage of capturing observed as well as unobserved environmental factors that are common to the siblings. The shared factors are, however, not only related to family background but can also include community factors. While this method is able to provide a lower bound on the family influence, it neither readily identifies the underlying factors themselves nor their relative importance.[6] One insight that can be gained from such an analysis is the comparison of the extent of correlation between countries. Björklund and Jäntti (2012) examine the influence of factors shared by siblings on various outcomes such as schooling or long-run income for Sweden and find they account for about 40 % to 60 % of the variation. Schnitzlein (2014), using the same approach, compares Denmark, Germany and USA with respect to permanent earnings, and analyzes education and willingness to take risks for Germany alone. He finds that sibling correlation is comparably high in Germany and the US (40 %) and significantly lower in Denmark (20 %). The results also suggest stronger correlations for brothers than for sisters. For education the correlation rises to even 55 % – 65 % in Germany. Mazumder (2008) examined the issue for the US and finds a sibling correlation of about 60 % for years of schooling.
The OECD has published statistics on intergenerational changes, the so-called mobility, in the distribution of education or wealth. Compared to other OECD countries, mobility in Germany is especially low at the tails of the distribution. That means the odds of attaining higher (lower) education are considerably lower for children whose parents have low (high) education than for children whose parents have high (low) education. This observation has been made on other data sets as well (Baumert et al., 2001; OECD, 2012a; Pfeffer, 2008).
Taken together, the sibling correlation studies and the descriptive analyses indicate that children's achievements depend on their origin. If one assumes the distribution of inborn ability to be independent of the origin, this implies inequality of opportunity which results in a waste of potential.

Peculiarities of the German educational school system are suspected to carry some responsibility for this educational inequality (Hanushek and Woessmann, 2006; Bauer and Riphahn, 2006). The relatively early stratification into different tracks, known as tracking or streaming, leaves little time for a school to activating a child's potential before they are assigned to one of the tracks. Another factor that might

[6] A lower bound is estimated because there exist factors which are indeed related to the family but are not shared by siblings. One example are genes, which are fully related to the family, but only about half their genes are shared by siblings (Björklund and Jäntti, 2012).

contribute to this situation is the duration of school days. Although the number of full-day schools is on the increase, half-day schools are still standard in Germany. Compared to other countries, a German child spends less time in school and more time in the home environment. This factor leads parents to take on a weighty role in the track choice because it partly depends on how parents have aroused their child's potential themselves. Thus, family background of all environments plays a major role when it comes to activate a child's potential and the formation of human capital. The main economic theme of this thesis revolves around the topic of how the family and the social environment affects the development and thereby the human capital development of children.

Given the importance of family background, the question arises what exactly family background is, as the term itself is rather unspecific. Nechyba et al. (1999) emphasize that the parents' pronounced role lies in the numerous influences on the child such as their choice of community, the degree of their school involvement, and their influence on the child's choice of peers. Haveman and Wolfe (1995) provide a comprehensive overview of different studies and approaches to explain educational attainment. They note that in virtually every study parental influences are partly regarded by their education or income. This comes at no surprise given the high correlation of these characteristics with a child's school success. Björklund et al. (2010) demonstrate, however, that such attributes may not sufficiently account for the family influence. Applying sibling correlation mixed-effects models on a Swedish data set, the authors find that parental income, education and occupation do not explain even half of the sibling correlation for long-run income. Yet, there are two main reasons for reducing the family background to these two characteristics. Haveman and Wolfe (1995) point out that one is data scarcity – there were no other indicators for family attributes available. Secondly, education and income might have effects on their own but are also potentially correlated with other beneficial aspects of family background, for instance certain parenting styles. In particular, the correlation between parental education and child's school achievement is substantial, as depicted in the mentioned OECD-statistics. Hence, parental education predicts a substantial amount of the variation in manifest school success of children.

But what causes this relation? Black et al. (2005) summarize the two main lines of argumentation. One is a selection issue, the other is a causal effect. The first line of argument states that those who opt for higher education differ systematically from those who do not. An intergenerational correlation in education is hence caused by certain attributes which influence both the propensity to obtain higher education and nurturing quality. On the other hand, a causal effect could emerge through education when the acquired knowledge improves the skills to raise a child. At this

point it is not relevant whether a direct influence through education is assumed or an indirect influence, e.g. education rising expectations which then improves parental nurturing quality as argued by Davis-Kean (2005). The literature on causal effects of education discloses only small or insignificant effects, though (Behrman et al., 2002; Black et al., 2005; Plug, 2004). This indicates that selection is highly likely part of the story, albeit its exact influence is unclear.

The question remains whether education suffices as an indicator for the family environment, i.e. is parental education truly a sufficiently good projection of traits like low aspiration and disadvantageous parenting styles? There are counterexamples. Pursuing the selection story, education as indicator will be of minor relevance if parents did not go for higher education in spite of having the (latent) traits of a high-aspirational group. In the fifties and sixties of the last century, for instance, such a refusal may have been caused by societal or family-related constraints. Such constraints manifested in terms like the "catholic daughter of a working class family living in the countryside"[7] (Dahrendorf, 1965) - a term which condenses demographic characteristics linked to structural disadvantages in education - and also found its way as a theme in literary works (Hahn, 2003). Geißler (2005) notes in this context that although the disadvantages for some characteristics have receded over time (e.g. being female), others have remained or were added (male offspring with migration background). The study by Björklund et al. (2010) supports the argument: While demographic factors cannot explain half the variance in long-run income (more concretely: between 13 % - 28 %), the inclusion of parental variables relating to involvement in schoolwork, parenting practices, attitudes and the number of books at home increased the amount by about 40 percentage points.

To structure the discussion about parental characteristics, the concept of primary and secondary disparities by Boudon (1974) is used. Schindler and Reimer (2010) describe primary disparities as differences in educational achievement that are due to a different social origin. Secondary disparities, by contrast, refer to the notion that certain educational choices are made within a social context and that these choices are independent of the observed performance of the child. Together, these terms refer to mechanisms of the reproduction of social inequality which are especially influential at transition stages like the end of primary and secondary school. While primary disparities provide an initial distribution in scholastic performance (Paulus and Blossfeld, 2007, p. 495) which can, for instance, depend on income or having a migration background, secondary disparities refer to mindset-related factors like aspirations, the pursuit to maintain a certain status and decisions

[7] Own translation of the German wording: "Katholische Arbeitertochter vom Land"

based on cost-benefit analyses on investment in education.[8] The crucial point is that secondary disparities differ across social contexts but are alterable. A mindset can be changed more easily than income, when dependencies on environmental factors are taken into account. One example is parental aspirations that vary between social contexts, even if the child's ability was the same. Parents from the educated middle-class may not consider anything else but academically oriented schools - all irrespective of the child's inherent abilities, while parents from the working class may shy away from sending their child to higher education (Paulus and Blossfeld, 2007, p. 492). In this example, primary and secondary disparities work in the same effect direction, but, as demonstrated, this need not be the case. Stated differently, there might be heterogeneity in the secondary disparities conditional upon primary characteristics.

The notion of secondary disparities extends the prior thoughts on family background indicators and is now able to describe the case of lowly educated parents who provide a fruitful background for their child as well as the other way around. It is, however, more difficult to argue for highly educated parents providing a disadvantageous background since the motive of avoiding downward status mobility is more relevant. This could be reflected in increased expenses for measures that enhance the child's school achievement. Moreover, uncertainties with regard to institutions of higher educational pathways might primarily play a role among lowly educated parents. Highly educated parents have experienced the curriculum themselves and are also more likely to have access to sources of information about it.

Taking secondary disparities into account may hence help to capture heterogeneity in the family environment, in particular when parents do not exhibit a high educational background. This view is strengthened by theories in the sociological literature. With reference to family environments, Teachman (1987) writes, "Family educational resources have a positive impact on educational attainment of children net of demographic indicators of family background.". Similarly, Corak (2013, p. 98) refers to subtle ways, such as the family culture, and attests it a significant influence.

One disadvantage of the extension to softer family background characteristics are less tangible definitions. It is not clear how a fruitful background or family educational resources are defined. Secondary disparities may be understood rather as a multidimensional conceptual term rather than a clearly measurable quantity. A

[8] By cost-benefit analyses it is referred to the model of Breen and Goldthorpe (1997). In this model, the decision regarding certain types of education is swayed by monetary restrictions but also to subjective factors such as the current social status, experience, risk aversion and the (subjective) belief in a successful completion.

natural point to start the data collection would be to inquire the parents' aspirations and plans for the child. Even though it would be possible to ask parents about their aspirations and assessments of their child's achievement, it can be ambiguous how they are related to the child's development as demonstrated by Stamm (2005). In her article, she analyzes two expressions of parental aspirations: stimulus and demand. Stimulus refers to private actions which attempt to foster the child's educational achievement, whereas demand refers solely to parental expectations. Her results suggest that high educational expectations in combination with high stimulus are associated with good school achievements of the child, while the lack of stimulus leads to a relatively worse development even when aspirations are on the same level. Aspirations alone will, hence, not cover the important facets of family background. Moreover, they are always at risk of being confounded by the observed achievement, which leads to endogeneity problems.

A different way to approach the topic is to consider a family in its social context. Since people interact with their social environment in many a way, the latter may influence their characteristics. For parents, the social context can therefore have a significant bearing on factors like aspirations or stimulus which makes this aspect worthwhile to consider. From a broader perspective, the approach is in line with social structure concepts, which are frequently used in sociology to provide a description of a society's composition. The cornerstone of these models is the identification of groups in a society. These approaches group individuals based on common characteristics. Through societal changes and developments within sociology, the groups have evolved from classes, over strata to milieus over time and thereby become increasingly fine-grained (Bremer, 2007, p. 26ff.). In their definition, milieus are classifications that describe groups of people living in similar circumstances and having similar norms and standards (Hradil, 2006, p. 3). The focus lies, therefore, not on the position in a society as in the class concept, but on everyday life culture. Nowadays, the milieu concept is commonly applied to explain behavioral differences in the population, e.g. regarding participation in further education (Bremer, 2007).

The milieu-related norms and standards may also encompass the educational resources of interest. Whether this is the case depends on the relationship between the characteristics which constitute the milieu affiliation and the characteristics that are relevant for a child's school success. If the two are closely related or even the same, milieus pertain as valid approximations for the educational resources of interest and can be used to capture the desired characteristics. On the other hand, if a milieu classification is based on factors which are uncorrelated with the traits of interest, the concept is of no use in the context of this work.

Although the milieu concept appears to be unique in theory, there are several

approaches to form them in practice. These approaches differ with regard to the characteristics on which the milieu assignment is based as well as the qualitative identification and naming of milieus. This calls back to mind that the milieu representation is only a simplified model of a society and may contain errors. Irrespective of the concept chosen and the property of whether the milieus are valid approximations to the factors of interest, there are typically some additional concerns in empirical work. One is data availability and it applies especially to the less popular concepts. If the data are available, the next issue is whether the assignment of individuals to milieus is correct. It might be based on individual characteristics, but sometimes regional indicators such as location of housing are used. Another important aspect concerns blurred boundaries between milieus. Individuals may have characteristics of several milieus at once, which can be depicted by probabilities or relative shares. Assigning an individual to the one with highest likelihood, however, threatens a milieu's homogeneity and possibly its meaningfulness in a statistical analysis. These points are a brief outline of the difficulties in using milieu concepts in practice.

The alternative approach pursued in this treatise avoids some of the issues with existing milieu concepts. While the key idea of similar mindsets is retained, the empirical implementation differs. Fundamental is the notion that people who differ in the factors that concern the attitude towards education differ also in other characteristics, such as personality, preferences or behavior. This corresponds to the milieu notion that life environments differ across milieus. To elucidate the approach, it is useful to first review the idea of using milieus in abstract terms: Milieus, were they suitably defined for this topic, might not be observable without undertaking efforts to obtain the necessary information on educational resources in home environments. They are hence unobserved and must be derived. Yet, those milieus are assumed to influence parental characteristics and the child's school achievement. Such relations can be described by a latent variable model.
A latent variable model consists of two main components: a set of few latent (unobserved) variables and a set of observed variables. The latent variables are assumed to shape the observed variables to a certain degree. In this context, milieus can be considered as latent variables and the child's school achievement as an observed variable. However, the child's school achievement is not the only observed variable that is affected by milieu-specific characteristics. Since milieus are marked by common views on life, they can be reflected in many other observed instances such as personality, norms and attitudes, time use indicators and demographic information. Those patterns in observed characteristics are conjectured to allow inference on the hidden latent variables.

The approach in this treatise exploits this argument and does two things in addition. Firstly, it explicitly considers observed family background characteristics in the model. These are viewed as noisy products of the latent variables, which means that they are influenced by the latent variables but do not feature a one-to-one relationship with them. Secondly, instead of defining latent variables as a priori defined milieus, the observed characteristics are used to infer the latent variables. The approach therefore retains the basic structure of the milieu-model, but generalizes the latent variables to being more flexible constructs. Denoting those constructs as milieus would be somewhat assumptive, however, because milieu concepts have deeper-rooted theoretical foundations and it is unclear to which degree they can be identified in a given set of observed characteristics. Instead, the latent variables in this version of the model are called facets, common patterns, or dimensions of family background and correspond to latent variables within the sub-population of parents.
In sum, the approach generalizes the principle of milieu concepts towards facets of family background. The main components of the model are the latent variables, which are assumed to influence both the set of observed characteristics and the child's achievement, and makes their identification of interest. To identify them, structures in the set of observed variables are exploited. Those variables are, however, measured with noise such that there is no one-to-one correspondence to the latent variables. Moreover, they can be influenced by several latent variables at once. An illustration for the model concept is given in figure 1.1, where relations between the elements of the model are indicated by the arrows' directions. In practice, the strength of relations varies and all latent variables are connected to all observed variables. For clarity, such details are omitted in this figure.

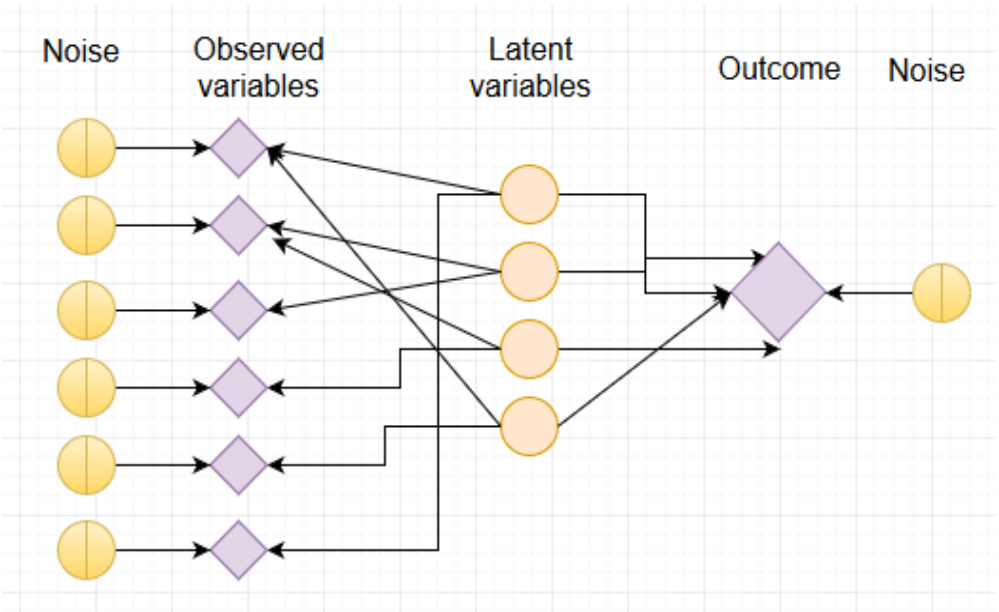


Figure 1.1: Latent variable model

Practically, the dimensions of family background are of key interest. However, they are not known beforehand because educational resources can be of manifold type and substitutable. Neither is it known, which observed variables are required to identify them. One example for this is the number of books at home that might well correlate with parental valuing of education; the latter could be viewed as a latent variable of interest. However, the number of books is typically measured with error and the same latent variable can find its expression in many other ways, too. As a given set of observed variables plays the major role in identifying latent variables of interest, it is crucial to have a theoretical foundation for their selection. Since the only interest lies in the family background, the rationale is to choose family background characteristics according to whether a child is in some way exposed to them. This ensures that the latent variables relate to the child as well. By focusing on family background characteristics, any school-related information, about the peer-group or gender-specific (dis-)advantages for instance, is not considered by definition.

Exposure to parental characteristics can be summarized in four (partly interrelated) categories: parental time use, parental attitudes and preferences, parental personality and general endowment characteristics, such as demographic indicators. Besides a possible direct exposure, there is a second way to affect the development of children. For a number of characteristics, the empirical literature has discovered a correlation between generations. This observation is interpreted as an intergenerational transmission. It extends the parental influence to an indirect effect when the child takes on traits that influence its educational achievement.
Based on the four categories, various characteristics are selected that function as observed indicators of family background. For some of them a direct or causal relation to the outcome has been detected in the literature. But, by virtue of modeling the problem as a latent variable model, every characteristic is primarily interpreted as stemming from latent variables. And although the selection is focused on parental indicators, the notion of a social environment's influence is also contained because it is likely reflected in various characteristics.

As a result of the rationale for selecting variables and the comprehensive data set at hand, the approach pursued in this treatise is based on a large number of observed parental characteristics. These are used to infer typical facets of family background or the parental mindset. Given the theme of heterogeneity within groups of education, the economic premise of this dissertation is to explore the patterns within discovered facets of family background and to evaluate whether, and how much, they contribute to explaining a child's school achievement. In the process of evaluating those facets, the importance of the standard demographic

indicators education and income can be assessed. This allows the identification of two cases: The first is to detect significant associations of facets that have hardly a relation to the standard indicators education and income. This case would confirm the hypothesis of remaining heterogeneity in levels of education and income. Otherwise, all relevant dimensions of family background relate to these indicators in some way, which would not support the stated hypothesis.

The modeling strategy that has been described up to here is expanded in chapter 2 which provides the theoretical foundation for the approach of this thesis. The chapter elucidates the issue why milieus are of interest in theory but may not suitable for the identification of educationally relevant facets of family background in practice. The outlined approach is explained more exhaustively, particularly addressing why it overcomes some of the disadvantages of using the milieu concepts. The chapter also discusses the variable selection rationale and elaborates on the four channels in greater detail. This discussion, in particular for the provided examples, takes place in deference to the data set at hand and is hence in anticipation of the empirical analysis at a later point.

One challenge that arises in an empirical treatment is that the set of selectable predictors can be large in relation to the number of observations. Transmission channels, like the time use/activity channel, can comprise many variables because there are many activities whose frequency could be measured. A further property of such a data set is the existence of clusters of correlated variables, which can be exploited and should not be ignored. Chapter 3 demonstrates the pitfalls of standard regression methods in such data-rich environments. The main argument is that the inclusion of too many variables leads to sample-specific models that can hardly be generalized. Choosing a set of variables beforehand is arbitrary, leads to unnecessary exposure to measurement error and would also not correspond to the argumentation of a latent variable model. The data structure requires using techniques that directly address the properties of the data set.
Reducing the dimensionality, which means to reduce the number of explanatory variables, is an approach to deal with such a data set. Such a reduction facilitates the interpretation of the resulting model because fewer variables need to be considered. In this dissertation, two classes of methods that achieve a dimensional reduction are examined in greater detail: Methods forming indices and a certain type of shrinkage (regularization) methods. As there are many candidates within the realm of these methods, the examination concentrates on the subset of suitable and popular ones, namely principal component regression types, factor regression types and regularization methods with variable selection.
An empirical depiction of the latent variable model theory is achieved most closely

by an index model. The basic idea is to combine the observed variables to indices, whereby the total number of indices is chosen to be smaller than the number of original variables. The resulting indices may be interpreted as the latent variables of interest and, hence, be subject to further analysis. Shrinkage methods, on the other hand, use all characteristics in a regression model but penalize a large effect size by pulling the coefficients towards zero. A variable selection takes place, as only the characteristics with a coefficient absolutely larger than zero are kept. Uncovering latent family facets is, however, difficult using these techniques. On the other hand, they are known for finding sparse models with good predictive performance and give a hint on the strength of the single predictors' association, an information that is not immediately apparent for index models. These arguments make it worthwhile to examine regularization methods. As it is unclear whether one of these methods is preferable to the others, a simulation study is conducted. In chapter 4, the selected methods are detailedly discussed regarding their dimension reduction feature and economic interpretability of the statistical output.
The simulation study presented in chapter 5 evaluates the performance of the introduced methods under conditions as described in chapter 3. Two models of data generation are studied in this context. One is the latent variable model, the other is a model in which the observed variables directly affect the outcome. Starting with baseline models which attempt to approximate the structure of the empirical data at hand, the model parameters are changed under different scenarios. This yields varying data structures, and thereby exposes the conditions on which a method's performance depends.

Using the results of the simulation study, the most promising method is used to examine the topic empirically with survey data. The empirical analysis is presented in chapter 6. The data set used for this analysis is the German Socio-Economic Panel Study (Wagner et al., 2007), short SOEP, which is a representative panel survey containing data about 11000 German households per year. Each year the DIW rolls out a comprehensive questionnaire asking for demographic aspects as well as life circumstances. Since 2000, the household head's children, when they turn 17, have been invited to the survey. This special 'Youth questionnaire' is useful in this context, as it asks for school-related information, such as the school type and the last grades obtained. Moreover, detailed information on the family environment during the childhood is gathered. In 2006, the DIW introduced a supplementary cognitive skills test. All those data can be matched to the parents' data from the normal questionnaires. Details on this matching process as well as on the generation of the endogenous variable(s) and the treatment of the explanatory variables is described in chapter 6.

This work adds to the literature in several ways. Motivated by the need for method selection specific to this data environment and the requirements, the simulation study additionally attempts to address several general research gaps. This is done by carefully examining methods whose properties fulfill the conditions of dimensional reduction and interpretability. Different data environments allow an evaluation of the conditions under which methods perform well or not. Owing to the interpretable model condition, rarely examined methods are included. Many simulation studies focus on predictive capability only, a fact that often excludes methods which aim to produce interpretable results. Another point not often addressed is the difference between in-sample and out-of-sample performance. This omission is examined here and gives an indication of the method's ability to produce sparse models. Moreover, two methods combining the strengths of index building methods and shrinkage models are proposed and examined for the first time. Since only general features of the data are defined, the insights from the simulation study can also be useful in other contexts where similar data structures and aims are present.

Concerning the implementation of the described theory, the following aspects add to the existing literature. Although it is not a novel idea to examine the relation between social milieus and the child's school success, there are, to my knowledge, only qualitative analyses on this topic. The connected idea of using characteristics beyond demographic ones is also not new, but the implemented approach is the first one that explicitly aims to detect facets of family background. Founded on the idea of a social milieu's influence on observed characteristics, the approach addresses the potential drawbacks while retaining the notion of identifying patterns in society by means of observed characteristics. Owing to the latent variable model implementation, the facets are based on a conglomerate of relevant indicator variables. Thereby the approach circumnavigates the reliance on single indicators and instead provides information on family background structures. While no causal relationship can be established, this analysis gives insights into the association of these structures with the child's school achievement. This may increase the understanding of different life environments and potentially also their substitutability. The approach also examines the relation of the standard demographic indicators to other, softer indicators. On the basis of these results, the approach additionally adds to the literature by identifying key indicator variables that support future research taking account of the heterogeneity in family environments. Since the set of included characteristics is relatively large, the results may be useful in other data sets than the SOEP, too.

Chapter 2

Modeling approach

2.1 Milieus

This chapter starts with a theoretical discussion of why milieus could approximate the family background characteristics that are relevant for the educational success of a child. Despite their theoretical appeal, drawbacks related to the definition of milieus occur in practice which motivate introducing a modified approach. Its starting point is to describe the problem as a latent variable model, in which milieus are interpreted as the latent variables. The proposed approach improves upon some of the drawbacks by generalizing the framework to a more flexible variant. After the concept of this approach has been elucidated, the importance of identifying the latent variables in this model is pointed out. As a consequence of the approach and the model, their identification depends on having suitable parental characteristics. A debate on how to select those follows upon that.
The selection rule emerging from this discussion requires that parental characteristics describe the family background to which the child is exposed. Since this rule does not limit the parental characteristics substantially, the chapter's remainder deals with the choice of variables in greater detail. Its focus lies on characteristics which are available in the data set at hand, so that the description occasionally anticipates section 6.1. The discussion is structured by grouping parental characteristics into four categories, which are personality, attitudes, time uses and demographic characteristics. Some parental characteristics can exceed the role of being an expression of a family background dimension. Then, they may involve a specific influence on the child's achievement. This may occur indirectly via intergenerational transmission of characteristics or directly by affecting issues concerning the child's scholastic performance. Relevant results from the literature on such matters is recapitulated during this discussion.

As argued in the introduction, the central idea is to find characteristics which indicate the parental (dis-)approval of a child's education. This is of interest, since the attitude towards education is influential for exploiting the scholastic potential a child has. Disapproval of education can concern the school achievement in unfavorable ways. It is unlikely to lead to the necessary encouragement, which leads eventually to a worse than possible outcome. A parental home, however, which provides a fruitful background, can pave the way for a good development of the child. The arguments given in the introduction point at possible deficiencies of demographic indicators. They are suspected to insufficiently represent the manifold family background influences on the child's school achievement. The factors of interest are arguably related to the parental mindset because it influences preferences, also towards education matters, which influences the choices made. These choices in turn affect the environment in which a child grows up and hence its development.[9] A parental mindset is, however, a vague term, which can have many manifestations. Since it is unclear which dimensions of the parental mindset are of interest, observations from sociology are used to approach the topic.

In the broader context of society, sociological studies point out that an individual's mindset bears resemblances with the social environment that surrounds the individual. This environment is called milieu. By definition, this term describes a group of people living in similar circumstances and having similar norms and standards (Hradil, 2006, p. 3). Those who belong to a certain social milieu interpret and shape the environment in similar ways and create a group affiliation by dissociating themselves from other milieus (Barth and Flaig, 2013, p. 12). Given this definition, it is not clearly defined in which respect people in milieus resemble each other. In principle, this notion permits likeness in aspects that are relevant for a child's school success. For example, there could be a milieu which comprises parents who value education and act accordingly. This could find expression in appreciating a child's efforts in school or the encouragement to achieve a good scholastic result. By contrast, another milieu, disapproving of education, might be marked by doing exactly the opposite. If these milieus were identified and quantified, they could be examined closer. For instance, their relation to demographic characteristics, such as parental income and education, could be scrutinized.

These thoughts demonstrate the power the milieu concept has in theory. In comparison to other models of society, it has several properties rendering it suitable for this analysis. First and foremost, it emphasizes the meaning of less tangible characteristics. Albeit rather generally defined, a milieu is demarcated from the

[9] For the time being, it is assumed that both parents share the same mindset and therefore offer a homogeneous family background. This assumption is made for simplicity's sake and is discarded in the empirical analysis.

concept of social classes in that it considers characteristics such as mentalities in addition to demographic ones such as income or occupational status (Hradil, 2006, p. 5). Moreover, milieus are often considered with regard to close surroundings, but this need not always be the case. Durkheim (1988, p. 44) distinguishes two dimensions of milieus. One is called the objective dimension, which is tangibly defined by close relationships, for instance, to relatives or colleagues; the other is the subjective dimension, which is based on common moral rules and specific bearings of a group of people for whom a close relationship is not required. One difference between the two dimensions is the degree of choice: The objective dimension can be taken as given, whereas the subjective one reflects decisions in the past. But since the two dimensions can greatly overlap, e.g. through self-selection, they are usually indistinguishable.

In view of these arguments, a mindset and a social milieu feature likely a close relation. However, a milieu always remains a simplifying abstraction. A mindset is an individual factor, whereas a milieu subsumes a larger group of individuals who resemble each other in some predefined criteria. Hence, a milieu is at best homogeneous only with respect to those criteria, while other (important) aspects of a mindset might nevertheless differ.[10] And it is indeed the definition of those criteria which causes problems with using milieu indicators for the purpose of this dissertation.

Before this issue is examined in detail, it is beneficial to address the question of the mutual dependence between milieu and mindset, i.e. to which degree a milieu shapes a mindset or the other way round. While a social milieu holds up certain norms and rules to which an individual is expected to conform, the individual needs to select this milieu beforehand. This choice is based on individual preferences and it seems plausible that a milieu is chosen according to the largest conformance with individual personality and opinions, true to the motto 'birds of a feather flock together'. However, an individual's preferences are not independently given. Not only are they malleable but they might also have been previously shaped by social interactions and can therefore not be regarded as exogenous. Barth and Flaig (2013, p. 14f.) name three factors influencing the milieu choice: current societal norms, individual inclinations and predispositions and the social environment of the family home. The habitus concept by Bourdieu (1987) supports dismissing exogenous preferences by emphasizing the importance of childhood for forming preferences. Despite arguing in the framework of social classes, the insights can be

[10] Some authors go as far as to reject the notion of a societal influence on the individual's behavior. In his individualization thesis, Beck (1986, Ch. 5) argues that one's existence has become increasingly independent of socio-structural aspects and core values have to be derived from one's own biography.

transferred to milieus. He explains the affiliation to a social class as the result of an individual's economic capital (wealth), social capital (relations to other people) and cultural capital (education, knowledge of the culture). According to Bourdieu, a certain habitus, i.e. a specific thinking and behavioral pattern, is unknowingly created through growing up in certain circumstances. Thus, one can conjecture that the magnitude of each type of capital has been significantly determined in childhood already (Vester, 2009, p. 39f.; Hradil, 2006, p. 5f.). A complementary note concerns time stability of preferences: Barth and Flaig (2013, p. 16) state that (milieu) preferences are relatively stable from adolescence on although changes are within the bounds of possibility. Hence, changing milieus as an adult is still possible.
Given these arguments, it is normally indistinguishable whether immanent individual or environmental factors cause preferences at a certain point in time. This, in turn, makes it difficult to establish whether oneself or the social milieu accounts for the observed characteristics. Are, for example, an individual's frequent cinema visits due to the milieu in which frequent visits are the norm or due to the individual's genuine interest in movies? Without additional information on the individual biography or friends and colleagues, both explanations are equally plausible, so is a combination of both. The same argument pertains for other characteristics, too, e.g. educational aspirations for the child or the time investment into the child.

The data used in this work do not provide any information about this issue either. However, for the theoretical justification and the approach that follows, this differentiation is no longer necessary. This is because the approach does not rely on predefined milieu indicators but on patterns in observed characteristics. Referring to the above example, only the fact that an individual frequently goes to the cinema is regarded as important. To develop the argument for this approach, it is nevertheless necessary to maintain the differentiation between the individual mindset and milieu influences.

Up to here, milieus have been outlined as groups within a society that live in similar circumstances and have similar views on life. Next to it, the concept of parental mindsets was introduced. For a statistical analysis, it is necessary to have a clear operationalization for a concept. This raises the question of measurement, which yields a twofold answer. A parental mindset can be considered too comprehensive, if it is not narrowed down to specific dimensions. Milieus, by contrast, are found to be defined quite clearly in practice since several scholars have dedicated themselves to this topic. Depending on the definition, there are different criteria along which individuals are assigned to milieus. These resulting groups are identified and qualitatively interpreted. Practical milieus concepts differ in the classification

criteria but also in the number of identified milieus and in their qualitative interpretation. There being available several approaches to milieus, the apparent uniqueness of this concept is undermined. Hence, practical milieu concepts differ from ideal milieus in the sense that they might not yield an optimal classification with respect to the child's school success.

In many cases, individuals are grouped according to two criteria: The social status and basic life-value-orientations. The criteria are used by the most popular milieu concepts, the Sigma milieus® (s. for instance Ascheberg, 2006) and the Sinus milieus® (s. for instance Sinus Sociovision, 2005). Hence, any distinction between milieus is attributable to a combination of these two criteria. Other concepts are put forward by Schulze (Schulze, 1992), called Experience milieus [11], and Vester et al. (Vester et al., 2001), called Agis milieus. A further alternative are Delta milieus® by Wippermann (Wippermann, 1998). Table 2.1 presents an overview of these milieus' characteristics. Details such as the number of milieus, their labels and descriptions are omitted owing to variation over time.

Table 2.1: Overview of renowned milieu concepts (for Germany)

Name	Author	Criteria
Sinus milieus®	Sinus Sociovision GmbH	Social Status / Basic life-value orientations
Sigma milieus®	SIGMA Gesellschaft für internationale Marktforschung und Beratung mbH	Social Status / Basic life-value orientations
Experience milieus	G. Schulze	Preferences for complexity / simplicity and order / spontaneity
Agis milieus	M. Vester	Habitus / principles of lifestyle
Delta milieus®	DELTA-Institut für Sozial- und Ökologieforschung GmbH	Social Status / Basic orientations

Notwithstanding the concepts can be similar in their criteria, they differ in their segmentation and their qualitative interpretation. The question of which concept to choose emerges as a consequence. This in turn requires to assess which concept qualifies best for this research topic. In an ideal concept, these milieus would separate family environments into different groups by their degree of appreciation of educational efforts of the child. In the real world, however, the grouping criteria are not necessarily related to this. The main objective of milieu concepts is to describe social environments of a society, sometimes the concepts are adduced for commercial use. The concepts group individuals who are similar in attitudes towards work, family, leisure and consumption (Schräpler et al., 2010, p.10). Whether these criteria coincide with parental factors of interest, is indeterminate

[11] Personal translation of "Erlebnismilieus"

a priori. Bremer and Kleemann-Göhring (2012) brought forward arguments in favor of it for Sinus Milieus but they remain of qualitative nature. The different objectives form the critique on using practical milieu concepts in this context.
Practically, there are precisely two ways that existing milieu concepts qualify as useful for this research. This applies if the criteria on which the segmentation is based either constitute the characteristics of interest or correlate strongly with them. But neither of the two is ensured. The first case can even be excluded, since the effort to enhance a child's school achievement is not a segmentation criterion in any of the milieu concepts. This has the simple reason that parents constitute a subset of society. As it is already challenging to determine nurturing quality for parents, determining it hypothetically for non-parents is even harder. Hence, the correlation option remains the sole possibility.
For Sinus milieus this implies that income and basic life-values must be able to explain variation in a child's school success. If the two criteria were unrelated to the child's outcome, the created milieus would be, too. One must hence conjecture that the factors of interest are systematically related to the classification criteria that form the milieus in order to make use of the actual milieu indicators. Were the classification criteria sufficiently good, however, one could pass on milieus and instead work directly with these criteria. The extensive qualitative descriptions are the only advantage milieu concepts offer in this respect.
A further critique of milieu concepts concerns the individual's assignment to a milieu. For some individuals boundaries between milieus may overlap, i.e. an individual might stand between two or even several milieus. Such overlaps could be reduced by expanding the number of milieus. However, a concept with countless milieus is not convincing, as it is after all a model intended to reduce reality's complexity. Overlaps are hence acknowledged and could be expressed by probabilities. If one wants to work with clear-cut assignments, in contrast to probabilities, though, an assignment has to be made. In such cases, a difficulty lies in imposing the correct milieu on an individual. The decision is likely to increase the heterogeneity within a milieu because the resemblance of individuals assigned to it decreases; in a statistical analysis, this may change the interpretation of the milieu indicators and thereby the results. Moreover, in the case of several classification criteria, whereof only one is suitable (i.e. correlating with the factors of interest), an error is made if the assignment is predominantly based on the unsuitable ones.
While the previous point relates to the way in which milieus are considered in statistical analysis, a further issue concerns the collection data and how probabilities are calculated. This is often based on large-level demographic features such as a street or a quarter, but not necessarily on an individual basis. Underlying this is

the assumption that residence and milieu correlate, which is not unreasonable, but neglects potential heterogeneity.
The final point of critique applies to the time-varying nature of milieus. As societies and preferences change, old milieus vanish and new ones arise. Many milieu concepts are updated to keep track with these developments by shifting the discriminating boundaries or relabeling the identified milieus. Yet, it is unclear how to compare different milieus over time if the data cover a longer period.

The previous arguments indicate that several limitations may occur by using milieu indicators. From a theoretical point of view, the first point referring to the classification criteria is the most crucial one: It cannot be ensured that the classification criteria are sufficiently correlated with the factors of interest. Practically, the availability of data in reasonable quality constitutes a second important hurdle.

The approach in this work extends the notion of milieus. The central idea is to view the problem in the context of a latent variable model. This model assumes the existence of latent variables, which impact on all observed characteristics including the child's school achievement. Owing to this mechanism, patterns between certain parental characteristics and the scholastic outcome arise. In the framework of milieu theory, the latent variables can be interpreted as the milieu indicators since they are thought to stand behind specific customs and attitudes which influence the child's school achievement. The patterns in parental characteristics would then reflect the existence of such milieus. To avoid the drawbacks of milieu concepts in practice, the model is generalized, however. By defining the latent variables as the (educationally relevant) facets of family background, which may include aspects of the parental mindset and social influences, a more flexible variant arises. On the downside, the latent variables are no longer accessible, so they have to be derived from patterns in the observed characteristics.
While latent variable models and milieus have not been linked in the literature, the modeling approach in this thesis explicitly assumes such a model. The open definition of the latent variables implies that they need to be derived instead of predefined. As their qualitative definition lacks, they are no longer milieus but instead labeled as facets of family background. They are assumed to have a bearing on both parental observed characteristics and the child's achievement. From a latent variable model view, the observed characteristics are interpreted as expressions of the latent variables.
As to the underlying rationale, the example from the introduction serves as a suitable case. The observed variable of the number of books at home might be a noisy proxy variable for a possible latent variable that relates to parental valuing of

education. Yet, it is not necessarily the number of books at home, but rather the latent variable that causes both the purchase of numerous books and the educational stimulus for the child. This is an example of a one-to-one correspondence, but a single latent variable can also influence multiple characteristics. Moreover, the configuration of the model allows the case of multiple latent variables influencing a single characteristic.

Describing the problem as a latent variable model addresses the drawbacks of the milieu approach for various reasons. Generally speaking, there is some similarity to the milieu approach because the model attempts to depict typical patterns occurring in society. But instead of focusing on the social status or basic life values alone, the approach is able to consider a variety of possible patterns in family background. This fact alone does not render the approach superior, however. Certain facets of family background could occur more often in certain milieus than in others. If these facets are significantly associated with the child's school success, milieu indicators are also useful. But as argued, this is not ensured. The proposed approach addresses the main disadvantages of the milieu approach as follows:
Firstly, the extraction of relevant facets is data driven; it is not predetermined (up to the general selection of input variables) which characteristics exert the discriminating function. This avoids prior commitment to using characteristics, which are possibly unrelated to the outcome, and ensures flexibility.
Secondly, the facets' relations to the child's achievement are explicitly taken into account, such that the facets are ensured to be relevant. Taken together, the two points address the argument that milieus are indeed good descriptions of societal groups but do not necessarily correlate highly with aspects that regard education. The proposed approach instead identifies typical patterns of family background that are relevant conditional on the data at hand.
Finally, there is no binary assignment to a group but rather a gradual propensity to a pattern or a facet. There may even be a vector of facets, representing the multiple dimensions of family background and possible manifold aspects of home environments. This provides finer-grained information than the dichotomous assignment to a milieu and could therefore yield additional insights.

In sum, the conjecture of this approach is that parents who differ in their observed characteristics, differ in their mindset and therefore also in factors that concern the child's educational success. In this case, the patterns in the observed characteristics can be used to draw conclusions about beneficial or disadvantageous family background factors. The linchpin of the approach is that the whole idea rests on the choice of inputs, i.e. the selection of parental characteristics which constitute the observed characteristics. If they are unsuitable, the discovery of relevant

facets fails. Bad characteristics either exhibit little variation or are independent of the latent variables of interest. An artificial example for little variation is the number of ingested meals per day, which can be expected to be roughly similarly across family environments. Choosing some bad characteristics is not problematic, however, as long as there are sufficiently many other useful characteristics, which provide the desired information. Since it is not readily clear which characteristics appertain to this, the following section's focus lies on this issue.

2.2 Family background indicators

The proposed approach requires finding observed characteristics which relate to the latent variables of interest in some way. As the latter are unknown, however, the selection must be based on a different reasoning. Here the idea is to gather as much information on the family background as possible. Once a characteristic is related to the family background, the child is likely to be exposed to this characteristic because it lives in this environment. From the exposure, an influence can be inferred. Whether this influence is decisive for the educational achievement, is not important at the selection stage - detecting relevant relations is done in a second step distinct from the theoretical considerations that follow. The characteristics in sum serve the purpose at finding significant patterns across families. On the face of it, one can divide parental characteristics into four broad categories:

- Personality traits
- Attitudes
- Time use indicators
- Demographic indicators

Based on the remarks on what characterizes milieus in theory, it becomes apparent that the last three groups of variables can be connected to milieus because they relate to similar attitudes and time uses. Additionally, milieus are linked to demographic indicators such as living in certain circumstances. However, it is not possible to clearly separate family-idiosyncratic and milieu characteristics. By gathering information on family background characteristics, possible milieu influences might be captured simultaneously. Before examples for the four categories are named, some general remarks about their scope are given.

While time use indicators and demographic indicators can be easily demarcated from the rest, delimiting personality from attitudes is sometimes hazy. One example of such a case is risk preferences. This name suggests they belong to attitudes, yet they are closely connected to personality and some researchers indeed view them as that (e.g. Checchi et al., 2014). A reason for potential ambiguity originates from the close relation between personality and attitudes. Some personality traits often accompany certain attitudes. Moreover, since personality is not yet physically measurable, indicative information drawn from attitudes is used to infer personality. However, the scope of attitudes goes beyond providing measures for personality. This work follows the elaborations by Ajzen (2005) who provides the following distinction between personality and attitudes on the first pages of his book:
Elements belonging to personality characteristics are categorized by drawing on the trait concept in social psychology. Personality traits "describe response tendencies in a given domain, such as the tendency to behave conscientiously, to be sociable, to be self-confident, and so forth." (p.6). They hence refer to general dispositions, i.e. tendencies to respond in particular manners to certain situations.
Elements of the second group are attitudes to which also norms and preferences count. The definition of attitudes has been subject to various changes over time. Nowadays, the evaluative character is considered to be the main attribute. This evaluation often consists of 'pro-con' or 'pleasant-unpleasant' statements directed towards a specific object or topic. Attitudes are thus heavily influenced by morals, (social) norms and beliefs. Like personality traits, the concept is not physically tangible and needs to be approached by overt or covert responses of the individual of interest.
The main difference between personality and attitudes is hence the specificity and evaluative nature; importantly however, Ajzen (2005) further notes that attitudes are typically seen as more malleable than personality traits. In case of new information about an issue, beliefs can be updated and attitudes change. The ability to change can be of economic importance if attitudes are to play a major role for the child's school success and justifies the differentiation between the two categories in this work. Also, the empirical results by Becker et al. (2012) indicate the complementary relation of personality and preferences. The authors scrutinize the relation between personality and preferences such as risk and trust preferences. There being some exceptions, their results generally indicate that personality and preferences are hardly related and so to be treated complementarily.
Time use characteristics are more easily defined, as they measure the frequency and/or the duration of activities. Therefore, the term 'time use' also encompasses behavioral aspects.
Demographic characteristics encompass factors such as age, family size or income,

which may or may not reflect certain latent traits but are often related to a social environment. Such information adds to the family background picture and is therefore regarded.

As indicated above, the categories cannot be considered as independent from each other – mutual influences are likely to occur. For example, it can be argued that personality traits influence attitudes which in turn may have a bearing on the type of activities. If the relation was sufficiently strong, details about attitudes and activities would be rendered redundant, as personality traits already contain the important information. However, the relation might not be as strong, because preferences exhibit a higher degree of malleability. An additional argument is provided by the theory of social milieus (Hradil, 2006): Attitudes and behavior can be significantly influenced by the social environment, depleting their dependence on personality traits. Personality traits, on the other hand, are not only considered more stable than attitudes but they are also less likely to be related to milieus. In sum, characteristics of these groups are expected to give a description of different facets of family background. They cover parental characteristics, their social environment as well as demographic indicators.

Up to here, the condition for considering a certain variable is that it describes the family background to which the child is exposed. The function is hence restricted to being a descriptive element for the facets of interest. For some characteristics, however, there is evidence of being more than a mere proxy variable. An example is parental risk preferences, which have been found to influence educational decisions (e.g. Checchi et al., 2014). When parents have to make educationally relevant decisions that contain a risky aspect, a direct influence from the parents' characteristics on the child's outcome is conceivable. There is an additional line of argument through which some parental characteristics exert an influence on the child. It has been observed that many parental traits correlate with those of their offspring (Duncan et al., 2005).[12] This observation is called intergenerational transmission and has been found, for instance, for personality traits (Zumbühl et al., 2013), attitudes (Dohmen et al., 2012) and time use pattern (for instance in employment: Couch and Dunn, 1997).
Different mechanisms for this observation have been suggested, whereby the mechanism typically depends on the specific trait. Some psychologists argue for a genetic origin in the case of personality traits. Duncan et al. (2005) name socioeconomic resources, parenting practices and role modeling as further mechanisms. The role

[12]It has to be noted that speaking of parental traits only is a simplification. Correctly, the traits should refer to the role model of the child. However, there are only few cases for which biological parents deviate from social parents in the empirical analysis of this work, which justifies the linguistic shortcut.

modeling argument is often used with regard to time use, for which it is called learning by imitation. In some of these instances, the role of learning to appreciate something is stressed, e.g. high cultural activities like theater or museum visits (comp. de Vries and de Graaf, 2008). Dohmen et al. (2012) scrutinized the topic for attitudes. Their findings emphasize the effects of socialization, i.e. by the child's parents and their local environment. The transmission becomes stronger, the more similar both parents' attitudes are to each other. The behavior of finding a partner who is similar to oneself, the so-called assortative mating, may be considered a strategy to pass one's attitudes more effectively towards the next generation. But not all children are equally strongly influenced. The results by Zumbühl et al. (2013) suggest that the extent to which preferences are transmitted depends on the parental time investment in common time with the child.
In sum, children are influenced by their parents traits in two ways: They are exposed to the parental characteristics in general, but might also take on their traits which in turn has an influence on factors like success in school. With this twofold influence in mind, traits of the four categories are detailedly described in the following, giving concrete examples and summarizing the relevant literature. The examples have been chosen chiefly with regard to availability in the data set at hand that is described in 6.1.

2.2.1 Personality traits

Parental personality traits can have a significant influence on the family environment, for they have a bearing on decisions like where to live, the organization of life and how much social contact is around – factors a child is directly exposed to. Their personality might also influence school-related decisions. These are examples of direct influences and they are complemented by the intergenerational transmission of traits to the offspring. Its relevance for the school achievement results from the child taking on those personality traits. This in turn might influence certain school-related behaviors and so affect its achievement in the long-run, resulting in two ways parental personality can affect the child's achievement.
Having pointed out the importance of parent's personality in abstract terms, it remains to substantiate its meaning and how it is operationalized. The impossibility of a physical measurement has led scientists to theorize and set up models. No model is left undisputed, but there are some more prominent ones. A common, probably the most often used taxonomy to measure human personality is the Big-5 scheme by Costa and McCrae (1985). The authors identify five broad dimensions or tempers, which are depicted in table 2.2. Examples of typical personality traits are used to describe each personality dimension (a dash denotes an untypical

Table 2.2: Overview of Big-5 personality traits.

Dimension	Typical attributes
Openness	is inventive
	likes to reflect, play with ideas
	values artistic/aesthetic experiences
Conscientiousness	tends to be disorganized (–)
	perseveres until a task is finished
	does a thorough job
Extroversion	is talkative
	is outgoing, sociable
	is sometimes shy, inhibited (–)
Agreeableness	is sometimes rude (–)
	is considerate and kind to almost everyone
	has a forgiving nature
Neuroticism	gets nervous easily
	is emotionally stable, not easily upset (–)
	is relaxed, handles stress well (–)

personality trait).[13] The Big-5 model has several beneficial traits: Firstly, it is able to account for different traits mostly without intersections. Secondly, it has been used frequently, and has shown stable results even in different cultural contexts (Dehne and Schupp, 2007, p.26).

Big-5 personality traits were found to be linked to many outcomes, some directly, others indirectly. Next to the Big-5 scheme there are also numerous other concepts, but Saucier and Goldberg (1998) supply evidence that this concept contains many others and is hence a suitable candidate to model personality. However, another model of importance is the locus of control framework by Rotter (1966). It is of interest here because of its implications for economically relevant decisions (comp. Becker et al., 2012). The concept emphasizes the (dis-)belief of having control over circumstances in life. In contrast to five-dimension model before, this concept has a single dimension with the internal locus of control on one side and the external one on the other. Being closer to the internal locus of control means a higher degree to which events are seen as the consequence of one's own behavioral actions, while the tendency to be on the opposite side is higher the more events are viewed as under control of others or due to luck/destiny/chance. The implications can be far-reaching and influence important attitudes. A high internal locus of control is, for example, associated with a higher intensity in job searching and higher

[13] For a more detailed overview, s. McCrae and John (1992, p. 178f.).

reservation wages (Caliendo et al., 2015).
Generally, both personality concepts are likely to exert an influence on the characteristics that are described in the three following subsections, i.e. attitudes/preferences, time use indicators and demographic characteristics. However, the evidence is not unequivocal as the mentioned study by Becker et al. (2012) suggests.

With the difficulties defining personality and its multiple influences, any identification of a causal effect is a difficult task in many respects. But since this is not the objective of this work – associations are sufficient here – it remains to introduce what the literature has found about the relation of personality traits to educational achievements. With the double-track influence in mind, this paragraph continues with the associations of personality with educational success. This part refers to the intergenerational transmission of traits, i.e. the indirect link between parent's personality and a child's school success. Subsequently, educationally relevant correlations on the child directly brought about by the parental personality itself are presented.

General results are provided by Cunha et al. (2010) who estimate the non-cognitive skills' share (under which personality traits fall) of variation in educational achievement at 12%. Komarraju et al. (2009) find positive associations of the Big-5 factors conscientiousness and extroversion with the college grade point average. The analysis by Anger (2013) distinguishes between children from low and high socio-economic status (SES) and two outcomes (University-entrance diploma and university degree). Her findings indicate that openness to experience is beneficial for males with low SES, whereas personality traits play no role for those with high SES. The results for young women are mixed. The analysis by Peter and Storck (2015) investigated the relevance of personality traits for the intention to study. Using the Big-5 indicators, they find openness to new experiences positively related to the decision of studying, while neuroticism was found to be negatively related. Also in this study, the openness trait is particularly decisive for children of families with lower (here defined as non-academic) socio-economic backgrounds.

The economic literature on parental personality traits and their direct relation with the offspring's school achievement is comparatively small which is likely on grounds of identification issues. Much of it revolves around parental risk attitudes that are viewed as a personality trait by some authors, but belong to attitudes under the applied taxonomy here. Literature from other fields has pointed out the relationship of personality traits to parenting styles. The meta-analysis by Prinzie et al. (2009) examines the Big-5 factors' relations to warmth, behavioral control and autonomy support. The authors find higher levels of all traits but neuroticism to be associated with more warmth and behavioral control. Moreover,

higher autonomy support is associated with higher degrees of agreeableness and lower levels of neuroticism. Such parenting styles in turn have an effect on the child's achievement (Kordi and Baharudin, 2010).

Considering parental personality traits in such an analysis cannot uncover whether possible influences are due to intergenerational transmission, a direct influence or a mix of both. The association must therefore be regarded as a total influence. One argument shifting the weight towards a direct influence is that the examined sample in this thesis consists of adolescents, where it has been observed that the intergenerational resemblance is relatively low at this age (Busch, 2013; Anger, 2011). This observation is attributed to personality developing over time and first stabilizing at a later age and because teenage years are particularly distempered.

2.2.2 Attitudes

Attitudes are informative indicators of the parental mindset and the social milieu and likely play a crucial role in identifying the facets of interest. Examples of such attitudes include core life values and norms, interests, opinions and (religious) beliefs. They can refer to manifold topics. However, in some cases an attitude's meaning can surpass being a pure expression of a milieu or a mindset. Evidence has been found that certain attitudes, for instance the parental one towards risk, have an effect on the child's school achievement. As with personality traits, a relation can be of direct or indirect nature. A direct one is marked by parental decisions on school-related factors which are significantly affected by their own attitudes. Whether the child is encouraged to go to a more demanding school type, for instance, may depend on parental risk attitudes. The indirect case would be one where the child adopts the parental attitude and acts accordingly. An example is the degree of interest in politics, which is linked to higher educational aspirations (Lange and Print, 2013, p. 73) and has also been found to be intergenerationally correlated (Shani, 2009, p. 229). This two-way influence is once more only highlighted with the purpose to show the possible effect channels – in the empirical analysis they cannot be distinguished from each other.

The remaining descriptions in this section focus on attitudes for which an influence has been attested in the literature and which can be found in the data set at hand. This is done because the number of attitudes for which a link to the child's school achievement has not yet been established is theoretically endless. And although one can hypothesize on the direction of association of such characteristics, attitudes are treated primarily as reflections of a mindset or a milieu. As with personality models, there are also specific value models, e.g. the Schwartz Values Inventory

(Schwartz, 2009). Data restrictions render a detailed description unnecessary, however. An exception is the value classification by Kluckhohn and Strodtbeck (1961) that considers the importance of certain life areas like family, career or altruism (Headey et al., 2013, p. 732).

Parental attitudes that have been linked to the child's achievement include gender roles, self-esteem and risk preferences. There is ample evidence for the transmission of gender role views (Cunningham, 2001; Fortin, 2005; Farre and Vella, 2013). The experimental study by Spencer et al. (1999) indicates that gender can work as a stereotype threat to math performance.[14] Time, risk and trust preferences are also likely to play a role for school success, because all have been linked to key economic outcomes (Becker et al., 2012).

One particular factor is self-esteem, which is the attitude toward the self. It is closely connected to personality, especially the locus of control (Judge et al., 2002) and the Big-5 dimensions (Amirazodi and Amirazodi, 2011). Following Kaplan et al. (2001), low parental self-esteem negatively relates to educational aspirations and through this channel harms the child's educational outcome. Closely connected to self-esteem and related to a high external locus of control is the attitude of status fatalism. It is the belief in social impermeability for oneself and constitutes an antithesis to aspirations. One can hypothesize that parental status fatalism is not only an expression of a certain milieu, but also has relevant ramifications when it leads to a lack of aspiration and encouragement.

Several scholars have analyzed the direct effect of parental risk aversion on indicators of education. Huebener (2015) examines the relation between parental, in particular the paternal, risk attitudes and a son's long-run educational achievement. Using a quasi-experimental setting he finds that lower levels of paternal risk aversion are associated with higher levels of son's education. Checchi et al. (2014) investigated the dependence of a child's college decision on parental risk attitudes when parents defray the cost of their child's education. The authors' results for Italian data suggest a negative link between risk aversion and the decision to attend a college. Brown et al. (2012), using the 1996 US Panel Study of Income Dynamics, find that parental risk aversion is inversely related to both college attendance, which is linked to achievements in adolescence, and early academic scores. While many an analysis deals with middle or long-term results, there are also studies focusing on early outcomes. Germany as a country with early tracking is a suitable object of study because parents often strongly influence the decision.

[14] Making people (subliminally) conscious about negative stereotypes linked to the social group they are in is called a stereotype threat. It is considered a threat because people are afraid to confirm these stereotypes which results in worse performance.

The results by Wölfel and Heineck (2012) for German data show that daughters of risk-averse mothers, in comparison to risk-neutral ones, have a higher probability of being enrolled in a lower secondary school track. Their results suggest that daughters are more strongly affected by parental risk preferences. In contrast to the results of Huebener (2015), however, they find the father's risk aversion less clearly related.

Another topic is the propensity to behave reciprocal. In a social context, reciprocity refers to the expectation of repaying something someone else has given to one. When this is something positive, failing to live up to this expectation in the eyes of others can lead to social isolation. On the other end, taking actions of reciprocity might strengthen the social network of a person. For these reasons, Putnam (1995) relates reciprocity to social capital, which in this case is to have access to a network of people. This network may be supportive, for instance provide the family with school-related information or even offer time for childcare when needed. Moreover, it may lead to increased social interaction which could improve the child's social skills.

2.2.3 Time use indicators

"Let's Read Them a Story!" is the title of an OECD publication about the parent factor in the educational success of children.[15] With the focus on reading books to children, the title highlights the importance of spending quality time with the child. In the context of this work, reading to children is an example of an observed variable that stems from a specific mindset but might also have some causal effect itself. In this treatise, time use indicators encompass a broad set of behavioral characteristics. These include habits, work and leisure activities and behavior. The underlying theory of this category is similar to the one behind attitudes, but these indicators measure concrete behavioral aspects whereas attitudes may remain hidden if they are not expressed. Time use indicators could also be separated according to whether parents include the child in their activities or not. However, in either case the child is concerned because not including the child has also an implication, which could, for instance, be less parental care.

Taking account of time use indicators in this thesis is done for several reasons. One of them is their function in the latent variable model where they aid in the description of family background and possibly also of milieu characteristics. Activities, in particular social ones, are often influenced by the social surroundings.

[15] OECD (2012b)

For this category of characteristics, the presumption is that the frequency of certain activities differs with the affinity to or the appreciation of education.

Drawing on the relation between time use indicators and the social environment, one can explicitly relate certain time use indicators to social capital: Since ties are more likely to be forged in social situations, joint activities may enhance a parent's social capital. This topic was examined by Büchel and Duncan (1998) who link extrafamilial parental activities, i.e. activities outside the family, to the notion of social capital. Extrafamilial activities allow parents to generate possibly useful ties outside the family and extend their network. For these reasons, the authors examined the hypothesis that parental social activity fosters the child's school success. Using a data set for Germany, they found particularly strong effects through the father's engagement in different activities on sons of low-income families. The direction of this effect depends on the considered activity. While exercise is linked positively, going out with friends has a negative association. A limitation of these insights lies in the way some variables are measured. Only the frequency but not the degree of social company is recorded.

Another function relates to the activities themselves. When parents are viewed as role models, whose behavior is adapted by their descendants, the type of time use can make a difference. This point relates to the educational value of this activity. An illustrating example is time spent on further education which signals the importance of education to the child. On the other hand, there might be activities which yield little or no educational stimulus, but this likely depends on how the activity is presented to the child.
The idea of the parental role model can be directly integrated into the framework of the intergenerational transmission of habits. Evidence for such is plenty. There is, for example, evidence on smoking (Loureiro et al., 2009), watching TV (Bleakley et al., 2013), volunteering (Bekkers, 2007) and high cultural activities (de Vries and de Graaf, 2008). Time use also refers to labor market activities: Morrill and Morrill (2013) have observed a correlation in labor force participation for mother-daughter pairs, the findings of Couch and Dunn (1997) suggest a correlation in annual work hours for father-son and mother-daughter pairs. A child does not participate in the labor market but the observed correlations give an idea of the extent of intergenerational transmissions in the long-run. Moreover, the literature points out that working mothers transmit a set of skills which benefits the children outside the home (McGinn et al., 2015). As mentioned before, the type of time use may be influenced by personality traits and attitudes. One example is the frequency of exercising for which researchers found a link to the internal locus of control (Cobb-Clark et al., 2014).

2.2.4 Demographic indicators

The fourth group of variables contains demographic indicators; they differ from the first three groups by relating to external life circumstances or endowment rather than to internal parental characteristics. Demographic characteristics encompass several aspects such as family type, household size, education or income. Such variables are likely to have a causal effect on the child's scholastic achievement. However, identifying these links by means of non-experimental survey data is usually not possible. The causal relation of education is discussed in the introduction, but there are also theories on the effect of parental income. Underlying a model in which the educational achievement can be impelled by purchasable goods. Examples include moving to a better neighborhood or purchasing private tuition for the child. Other studies emphasize the psychological effects, for instance less stress, especially in low-income backgrounds (e.g. Duncan et al., 1998). The sociological view of this goes under the term economic capital. In a non-experimental setting, however, income stemming from different sources cannot be treated the same and added up since selection issues arise from unobserved confounding factors. Parents who receive income from public transfers, for example, are likely to be different from those who gain their income through work.[16] Depending on the source of public transfers, a high receipt typically either indicates transfers due to parental unemployment or due the number of kids in the household. This can lead to the observation that in spite of additional disposable income through public transfers, income is negatively related to the child's school achievement. In case of parental unemployment, a negative association can be attributed to the effects of unemployment (comp. e.g. Gregg et al., 2012) or latent characteristics which influence both the probability of being employed and the child's school success. The latter reason would at hint a pattern which is located in a certain milieu. High public transfers owing to many kids in the household are a different matter, as negative effects could arise through less parental care time and income that is left per child.

Similarly, private transfers often consist of alimony payments, their receipt hence indicates a household with separated parents. Based on these arguments, it is necessary to differentiate types of income according to their source.

The demographic characteristics considered here are likely to reflect underlying latent factors as well, which is one reason why they are included in this analysis. The second reason is that the added value of explaining the child's school achievement

[16] Next to different propensities there can also an economic reason to treat separate income sources differently. Zhan (2006) argues that properties of income, such as the flow or stock character can play a role which is why one should differentiate between labor income and asset income.

with indicators beyond demographic ones can only be evaluated if the latter are regarded.

The term endowment can be broadened in definition as to cover family circumstances like parental separation or divorce. An obvious reason to include such variables is the varying propensity for such events across milieus. The likelihood of a divorce can be assumed to be much lower in a religious or conservative milieu than in other, more liberal milieus. In the style of this argumentation, the number of offspring and the area of habitation should also be regarded. Other important aspects are parental age, a migration background and living in former East Germany. While the latter two aspects could hint at cultural differences, parental age is owed to the passing of time and societal changes which influence attitudes but also the time spent on certain activities.

Chapter 3

Methodological approach

The previous chapter dealt with linking observed parental characteristics to a child's school achievement via latent variables. Since identifying latent variables depends on the input of observed variables, a selection rationale for the latter was proposed and four non-independent categories established. Although each category contained some concrete examples, their definitions are comprehensive enough to allow for a large number of suitable choices. An example for such a category is the time use and activity category. Depending on the data set at hand, there could be many parental activities whose frequencies are recorded and possibly contribute to drawing a more detailed picture of family background.
If no further processing takes place, this can result in a large (in the sense of wide) data set and may cause difficulties in interpretation due to its sheer complexity. Limiting the choice of characteristics to those that approximate the aspects of interest best is, hence, expedient. However, in a latent-variable model setting it is not obvious which ones to choose. This is primarily the case because the characteristics are conceived as proxy variables with no exact causal interpretation while the aspects of interest are presumed to be latent. Limited data might make the decision in cases where one has a prior idea about the latent variables, but doing so would contradict the rationale of the proposed approach.
Discussing the treatment of such a data set, it should be regarded that some variables might correlate strongly with each other. Such overlaps in variation are not restricted to a specific category of variables – they can also occur across them if, for instance, certain personality traits correlate with certain types of activities.

The data set used in this dissertation indeed provides many potential candidate variables. In the empirical part of this dissertation, the main task is, thus, to find a method which filters out the important bits of information, which are the variables that significantly contribute to explaining variation in the child's school achievement.

One could stop here and find a method producing a model which explains the child's school achievement best according to some criterion. Insights and subsequent use of this model might be limited, however. To ensure a further use, methods dealing with variable selection ideally have two favorable traits in addition: First, they produce a result which can be (economically) sensibly interpreted to be able to draw conclusions. Here, it is particularly expedient to allow inference on the emphasized latent structures. Parsimony by variable selection is a less preferred criterion, when this is not possible. This simplifies interpretation by reducing the dimensionality of the model parameters. On the downside, pure variable selection shows a deficit by not indicating latent structures in the data. Hence, it is less preferable than a model identifying latent structures in this context. Producing a sufficiently general model is the second property a method should exhibit. General is to be understood here as not sample specific, so a resulting model holds approximately also in others than the drawn sample. This property can be examined by evaluating the predictive capability of the model on new data. Hardly general findings are portended if the predictive power breaks down in comparison to the model fit. In the context of method evaluation later, the first property is called (weak) interpretability and the second one generalization ability or stability. The remainder of this chapter approaches and deepens the topic of variable selection, in which a larger set of correlated predictor variables with relatively few observations is assumed to be at hand. Moreover, the dependent variable is assumed to be continuous.

In the search for a suitable method, it is insightful to begin with the analysis of a frequently used one. This is a linear regression model containing all predictors whose parameter estimates are obtained by Ordinary Least Squares (OLS). As the drawbacks of the approach in such a data environment are pointed out, the concepts of multicollinearity and the bias-variance trade-off are expounded. Those two play an important role for attaining the desired properties. With this in mind, using more involved methods addressing these shortcomings is motivated subsequently. Regarding terminology, the names predictor(s), input variable(s), feature(s) regressor(s), explanatory and right-hand-side variable(s) are used interchangeably. The dependent variable is sometimes also called left-hand-side variable or outcome.

When developing a model, a useful and often suitable simplification is to assume a linear relationship between the dependent variable and the predictors. Defining a different functional form requires a theoretical justification or will be tedious if the data set contains numerous right-hand-side variables. Moreover, linearity only refers to linearity in parameters, a linear model's flexibility can always be

enhanced by considering transformed features, e.g. through squared terms. In the full linear regression model all available predictors plus selected transformations serve as regressors. Under the mean squared error loss function, OLS minimizes the loss and yields the parameters of the best linear predictor (Hayashi, 2000, p. 139).[17] The OLS estimate is unbiased with respect to estimation bias.[18]
When no variable selection occurs, the resulting model will be difficult to interpret for its complexity alone. If there are more than a hundred variables, it becomes challenging to analyze each coefficient. Assuming the usual conditions being fulfilled, two additional drawbacks may occur in a regime with many, potentially good regressors. The first refers to inflated standard errors and emerges when some regressors are highly correlated, i.e. there are large overlaps in explaining variation. The other is called overfitting and becomes apparent when a prediction based on the estimated parameters is made, i.e. the generalization ability of the model is checked. Overfitted models lead with high likelihood to a much worse model fit when predictions on new data are made.

The first phenomenon is called partial multicollinearity or, simply, multicollinearity. As opposed to full multicollinearity which violates the full-rank condition models with partial multicollinearity among the predictors can be estimated by OLS. The phenomenon can yet cause misleading interpretations.
To reconstruct the effects on the parameter estimates when two regressors are strongly correlated, it is worthwhile to recapitulate how their parameters are calculated. A multiple regression of y on three explanatory variables x_1, x_2 and x_3 shall serve as an example. The model with the residual term denoted as u can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u. \tag{3.1}$$

Let the parameter estimate of interest be β_1. One way to obtain the multiple regression coefficient of variable x_1 is to first regress x_1 on a constant, x_2 and x_3 and obtain the corresponding residual vector $\widetilde{u}$ from

$$x_1 = \widetilde{\beta}_0 + \widetilde{\beta}_2 x_2 + \widetilde{\beta}_3 x_3 + \widetilde{u} \tag{3.2}$$

.

Then conduct a regression of y on $\widetilde{u}$. The estimated coefficient of $\widetilde{u}$ will equal β_1

[17] Taking into account other loss functions such as absolute loss or asymmetric loss in the context of school achievement may be interesting for special purposes but lacks a justification here.

[18] This is a statistically controllable bias, which is, however, not necessarily free from model bias which originates from a wrong specification. Model bias is different, as it rests on exogeneity of the predictors. Hence, when it is referred to the unbiasedness of Ordinary Least Squares, estimation bias is meant.

from 3.1

$$y = \overline{\beta}_0 + \beta_1 \widetilde{u} + \overline{u}. \tag{3.3}$$

This is called 'partialing out' or 'netting out' the effect of other variables. In the case of x_1 being highly correlated with the other two variables, $\widetilde{u}$ will be close to zero and its coefficient will be unstable, i.e. sensitive to minor changes in the data set accompanied by a large standard error (Hastie et al., 2009, p. 55; Dormann et al., 2013, p. 28).
This can also be seen by considering the variance formula

$$Var[\beta_1|X] = \frac{\sigma^2}{(1 - R_1^2) \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)^2}, \tag{3.4}$$

where R_1^2 is the multiple coefficient of determination from 3.2. Hence the higher the correlation of x_1 with the other variables, the higher the R_1^2, and the larger becomes the variance. In the limiting case, where x_1 is a linear combination of the other variables, the variance becomes infinitely large (Greene, 2012, p. 129ff.).

The presence of multicollinearity drives up the concerned variables' standard errors, so they become less statistically significant. Therefore, too often the null hypothesis of a coefficient equaling zero will fail to be rejected. Methods trying to yield parsimonious models based on the predictors' significance can fail because inflated standard errors conceal true relations.
On the other hand, multicollinearity does not lead to systematically biased coefficients. Large variances, however, will cause the coefficients to be unstable especially in small samples (Mittelhammer, 1996, p 457f.). This has also an implication for prediction. While there is no problem in "predicting" the data in the estimation sample, true predictions, i.e. for yet unknown y, can be poor (Dormann et al., 2013, p. 29). In real-world data, explanatory variables are seldom if ever orthogonal to each other, there is hence always some slight effect induced by multicollinearity, so that the passage is a fluent one to when the consequences become more severe. Different ways to judge the degree of multicollinearity are described in Dormann et al. (2013, p. 30ff.).

In a data-rich environment, meaning that a data set is wide, often a second issue called overfitting occurs. As with multicollinearity it is not a dichotomous property but a gradual phenomenon. The term describes cases in which a model is too complex, which means it has few degrees of freedom. This is a result from having (too) many explanatory variables relative to the number of observations. For a given sample, this is not an immediately visible problem: Overfitting makes a model more precise for known data, since adding an additional predictor to the

model never decreases the explained variation. However, it can adversely affect the model's predictive ability since higher precision comes at the price of higher variance. Variance indicates how strongly the model parameters vary in different samples. In other words, an overly precise model is typically custom-tailored for the specific sample at hand but this limits the model in its generalization ability. This property originates from sample-specific noise in the data. With increasing complexity of the model, noise instead of information is explained. This is of limited or no use in other samples.

The formal concept to grasp this relation is the so-called bias-variance trade-off. The following description is based on Lebanon (2010). The bias-variance trade-off stems from an analytical decomposition of the mean squared error (MSE). The MSE is defined as the expected squared difference between the true parameter vector $\beta \in \mathbb{R}^q$, and its estimate $\widehat{\beta}(x_1, ..., x_q)$ which depends on the estimator as well as the sample observations at hand and is thus a random variable.
In order to increase readability, the estimator's dependence on the sample is omitted in notation, such that $\widehat{\beta} \in \mathbb{R}^q$. In the following, it can be shown that the mean squared error is composed of the sum of the squared bias, i.e. the squared deviation of the estimated coefficients from their true values, and the variance. As the estimate is a random variable, the expected value needs to be calculated: Writing the MSE in expectation notation and extending by $E\left[\hat{\beta}\right]$ yields:

$$MSE\left(\hat{\beta}\right) = E\left[\sum_{j=1}^{q}\left(\hat{\beta}_j - \beta_j\right)^2\right] = E\left[\left(\hat{\beta} - \beta\right)^2\right] = E\left[\left(\hat{\beta} - E\left[\hat{\beta}\right] + E\left[\hat{\beta}\right] - \beta\right)^2\right] \tag{3.5}$$

Expanding gives:

$$E\left[\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)^2 + 2\left(\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)\left(E\left[\hat{\beta}\right] - \beta\right)\right) + \left(E\left[\hat{\beta}\right] - \beta\right)^2\right] \tag{3.6}$$

Using the rules of the expectation operator on sums and products obtains:

$$E\left[\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)^2\right] + 2\left(\left(E\left[\hat{\beta}\right] - E\left[\hat{\beta}\right]\right)\left(E\left[\hat{\beta}\right] - \beta\right)\right) + E\left[\left(E\left[\hat{\beta}\right] - \beta\right)^2\right] \tag{3.7}$$

As the middle term cancels out, the following is left:

$$E\left[\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)^2\right] + E\left[\left(E\left[\hat{\beta}\right] - \beta\right)^2\right] = Var\left(\hat{\beta}\right) + Bias\left(\hat{\beta}, \beta\right)^2. \tag{3.8}$$

Although this decomposition holds, it does not imply the existence of parameter values which exploit this trade-off. It depends on the relative gains and losses in

MSE induced by modifying the parameter values. Moreover, it is also possible to increase both bias and variance, for instance by enlarging unbiased parameter estimates. In such a case, there is no trade-off.

While overfitting refers to strong reactions in the predicted outcome through small changes in the input data, one can also establish a model too general. The extreme case for a model with low variance is a constant value function which is supremely robust to unsystematic changes in the data set. For different (large enough) random samples, the mean value hardly varies. But it is also biased as the predicted value is the same for every data point. Such a model is rather uninformative. For the researcher, the drawbacks of low precision and few insights arise. Analogously, one could speak of underfitting in this case.

The points made so far can be illustrated for the two-dimensional case in which a dependent variable, whose values are denoted on the ordinate, is explained by a variable, whose values are denoted on the abscissa. The following two illustrations show two small samples drawn from the same distribution. The points denote the observed sample values and the curve is the predicted dependent variable from a model containing polynomials up to 6th-order in the explanatory variables. Hence the model consists of seven explanatory variables in total. It is observable that the

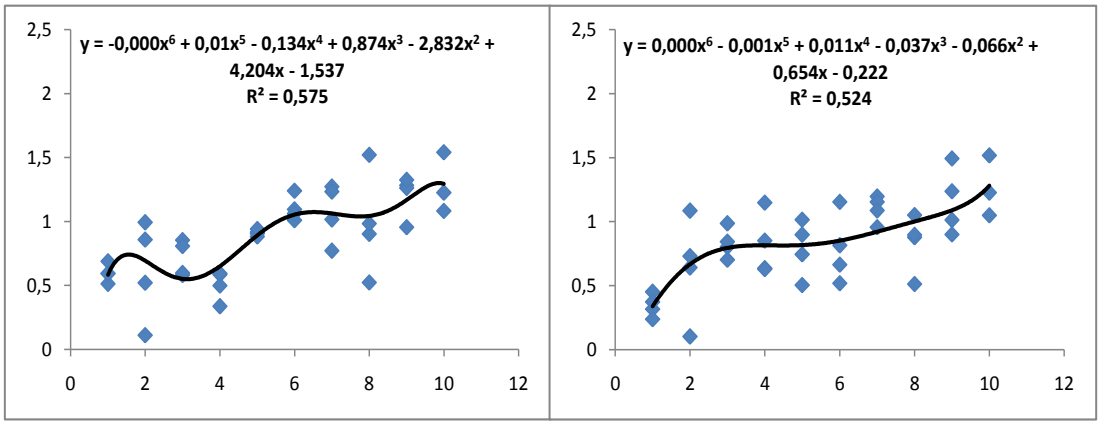


Figure 3.1: 6-th order polynomial approximation of the data.

coefficients and hence the shape of the predicted polynomial are quite different between the samples.
The figures shown next present the results of a linear approximation, a polynomial of order 1 which is an intercept and a slope parameter, for the same data points as in the previous cases.

As expected, the more complex model is much more precise for the given data, indicated by the larger R^2. The linear model, however, is relatively more robust across different samples. The slope coefficient and the intercept change only slightly,

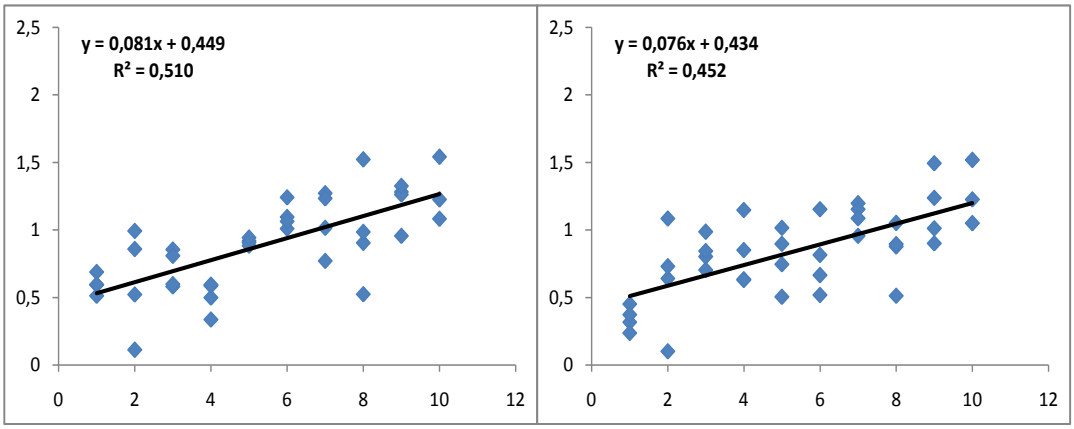


Figure 3.2: Linear approximation of the data.

whereas the coefficients vary wildly in the more complex model.
The robustness property itself is of value, but what is its relation to prediction? In this context, prediction is forecasting the value on the ordinate given a value on the abscissa. For example, a prediction of the values in the right-hand-side figures based on the estimated models of the left-hand side figures means the following: The respective line or curve of the left-hand-side figures has to be copied into the right-hand-side diagrams and its fit to those data points needs to be evaluated. The linear model would perform only slightly worse in the right-hand-side sample data, while the more complex model would perform considerably worse. This difference in model fit between estimation and prediction sample for the high-polynomial model displays overfitting.

The previous observations indicate that a trade-off between precision and generality can occur. Its extent depends on the complexity of the true model, the model specification and the noise in the data. In general, however, a relation as depicted in figure 3.3 holds. It exemplifies the magnitude of the mean squared forecast error in dependence of the model complexity. The dotted line denotes the squared bias which decreases with increasing model complexity. The variance, the dashed line, increases on the other hand. The total error is minimized at about the middle.
Finally, the permitted degree of complexity also depends on the number of observations at hand. More data points for a fixed degree of complexity yield more degrees of freedom and the model comes closer to attaining asymptotic properties.

The trade-off between precision and variance as well as the value of sparse models for interpretability and generalization ability motivate the need for methods which manage to find such models.

In this thesis two strands of methods are deemed suitable for this goal and are examined more closely for this reason. The first one achieves weak interpretability

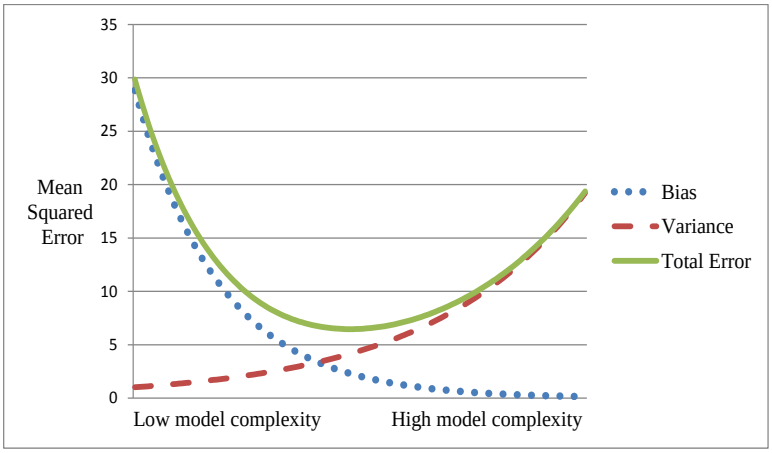


Figure 3.3: Visualization of Bias-Variance Trade-Off

and possibly also generalization ability by condensing common variance of the predictors into a smaller number of new predictors, so-called indices. Certain methods in this category even achieve interpretability, as the created indices can often be meaningfully interpreted.

The other group is motivated by the bias-variance trade-off observation. This strand embarks at the full model and deliberately induces bias by shrinking all coefficients towards zero. Thereby variance is reduced, yielding a model that has improved predictive capabilities. Such methods are called shrinkage or regularization methods. They are expected to yield a sparse model with good generalization ability while interpretability usually only emerges from variable selection and is hence weak. An exception is a combination of shrinkage methods with index-building methods. Two variants are proposed in this thesis, which differ in the type of shrinkage method applied but not in the method to create indices.

Chapter 4

Overview of empirical methods

This chapter presents a selection of methods that can be applied to address the issues related to having too many predictors. It is divided into four sections. The first deals with procedures which are fundamental ingredients to more involved methods. This section comprises the description of two procedures which condense information in the explanatory variables, called Principal Component Analysis and Factor Analysis. Moreover, the technique of rotation is presented which can aid in the interpretation of the output of information-condensing procedures. The final method described is Cross-Validation which is a particularly useful technique for model selection. The latter can cause problems in small samples since splitting the data into a training, a validation and a test sample results in too small sample sizes. Cross-Validation supports the selection of a model under these conditions by repeatedly splitting the data.

The next section is devoted to subset selection methods. These methods choose a subset of the original variables by repeated testing of the same sample. Subset selection methods are introduced prior to the more involved methods because they are often used for variable selection. Although being intuitive and easy to apply, they generally have some undesirable properties.

The idea of a factor or component model which exploits overlapping explanatory power within the set of predictor variables to reduce the dimensionality is pursued in the following section. The methods in this section produce indices, which are linear combinations of the original explanatory variables. Following the terminology of Tu and Lee (2012), such procedures are classified into being supervised or unsupervised. This differentiation refers to the manner by which the indices are created. Tu and Lee differentiate between methods which create indices by explicitly regarding their relations to the dependent variable, in which case they call them supervised, or if the creation is independent of the dependent variable, unsupervised.

Subsequently, a family of methods is presented which exploit the trade-off between bias and variance described in chapter 3 in order to improve the quality of a forecast.

The fundamental idea is to shrink the coefficients of the full linear model towards zero to reduce the overall variance of the model. In contrast to the dichotomous decision of purely selective methods on whether to include a variable in the model or not, the methods in this family gradually shrink their influence. The presented methods cover several variations as well as a procedure that combines linear index models with the idea of regularization.

Each approach is discussed with regard to its accomplishment of yielding a sparse and interpretable model. Moreover, a method has usually additional setscrews which can be subsumed under model selection criteria. They can be of high importance, in particular in small samples. These criteria are discussed for each method separately. Based on the results of the theoretical arguments, promising or frequently used methods and their model selection criteria (algorithms) are selected. The corresponding evaluation of their performance takes place by means of a simulation, which is presented in chapter 5.

4.1 Basic procedures

This section portrays techniques which often serve as fundamental parts in more involved methods. One is Principal Component Analysis and its underlying spectral decomposition as a key technique of dimensional reduction and Factor Analysis as a related procedure. After that, rotation techniques are introduced and Cross-Validation as a method used for model selection is presented.

4.1.1 Principal Component Analysis

A basic method for many dimensional reduction techniques is Principal Component Analysis (PCA) which dates back to Pearson (1901) and Hotelling (1933). Because of some similarities, PCA is sometimes confused with Factor Analysis; there are important differences, however. To avoid misunderstandings, the two methods are strictly separated in this thesis. This also concerns the naming of the constructs that are created in these methods: In PCA and methods that rely on its technique, they are called components, whereas in Factor Analysis they are called factors. This naming is done despite the use of the same notation. Another difference lies in the name of the vectors and matrices that connect the original variables with the constructs. In PCA and related techniques they are called weights, in Factor

Analysis loadings. The following description of PCA follows Timm (2002), Jolliffe (2002) and Jackson (1991) if not stated otherwise.

PCA is a descriptive technique with no causality structure assumed in the data. The method basically transforms a set of q variables, which are here assumed standardized and not pairwise orthogonal, into a set of q new variables, called components. These components are linear combinations of the original variables. In contrast to the original variables, however, the new components are not only uncorrelated with each other but also contain differing amounts of variance. While each original variable has a variance of one owing to standardization, some components will have higher variance, others lower. This is exploited to achieve the standard objective of PCA of finding fewer components than original variables ($k << q$) that comprise as much of the original variance as possible. Dimensional reduction is induced by selecting the components that contain the highest amount of variance and discarding the components with little variance. Since the components are uncorrelated, PCA can also be used to overcome problems in the original variables owing to multicollinearity.

Let $x_1, \ldots, x_q \in \mathbb{R}^N$ be q standardized predictor variables. A matrix of weights W is searched for in order to construct k components $f_1, \ldots, f_k \in \mathbb{R}^N$ which are linear combinations of the predictors. Component-wise, this can be formulated as follows

$$\begin{aligned} f_1 &= w_{11}x_1 + w_{21}x_2 + \cdots + w_{q1}x_q \\ &\vdots \\ f_k &= w_{1k}x_1 + w_{2k}x_2 + \cdots + w_{qk}x_q. \end{aligned}$$

Or more compactly for a specific component j

$$f_j = Xw_{\cdot j} \tag{4.1}$$

where $w_{\cdot j}$ $(q \times 1)$ denotes the j-th column of the weight matrix W and represents the weights vector belonging to j-th component. For all components

$$F = XW. \tag{4.2}$$

Let $\mathrm{Cov}[X]$ denote the empirical covariance matrix of $X = [x_1, x_2, ..., x_q]$. Due to prior standardization of the predictors, it is a correlation matrix in this case, hence $\mathrm{Cov}[X] = Corr[X] = Var[X] = E\left[X^T X\right]$; it plays a key role in PCA. When the original variables are unstandardized, PCA can also be conducted on the covariance matrix. One has to bear in mind, however, that PCA is only invariant under orthogonal transformations. Therefore, the result is sensitive to

the units of measurement and not necessarily equivalent using standardized or non-standardized variables.
The basic procedure of PCA can be formulated iteratively, that is to find the principal components one after another. To begin with, the goal is to find the first principal component, f_1, of X being the linear combination of the predictors that has maximal variance. Since X is given by the data, maximizing the variance of f_1 implies finding the vector $w_{\cdot 1}$ that maximizes

$$Var\left[f_1\right] = E\left[f_1^T f_1\right] = E\left[w_{\cdot 1}^T X^T X w_{\cdot 1}\right] = w_{\cdot 1}^T \mathrm{Cov}\left[X\right] w_{\cdot 1} \tag{4.3}$$

subject to the normalization restriction that $w_{\cdot 1}^T w_{\cdot 1} = 1$. This restriction is necessary for the principal component to be unique up to the sign. The rearrangement in 4.3 is exploiting that $Var\left[f_1\right] = E\left[f_1^T f_1\right] - \left(E\left[f_1\right]\right)^2 = E\left[f_1^T f_1\right] - 0$. This relation holds because the predictors were assumed to be centered. The problem can be solved using the Lagrangian Multiplier method, such that the maximization problem for the first component can be formulated as

$$\max \; w_{\cdot 1}^T \mathrm{Cov}\left[X\right] w_{\cdot 1} + \lambda \cdot (1 - w_{\cdot 1}^T w_{\cdot 1}). \tag{4.4}$$

Partial derivation with respect to $w_{\cdot 1}$ and setting to zero yields

$$\left(\mathrm{Cov}\left[X\right] - \lambda I_q\right) \cdot w_{\cdot 1} = 0. \tag{4.5}$$

To solve the system of homogeneous equations, the q eigenvalues of $\mathrm{Cov}\left[X\right]$ need to be calculated. Knowing the eigenvalues, the orthonormal[19] eigenvector υ_1 which corresponds to the largest eigenvalue of $\mathrm{Cov}\left[X\right]$, λ_1, can be calculated. This eigenvector is the weight vector $w_{\cdot 1}$ that maximizes 4.3, i.e.

$$f_1 = X w_{\cdot 1} \tag{4.6}$$

where $w_{\cdot 1} = \upsilon_1$. The value of the largest eigenvalue contemporaneously represents the variance the first component comprises. Unless the original variables are pairwise orthogonal, the eigenvalue is larger than 1 so that the first component binds more variation in X than a single original variable. The individual observations on the components are called scores. Like tall people are more inclined to the variable measuring height than small people, scores measure how inclined someone is to a certain principal component. Scores play an important role when the components are further utilized.

[19] Orthonormal: Unit length and uncorrelated to the other eigenvectors.

The second component can be obtained in the same fashion; however, the additional restriction of being orthogonal to the first component is imposed, that is $f_1^T f_2 = 0$. This means the second component will account for as much variance as possible from the variance not already covered by the first component. Alternatively formulated, one could subtract the explained variance of component 1, so that $\text{Cov}[X] - v_1 v_1^T = \text{Cov}[X]^1$, and apply the same maximization procedure on $\text{Cov}[X]^1$ as for component 1 in 4.3, i.e. without the orthogonality restriction. From here on the procedure continues in one or the other fashion until the last component f_q is reached.
However, this iterative calculation is not necessary. A useful generalization can be carried out by a spectral decomposition (SD) of the symmetric matrix $\text{Cov}[X]$. Obtaining the eigenvectors by a spectral decomposition factors $\text{Cov}[X]$ in the following way

$$\text{Cov}[X] = V \Lambda V^T. \tag{4.7}$$

Λ is the diagonal matrix of eigenvalues and V is an orthogonal matrix of dimension $q \times q$ containing the standardized eigenvectors. Then, arranging the eigenvalues by size $[\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_q]$ and their associated eigenvectors $v_1, \ldots, v_q$ in order, the complete weight matrix W is given by $V = [v_1, \ldots, v_q]^T$.

Having obtained the eigenvectors and keeping all components ($k = q$), the $N \times k$ matrix of component scores is calculated using the relation in 4.2.

PCA only leads to a new coordinate system through a principal axis rotation. One implication is that the sum of the original variable's variances equals the sum of the eigenvalues. In the case of standardized variables and extracting all components, this implies $tr[\text{Cov}[X]] = tr[\text{Cov}[F]] = \sum_{i=1}^{q} \lambda_i = q$.

It has to be noted that PCA is sensitive to outliers which may cause the existence of single components. Timm (2002) recommends detecting outliers a priori or using robust estimates of the correlation matrix.

As mentioned, dimensional reduction is often the primary reason why PCA is applied. This can be achieved by only using k principal components with the highest variance and discarding the remaining. Then some original variance will be left unexplained, leading to a loss of information. But simultaneously a dimensional reduction of the data takes place. The criterion of which or how many components to keep has been discussed extensively in the literature without a final conclusion, one reason being that the optimal choice depends on the researcher's aim. At this point the discussion is deferred to section 4.3.1 on Principal Component Regression because it is the regression procedure for which PCA plays the key role and there are additional aspects, such as rotation, that need to be considered.

4.1.2 Factor Analysis

A related method to PCA is Factor Analysis (FA), sometimes also called Classic Factor Analysis or Exploratory Factor Analysis. It is the foundation for the methods which correspond best with the theoretical ideas given in the previous chapters. This section follows the elaborations of Überla (1968), Timm (2002, p. 496–510) and Jackson (1991, p. 388–423).
The development of Factor Analysis goes back to Spearman (1904) and Thurstone (1931). Typically, its objective is described as "explaining" the correlation structure amongst the set of (noisy) observed variables by means of a few common factors. These common factors are interpreted as latent variables and thought of as influencing the observed variables so that they become correlated. Because the common factors are unknown a-priori, they have to be derived from the observed variables. The observed variables, however, are not completely dependent on the common factors: A differing share of variation in the observed variables is left unexplained. This remaining share is often interpreted as measurement error or noise and is depicted by an error term.
One useful feature in Factor Analysis is that the common factors are often meaningfully interpreted. The model formulation makes it appealing for the analysis of the economic problem whose structure was formulated similarly. In contrast to PCA, which has the goal to compress the variance in the predictors, the resulting components' meaning might be recondite. Why the procedures differ in interpretation is grounded in the formulation of the model and accordingly how the indices are obtained: In PCA, the components are linear combinations of the original variables, whereas in Factor Analysis the original variables are linear combinations of the latent factors. Nevertheless, the procedures can often yield similar results but this depends, as will be demonstrated, on the data structure and on the manner of conducting Factor Analysis. In contrast to PCA, there is no single way to do Factor Analysis so some confusion might arise. Thus, this section attempts to carefully point out the differences between the main procedures of Factor Analysis.

Assuming all variables are standardized, the basic model of Factor Analysis can be formulated as the observed variables x being a weighted linear combination of the latent factors f and an error term ϵ:

$$\begin{aligned} x_1 &= l_{11}f_1 + l_{12}f_2 + \cdots + l_{1k}f_k + \epsilon_1 \\ &\vdots \\ x_q &= l_{q1}f_1 + l_{q2}f_2 + \cdots + l_{qk}f_k + \epsilon_q. \end{aligned}$$

The loadings link the factors and the original variables. The above system can be compactly written as

$$X = FL^T + E. \tag{4.8}$$

In order to obtain estimates for the unknowns in the model, some assumptions have to be made. The error term and the factors have zero mean on expectation, so $E[\epsilon_i] = 0$ and $E[f_j] = 0$. Moreover, the latent factors are uncorrelated with each other (strictly speaking, this assumption is not necessary) and have variance of one, so that $\text{Cov}[F] = I_k$. This condition leads to non-unique solutions, a property particular to Factor Analysis which makes it different from PCA.
The variable-specific errors are assumed to be uncorrelated with each other and are allowed to have different variances, $\text{Cov}[E] = \Psi = \text{diag}[\psi_1, \psi_2, ...\psi_q]$. The variance ψ_i is referred to as the uniqueness of variable i and stands for the variance in variable i that is not explained by the factors. The errors are also assumed to be uncorrelated with the common factors, $\text{Cov}[F, E] = 0_{j,i}$.
Based on these assumptions, the following important relations, which are part of the fundamental theorem of Factor Analysis, are shown:

$$\begin{aligned}\text{Cov}[X] &= E\left[X^T X\right] = E\left[(FL^T + E)^T(FL^T + E)\right] \\ &= E\left[LF^T FL^T\right] + E\left[LF^T E\right] + E\left[E^T FL^T\right] + E\left[E^T E\right] \\ &= E\left[LL^T\right] + E\left[E^T E\right] = LL^T + \Psi \end{aligned} \tag{4.9}$$

$$\text{so: } \text{Var}[x_i] = l_{i1}^2 + l_{i2}^2 + ... + l_{ik}^2 + \psi_i \equiv h_i^2 + \psi_i \tag{4.10}$$

This shows that the variance of an observed variable i is split into the sum of its squared loadings on the factors, called communality h_i^2, plus its uniqueness, ψ_i. Communality corresponds to the explained variance, while uniqueness describes the unexplained part. When the variables are standardized, the communality and specific variance must hence sum up to 1.
The identification of the factor loadings does not lead to a unique solution, since the correlation matrix $\text{Cov}[X]$ can be reproduced equally well under different loading structures. Let T be an orthogonal matrix of dimension $k \times k$, by plugging $T^T T$ into equation 4.8, a transformed factor model arises:

$$X = FT^T TL^T + E \tag{4.11}$$

Redefining $F^* = FT^T$ and $L^{T*} = TL^T$ and replacing L in 4.9 by L^*, gives:

$$\begin{aligned}\text{Cov}[X] &= L^* L^{T*} + \Psi = (LT^T)(TL^T) + \Psi \\ &= LT^T TL^T + \Psi = LL^T + \Psi \end{aligned} \tag{4.12}$$

This shows that any orthogonal transformation is also a permissible solution. Since this is an orthogonal axis rotation, the loadings will be different, but the factors remain orthogonal to each other and the communalities stay constant, too. The indefiniteness seems to be a drawback at first, but can be exploited to make the common factors better interpretable.

If the assumption of uncorrelated factors is given up, such that $Cov[F] = \Phi$, where Φ is any valid covariance matrix, it holds that

$$\text{Cov}[X] = E\left[L\Phi L^T\right] + E\left[E^T E\right] = L\Phi L^T + \Psi. \tag{4.13}$$

The factors are now said to be oblique.

Finally, for reasons of interpretation, the relation between the variables and the factors is examined which is the second part of the fundamental theorem of Factor Analysis

$$\begin{aligned} \text{Cov}[X, F] &= E\left[X^T F\right] = E\left[(FL^T + E)^T F\right] && (4.14)\\ &= E\left[LF^T F\right] + E\left[E^T F\right] = L && (4.15) \end{aligned}$$

for which some of the above assumptions were exploited. The equation shows that a loading l_{ij} equals the correlation coefficient between factor j and variable i.

Having discussed the basic properties of a factor model, it remains to elucidate how the parameters are obtained. At first, one attempts to estimate the loading matrix L in 4.8. For L cannot be obtained by the basic equation, the relation in equation 4.9 is exploited for this purpose. This process is the so-called extraction of factors to which end different methods exist. Table 4.1 names several possible procedures to obtain L. Some of them will be explained in more detail. In order to emphasize that the loading matrix is non-unique, even irrespective of any rotation, it is tagged with a hat.

Table 4.1: Procedures to extract factors

Extraction of $\widehat{L}$
Principal Factors*
Principal Component Factors
Iterated Principal Factors*
Maximum Likelihood
Image Analysis
Rao's Canonical Factoring

The typical algorithm used for extracting the loading matrix is called Principal Factors (PF). Another frequently used algorithm is Iterated Principal Factors (IPF). Apart from those two, two others will be introduced briefly but not further pursued for reasons explained later on; Principal Component Factors, which constitutes a bridge between PCA and Factor Analysis and the maximum likelihood approach. The following description of the factor extraction methods is based on the work by Rencher (Rencher, 2003, p. 415-430).

A property that a certain class of factor extraction algorithms shares is to use 4.9 to obtain an estimate for the loadings. Since $\mathrm{Cov}\,[X]$ is given by the data, only L and Ψ need to be specified. These factor extraction procedures make an assumption about the uniqueness matrix Ψ and then L is sought such that the relation in 4.9 is approximated as closely as possible. Let $\widehat{\Psi}$ be the assumed matrix of uniqueness, then

$$\mathrm{Cov}\,[X] = LL^T + \widehat{\Psi} = \widehat{L}\widehat{L}^T \tag{4.16}$$

where the last term has absorbed $\widehat{\Psi}$. To obtain $\widehat{L}$, $\mathrm{Cov}\,[X]$ can be factored which can be done by means of a spectral decomposition, as described in 4.1.1:

$$\mathrm{Cov}\,[X] = V\Lambda V^T \tag{4.17}$$

To end up in the desired form, it is exploited that Λ can be written as $\Lambda = \Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}$,[20] such that

$$\mathrm{Cov}\,[X] = V\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}V^T. \tag{4.18}$$

But instead of defining $V\Lambda^{\frac{1}{2}} = \widehat{L}$, which would be a $q \times q$ matrix, a dimensional reduction should take place. In the established manner only k eigenvectors, for example those with the largest corresponding eigenvalues, are selected, such that $V_1 = [v_1, v_2, ...v_k]$. Then the loading matrix is calculated as

$$\widehat{L} = V_1\Lambda^{\frac{1}{2}} = (\sqrt{\lambda_1}v_1, \sqrt{\lambda_2}v_2, ..., \sqrt{\lambda_k}v_k). \tag{4.19}$$

Independent of the selected number of factors, it becomes clear that the result depends on the initial uniqueness estimate, and it is indeed what often leads to differences across the extraction methods.

Principal factors uses the multiple coefficient of determination R^2 of variable x_i on the other $q-1$ variables to obtain a uniqueness estimate of variable i

$$\widehat{\psi_i} = 1 - R_i^2. \tag{4.20}$$

[20] A matrix to the power of $\frac{1}{2}$ denotes a square root of a matrix. A matrix Z is defined to be a square root of a matrix Y if $ZZ = Y$ holds (Higham, 1986). Since Λ is a diagonal matrix with positive or zero values, the relation holds.

It has been shown that R_i^2 constitutes a lower bound for a variable's communality and that with an increasing number of variables and a constant number of factors, the lower bound holds with equality (Überla, 1968).
In the next step, the reduced correlation matrix $\mathrm{Cov}\,[X]^R$ is calculated as the difference of the empirical correlation matrix and the estimated uniqueness, $\mathrm{Cov}\,[X]^R = \mathrm{Cov}\,[X] - \hat{\Psi}$. Since $\widehat{\Psi}$ is diagonal, the off-diagonal elements stay the same, while the diagonal elements are reduced by the ψ-estimates. Thereby, the total variability is reduced to the variability that the original variables have in common. The variance which is unique to each variable is discarded.
In the next step $\mathrm{Cov}\,[X]^R$ is factorized by means of a spectral decomposition. $\mathrm{Cov}\,[X]^R$ is almost always no longer positive semi-definite, so negative eigenvalues will emerge. By first deleting the eigenvectors with the negative eigenvalues, and then applying 4.19, one obtains the final loading matrix. The goal of a dimensional reduction has already been achieved by discarding the eigenvectors with negative eigenvalues. In practice, however, the number of factors might still be quite high, motivating a further selection thereafter.

Iterated Principal Factors is related to Principal Factors but seeks to improve the communality estimate. The method continues where Principal Factors stops. Having obtained the loading matrix one can calculate the communalities anew. They are different because negative eigenvectors have been discarded. Iterated Principal Factors uses these quantities to improve the prior estimates. It replaces the diagonal elements of the reduced correlation matrix $\mathrm{Cov}\,[X]^R$ by the newly calculated communalities.
This updated reduced correlation matrix undergoes a factorization in the established manner whereupon the loading matrix is calculated anew. As indicated by the name, the algorithm proceeds iteratively until the changes become very small. Rencher (2003) notes that the procedure can be similar to Principal Factors if either the number of variables or the correlations between them are large. Moreover, there is a tendency to end up in the Heywood case in which communalities are falsely estimated to be larger than one.

Not pursued in this work are *Principal Component Factors* and estimation by *Maximum Likelihood*. The first carries a similarity to PCA not only by its name: Principal Component Factors sets all communalities to one, so $\hat{\Psi}$ becomes a matrix of zeros. The resulting spectral decomposition hence yields the same results as in PCA. It also implies that q factors can be extracted. The only difference is that PCA uses so-called "raw eigenvectors" as loadings, while in Principal Component Factors formula 4.19 is used to calculate the loadings. The difference is merely one of scaling. With PCA already considered, little is lost by omitting this algorithm.

With regard to estimation by Maximum Likelihood, the assumption of a multivariate normal distribution of the predictors is made. The next step is to formulate the (log-)likelihood and numerically maximize it. However, the initial tests in the simulations lasted long and disclosed unstable, sometimes far-off results. Moreover, the method seems prone to end up in the Heywood case and is therefore omitted. The solution by Maximum Likelihood was shown to be equal to the one Rao's Canonical Factoring gives (Timm, 2002, p. 505). Having decided upon a method to extract the factors, one can rotate the system for convenience in interpretation. Rotation uses the indefiniteness property depicted in 4.12 and is discussed in detail in the next section.

The last missing part of the system is the factor scores F. These cannot be calculated immediately. The factor model is different to PCA as it requires an inversion of equation 4.8. So far only the q original variables, the (rotated) loadings matrix and the uniqueness-estimates are available, but there are k unknown factors f and q unknown unique factors ϵ remaining. Hence, the system is underdetermined and inverting the relation in 4.8 is infeasible. Therefore estimation or some other approximation method is required.

DiStefano et al. (2009) partitions the methods into refined and non-refined methods. One example for the latter is the surrogate variable technique which chooses the variable that loads highest on a factor in absolute terms as a representative for the factor. Hereby, several disadvantages occur. The surrogate variables can be highly correlated with each other and are also misleading in interpretation. If unique variance is interpreted as measurement error, then relying on a single variable unnecessarily exposes the model to measurement error and potential advantages of using Factor Analysis remain unexploited. There are other methods which are unclear with respect to their properties, for example the creation of sum scores, which for each factor forms the loading-weighted sum of those variables that load highly on a factor.
Refined methods, on the other hand, create scores by estimating them. The estimation of factor scores stands in contrast to calculating the component scores in PCR. There are several techniques available to this end. Table 4.2 summarizes the most common options to obtain F. If a procedure is marked by a star, a more detailed explanation is provided in this section.

Among the presented extraction methods, a general point concerns the non-existence of an unbiased and correlation-preserving estimator. That means there exists no estimator for factor scores which is unbiased and simultaneously preserves the correlation pattern of the factor model (McDonald and Burr, 1967). Embarking from an orthogonal factor model this implies that factor scores are either correlated

Table 4.2: Overview of factor scores estimation methods

Estimation of F
Bartlett*
Thurstone*
Anderson-Rubin
Non-refined methods

or biased.
One estimation approach views the factor model as a regression model. Bartlett (1937) proposed a weighted least squares estimate

$$\hat{f}_j = (\hat{L}^T \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^T \hat{\Psi} X^T \tag{4.21}$$

which for all factors becomes

$$\hat{F} = X \hat{\Psi}^{-1} \hat{L} (\hat{L}^T \hat{\Psi}^{-1} \hat{L})^{-1}. \tag{4.22}$$

The estimate reduces the influence of variables with high uniqueness and is unbiased. It requires Ψ to be nonsingular. Another approach is regression scoring (Thomson, 1951)

$$\hat{F} = X \operatorname{Cov}[X]^{-1} \hat{L} \Phi \tag{4.23}$$

where $\Phi = I_q$ if the common factors are orthogonal. It requires that $\operatorname{Cov}[X]$ is nonsingular. The method is biased but has a low mean squared error.
A third strand of estimators attempts to preserve the correlation of the common factors. The topic was scrutinized by Ten Berge et al. (1999). The first estimator in this area was due to Anderson and Rubin (1956) which, however, only works if $\operatorname{Cov}[F] = I_k$ and Ψ is nonsingular. McDonald (1981) expanded the solution to oblique factor models. This class of estimators is omitted here for two reasons: The correlation structure between factors is not substantially changed by both the Bartlett and Thurstone method, moreover, in oblique factor models, the correlation structure does not play an important role.

4.1.3 Rotation

When conducting Principal Component Analysis and Factor Analysis, interest lies in being capable of interpreting the indices meaningfully. To this end, the patterns in the loading or weight matrix are examined. The single indices are labeled according to what the original variables that load highest in absolute terms on a

component have in common. This is a subjective aspect where different researchers might emphasize different aspects and therefore yield different conclusions. In general, however, any interpretation of the indices as they come out of the model process is often challenging because they usually carry a high degree of ambiguity. A clearer picture arises if a so-called simple structure (Thurstone, 1931) is present. Loosely speaking, such a structure characterizes a loading or weight pattern where each single original variable loads either highly or low on an index in absolute terms, rarely with intermediate values. Thereby, only few variables with high loadings need to be considered for interpreting the constructs. Rotation techniques can aid in achieving this goal.

Through the formulation of the models, there are differences concerning Principal Component Analysis and Factor Analysis when it comes to rotation. Therefore, the discussion continues with general principles of rotation and is detached from the two methods. Aspects linked to rotation within PCA and FA are elucidated subsequently.

The general principle of rotation can be illustrated best by using diagrams and considering a two-dimensional space of variables. A simple structure is to be obtained with respect to the points in this system, where the points are an ordered pair of values. A rotation towards a simple structure means to rotate the axes, which are the constructs for which an interpretation is required, such that one value of the pair gets either small or large in absolute terms. This is visually exemplified in figure 4.1.
The initial situation corresponds to one, where both values of the pair exhibit similarly high values on both axes. Their use for giving a meaning to the axes, which are the factors, is hence limited. The rotation is indicated by arrows in this picture, whereby the axes $a1$ and $a2$ are rotated orthogonally. After rotation, giving an interpretation to the axes is simplified.

Rotation methods are classified into orthogonal rotations, which turn the axes but keep them orthogonal to each other, and oblique rotations, which are often closer to a simple structure at the expense of the orthogonality property. Supposing L denotes the matrix of weights or loadings, then rotated loadings are given by $\tilde{L} = LT$, where T is a rotation matrix. The elements of L are found by choosing T such that a specific criterion is maximized.
Starting with orthogonal rotation methods, the most commonly used is the Varimax procedure (Jolliffe, 2002, p. 154). Let $\tilde{l}$ denote the loadings after rotation and k

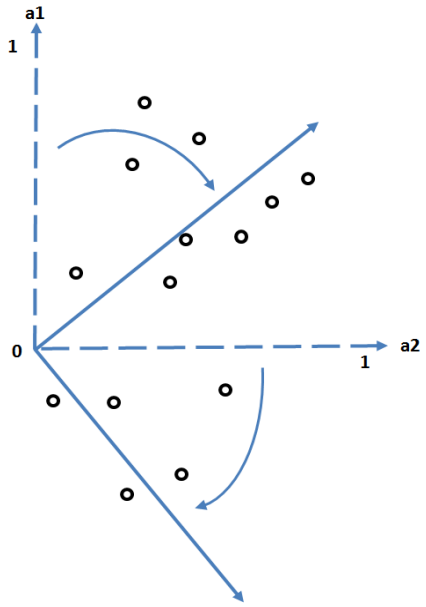


Figure 4.1: Example for rotation.

the number of axes, the criterion to maximize is

$$Q_{VM} = \sum_{j=1}^{k} \left[\sum_{i=1}^{q} \tilde{l}_{ij}^4 - \frac{1}{q} (\sum_{i=1}^{q} \tilde{l}_{ij}^2)^2 \right]. \tag{4.24}$$

The solution is obtained by numerical procedures and maximizes the sum of squares of L column-wise. Noting the fact that loadings are $0 \leq \tilde{l} \leq 1$, it is apparent that Q_{VM} becomes larger as the difference between a single loading, the first term, and the average loading, the second part of the sum, on a factor becomes larger. The maximum is hence attained when the loadings are either 1 or 0. There are also other orthogonal rotation methods but Varimax is the most popular one.
An often-used oblique rotation is the Promax method (Hendrickson and White, 1964). The algorithm starts with a loading matrix and applies at first a Varimax-rotation to it. This result is then attempted to be improved (in terms of attaining a simple structure) by considering the matrix Q, which is defined element-wise as

$$q_{ij} = |l_{ij}^{g+1}|/l_{ij}, \tag{4.25}$$

where $g > 1$ and l_{ij} referring to the Varimax-rotated loadings. The consequence of raising the loadings to a higher power is that small loadings are driven further to zero, while large loadings are only slightly diminished. In the next step the Promax-rotated loadings are obtained by the least-squares fit formula

$$\tilde{L} = (L^T L)^{-1} L^T Q. \tag{4.26}$$

The last step consists of normalizing L, such that each column's sum of squares

equals unity. A value between 2 and 4 is typically recommended as value of g. A consequence of an oblique rotation is the increased focus on large correlations. It leads to more pronounced loadings, which reduces the likelihood of factors to encompass numerous variables.

Having discussed two important rotation procedures, it remains to illumine the application of rotation in Factor Analysis and Principal Component Analysis. First of all, and this holds in both procedures, the number of factors must be smaller than the number of original variables. Otherwise the simplest structure of all is attained, i.e. the components equal the original variables up to the sign. Another point that arises from the rotation idea is that the result depends on the number of considered indices.

Rotation in PCA leads to the following changes (following the elaborations by Jolliffe, 2002, p. 272ff.): Rotating k eigenvectors orthogonally does not only change the loadings but also a component's (co-)variance. Thereby, rotation destroys maximum variance characteristics while the total variability of the concerned components remains constant. But a caveat arises through the fact that the components are no longer uncorrelated – even under orthogonal rotation. This stems from the properties of PCA, which include orthonormal component weights and orthogonal components. After rotation one of the two properties will be lost. The first k components' scores can be calculated as $F_k = XW_k$. Rotation implies to multiply this equation by a $k \times k$ rotation matrix T so that

$$F_k T = XW_k T = X\tilde{L} = F_k^R. \tag{4.27}$$

Because of orthogonal weights $\tilde{L}^T\tilde{L} = T^T W_k^T W_k T = T^T T$, they keep this property in case of an orthogonal rotation. However, with unrotated components it holds that $F^T F = W^T \mathrm{Cov}\,[X]^T\, \mathrm{Cov}\,[X]\, W = W^T W \Lambda W^T W \Lambda W^T W = \Lambda^2$ which is diagonal, whereas with any rotation $F_k^{RT} F_k^R = T^T \Lambda_k^2 T$ which implies non-orthogonality of the components. This observation is of particular importance when the scores are used in regression analysis.

Rotation in Factor Analysis is uncomplicated with regard to these issues. After having decided on the procedure and the number of retained factors, formula 4.9, which shows that any orthogonal rotation is a permissible solution, is exploited. This is because rotation refers to the k-dimensional space of retained factors and not the q-dimensional one of the original variables. When it comes to the interpretation of the indices created by the two methods, factors are suspected to have advantages over components.

4.1.4 Cross-Validation

This section describes a useful algorithm to do model selection in light of limited data and threat of overfitting. The description mainly follows the explanations provided by Arlot and Celisse (2010), James et al. (2013) and Hastie et al. (2009). As mentioned in chapter 3, a way to protect oneself from overfitting models is to conduct model selection on a data set that is split into three independent sub samples. Using these three parts, a training, a validation and a test data set, correctly, reduces the likelihood of using too optimistic models. Yet this procedure requires sufficiently many observations, which is not always the case in practice. For such situations, cross-validation may constitute a well-working alternative.
Given a model, the idea of cross-validation is to omit the validation and test sample and instead work with splits of the training data. In k-fold cross-validation the training data N are randomly split into k mutually exclusive, roughly equally sized chunks denoted as $n^1, ..., n^k$. In the next step, the first chunk n^1 of data is separated from the remaining data $n^2, ..., n^k$. Using only the observations $N \backslash n^1$, a model is estimated and used to predict y_{n^1}, the outcome observations in chunk n^1. These predictions $\hat{y}_{n^1}$, which are out-of-sample, are recorded and are estimates of the test error because the samples are independent of each other. To increase the number of observations with which the test error is calculated, the procedure continues with excluding chunk n^2. The chunks $n^1, n^3, ..., n^k$ are used for fitting the model anew. Based on this model the outcome y_{n^2} in n^2 is predicted and also recorded. The procedure follows in this fashion until all splits have been excluded once. Thereby, the structure of the model must stay the same over the folds to ensure comparability, only the parameter estimates are allowed to differ. Then:

$$CV_k = \frac{1}{N} \sum_{i=1}^{k} L(\hat{y}_{n^i}, y_{n^i}) \tag{4.28}$$

denotes the cross-validated prediction error, where $L(\bullet)$ is a custom loss function for the discrepancy between the true values and the prediction - often the mean squared error, which is a consistent estimate of the test error. To use CV for model selection the same procedure as before is conducted for every other rival model. The model with the smallest cross-validation error should be favored.
Of interest in the current context are the questions of which properties CV has and, more importantly, how k should be chosen; arguably the properties of CV are likely to depend on the choice of k as well as the total number of observations. It is insightful to consider the extreme cases. One case is the leave-one-out cross-validation (LOOCV), where $k = N - 1$, which constitutes the most exhaustive type where only one observation per fold is taken out. Its use can be costly in

terms of calculation time, but there is a shortcut for OLS type problems (Zhang, 1993, p. 301ff.). The other extreme is called hold-out or simple validation, where $k = 2$, and the sample is hence split in half. One risk here are too few observations in one or the other half. In addition, there is a risk of having an unfortunate split. Compromises are found in the range $2 < k < N - 1$.
James et al. (2013) note that the simple validation might turn out to be quite variable if repeated several times. The LOOCV, however, virtually always gives similar results. The authors further argue that there is also a bias-variance trade-off for the choice of k with regard to the estimation of the prediction error. Approximately unbiased estimates of the test error are obtained by applying LOOCV - on the other hand, the single training sets are highly correlated which tends to increase the variance by the amount of covariance (Clarke et al., 2009, p. 594). In analogy to the problem of variable selection, it can be argued that the bias for the estimated prediction error is low for the sample at hand but this error will hardly be generalizable for different samples. When k is reduced, bias is increased while variance lowered. However, due to the smaller fold sizes, the variance can also increase if the training samples exhibit sufficiently variable structures. This danger alleviates when the total number of observations in the sample increases. A typical choice, independent of the sample size, is $5 \leq k \leq 10$ (Hastie et al., 2009, p. 243).

4.2 Subset selection methods

This section summarizes the most important types of subset selection methods. The common feature of these approaches is to choose a subset from the set of original predictor variables and so mitigate the problems of having too many (similar) predictors. The resulting model should be parsimonious and have desirable properties, i.e. be interpretable and to accurately predict the response variable. Since the procedures differ in obtaining a subset, the results often differ. To avoid naming ambiguities among the procedures, this paragraph follows the terminology of James et al. (2013).

Best Subset Selection

The most comprehensive method is the Best Subset Selection method. Starting with the baseline model of a constant only, it compares the fit of all possible combinations of predictors while proceeding from a model with one predictor to

the model with all predictors. Since the R^2 in training samples never decreases, the evaluation of the model battery should be done by means of an information criterion (AIC, BIC) or an appropriate statistic (C_p, Adj.R^2). Computational difficulties arise when the number of variables increases since 2^q models have to be estimated in this procedure.
Alternatively, one can consider optimizing an information criterion directly. The typical structure of an information criterion consists of some measure of goodness of fit and a term that penalizes the model complexity. The minimization criterion for an information criterion in general can be expressed as follows (Savin and Winker, 2013, p. 169)

$$\widehat{\beta}_{IC} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|_2^2 + \lambda\|\beta\|_0). \tag{4.29}$$

The first bit seeks to minimize the discrepancy between the observed and predicted responses via $\widehat{\beta}$. The second part is the ℓ^0 norm on $\widehat{\beta}$, also representable as $\sum_{i=1}^{q} |\beta_i|^0$; it is the count of non-zero elements in the coefficient vector (strictly speaking, zero-coefficients are removed from the vector prior to summing up). Through this optimization criterion the number of variables with non-zero coefficients is penalized. The penalty's strength is determined by λ. The ℓ^0 norm implies, however, that the optimization takes place in a discrete space of models which means that standard gradient methods cannot be used (Ibid.). Thus, estimating 2^q models cannot be avoided. The practical difficulties for this method in data-rich environments has led to alternatives such as Backward or Forward Stepwise Selection.

Backward Stepwise Selection

A faster procedure is Backward Stepwise Selection regression (BSSR). The algorithm is depicted in table 4.3.

A variation of BSSR is to ignore the threshold α_1 and continue elimination until no predictors are left in the model. At each elimination step, the current model is evaluated with respect to the information criteria above and the best is picked.

Forward Stepwise Selection

Forward Stepwise Selection regression (FSS) works the other way round. Its algorithm is described in table 4.4.

Table 4.3: Algorithm for Backward Stepwise Selection

Initialize:	Determine a minimum significance level α_1, whereby significance is the variable-individual probability of rejecting the null hypothesis of a zero coefficient when it is true. It specifies the minimum p-value that each variable in the final model has to meet or to be beneath.
Step 1:	Include all input variables in the model, obtain fitted parameters by OLS and eliminate the least significant variable if its p-value is larger than α_1.
Step 2:	The remaining predictors are fitted anew and, as before, the least significant variable is removed from the set.
Step 3:	Step 2 is repeated until all variables in the model are at least as significant as α_1.

Table 4.4: Algorithm for Forward Stepwise Selection

Initialize:	Determine a minimum significance level α_1. It specifies the level of significance which all variables in the final model have to meet or to exceed.
Step 1:	Starting with constant-only model, the most significant predictor is added to the subset.
Step 2:	Exit, if the predictor is not at least as significant as α_1, otherwise continue.
Step 3:	Given the constant and the first predictor, the next most significant predictor, is added.
Step 4:	Exit, if the predictor is not at least as significant as α_1, otherwise continue.
Step 5:	Step 3 and 4 are repeated until all variables in the model are at least as significant as α_1.

As in the previous procedure, there is a variation of the method. The algorithm continues until all variables are in the model and evaluation takes place as for BSSR.

A hybrid approach of BSSR and FSSR is also possible. In FSSR this boils down to removing non-significant variables from the present model after having added a new variable to the model.

Subset selection methods typically have several drawbacks. One occurs in the presence of multicollinearity where t-statistics are inflated and so disguise a predictors' importance. Good predictors might be dropped too quickly or not added. As mentioned, multicollinearity per se does not lead to biased parameter estimates, stepwise selection methods can nevertheless do so when a variable from the set of multicollinear variables is removed. Moreover, small changes in the data can cause different selection choices which may affect the subsequent choices

substantially (Hesterberg et al., 2008, p 65). More technically, there are problems to repeated testing of the same sample. This includes R^2 values which are biased upwards, while standard errors of the parameter estimates are too small (Flom and Cassell, 2007). It resembles the custom-tailored properties mentioned in the context of overfitting - the chosen model is likely too optimistic.

Incremental Forward Stagewise Regression

The last method presented in this section is Incremental Forward Stagewise Regression which differs fundamentally from the previously described selection methods. This paragraph follows the descriptions by Hastie et al. (2007) and Tibshirani (2014).
The main difference to BSSR and FSSR lies in an r-step algorithm which sequentially adds a small amount ε to the coefficient of the variable that has the largest inner product (or correlation, if variables are standardized) with the current residual of the outcome. Instead of adding or removing a variable completely, only its coefficient is changed slightly. Assuming the outcome y and the predictors $x_1 - x_q$ are standardized, the algorithm proceeds as depicted in table 4.5:

Depending on the choice of ε and the number of variables it will take many steps to obtain reasonable estimates for the coefficients. As demonstrated later in section 4.4, this algorithm has many similarities with the solving algorithms for specific regularization methods.

Table 4.5: Algorithm for Incremental Forward Stagewise Regression

Initialize:	Set the step size $\varepsilon > 0$, but small, and let the vector of coefficients $\widehat{\beta}$ initialize with zeros.
Step r=0:	Choose the variable x_i which has the largest absolute correlation with the residual outcome, i.e. find the i that maximizes $Corr\left[x_i, y - X\widehat{\beta}\right]$. Since the vector $\widehat{\beta}$ contains only zeros in the first round, this reduces to $Corr\left[x_i, y\right]$. Then update the coefficient value of $\widehat{\beta}_i$, which is the coefficient belonging to x_i, by: $\widehat{\beta}_i^{\,r} = 0 + \varepsilon \cdot Sign\left[Corr\left[x_i, y\right]\right]$. The upper index of $\widehat{\beta}_i$ refers to value of $\widehat{\beta}_i$ at step r. Hence the coefficient value of the variable with the highest correlation to the outcome is increased or decreased by the amount ε.
Step r=1:	Choose the variable x_i which has the largest correlation with the residual outcome, i.e. maximize $Corr\left[x_i, y - X\widehat{\beta}^{r-1}\right]$. The latter term no longer cancels out, since the vector $\widehat{\beta}^{r-1}$ no longer contains zeros only. Again, update the coefficient of x_i by the formula: $\widehat{\beta}_i^{\,r} = \widehat{\beta}_i^{\,r-1} + \varepsilon \cdot Sign\left[Corr\left[x_i, y - X\widehat{\beta}^{r-1}\right]\right]$. This takes the coefficient's value of x_i from the previous step, $\widehat{\beta}_i^{\,r-1}$, and adds or subtracts ε to it. Hence, if x_i from $r = 0$ equals x_i from $r = 1$, then $\widehat{\beta}_i^{\,r-1} \neq 0$, otherwise $\widehat{\beta}_i^{\,r-1} = 0$.
Further steps:	Continuing in the fashion of the previous step, such that for each step one element of the vector $\widehat{\beta}$ is updated. This is done until some termination criterion is reached. All variables having zero correlation with the residual is the typical criterion.

4.3 Linear index methods

Index models, sometimes also called factor models, are methods in which a set of predictors is combined in some fashion into new variables, called indices (components, factors). In many applications and also here, these indices are weighted linear combinations of the predictors, constructed in a way that they condense the original variables' information with as little loss as possible. In the next step, the derived indices instead of the original variables act as predictors in a regression. Practically, the reduction is conducted by exploiting linear relationships, e.g. correlations, between the predictor variables. Intuitively, if several predictor variables are highly correlated with each other, they have overlapping explanatory power. An artificial variable which summarizes this shared variance will bundle a lot of the variation and can hence function as a representative for it. Using this artificial variable as an explanatory variable instead of the original ones reduces the dimensionality of the model with, ideally, little loss of information. The magnitude of the information loss depends on the covariance structure of the original predictors and the number of selected indices.

It has to be noted that in all cases in which the predictor's matrix has full rank, some information in the data will be discarded when the number of indices is chosen to be smaller than the number of original variables. The disregarded variation is sometimes interpreted as noise. On the other hand, if all predictor variables are already orthogonal to each other, no dimensional gains are feasible. As this rarely occurs, problems induced by having too many predictors are suspected to be effectively addressed by dimensional reduction techniques.

Apart from pure dimensional reduction, the analysis can have the goal of inspecting and interpreting the indices. This concerns particularly the weights or loadings, in order to obtain an overview of the data structure and patterns in it. Corresponding to the goal of interpretability, the indices might be meaningfully viewed as latent variables. This corresponds to the interpretation of the predictors' correlation structure as stemming from latent variables. There is, however, no proof for the validity of any interpretation and it is up to the researcher to come up with a plausible and credible interpretation.

Not only relationships within the set of predictor variables can be exploited. If the indices are eventually used as regressors, it may be beneficial to regard associations between the predictors and the dependent variable in the index-building process. On grounds of these methodical differences, methods can be classified accordingly. If the indices are created independently of the dependent variable, the procedure is called unsupervised. Principal Component Regression and all types of Factor Regression fall into this category. If, on the other hand, the dependent variable is systematically regarded in this process the analysis is termed supervised. In other words, regardless what the dependent variable is, the unsupervised analysis always yields the same indices as regressors, while there would be different ones in supervised index models – possibly to fit the dependent variable better.

This definition uncovers one main weakness of unsupervised methods: The explanatory relevance of the indices for the dependent variable is not directly controllable. Only a good prior choice of predictors is able to yield improvements. Supervised models draw on this disadvantage. Such approaches seek to achieve a balance between reducing dimensionality and good prediction (Principle Covariates Regression) or use regression techniques to find the weights in the indices (Partial Least Squares).

4.3.1 Unsupervised index models

Principal Component Regression

Principal component regression (PCR) is a procedure which is based on the results of a Principal Component Analysis. The extension is straightforward. The scores that have been obtained by equation 4.2 (or a rotated version of them) measure an individuals' propensity toward each component. These scores are consequently used in a regression on the dependent variable. Given the components, the minimization criterion reads as follows

$$\widehat{\gamma}_{PCR} = \underset{\gamma}{\operatorname{argmin}}(\|y - F\gamma\|_2^2). \tag{4.30}$$

As indicated by the 2-index, its core is the ℓ^2 norm, which in this case can also be written as $(\sum_{n=1}^{N} |y_n - F_n\gamma|^2)^{1/2}$. In order to remove the effect of the square root, the norm is squared.
In the context of this work, PCR becomes meaningful only if fewer components than original variables are chosen, i.e. $k << q$, for otherwise no dimensional reduction takes place. It triggers the questions of which and how many components should be retained.
This decision partly depends on whether the components were rotated or not, since rotation changes the properties of the components. The conclusions drawn in the section on rotation were the following: Without rotation the components are uncorrelated but normally difficult to interpret. With rotation, the components are correlated and probably better to interpret. The result of the rotation, however, depends on the number of selected components. The maximum number of selectable components is $q - 1$, because for q selected components the original variables are retrieved through rotation (up to the sign). One implication of this relationship is that a selection of components cannot be based on the set of rotated components but must refer to the unrotated ones. Rotation can therefore only be the second step after having made a decision on certain unrotated components. Additionally, the variance distribution in a subset of components changes through rotation while the total variance remains constant. Even if no rotation is intended, at least one component must be discarded, otherwise $q = k$. This is different to methods which are based on Factor Analysis where $k < q$ by procedure.

One popular criterion developed by Kaiser and Guttman (Guttman, 1954) is to use all components whose corresponding eigenvalue is larger or equal to one $\lambda_j \geq 1$. The motivation is that only those components compress variance more

efficiently than a single variable. Another way is to look at the size-arranged plot of eigenvalues and visually identify the "elbow" in the plot, the so-called scree plot test. This criterion can be subjective and hard to formulate mathematically which renders it unsuitable for evaluating a battery of models as in a simulation. Another suggestion, Parallel Analysis (Horn, 1965) has the rationale to compare the original data with random data and in this way to infer on the share of noise in the former. The decision rule is to keep those components from the original data whose eigenvalue is larger than the average corresponding eigenvalue from the random data set. Additional propositions, like the Bartlett's test on the equality of roots (Bartlett, 1950) or the average root procedure, can be found in chapter 2 in Jackson (1991).

One disadvantage all these approaches share is that it is unclear how well the components explain the outcome. Caused by the PCA's pure orientation towards dimensional reduction, the goal of explaining variation in the predictors is indeed achieved but this happens independently of the outcome. Unfortunate situations arise when important variation for explaining the dependent variable is excluded by selecting the wrong components. The extreme case would be where the component with the smallest eigenvalue contains all the relevant variation (Jolliffe, 2002, p.174). This issue is called here *last-factor-phenomenon*. Such a situation seems unlikely to occur in practice where a sensible choice of predictors minimizes this risk but it can not be expected in general that the components' explanatory power correlates sufficiently with the size of the eigenvalue. Hence, an ad hoc selection that takes account of the results of PCA and the components' relation to the dependent variable is needed. The procedure obtains a tinge of supervision in this way.

One idea to tackle this problem is to use cross-validation. Here the first step is to calculate the scores for all q components. Starting with a one-component model, where the component is the one with the largest eigenvalue, cross-validation is conducted for this model. In the next step, the component with the second largest eigenvalue is added to the model, which is also cross-validated. This is done until the model with all components is cross-validated. The model in the step which has the lowest test average error will constitute the choice.
This is a costly procedure, since many cross-validations have to be conducted. Moreover, using the eigenvalue-ordering to sequentially add components to the model has the possible drawback of picking up bad predictors on the way: if, for instance, components $1 - 10$ are bad predictors and first number 11 is relevant, then the optimal choice due to CV is larger or equal to 11, although the first ten components could be dropped.

This argument points back in the direction of Best Subset Selection, which removes bad predictors. Given the computational efforts, the orthogonality property of the (unrotated) principal components is exploited by conducting a selection on values of t-statistics. Adding or removing a component from the model does not change the t-statistics of the other components. Therefore one can state a certain threshold level of a t-statistic and keep all components which are good enough at the cost of one regression. The threshold level itself could be determined using CV, i.e. searching through different levels of significance. One suggestion is to start at a level of $\alpha = 0.01$ and incrementally increase the level until 0.15 is reached. However, Jolliffe (2002, p. 175) notes that t-tests for components with small variance have low power and are therefore less likely to be retained. Moreover, including components with high predictive power but low variance can lead to instability, which negatively affects the generalization ability. The compromise the author suggests is to start eliminating components with the lowest variance onwards until a component is found which is sufficiently significant. As with selection by CV, this approach has the drawback that bad components with higher variance but little explanatory power remain in the model. In consideration of this, the risks of selection by p- or t-value seem acceptable.

When it comes to interpretation, the weight structure of the included components is relevant. An interpretable structure is important because the components' meanings are traced back to it. A simple way to obtain the loadings of a specific component is to conduct a regression of the component on the set of predictor variables. The resulting coefficient estimates equal the weights. This relation holds for both unrotated and rotated components.
Rotation in the context of the suggested model selection algorithms means for the Kaiser-Guttman approach that after rotation some components with variance less than 1 are included. For sequential extraction methods, it implies that a components' meaning depends on the number of extracted components. In both methods, the total variability and thereby the model fit stays constant. The same also holds for selection on t-values. However, it has to be noted that rotation of hand-picked components is typically not implemented in statistical software. Usually one is only able to rotate k sequential components.

In the simulation presented in chapter 5, two different options of using Principal Component Regression are examined:

- Due to its popularity and calculation speed: The Kaiser-Guttman criterion keeping all components with $\lambda \geq 1$.
- Significance level cut-off. Threshold selection by CV, using the levels 15 %,

10 %, 5 % and 1 %.

Factor Regression

A method related to PCR is Factor Regression (FAR). The general procedure is similar, apart from FAR relying on Factor Analysis instead of Principal Component Analysis. Therefore some issues that occur in PCR also occur in FAR. Yet it is worthwhile to take a closer look at the approach, because its theoretical foundations correspond well to the encountered economic problem.

Having discussed the basic properties of a factor model, one proceeds as follows in Factor Regression: After the decision which procedure is used to extract the factors, the extracted loading matrix is typically rotated to facilitate interpretation. Rotation in FAR is distinctly less problematic than in PCR, as the factor scores can still be estimated as being (nearly) orthogonal to each other. Having estimated them, the factor scores, marked by hat due to their origin, are utilized as regressors. The optimization criterion is as follows:

$$\widehat{\gamma}_{FAR} = \underset{\gamma}{\operatorname{argmin}}(\|y - \hat{F}\gamma\|_2^2). \tag{4.31}$$

The general points made in PCR also apply mostly here. In particular the point, that the outcome from rotation depends on the number of extracted factors. There are two important differences, however: One is that the number of factors will be lower than the number of original variables in FAR if the factor extraction is one that works on the reduced correlation matrix. It is hence possible to consider all factors and still achieve a dimensional reduction.[21] Despite this initial dimensional reduction, it is not guaranteed that it is enough to yield an interpretable model. Due to the construction of factor models, there arises a second difference which concerns cut-off criteria such as the one of Kaiser-Guttman. Factor Analysis seeks to explain correlations - variables that do not correlate highly with others often end up on factors with small eigenvalues. Although the theoretical basis of this work assumes correlating structures within the set of explanatory variables, this need not be the case for each single variable. A single variable which is important and only weakly correlated to other variables is presumably disregarded when cut-off criteria are applied. But even if the variable showed substantial correlations to other variables, it might still be outnumbered if larger clusters of correlated variables existed.

[21] Strictly speaking: Factors with eigenvalues larger than 0, for statistical software sometimes displays negative eigenvalues.

Rotation of factors is omitted in the simulations, although it is likely to have an effect on the Kaiser-Guttman approach. There is a difference between first selecting factors with variance larger than 1 and then rotating them or first rotating all factors and then selecting factors according to Kaiser-Guttman. Usually, there are fewer factors in the first variant because variance is maximized for each factor according to the procedure described in section 4.1.1. Common practice seems to be the first variant.

Factor Regression models will be evaluated with respect to the following variations both for Principal Factors and Iterated Principal Factors as factor extraction methods and the Bartlett Method as factor scores estimation method:

- Due to its popularity and calculation speed: The Kaiser-Guttman criterion keeping all components with $\lambda \geq 1$.
- Using all available factors, i.e. $\lambda > 0$. Dimensional reduction occurs through factor extraction.

Selection on t-values is omitted here because oblique rotations are particularly appealing in Factor Analysis. But with correlated factors, selection on t-values becomes at least as computationally expensive as Best Subset Selection. If the minimum level of significance shall be additionally determined by cross-validation instead of using an information criterion, a multiple of 2^k models has to be computed. Instead one specification will combine FAR with regularization methods. Because this variant requires an additional technique, it is introduced in section 4.4.

Comparison between PCR and FAR

This comparison section mainly follows (Jolliffe, 2002, p.158-161). Comparing PCR and FAR essentially boils down to two important differences between PCA and Factor Analysis. One is the different model structure which results in estimating (Factor Analysis) or calculating (PCA) the scores. The second point concerns the procedure to obtain the loading/weights matrix. Both procedures apply an SD on a correlation matrix. However, the diagonal elements of $\mathrm{Cov}\,[X]$ remain unchanged in PCA whilst they are typically reduced in Factor Analysis. In PCA the objective is to account for as much variance as possible, which often also explains the off-diagonal elements well because covariance is common variance. In Factor Analysis it depends on how the regressors are interrelated. If the variables are strongly correlated, the uniqueness share is low; if they are virtually independent the uniqueness will be almost as large as the variance. It is for these reasons

that Factor Analysis is known for explaining covariance and PCA for explaining variance.
These notions become clearest through the following example of a variable which is uncorrelated to the other variables: it will get its own component in PCA but not a common factor in Factor Analysis. To be considered by Factor Analysis, it must be in a group of at least two correlated variables. Results of Schneeweiss and Mathes (1995) point in this direction, suggesting that PCA and Factor Analysis are most similar, if the number of predictors is high and their uniqueness low. Drawbacks for Factor Analysis arise through the somewhat arbitrary choice of factor extraction, rotation and score estimation. On the contrary, rotation as a medium of easing interpretation has less disadvantageous implications.

4.3.2 Supervised index models

Principal Covariates Regression

Principal Covariates Regression (PCovR) (De Jong and Kiers, 1992) is another index-creating method and can be motivated by the drawbacks of PCR and multivariate regression. The following description of the method rests mainly on the works of De Jong and Kiers (1992), Kiers and Smilde (2007) and Vervloet et al. (2013).

If dimensional reduction of the predictors and good prediction of some criterion are not concurrent, owing to the mentioned last-factor-phenomenon for instance, a recourse may be to find a compromise between the two goals. Hence a lower-dimensional subspace is searched for, which seeks to balance variance compression in the predictor variables with explaining or predicting the outcome. This is done by extracting certain linear combinations of the predictors, the principal covariates (or more general: components). The core idea of PCovR is to minimize a weighted sum of dimensional reduction error and prediction error.
The model is explicitly formulated by the following three equations (De Jong and Kiers, 1992)

$$\begin{aligned} F &= XW \\ X &= FL^T + E \\ y &= F\gamma^T + u. \end{aligned} \tag{4.32}$$

The first equation is the dimensional reduction feature, where F is a $N \times k$ index matrix, X is a $N \times q$ vector of predictors and W is a $q \times k$ weight matrix. Dimensional reduction is achieved when $k < q$, which implies that some of the original variance in the X is cast away unless perfectly collinear variables were

among them. In this formulation the notion of PCA is depicted. The second and the third equation amend the relation in a latent variable model style. X is explained by the index matrix via L, a $k \times q$ loading matrix. Its unexplained variance is denoted as E. The aim is to keep it low. The third equation describes the relation of the outcome to the components, where γ denotes the regression coefficients and u the unexplained variation in y. The aim at this step is likewise to minimize the residuals but now of the outcome. Since the minimization goals are unlikely to coincide, a weighted average may be a compromise.

Given the two minimization goals and a quadratic loss function, the following minimization criterion is formulated in PCovR (assuming all concerned variables to be standardized) as

$$min\ \Xi(W, L, \gamma) = \frac{\alpha \left\| X - XWL^T \right\|_F^2}{\|X\|_F^2} + \frac{(1-\alpha) \left\| y - XW\gamma^T \right\|^2}{\|y\|^2}, \tag{4.33}$$

where $\|\bullet\|_F$ is the Frobenius matrix norm. For identification purposes, the component scores are often constrained to being column-wise orthonormal, i.e. $F^TF = I$. The first term is the reduction term seeking to minimize the discrepancy between the original predictors and the components. The second term is the prediction term, minimizing the residual sum of squares u. The parameter α regulates the weight that is given to either and eventually needs to be given a value by the researcher; convention is a value on the interval between 0 and 1.
Here, it is worthwhile to analyze the limit cases $\alpha = 1$ and $\alpha = 0$, because they turn out to lead to familiar procedures. In the first case, $\alpha = 1$, all weight is put on dimensional reduction, none on reduction in prediction error. This leads to PCR. When $\alpha = 0$ the emphasis is completely on reducing prediction error, then PCovR will produce results equal to multivariate regression or, when F is of lower rank than both y and X, reduced rank regression (for instance described in Aldrin, 2006).
An expedient aspect of PCovR is that it has a closed-form solution for a specific α and a specific number of components k (Vervloet et al., 2013, p. 37). Hereby F is set equal to the first k normalized eigenvectors of the matrix G given by

$$G = \alpha \frac{XX^T}{\|X\|_F^2} + (1-\alpha) \frac{H_x y y^T H_x}{\|y\|^2} \tag{4.34}$$

where $H_x = X(X^TX)^{-1}X^T$. Then W, L and γ are calculated as

$$\begin{aligned} W &= (X^TX)^{-1}X^TF \\ L &= X^TF \\ \gamma &= y^TF. \end{aligned} \tag{4.35}$$

The final equation in terms of the original predictors that can also be used for prediction is

$$y = XW\gamma^T \tag{4.36}$$

How to choose α and the number of components k has only been addressed for the limit cases so far. However, neither PCR nor OLS/RRR are desirable. De Jong and Kiers (1992) seem most convinced of using cross-validation, which can be justified by the general applicability of the procedure. This implies running through two nested loops, one with different values of alpha, the other with the number of components. This approach is called simultaneous selection procedure by Vervloet et al. (2015). Choosing this procedure can result in long calculation time; the magnitude depends on the size of the data set, how fine the grid is chosen (i.e. how many combinations of α and k are considered) and the number of cross-validation folds. An alternative adapted by Vervloet et al. (2013) is the so-called sequential procedure. Its principle is to first determine α based on assumptions on the errors E and u in equation 4.32 which make estimation by maximum likelihood feasible. These assumptions concern the distributions, which are taken to be normal with mean zero. Variances are either predetermined or estimated through PCA and multiple regression respectively. Once a value for α has been found, the number of components can be determined by either cross-validation or an information criterion. With the results of number of components at hand, one could optionally return to α and check for optimization potential.

Vervloet et al. observe that the estimate for α moves towards 1 when the predictor data become noisier relative to the outcome and when the number of predictors increases. Choosing α can hence be statistically approached, but there are also theoretical considerations which may be taken into account. The authors further note that the choice matters especially little when compression in X and predicting y are concurrent. This would be the same case where PCR is expected to perform well and a low risk for the last-factor-phenomenon predominates. The advantages of PCovR become therefore relevant, when compression and prediction are not concurrent or there even is a trade-off between them.

Empirical results on this topic are mixed, but tend to favor a larger to a smaller α. The study by Heij et al. (2007) finds the optimal α between 0.9 and 1 but also dependent on the amount of error in the predictor variables. Also the study by

Kiers and Smilde (2007) indicates the optimal value larger than 0.5. Vervloet et al. (2013) find that choosing α by the sequential procedure performs well in case of having precise information about the error in the regressors. If such information is not available, they advocate the use of cross-validation.

In this dissertation, the sequential procedure with subsequent cross-validation on the number of components is applied. Prior checks did not indicate large differences between the procedures and their modifications, so that the speed argument gains weight in simulations with many replications.
A note concerns the properties of the components in the context of prediction on new data. Since future/out-of-sample outcomes are assumed to be unknown, the procedure cannot be applied to these data. Merely the right-hand-side variables are known. An out-of-sample prediction, comp. equation 4.36, therefore implies combining coefficients belonging to the "old" sample with predictor data from the unused sample. One consequence is that the predicted component scores will no longer be orthogonal to each other. This is a property that supervised methods share and which does not apply to unsupervised methods unless one separates the training data from the new data also in the process of index building.

The last point in the context of desired properties is to be able to interpret the components. As seen by equation 4.36, both γ and W are of importance. The rotation suggested by Vervloet et al. (2015) refers only to the loading matrix L and hence does not affect the other two matrices. Since the matrices stem from an optimization criterion directed to mathematical distances, interpretation of components might turn out to be as difficult as in the case of PCR.

Univariate Partial Least Squares

Partial Least Squares (PLS) is a method accredited to Herman Wold (Wold et al., 1984) and has gained considerable popularity in fields like chemometrics, sociology and marketing. Although PLS was originally developed for several endogenous variables, the method is described for a single one only as it is sufficient and simplifies the description. PLS features a strong similarity to Canonical Correlation Analysis, because the comovement of two sets of variables is maximized. In Canonical Correlation Analysis, the correlation is maximized, while PLS maximizes the covariance in the case of non-standardized variables. The relevant difference lies in the number of components that are created. In the case of a single endogenous variable, the maximum number of components for Canonical Correlation Analysis is one, whereas PLS can have more. If not

otherwise stated, this section follows the descriptions by Boulesteix and Strimmer (2007), Wang et al. (2005) and Mevik and Wehrens (2007).

Partial Least Squares is based on the factor model

$$\begin{aligned} X &= FL^T + E \\ y &= F\gamma^T + u \end{aligned} \tag{4.37}$$

where F denotes the component scores, L the loading matrix and γ the vector of regression coefficients. The residual term in the regression equation is denoted by u.
The components themselves stem from the observed variables

$$F = XW \tag{4.38}$$

where W denotes the loading weights. To obtain the components, the matrix of loading weights being of dimension $q \times k$ needs to be calculated. Once W is obtained, F can be derived and the elements of the vector γ estimated by an OLS-regression.
The approach in PLS to find W is such that the covariance between y and F, which is the linear combination of the variables in X, is maximized. The scaling of variables affects the solution, so standardized variables are assumed here for simplicity and comparability to other methods that encourage standardization.
The maximization criterion for the loading weights vector $w_{,1}$ of dimension $q \times 1$, abbreviated as w_1 to unclutter notation, belonging to the first component can be written as follows (Boulesteix and Strimmer, 2007)

$$\begin{aligned} &w_1 = \underset{w}{\operatorname{argmax}}\, w_1^T X^T Y Y^T X w_1, \text{ subject to:} \\ &w_1^T w_1 = 1 \text{ and } w_1^T X^T X w_1 = 0. \end{aligned} \tag{4.39}$$

Once w_1 has been found, the vector is used to calculate the first component f_1 which is, in turn, put into action as a regressor in the regression models

$$\begin{aligned} X &= f_1 l_1^T + X_1 \text{ and} \\ y &= f_1 \gamma_1^T + y_1 \end{aligned} \tag{4.40}$$

where $l_{,1}$ and γ_1 denote the first component's loadings and the regression coefficient respectively. X_1 and y_1 denote the residuals. In the subsequent step, these residuals substitute X and y in equation (4.39) and the above procedure is repeated to obtain $w_{,2}$ and f_2. In this fashion, the procedure continues until the desired number of components k with f_k is reached.

For a final model, the dimensionality of W with respect to number of components, k, must be determined. The number of regarded components can lie between 1 and $min(N, q)$. Contrary to PCA where the eigenvectors are calculated through a SD, the maximization problem in PLS is typically solved by iterative algorithms, of which several exist. In this dissertation, the NIPALS-algorithm is applied. With a single dependent variable, the NIPALS-algorithm leads to the same results as the SIMPLS-algorithm (Kiers and Smilde, 2007, p. 200). Table 4.6 presents the NIPALS-algorithm after Mevik and Wehrens (2007).

Table 4.6: NIPALS-algorithm for Partial Least Squares

Initialize:	Standardize all variables.
Step 1:	Calculate $w_{,1}$ as the cross-product $X_1 y_1$, where the number in the subscript denotes the iteration step.
Step 2:	Normalize $w_{,1}$ by $w_{,1} = \frac{w_{,1}}{\sqrt{w_{,1}^T w_{,1}}}$, and calculate the first component as $f_1 = X_1 w_{,1}$
Step 3:	Conduct a regression of y_1 on f_1 and save the vector of residuals in y_2.
Step 4:	Conduct regression of each variable in X_1 on f_1 and save the residuals in X_2.
Further steps:	Iteratively repeat Step 1-4, thereby increasing the iteration step by one each time, i.e. using y_2 instead of y_1 and X_2 instead of X_1 to calculate $w_{,2}$ in the second iteration, until the last iteration, i.e. $min(N, q)$, is reached.

The issue of how many components to choose can be approached with cross-validation. Moreover, in the special case when the variables in matrix X are orthogonal, PLS yields the OLS estimates after one step (Hastie et al., 2009, p. 81).

An important aspect, in particular when compared to other supervised methods, such as Principal Covariates Regression, is how the procedure implicitly balances variance compression in comparison to the correlation with the dependent variable. This aspect could be operationalized by the percentage of explained variance in y and X to get a feeling for the goodness of fit and the amount of compression. Hastie et al. note that the compression of variance tends to dominate the correlation to the dependent variable.

When it comes to the interpretation of the results, it begins with the analysis of the significance of the latent components. Typically, it will be of descending order from the first to the k-th component. In the second step, the predictor's weights on those latent components have to be examined. Interpretability will depend on whether it is possible to determine the most important original variables for the most important components. Due to the deflation of the original variables, the

interpretation has to be conducted carefully for other than the first component. For interpretation, in PLS similar issues as in PCovR arise. Rotation methods, such as proposed by Wang et al. (2005), refer to the loading matrix, which is of minor help here. It may therefore be difficult to obtain meaningful interpretations from this method.

4.4 Regularized models

This subsection is about regularized linear models, which are marked by two characteristics. They assume a direct linear functional relationship of the predictor variables' parameters to the outcome and the parameters are subject to regularization, which means to introduce a penalty on their size. While the methods presented in the previous section also belong to the category linear models, their predictors are linear combinations of the original variables. For the methods of this subsection the set of original variables enters the model directly as in a classic regression model.
The arguments given in chapter 3 highlight the risk of applying OLS on a linear model with all available regressors. This chapter introduces modifications which make the linear model more suitable for (generalizable) predictions.

As with OLS, the starting point of regularized linear models is to minimize the residuals of a given linear model. The difference lies in an additional constraint (the penalty) which is taken account of in the estimation of the model. This constraint refers to the values of the parameters and stipulates an upper bound for their sum of their values. The constraint ensures that the upper bound cannot be exceeded. If the value is chosen small enough, the solution becomes one in which parameters have been dragged away from their OLS estimates towards zero. If the upper bound is too large, i.e. the constraint is not binding, the OLS solution will result. The introduction of estimation bias constitutes an immediate consequence of a binding constraint because the coefficients no longer equal the OLS estimates. But since the coefficients are reduced in absolute size, the variance of the model is reduced.

The methods in this group differ conceptually in the transformation applied to the coefficient values before they are summed up. Among the best-known methods in this area is Ridge Regression (Hoerl and Kennard, 1970). Its constraint consists of the sum of squared coefficient values, the L^2-norm, which must be lower than or equal to some threshold value. The Lasso (Least Absolute Shrinkage and

Selection Operator) introduced by Tibshirani (1996) uses the sum of absolute coefficient values, the L^1-norm. There are also hybrid methods such as the Elastic net (Zou and Hastie, 2005). One feature they all share is to shrink parameters gradually. By contrast, the general minimization criterion for information criteria (comp. equation 4.29) can also be viewed as a type of regularization. But the L^0-norm leads to the fact that a parameter is either not shrunk at all or set to zero. Generally, the type of constraint influences two important aspects. One is feasibility of estimation and with that the uniqueness of the estimates, the other factor is interpretability. In the next section, it is demonstrated that the Lasso, in particular, has some beneficial traits which other regularization techniques do not have.

A general disadvantage of regularized linear models is their lack of interpretability in a latent variable model sense. There are no derived factors, only the set of original predictors, which consists of proxy variables. On the other hand, one gets to see the strength of each predictor's association with the outcome, something that is not immediately evident in index model regression types. Moreover, there are indications of a superior predictive ability of regularized linear models, which provides a useful reference point to compare the performance of linear index models.

Lasso

The Lasso is a technique that seeks to minimize the residual sum of squares of the linear model under the constraint that the sum of absolute coefficient values is lower than a specified value. If not otherwise stated, the descriptions in this paragraph follow Hesterberg et al. (2008), Savin and Winker (2013) and Tibshirani (1996).
As pointed out, the difference between estimation by the Lasso and OLS lies in the additional constraint on the parameters. The optimization problem can be written as

$$\widehat{\beta}_{Lasso} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|_2^2), \text{ subject to: } \sum_{i=1}^{q} |\beta_i| \leq t.$$

The first part in the optimization argument is the usual criterion to minimize the distance between the outcome and a linear combination of the predictors. The part in the constraint says that the sum of β in absolute terms shall not exceed t.

Bringing the two parts together yields the Lagrangian formulation

$$\widehat{\beta}_{Lasso} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{q} |\beta_i|).$$

The complexity parameter $\lambda \geq 0$ and t feature a one-to-one correspondence. The smaller the lambda, the stronger the shrinkage. The scale of the predictors generally affects the solution which is why they are standardized. Moreover, the constant in the model, whose coefficient is non-zero if the outcome is not centered, is not subject to shrinkage. Otherwise the solution would depend on the origin.
To obtain the estimates for the βs, one cannot resort to a closed-form solution because the constraint makes the solution non-linear in y. But since the problem is convex (when X is of full rank) it can be solved for a given λ. There are two acknowledged options for this. One is numerical optimization using the coordinate descent algorithm (s. Friedman et al. (2010) for details). Another possibility is using the Least angle regression (LARS) algorithm in the Lasso modification (Hastie et al., 2009, p. 76), which offers computational advantages and is therefore used in this dissertation. The algorithm is described in detail in table 4.7.

The practical aspect of this algorithm lies in its calculation time. It has been demonstrated that with one run only it produces the entire path of Lasso solutions, that is as the shrinkage parameter λ ascends from zero to infinity. The algorithm features strong similarities to the one for Incremental Forward Stagewise Regression. This is because the latter is also a modification of the LARS-method (Efron et al., 2004). Both variations can deliver similar or under certain conditions even equal results (Tibshirani, 2014). Hence the added value of Incremental Forward Stagewise Regression is expected to be low and therefore not regarded in the simulation.

The Lasso is not the first method to apply a shrinkage approach. Precedent to it was Ridge Regression whose drawbacks are addressed by the Lasso. To elucidate the Lasso's advantages, it is insighful to consider Ridge Regression in more detail. Given the quadratic constraint, the minimization criterion for Ridge Regression is

$$\widehat{\beta}_{Ridge} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|_2^2), \text{ subject to: } \sum_{i=1}^{q} \beta_i^2 \leq t,$$

and its Lagrangian formulation is hence

$$\widehat{\beta}_{Ridge} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{q} \beta_i^2).$$

There is no need for an iterative algorithm, because this problem has a closed form

Table 4.7: Algorithm for the Lasso

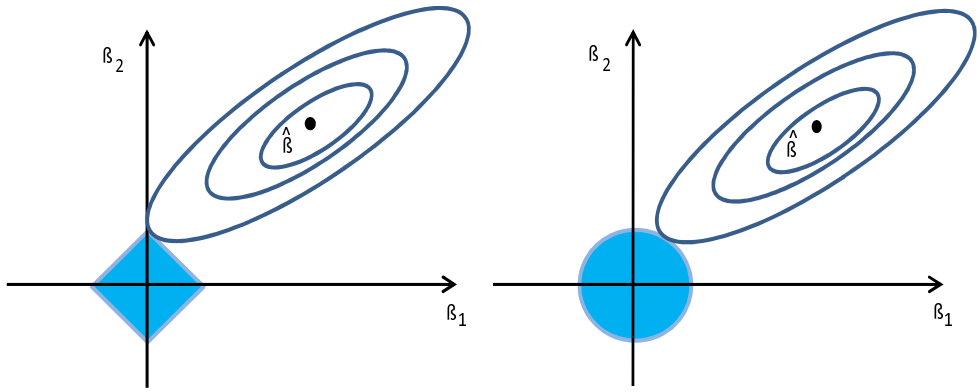
Initialize:	Center the outcome and standardize the predictors. Let the vector of coefficients $\widehat{\beta}$ initialize with zeros.
Step 1:	Choose the variable x_i which has the largest absolute correlation with the residual outcome, i.e. find the i that maximizes $Corr\left[x_i, y - X\widehat{\beta}\right]$. Since the vector $\widehat{\beta}$ contains only zeros, this term reduces to $Corr\left[x_i, y\right]$. Define a set A which denotes the set of active variables in the model and add x_i to A.
Step 2:	The coefficient of x_i is gradually moved from zero towards its univariate least squares estimate. Along the way the residual outcome is steadily updated, i.e. $y = y - x_i\widehat{\beta}_i$. $\widehat{\beta}_i$ is moved as long as no other predictor has a higher correlation with the current residual outcome. This procedure results in a steadily decreasing correlation between x_i and the current residual of the outcome.
Step 3:	Once another variable, say x_g, has as much absolute correlation with the current residual, x_g is added to the model and to the set A. Now both x_i and x_g are in the model and equally correlated with the current residual. Then a regression of the current residual of the outcome on the set A is conducted. Based on these results, the coefficients of the variables in set A are moved into the direction of their joint least squares estimates. As in step 2, the current residual is steadily updated during the movement.
Further steps:	The updating takes place until another variable has as much correlation and step 3 is repeated. If a non-zero coefficient becomes zero again in this process (e.g. by changing it's sign) then the connected variable is dropped from the set A and the joint least squares estimates are obtained anew. The algorithm stops if the correlation is zero for all predictors.

solution which is

$$\widehat{\beta}_{Ridge} = (X'X + \lambda I)^{-1}X'y.$$

The difference between Ridge Regression and the Lasso finds expression when the constraint binds sufficiently strongly. Such a situation is illustrated in figure 4.2. It shows the following: The picture on the left side belongs to the Lasso, the picture on the right side to Ridge Regression. Depicted is the case of a linear model with two predictors whose parameter values are displayed on the axes. The point $\widehat{\beta}$, which is the same in the two pictures, is the parameter pair emerging under estimation by Ordinary Least Squares where no shrinkage occurs. When these parameters are changed, the mean squared error for the given sample rises, because OLS provides the unbiased estimates. Around the point $\widehat{\beta}$, contours of the error function are displayed. A contour illustrates combinations of parameters with a constant mean squared error for the estimation sample. They are elliptical due to the quadratic first part of the minimization criterion and are hence the same in both illustrations. The solid areas around the area where the axes cross

Figure 4.2: Contour plot of optimization rationale



are the constraint functions. The parameters have to lie within or at the margin of these areas. When optimization takes place, both methods will find the first coordinate where a contour meets the constraint. In the graphical example for Ridge Regression, this point will be one where both parameters are non-zero. For the Lasso, however, one parameter will be zero.

This observation has an implication for the interpretation of a model. Because the reduction in the coefficients proceeds proportionally, no parameter estimate is exactly zero in Ridge Regression estimates, which means that all variables have to be taken account of in the interpretation. The only advantage over OLS estimates is the better predictive ability caused by exploiting the bias-variance trade-off. The Lasso, on the other hand, performs an actual variable selection in addition. In comparison to unbiased methods such as OLS, however, an important disadvantage occurs. Standard errors and therefore confidence intervals cannot be calculated easily, even using bootstrap. Goeman et al. (2016) even argue that it is almost always impossible in practical cases to obtain standard errors. Their argument being that one needs to obtain unbiased estimates of the bias since it is part of the mean squared error. For shrinkage methods and their associated modification of the bias, these are normally not available, however. Reporting standard errors would ignore the bias in the coefficients and is therefore misleading. However, Lockhart et al. (2014) have recently provided an approach to significance testing in the Lasso-framework. However, it is not undisputed (for instance in Fan and Ke (2014) and its corresponding implementation in R did not show convincing results under commonly applied thresholds. Another approach is to use the Lasso-selected regressors in a subsequent OLS-regression. However, significance tests there do not account of the prior selection procedure and therefore type-I errors will be overstated. A third alternative is calculating credible intervals with a Bayesian approach, but this is not pursued here. It seems therefore most appropriate to

state the coefficient size only.

The remaining task in the estimation of a model is to determine the shrinkage parameter λ. Since the Lasso algorithm computes the path of solutions for all λ, cross-validation can be easily applied. In addition to cross-validation, a popular choice is to use a model selection criterion, often the minimum of a Mallows's C_p-type statistic Zou et al. (2007). Similar to other criteria such as the AIC, it becomes more favorable for models with a good fit or a number of observations and sanctions a high number of regressors, in particular weak ones. The authors deliver the following definition

$$C_p(\hat{y}) = \frac{\|y - \hat{y}\|^2}{n} + \frac{2df(\hat{y})}{n}\sigma^2. \tag{4.41}$$

One challenge applying this formula in the context of the Lasso arises with the degrees of freedom $df(\hat{y})$. They cannot be determined in the conventional way, as all variables remain in the model even though with shrunken coefficients. The authors, based on both theory and simulations, propose to estimate the degrees of freedom by the number of non-zero coefficients in the model. A model selection can then be done by choosing the λ for which the C_p statistic has its minimum. Advantages of this statistic may occur when sample sizes are small or the sample is particularly homogeneous, as CV, depending on the number of folds, may yield unstable results in such situations.

In regard to the Lasso's properties, the ability to yield parsimonious models could often be confirmed. Savin and Winker note a drawback, though: in a setting of pairwise highly correlated predictors, Lasso fails to select the complete set of 'true' predictors. If there are one or more groups of highly correlated variables, the Lasso tends to randomly pick only one of each group. Methods such as Elastic Net and Adaptive Lasso deal with the potential drawbacks.

Adaptive Lasso

Adaptive Lasso introduced by Zou (2006) is an extension of the Lasso. The author motivates the use of the Adaptive Lasso by emphasizing potential variable selection inconsistencies occurring in Lasso. To point this out, he introduces new terminology. If a procedure asymptotically achieves correct identification of variables with zero coefficients, and the difference between the true and the estimated parameters converges in distribution to a normal distribution, it is said to have the "oracle property". Accomplishing this goal also improves the estimation of the nonzero

coefficients (Fan and Li, 2001). The improvement can hence be seen as yielding a sparser model.
The basic idea is to include prior information about the importance of the variables into the estimation process of a Lasso. This information is regarded as a weight in the estimation. Important predictors will be shrunk less than unimportant predictors. This results in a re-weighted Lasso. Prior information can, for example, stem from sources, such as univariate OLS regressions or a Lasso estimation. The optimization criterion is

$$\widehat{\beta}_{AdaLasso} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{q} \omega_i |\beta_i|),$$

where $\omega_i = 1/|\widehat{\beta}_{init,i}|^{\xi}$ for $\xi > 0$ is the additional weighting term in which the prior information enters as $\widehat{\beta}_{init,i}$.
The computation of the Adaptive Lasso does not involve many extra steps compared to the Lasso. As soon as the predictors' individual weights are obtained, the predictors are divided by their corresponding weight and afterwards used in a Lasso estimation. In a final step the coefficients are divided by the corresponding weight.
In this thesis the individual weights are obtained by Lasso estimates using a 5-fold Cross-Validation. Also the shrinkage factor λ is retrieved by using Cross-Validation.

Elastic Net

The Elastic Net by Zou and Hastie (2005) is a method that, amongst others, addresses the Lasso's issue with groups of highly correlated right-hand side variables. The description of this method follows Zou and Hastie (2005) and Hesterberg et al. (2008). The starting point is the observation that the Lasso selects one predictor in a group of highly correlated variables, which is called sparsity property, while Ridge Regression shrinks the coefficients of these variables toward each other, the so-called grouping effect (Hastie et al., 2009, p 662). In the extreme case of perfectly collinear predictors, a desirable property of an estimator would be identical coefficients for the concerned variables.

The Elastic Net approach connects both shrinkage types in an optimization criterion and so exploits the sparsity property of the Lasso and the grouping effect of Ridge Regression. In general, the method tends to keep or drop groups of highly correlated predictors when Lambda grows, whereas the Lasso, by contrast, tends to drop

smaller groups or single predictors.
The optimization criterion for the Elastic Net is as follows

$$\widehat{\beta}_{ElasticNet} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|_2^2 + \lambda_1 \sum_{i=1}^{q} |\beta_i| + \lambda_2 \sum_{i=1}^{q} \beta_i^2).$$

Determining the pair of tuning parameters can be carried out by running CV over a predefined grid of value pairs for λ_1 and λ_2.
The literature emphasizes that a corrective calculation should be applied. Because the Elastic Net performs a double shrinkage, it is likely to introduce too much bias. This is corrected through rescaling the estimated parameters by

$$\widehat{\beta*}_{ElasticNet} = \frac{1}{\sqrt{1+\lambda_2}} \widehat{\beta}_{ElasticNet}.$$

This last step, however, plays a more important role in the simulation study than for the interpretation of a model. For making sense of the predictors, only the relative importance of variables with non-zero coefficients is considered. Rescaling them by a constant term does therefore not change the picture.

Combing Factor Regression with regularization

An approach that combines the strengths of latent variable models and regularization methods is what I call the Factor Regression with Lasso variable selection, short FarLasso. The starting point is to extract all (rotated) factors, but instead of an OLS-regression, apply the Lasso technique to them. The optimization criterion is hence defined by

$$\widehat{\gamma}_{FarLasso} = \underset{\gamma}{\operatorname{argmin}}(\|y - \hat{F}\gamma\|_2^2 + \lambda \sum_{i=1}^{k} |\hat{\gamma}_i|).$$

It can be observed that the shrinkage is applied directly on the factors, which yields an additional dimensional reduction provided the chosen shrinkage parameter is large enough. Interpretability of the factors is ensured within the boundaries of the Factor Analysis model, because factor scores can be computed from the rotated loading matrix. Since the Lasso is expected to perform satisfactorily when the regressors are moderately correlated, factors arising from oblique rotations should not be problematic. In the case of high correlations between factors, the Elastic Net or the Adaptive Lasso could pose alternatives to the Lasso.
To the best of my knowledge, combining Factor Regression and the Lasso is a new

approach. Compared to the approach of using all factors in the regression model – thereby exploiting merely the inherent dimensional reduction of Factor Analysis – FarLasso will shrink some factors' influences to zero and therefore likely produce a more parsimonious model.
With regard to interpretation, the same remarks as for Factor Regression hold. Loadings indicate the correlation between the original variables and the factors, and the (shrunken) coefficient size expresses a factor's relative importance in the model. A drawback from the viewpoint of interpretation can arise, if only few factors attain a zero coefficient, and many others are equipped with a small but non-zero coefficient. While the induced bias might have improved the predictive capability, the model is still hard to overview on grounds of a large number of retained factors. Such a situation can occur if coefficients vary strongly in size. Using different shrinkage factors, such as the Adaptive Lasso does, has the potential to mitigate this issue since weak predictors are penalized beforehand. For that reason, the FarAdaLasso is also considered. Its optimization criterion is

$$\widehat{\beta}_{FarAdaLasso} = \underset{\beta}{\operatorname{argmin}}(\|y - \hat{F}\gamma \|_2^2 + \lambda \sum_{i=1}^{q} \omega_i |\gamma_i|),$$

where the parameter meanings and the procedures to obtain them are defined as before.

Chapter 5

Simulation

The earlier chapters revolved around the topic how a latent variable model can be used to describe the family background's influence on a child's school success. In this framework, observed variables are interpreted as manifest variables which are expressions of latent variables, but do not necessarily carry a causal effect on the outcome. Choosing them according to whether they are an expression of family background led to having many potentially powerful regressors. With a data set rather small in observations and multicollinearity structures among the predictors, the need for a dimensional reduction was highlighted. Several methods that could achieve this were proposed and their potential advantages and caveats discussed. In particular, performance dependencies on the data structure were debated. However, the discussion so far could only allude to the important aspects of fit and generalization ability. To answer whether one method is preferable to others, a simulation study is conducted. By means of synthetically generated data, the researcher can control the data environment and obtain insights into the performance of methods within particular data structures. The objective of this exercise is to deliver a further basis for the decision on the method applied on real data.

Two models are proposed as simulation frameworks in this dissertation: One is a latent variable model, the other a regression model. These models were chosen owing to theoretical considerations and specify the data generating process (DGP) by their structural model equations. These structures are fixed, but they contain various parameters that can be altered. Changing those, changes the DGP and thereby the data structure. The data can, for instance, exhibit different degrees of multicollinearity or let the regressors vary in their explanatory contribution to the outcome. Additional setscrews concern general settings - for instance the number of observations a model is estimated on. If small-sample and large-sample properties are different, a simulation study will show.

A specific parametrization and setting configuration is called scenario. In this dissertation, there is one main scenario that serves as a baseline. It is the most important one, for it resembles the real data structure closely. The remaining scenarios differ to this baseline scenario in only one particular property at a time. Examining the performance in different data environments facilitates the identification of strengths and weaknesses of the proposed methods. Insights into whether a method's performance is affected by structural changes in the data set or a setting will support the overall assessment.

The measure of performance applied here is the average fit, i.e. how close the model predicts the observed values. On observation basis, a quadratic loss function is applied. Since the amount of explicable variation differs across scenarios and replications, relative performance differences between the methods constitute the final measure of performance. Relying on the fit as a performance measure has the advantage that different ideas, such as regularization and index models, can be easily compared. A disadvantage of this measure is that it does not take into account a model's sparsity in terms of the number of retained variables/factors.

An alternative goal to measuring the fit is parameter recovery, that is to calculate the difference between the true and estimated parameter values. Achieving this goal is not of primary theoretical interest, however. As stressed in the theory section, the explanatory variables are interpreted as proxy variables and not as structural determinants causally affecting the outcome. However, parameter recovery and fit are correlated since close recovery of true parameters means that a method is not significantly influenced by error in variables. In addition, out-of-sample predictions are good when the estimated model parameters are close to the true ones (Vervloet et al., 2013).

The fit as performance measure makes checking the accuracy of out-of-sample predictions intuitive. In this simulation, data are randomly split into in-sample data (training data) and out-of-sample data (test data). To obtain the model and its corresponding parameter estimates, only the training data are used. The test data are held back and are used exclusively for comparing the predicted with the actual dependent variable. Since in-sample and out-of-sample data differ, performance is measured separately for each data set.

Partitioning the data in this fashion has several reasons. The most important is to check the generalization ability of a model. If a model fits the training data well but fails in the test sample, doubt is cast on the model's validity as a whole (Kiers and Smilde, 2007). Ideally, the model would perform on average only slightly worse on test data, otherwise there could be an indication of overfitting (comp. explanations on the MSE in chapter 3). A property closely linked to this aspect is the stability of a model: small changes in the data set should not cause severe

changes in the model. To be generalizable, a model must be stable. Considering only in-sample performance for the quality of a model may be misleading, as it is difficult to ascertain how it performs in other data sets, even if a model selection criterion was applied. Evaluating out-of-sample prediction accuracy is, therefore, a natural and fair way to compare different methods.
Another reason for partitioning the data in this manner is practical advice; the intention of this approach is to mimic real situations. When observations are scarce, the researcher merely has sufficiently many observations to employ a training data set. There is no possibility to divert data for a validation or even a test set to do sophisticated model selection. In order to obtain a sparse model nevertheless, model selection must take place on training data set as well, e.g. choosing the λ in the Lasso or the number of factors in Factor Regression. Hence, model selection criteria are additional elements that influence a method's performance and so add an algorithmic flavor to the procedure.

The remaining chapter is structured as follows. The next section presents the results from related research to point out the gaps this inquiry addresses. After that, the two simulation models are introduced, followed by a detailed discussion of the implemented specifications, the scenarios, and remaining technical aspects. Thereupon the replications and evaluation are described, followed by the results. Finally, a conclusion is provided.

5.1 Related Literature

Relevant literature has been selected with respect to performance comparisons which involve the presented methods. There being many ways to select a model given a method, it is not required to have the same choice on model selection algorithms. This section gives an impression of the proposed methods' performances and presents the reasons for the need to conduct a simulation study as proposed.
Heij et al. (2007) compare the performance of Principal Component Regression and Principal Covariates Regression for time-series data in which one-step ahead out-of-sample forecasts are used to evaluate the predictive capabilities. For component selection in PCR the authors use different types of information criteria. Applying PCovR, the number of components is chosen ad hoc and the weighting parameter α is varied over five values on the interval between $0+\epsilon$ and 0.9. Instances of dynamic factor models are used as underlying data-generating processes. Those models are based on different parameter values to examine the performance under various data structures. Among others the absolute number of regressors and

their correlation structure is subject to change. The authors' results suggest that PCovR can exhibit superior predictive power when there are many underlying factors in the data. However, the method is prone to overfitting if the number of factors and/or the α is chosen inadequately. Another insight is that the BIC as selection criterion for the number of components in PCR outperforms related ones such as the AIC.

Another study set in a time-series framework is by Savin and Winker (2013) who examine and compare the performance of the Lasso, the Adaptive Lasso and the Elastic Net to a modified version of the BIC on OLS. Their data-generating processes are based on Autoregressive Distributed Lag models and contain variations of a larger number of predictors with varying degrees of collinearity situated in time-series of short or middle-length with different number of observations. Moreover, they add few noise regressors, which have no relation to the outcome, into the set of predictors. The authors conduct the evaluation on out-of-sample data once with regard to the correct choice of predictors, i.e. possibly discarding all noisy regressors, and once in terms of accuracy by means of the mean squared error. Their results show that modified information criteria dominate the Lasso-type methods in medium or larger sized samples with low noise both in terms of accuracy and selection. When observations are scarce, however, Lasso-types perform better. Among the Lasso-type methods, the Adaptive Lasso often provides relatively better results.

The simulation study that is included in a work by Bair et al. (2006) deals among others with variants of Principal Component Regression, Partial Least Squares and the Lasso. Particular to this simulation study is that the number of regressors exceeds the number of observations by far but only a sliver of them actually contributes to the outcome. This setting emphasizes the danger of overfitting and excludes methods that rely on having more observations than variables such as OLS. To compare the methods, the authors define two data-generating processes. The first has a rather simple structure, which makes the identification of important variables and thereby the prediction not too difficult. The second process is harder since clusters of relevant information correlate substantially with noise. The methods' performances are compared by means of measuring cross-validated and out-of-sample error. The author's results suggest that the fit differences are not particularly large. One exception is Principal Component Regression using only the first component which performs particularly badly. However, due to the special data structure, merely careful overall conclusions can be drawn from this study.

Multicollinear data structures are explicitly addressed in a study by Kiers and Smilde (2007). The authors compare among others Principal Component Re-

gression, Partial Least Squares and Principal Covariates Regression. Their work pursues a two-track approach to evaluation by not only considering predictive ability but also recovery of the true parameter values from the data-generating process. Both measures are evaluated by means of average absolute deviations. The simulation designs applied in their work involve manifold data structures. Relevant conclusions – from the viewpoint of this dissertation – include that in settings with many variables and high collinearity, PCR and PLS perform poorly in terms of predictive power. PCovR, on the other hand, produces reasonable results in many settings. For all methods the authors analyzed they found that the performance with respect to prediction works better than with respect to recovery. Also, the work by Dormann et al. (2013) deals with the issue of multicollinearity. The group of authors examine several methods they consider to be suitable to address the drawbacks that come with clusters of highly correlated data. Amongst the examined methods are the Lasso, types of Principal Component Regression and Partial Least Squares. The author's simulation setting is one where the number of variables is relatively low in comparison to the number of observations. Variations concern the complexity of the relationship between the predictors and the outcome as well as correlation patterns between the predictors. The latter are designed as having clusters of predictors in which the degree of correlation varies across specifications. The prediction error in test data is taken as a measure of success here, too. The results portend that rather the collinearity structure of the data than the complexity of the model underlying the DGP plays a role. In more detail, the authors found the Lasso performing well in settings with no multicollinearity while PCR and PLS perform worse. In settings with moderate collinearity, all three methods perform sufficiently well. In all settings, however, the predictive ability of the Lasso is at least slightly better than that of PLS and PCR.
Another relevant study in this context is by Vigneau et al. (1997). The authors compare Ridge Regression, PCR, a combination of the two, which has some similarities to the one proposed in section 4.4, and PLS to OLS. The performance is evaluated by considering the cross-validation error over two data sets with different degrees of collinearity among the predictor variables. With a special focus on the way components are selected, their results portend that PLS and PCR can yield unstable results, while Ridge Regression as well as the combination of Ridge Regression and PCR exhibit stable and relatively good results.

Most of the research presented in this section concentrates on forecasting performance, ignoring the property of (non-)interpretability. This explains why models based on Factor Analysis are rarely considered; their main pro, and the reason they play a significant role in the methods proposed, is the promise of the factors' superior interpretability. A better predictive performance than models which are

based on Principal Component Analysis can hardly be expected.
Unfortunately, many studies do not explicitly state the applied model selection criteria. This is problematic – in particular for methods like PCR where a variety of different criteria could be used. As to the remaining methods, different data-generating processes render comparisons between the simulation studies difficult. Hence, at most careful conclusions in the direction of general performance tendencies can be drawn. Regularized linear models typically perform as least as good as index models. Within the realm of the latter, the results are ambiguous. According to my best knowledge, there exists no simulation study which analyzes the performance of the suggested methods in a data environment as it has been described here.
In addition to be able to adjust the parametrization, a customized simulation study has two crucial advantages: The first is to allow a comparison between methods that yield highly interpretable models and methods with less interpretable results but presumably good generalizing capabilities. The insights gained help in the decision process for the most suitable method. Secondly, the newly proposed methods combining Factor Analysis with Lasso-type factor selection methods can be evaluated directly. Theoretically, these methods are particularly promising for fulfilling the purpose from a theoretical point of view, conducting a separate simulation study is hence beneficial at this stage.

5.2 Models

5.2.1 Latent variable model

One model is the latent variable model, as for instance described in Timm (2002), which corresponds to the theoretical considerations in this work. In this model the existence of a set of latent variables, so-called factors $f_1 - f_k$, is assumed. They shape both the observable outcome y and the set of observed variables $x_1 - x_q$. In the context of this work, these factors are interpreted as dimensions of family background, the outcome as the child's school achievement. If the observed variables are allowed to have a certain share of idiosyncratic variance ϵ_i which cannot be traced back to the latent factors, they can be interpreted as noisy proxy-variables for the factors.
As family background factors do not explain school achievement completely, there is a specific factor f_s in this model which captures sources of remaining variation; this could be attributed to community factors outside the family environment,

school, teachers or peers. This factor may or may not be correlated with the other factors but is assumed not to affect the parents' observed explanatory variables. In its general formulation, the underlying latent variable model for the simulations is described by the following two equations

$$y = f_1 b_1 + ... + f_k b_k + f_s b_s \tag{5.1}$$

$$x_i = f_1 l_{i,1} + ... + f_k l_{i,k} + \epsilon_i. \tag{5.2}$$

To obtain realized values for the simulation, one needs to specify the joint distribution of the elements in the model. It determines the distributions from which both $f_1 - f_k$ and y are drawn in the simulation. Moreover, it stipulates how the factors relate both to the dependent variable by specifying the vector b, and the set of explanatory variables by defining the loading matrix L. These parameters are the principal setscrews with which the model can be altered to obtain different data structures.
In the simulations, the set of observed variables $x_1 - x_q$ function as the input variables for the candidate methods, which have to somehow select the most useful ones. From the model equations, it is apparent that an observed variable's relevance for the outcome depends on two elements: the parameter vector b and the loadings matrix L. For variable x_i to be an important predictor, the loading $l_{i,j}$ must be large but also the parameter b_j belonging to factor j must be large relative to the other parameters. The interpretation is that an observed variable is relevant for the outcome if it is highly correlated with a latent factor which is sufficiently relevant for the outcome. If, on the contrary, a predictor has high correlations with irrelevant factors only or has a particularly large share of specific variance, then it could turn out to be a poor predictor.

5.2.2 Regression model

The second model is a regression model in which the observed variables cause the endogenous variable and consequently involve no intermediate factors. As in the latent variable model, however, an additional factor that leaves some unexplainable variance in the dependent variable is present.

$$y = x_1 b_1 + ... + x_k b_k + f_s b_s \tag{5.3}$$

Specifying the parameters defines the joint distribution of the outcome, the regressors and the factor. In comparison to the previous model, solely the parameter b

determines an observed variable's relevance for the outcome.

5.3 Specifications

Underlying all specifications is a so-called *basic scenario* which provides the baseline configuration of the model. Alternations are introduced by changing a single setscrew at a time. Starting with a detailed description of the basic scenario, table 5.3 summarizes the other scenario's specifications and how they are realized. An initial remark concerns the random elements described in the following. At first, and thereby prior to the replications, the true model is created. This creation is based on drawing random numbers from prespecified distributions, which differ across the scenarios. The resulting model is fixed for the replications, which means that at the beginning of each replication the same data set is accessed. Stochastic elements enter the process only by disturbing the observed variables in the replications. To sum up, although the true model is also based on random draws from certain distributions, the difference is that it only occurs once per scenario and before the replications.

Latent variable model - Basic scenario The simulations are based on $N = 1000$ observations, whereof $n_1 = 800$ serve as the training sample and the remaining observations $n_2 = 200$ as the test sample (deviations occur in the *small sample size scenario*). The data set will consist of $k = 15$ unobserved factors and $q = 100$ observed variables.

Generation of $f_1 - f_k$: In the first step the factors $f_1 - f_{15}$ and f_s are drawn from a standard normal distribution, i.e. with zero mean and variance one.[22] In addition, it is ensured that the factors are independent of each other, so that

$$f \sim N(0, I_k). \tag{5.4}$$

Generation of y: The values of the parameter vector b, which define the single factors' impacts on the outcome are likewise drawn from a standard normal distribution. Based on these data and added normally distributed noise, the outcome y is calculated.

[22] For information on how actual values are obtained consult the article on random number generation by L'Ecuyer (1998) and the description of obtaining specially distributed numbers by Cheng, R.C.H. (1998).

Generation of $x_1 - x_q$: In the next step the factors $f_1 - f_{15}$ have to be related to the observed explanatory variables. To avoid scaling effects through single variables, they should have the same variance, which is set to one here but could be any other number. It corresponds to the usual standardization of variables, though.
By specifying the loading matrix the connections between factors and variables is established. One condition for the basic scenario is that it resembles the data structure encountered in the real data set. The arbitrary assumptions on the factors so far are permissible since the true factors are unknown and the dependent variable can always be rescaled. The loading matrix L, however, cannot be chosen completely arbitrarily for it affects the correlation matrix of the observed variables. For this matrix a real counterpart exists. Designing the loading matrix is done by inspecting Varimax-rotated loading matrices in the data. The implications of Varimax-rotations and eyeballing suggest that the observed variables often load highly on only one factor, seldom on two or three; rarely, variables do not load highly on any factor at all but instead exhibit a high degree of specific variance. A synthetic loading matrix should hence reflect these aspects. Hand typing the entries of a matrix of dimension 100×15 is costly, therefore random numbers from specific distributions and under certain restrictions are drawn here. While there are multiple ways to achieve a certain structure of the loading matrix, two other conditions have to be met in addition: For a start, a single loading must not exceed one in absolute terms. Moreover, the squared loadings for a variable must sum up to smaller than or equal to one. If one of the conditions fails, an observed variable has a variance larger than one, which, as defined above, must not happen. The following two-step procedure ensures that the necessary properties hold. In a first step only the desired structure is set up, which is depicted by a preliminary loading matrix. In a second step this structure is transformed to fulfill the imposed restrictions.
Let the entries of the preliminary loading matrix L^* be generated by drawing random numbers from a ratio-distribution that has the potential to cover a large spectrum of numbers

$$L^* \sim \frac{N(0,1)}{(2 \cdot U(0,1) - 1)}. \tag{5.5}$$

Table 5.1 indicates the conceptual structure of the preliminary loading matrix L^*. Drawing from the above distribution often yields values that exceed 1 by far, therefore the row sum of squares also (almost) always exceeds 1. However, since the final loading matrix retains the relative shares on loadings, only the ratio of a single squared loading to the sum of squared loadings in a certain row is important at this stage. To assess whether the desired loading matrix structure has been

Table 5.1: Conceptual sketch of intermediate loading matrix L^*

	f_1	f_2	$\dots$	f_{15}	Row sum of squares
x_1	$l^*_{1,1}$	$l^*_{1,2}$	$\dots$	$l^*_{1,15}$	$\sum_{j=1}^{15} l^{*2}_{1,j}$
x_2	$l^*_{2,1}$	$\vdots\dots$	$\vdots\dots$	$\vdots\dots$	$\sum_{j=1}^{15} l^{*2}_{2,j}$
$\vdots$	$\vdots\dots$	$\vdots\dots$	$\vdots\dots$	$\vdots\dots$	$\vdots$
x_{100}	$l^*_{100,1}$	$\dots$	$\dots$	$l^*_{100,15}$	$\sum_{j=1}^{15} l^{*2}_{100,j}$

attained, a row-wise consideration of the loadings is undertaken. It turns out that single loadings are often relatively large compared to others in the same row. This resembles the empirical observations that a single variable often loads highly on one factor after rotation and that several variables load highly on the same factor. The next step is to rescale the loadings so that their squared sum is smaller than one; preceding this, however, the variable specific share of communality (or uniqueness respectively) on total variance has to be defined. Determining this share limits the absolute influence of the factors on an observed variable. The communality share for each variable is independently drawn from a distribution that is calibrated such that it roughly matches the empirically observed pattern of communality. The beta-distribution is a suitable candidate for its support lies between 0 and 1 and its shape is flexible. Let $0.95 \cdot (1 - Beta(2,9))$ be the underlying distribution from which the variable-specific share of communality c_i for variable x_i is drawn. This distribution gives an average communality of about 0.77 with a variance of 0.011 and a median at about 0.79.
The scaling equation which preserves the loading rations and transforms an unscaled loading $l^*_{i,j}$ to the scaled loading $l_{i,j}$ is given by

$$l_{i,j} \equiv \frac{l^*_{i,j}}{\sqrt{\sum_{j=1}^{k} l^{*2}_{i,j}}} \cdot \sqrt{c_i} \tag{5.6}$$

such that: (5.7)

$$c_i = \sum_{j=1}^{15} l^2_{i,j} < 1. \tag{5.8}$$

As a result, table 5.2 conceptually indicates the final loading matrix:

In a final step, the random noise vector ϵ has to be specified. Given the model assumptions, the observed variables' individual variances need to equal one. So far, the sum of an observed variable's squared loadings yields the variable's individual

Table 5.2: Conceptual sketch of final loading matrix L

	f_1	f_2	$\dots$	f_{15}	Row sum of squares
x_1	$l_{1,1}$	$l_{1,2}$	$\dots$	$l_{1,15}$	$\sum_{j=1}^{15} l_{1,j}^2 = c_1$
x_2	$l_{2,1}$	$\vdots\dots$	$\vdots\dots$	$\vdots\dots$	$\sum_{j=1}^{15} l_{2,j}^2 = c_2$
$\vdots$	$\vdots\dots$	$\vdots\dots$	$\vdots\dots$	$\vdots\dots$	$\vdots$
x_{100}	$l_{100,1}$	$\cdots$	$\cdots$	$l_{100,15}$	$\sum_{j=1}^{15} l_{100,j}^2 = c_{100}$

variance. This sum yields values on the open interval between 0 and 1. To obtain a variance of one, the noise component must consequently fill in the remaining variance for each variable. On variable level, this amounts to $1-c_i$ and is generated according to

$$\Sigma_\epsilon \sim N(0, (1 - \sum_{j=1}^{15} l_{i,j}^2) \cdot I_q). \tag{5.9}$$

The specific variances are hence normally distributed and independent of each other which reflects the interpretation of variable-specific variance, for instance induced by measurement error.

Regression model - Basic scenario The simulations are based on $N = 1000$ observations, whereof $n_1 = 800$ serve as the training sample and the remaining observations $n_2 = 200$ as the test sample (deviations occur in the *small sample size scenario*). The simulated data set consists of $q = 100$ observed variables.

Generation of $x_1 - x_q$: The regression model is based on similar calculations as the latent variable model. Here, no specific data structure, as for instance clusters of correlated variables, is assumed. To compare the results to the ones of the latent variable model it is useful to have a similar data structure nevertheless. To generate the true model data, the first step consists in generating 15+1 factors as in the latent variable model. The next step, which is to calculate the dependent variable, is skipped, however. Instead the observed variables $x_1 - x_{100}$ are created by proceeding in the same manner as in the latent variable model. The variables therefore result from the factors and the defined loading matrix.

Generation of y: In the third step, the observed variables and factor f_s are used to calculate the outcome. Here, another difference to the latent variable model occurs. Instead of drawing the parameter coefficients from a standard normal distribution, a mixture distribution $N(0,1) \times Beta(2,22)$ is used. This is done because the differences in the coefficients would have been otherwise too small to

pose a variable selection challenge. Lastly, normally distributed noise disturbs the outcome.

Common manipulations in all models and scenarios

Prior to the replications, some of the observed variables are discretized by quantile splits. Hereby, ten randomly chosen regressors each are recoded into dummy (two-category), three-category, five-category and ten-category variables.

5.4 Scenarios

Since the two data-generating processes only differ in the manner the endogenous variable is generated, the scenarios are the same for both simulations except for slight changes the noisy regressors scenario.

Dispersed loadings

This variation alters the distribution of factor loadings from a ratio distribution with fat tails toward a *Beta(2,2)*-distribution. The loadings become hereby more evenly distributed over the factors. Its implication is that the variables load similarly highly on more factors than before, yielding a data structure that aggravates selecting good predictors.

Small sample size

One setscrew that does not manipulate the joint distribution of the variables is the sample size. This scenario will examine the effects of reducing the sample size significantly and thereby emphasize the pitfalls of having too many variables in relation to observations. In this scenario the number of observations remains at $N = 1000$, but the training sample size is reduced to $n_1 = 400$, so that $n_2 = 600$. The remaining parameters are the same as in the basic scenario.

School factor

In the basic scenario the school factor is uncorrelated with the other factors. This scenario defines the school factor to be a linear combination of some of the other factors plus some normally distributed noise. Also, the weights for the linear combination are drawn from this distribution. The school factor is still unobserved in the simulation, but is now correlated to the regressors through the other factors.

Noisy regressors

In this variation, 15 random (25 in the regression model specification) regressors

are picked after they have been disturbed stochastically in a replication. These regressors are replaced by a weighted sum of the disturbances of 6, randomly chosen variables plus normally distributed noise. The weights are randomly obtained in each replication by drawing from a *Gamma(2,1)*-distribution. The procedure thereby generates artificial regressors which are marked by being essentially uncorrelated to the outcome but at the same time correlated to 6 of the other regressors.

High uniqueness

This variation modifies the distribution of communality. c_i is now drawn from a *Beta(4,10)*-distribution which results in a lower average communality (0.67). This increases the share of a predictor's specific variance. The immediate consequence is a weaker link to the factors, which lowers their predictive ability in the latent variable model DGP. In addition, correlations between the regressors are attentuated, because the specific variances are uncorrelated across the regressors.

Table 5.3: Overview of scenarios

Name	Communality distribution	Loading distribution
Basic scenario	c_i~0 $95 \cdot (1 - Beta(2,9))$	L^*~$\frac{N(0,1)}{(2 \cdot U(0,1)-1)}$
High Uniqueness	c_i~$Beta(4,10)$	L^*~$\frac{N(0,1)}{(2 \cdot U(0,1)-1)}$
Spread loadings	c_i~0 $95 \cdot (1 - Beta(2,9))$	L^*~$Beta(2,2)$
Name	Distinctive features	
Noisy regressors	Partial replacement of regressors which are noisy	
School factor	Correlation between factors and school factor induced	
Small sample size	Sample size reduced by half	

5.5 Replications and evaluation

The previous section described the generation of the data that serves to generate specific data structures. In order to have variation in the replications, the observed variables and the outcome are stochastically disturbed in each replication. This can be interpreted as adding measurement error to the data. The disturbance enters in the following way:

In the latent variable model, each standardized variable is disturbed by normally distributed noise with zero mean and variance ranging randomly from 0.09 to 0.36. In the noisy regressors scenario, the difference of the undisturbed and disturbed variable also serves as the basis for generating the noisy regressors. Those differences are disturbed by normally distributed noise with mean zero and

variance 0.0025.
Except for the noisy regressors scenario, the same disturbance pattern applies to the regression model specification. In noisy regressors scenario the original variables are disturbed by standard normally distributed noise, i.e. having a mean of zero and a variance of one. The subsequent difference is disturbed by a normal distribution with mean zero and a variance ranging randomly between 0.000001 and 0.000004.
Since the noise was drawn from a continuous distribution, the discrete variables are subsequently recoded by means of quantile splits.
Finally, the standardized endogenous variable is also disturbed using standard normally distributed noise.
Each scenario undergoes $g = 100$ replications. In each of them the predicted residual sum of squares of the endogenous variable, the school achievement indicator, divided by the number of observations is calculated - split by method and dataset (in-sample, out-of-sample). These results serve as base to compute the overall statistics for evaluation.
The evaluation of the simulation study is conducted as follows: As the goal is to compare different methods, it is of interest to calculate differences in relative performance. Due to different shares of noise in the data, the amount of explainable variation differs from replication to replication, which makes the comparison of absolute values misleading. Relative values, obtained by dividing the predicted residual sum of squares of a method by the predicted residual sum of squares of the method with the lowest prediction error in a replication, are hence calculated. This is done for each replication, separately for in-sample and out-of-sample data. Thereby, it relates all methods' performances to the best performance. The best method is consequently allowed to differ by scenario, sample and replication. However, since OLS is the BLUE estimator and no considered method transforms features to gain predictive power, its predicted residual sum of squares must constitute the reference in all in-sample data sets.
The calculated relative deviations are the results. They are presented in two ways. One is boxplots, which provide a quick overview of the ranking and performance differences between the methods. To facilitate interpretation, the deviations in the boxplots are expressed as percentage values. If, for instance, a method has a 1.31 higher MSE than the best method in a specific replication, the data point is visualized as 31 %. The boxplots are sometimes cut off from above for some methods. This happens when the performance differences for single replications become too large. Trimming the boxplot helps to focus on the differences between the methods that perform reasonably well at the low cost of masking severely poor performances.

The second way to present the results is by means of tables which provide detailed numeric information about the mean, median, standard deviation and mean absolute deviation of a method's relative deviations.[23] The table is more objective than a boxplot in the sense that the latter's scaling can influence the impression. Moreover, the dispersion is not necessarily discernible in a boxplot.
The reason to also include the median and the mean absolute deviation is grounded in the occurrence of outliers. Some methods produced poor models in some of the replications while they worked out fine in the majority of others. Hence, there are some cases in which their mean squared error skyrockets. In these cases, the trimmed boxplots show a too optimistic picture by not displaying extreme values. Since failed models can be identified relatively easily, it would also have been possible to exclude the concerned replications. Doing so, however, could induce bias through selection – likely in favor on this method. Another point connected to outliers is that deviation values larger than 9, corresponding to a deviation of more than 900 %, are short-coded by > 9.

5.6 Results

This section contains the results of the simulation study. For an overview, the examined methods, including their model selection criteria, are presented in table 5.4.

Before turning to the results, one note concerns the manner of forming indices in unsupervised methods. While in-sample data and out-of-sample data are strictly separated in supervised methods, access to the independent variables of the out-of-sample data set is granted to unsupervised index methods for the calculation of indices. This can be justified by arguing that the explanatory part of the out-of-sample data is known, since it is used for prediction. The statistical advantage lies in a larger sample size for the creation of the indices. The technical advantage is that the same factors/components appear in both samples without extra calculations.

[23] To facilitate readability both the standard deviation and the mean absolute deviation are multiplied by 100.

Table 5.4: Overview of methods and algorithms

Base Method	Model Selection	Abbreviation
OLS	None	OLS
Factor Regression	Extraction by Principal Factors, estimation by Bartlett-method, no further selection	FarPfBar
Factor Regression	Extraction by Iterated Principal Factors, estimation by Bartlett-method, no further selection	FarIpfBar
Factor Regression	Extraction by Principal Factors, estimation by Bartlett-method, minimum eigenvalue 1	FarPfBarM1
Factor Regression	Extraction by Iterated Principal Factors, estimation by Bartlett-method, minimum eigenvalue 1	FarIpfBarM1
Factor Analysis Lasso	Extraction by Principal Factors, λ determined by Lasso's minimum C_p-statistic	FarLasso
Principal Component Regression	Components having minimum eigenvalue 1	PCRM1
Principal Component Regression	OLS-regression – Components having p-values below {0.15, 0.10, 0.05, 0.01}, 5-fold CV	PCRT
Forward Stepwise Regression	Regressors with p-value<0.10	FSTP
Backward Stepwise Regression	Regressors with p-value<0.10	BSTP
Lasso	5-fold CV to determine λ	LassoCV
Partial Least Squares	5-fold CV to determine k	PLSCV
Principal Covariates Regression	Sequential procedure with component cross-validation to determine α and γ	PCovR
Adaptive Lasso	5-fold CV to determine both λ and ω using the Lasso to determine initial weights.	AdaLasso
Factor Analysis Adaptive Lasso	5-fold CV to determine both λ and ω using the Lasso to determine initial weights.	FarAdaLasso
Elastic Net	Grid search (density 0.1) with 3-fold CV to determine both λ and α	Elastic Net

5.6.1 Latent variable model

Basic scenario

For the data-generating process under the parameters of the basic scenario, the results presented in figure 5.1 and table 5.5 emerge for the training data. Confirming the theoretical expectations, OLS as the BLUE estimator has, on

average, both the lowest mean squared error and variance. Such an in-sample performance may be an indication of overfitting the data, though. Whether this hypothesis is correct can be scrutinized out-of-sample.
With regard to outliers, only Partial Least Squares is concerned. The results table discloses the existence of outliers, as the mean and the variance lie substantially above the values of other methods. This observation also holds for the remaining scenarios, where the deviations are often even bigger. When outlier-robust measures are considered instead, i.e. the median and mean absolute deviation, PLSCV performs similarly to OLS and FarIpfBar. Ignoring the performance outliers, similar results of the three methods can be detected throughout the examined scenarios in both specifications. For that reason, PLSCV does not receive special attention in this simulation study. As PLSCV is the only method with occasional outlier problems, the question of its cause arises – especially in light of an otherwise undisturbed performance in most of the replications. Scrutinizing the details exceeds the scope of this work but two explanations are conceivable. One is that PLSCV may be particularly sensitive to unfortunate cross-validation splits. Another may lie in occasional computational difficulties of the cross-product. An unfortunate consequence of bad in-sample performance is that out-of-sample predictions for the dependent variable will also be poor in the concerned replications.
With regards to the other methods, the examination continues with the ones related to Factor Regression. Those with unsupervised factor selection criteria (FarPfBarM1, FarIpfBarM1) perform worse both in regard to error and variance than those relying on Factor Analysis's inherent dimensional reduction (FarPfBar, FarIpfBar). In particular, using Iterated Principal Factors as the extraction method leads to almost the same performance as OLS, indicating that hardly any dimensional reduction took place. For the methods applying the Kaiser-criterion (FarPfBarM1, FarIpfBarM1), it is remarkable that the performance differences to OLS are not larger. After all, an eigenvalue of 1 is a data-independent choice and hence a somewhat arbitrary threshold. The most likely explanation for the performance at this stage is the particular DGP which creates a data structure that contains a lot of informative variance on few factors. This selection criterion is benefited by such a data structure.

The results of the FarLasso lie between the two aforementioned groups of Factor Analysis methods. Since its set of regressors is the same as for FarPfBar, and regularization leads to a higher mean squared error in-sample, FarLasso cannot perform better than FarPfBar on the in-sample dataset. The same notion holds for AdaFarLasso, which performs similarly, but slightly more variable compared to the two Kaiser-criterion based factor regression methods, FarPfBarM1 and FarIpfBarM1. Together with the two PCR-methods, in which selection by p-values

Figure 5.1: Results basic scenario (LVM): In-sample.

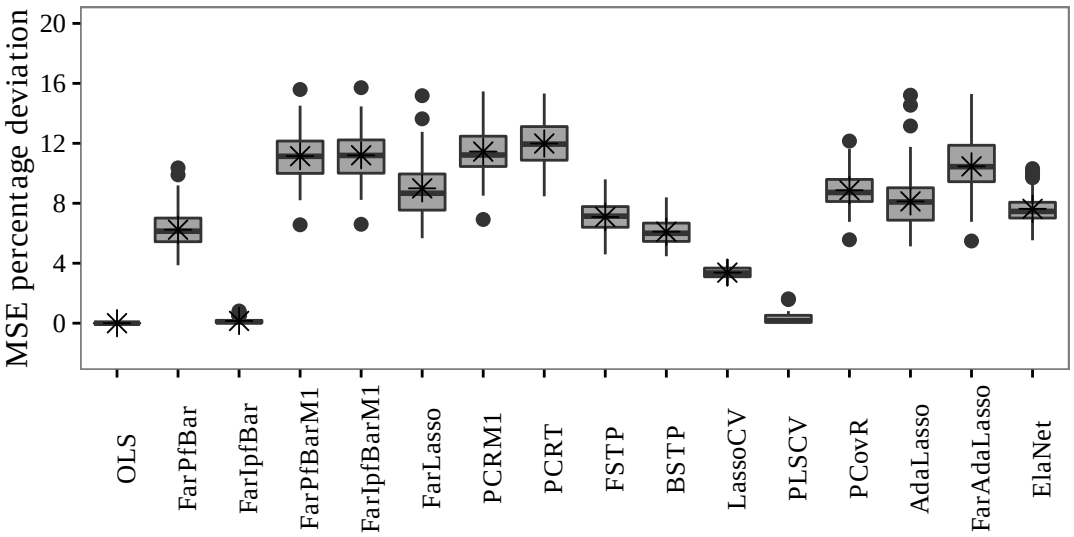


Table 5.5: Simulation results: Basic scenario in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.06	1	1.11	1.11	1.09	1.11	1.12	1.07	1.06	1.03	4.58	1.09	1.08	1.1	1.08
Median	1	1.06	1	1.11	1.11	1.09	1.11	1.12	1.07	1.06	1.03	1	1.09	1.08	1.1	1.07
STD*100	0	1.22	0.17	1.57	1.59	1.95	1.55	1.61	1.09	0.88	0.41	>9	1.15	1.83	1.81	1
MAD*100	0	1.28	0.11	1.57	1.62	1.8	1.4	1.64	1.06	0.84	0.39	0.23	1.06	1.61	1.66	0.74

is less accurate than factor selection by eigenvalue size, the five methods have the highest MSE-deviation relative to OLS (disregarding the outlier affected results for PLSCV).

A better performance is delivered by methods grounded on stepwise variable selection. They yield a relatively stable performance deviation and are practically identical from each other in terms of MSE deviation. Backward stepwise selection is, however, less variable in its performance. The observation of a mediocre performance of those methods in-sample is stable as it recurs in the other scenarios of this specification.

Amongst the remaining methods, LassoCV protrudes with an outstanding performance in terms of MSE and variance, while the Elastic Net and Principal Covariates Regression are on about the same level as FarLasso and AdaLasso and thereby slightly more accurate than FarAdaLasso. Through its comparably low volatility both in terms of variance and mean absolute deviation, the Elastic Net additionally shows a stable performance.

The summary for the relative error in predicting the dependent variable out-of-

sample, wrapped up in figure 5.2 and table 5.6, highlights reverse patterns to the in-sample results: Methods performing well on training data show signs of overfitting, whereas sparser models predict unknown data relatively well. The observation of overfitting concerns especially OLS, FarIpfBar, PLSCV and the two stepwise selection algorithms. These patterns recur in many scenarios.
Beside these general results, there are some noteworthy details about the other procedures. These concern the methods based on Factor Analysis with Kaiser-criterion selection and the Elastic Net and PCovR, which all perform well out-of-sample. Moreover, FarLasso confirms its hypothesized sparsity property and ranks shortly behind FarPfBarM1 and FarIPfBarM1. Those methods are only outperformed by the Elastic Net and FarAdaLasso, whereby the latter also performs best in terms of median, variance and mean absolute deviation. The results of FarPfBar, LassoCV and AdaLasso, on the other hand, show a clear gap in explained variance compared to the methods mentioned before. Together with the results from the in-sample data at hand, this indicates overfitted models. Such results are particularly unexpected for AdaLasso and LassoCV as their regularization feature is supposed to reduce the risk of overfitted models.

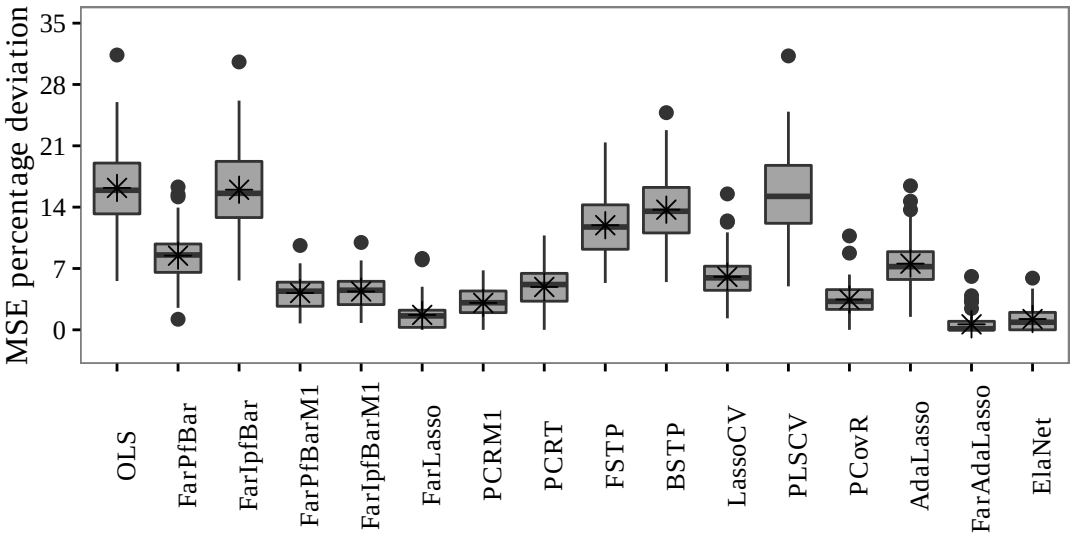
Another notable result is that the unsupervised Factor Analysis based methods (except for FarIpfBar which practically behaves like OLS) perform similarly to methods based on Principal Component Analysis in terms of mean squared error. In particular, factor selection by eigenvalue size is successful in prediction. Therefore, comparing PCR with Factor Regression, the latter exhibits better interpretability with a similar numerical performance in this scenario. This observation leaves little reason to use PCR. PLSCV performs similarly to OLS when outlier-robust measures are considered. It confirms the interpretation concerning overfitting that was derived from the results of the training data set.

Table 5.6: Simulation results: Basic scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.16	1.08	1.16	1.04	1.04	1.02	1.03	1.05	1.12	1.14	1.06	4.9	1.03	1.08	1.01	1.01
Median	1.16	1.09	1.16	1.04	1.05	1.02	1.03	1.05	1.12	1.14	1.06	1.15	1.03	1.07	1	1.01
STD*100	4.49	2.77	4.48	1.73	1.76	1.62	1.66	2.29	3.36	3.79	2.32	>9	1.73	2.65	1.03	1.33
MAD*100	4.55	2.7	4.76	1.85	1.92	1.65	1.88	2.15	3.77	3.84	2.01	4.6	1.72	2.37	0.18	1.28

Since the results of the basic scenario are of major importance for the choice of the final method, an interim conclusion is drawn at this stage. When the results pf both in-sample and out-of-sample data are taken into consideration, five to six considerable candidates are identified. As indicated by the result tables, PLSCV

Figure 5.2: Results basic scenario (LVM): Out-of-sample.



yields good results in-sample when successful but the failures in some instances constitute a severe drawback. To exclude such failures in practice, it would be necessary to compare the PLSCV model's performance with that of a robust method such as OLS. When it works, the method is additionally prone to generate overfitted models, which can be observed in figure 5.2. Neither stepwise methods nor the two Lasso types, Lasso and AdaLasso, yield a convincing performance in this scenario. From the perspective of predictive ability, AdaFarLasso, FarLasso, PCovR and the Elastic Net, but also the two Kaiser-criterion based Factor/Component models make it on the short-list.

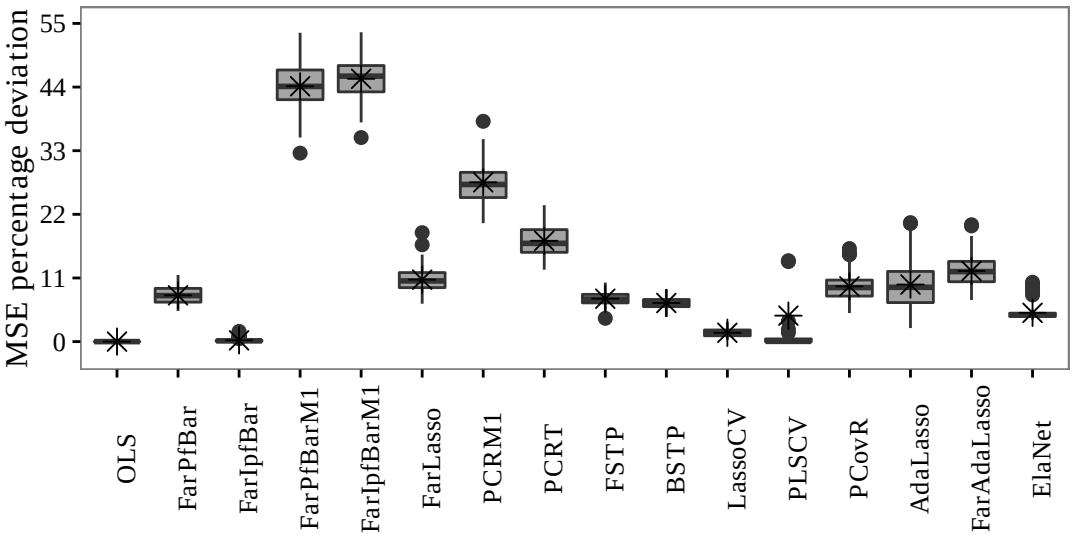
However, it cannot be ensured that the synthetic data generated in this scenario meets the real data in all its facets. For an overall assessment, it is therefore of interest to consider how the methods compare in unwieldy data environments. Bad performance there may further narrow down the candidate list of potential methods.

Dispersed Loadings

The results for the dispersed loadings scenario are presented for in-sample data in figure 5.3 and table 5.7.

Some patterns that could be observed in the basic scenario emerge once again but differences between methods are significantly more distinct in this scenario. Factor regression methods using the Kaiser-Criterion perform especially poorly. These results occur because the specification of the DGP leads to many relevant factors of which some are likely to have an eigenvalue smaller than 1. But these factors

Figure 5.3: Results dispersed loadings (LVM): In-sample.



are excluded despite their importance. The results demonstrate the caveats of a data-independent rule on the number of factors. It does not take a hugely different DGP to end up in this situation. A selection which is based on other criteria, for instance the elbow in a scree plot, could yield better results in this case. A lower MSE is achieved by PCA-based methods, by contrast; but the explanatory power is still considerably worse than for the remaining methods.

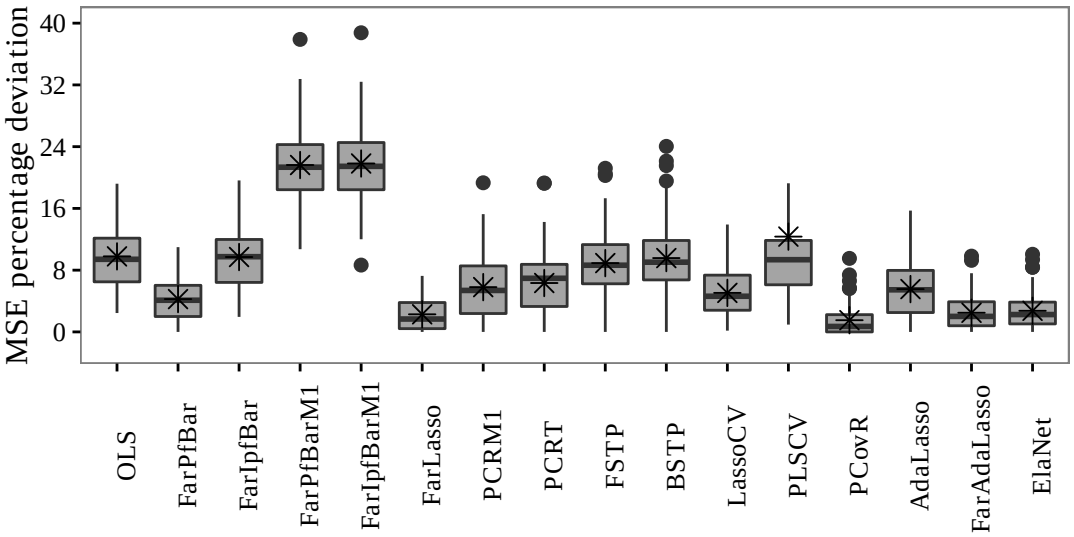
Among the methods with better performances, distinct patterns occur. OLS and FarIpfBar share the top position and LassoCV brings about an only slightly worse performance, both in terms of mean squared error and median. Following are FarPfBar, the Elastic Net and the stepwise methods, compared to which FarLasso, FarAdaLasso, AdaLasso and PCovR yield a less accurate model in-sample in this scenario. Moreover, the Elastic Net produces once more stable results, while the AdaLasso tends to fluctuate strongly.

Table 5.7: Simulation results: Dispersed loadings scenario in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.08	1	1.44	1.45	1.11	1.28	1.17	1.07	1.07	1.02	1.04	1.1	1.1	1.12	1.05
Median	1	1.08	1	1.44	1.46	1.1	1.27	1.17	1.07	1.07	1.01	1	1.09	1.09	1.12	1.05
STD*100	0	1.52	0.29	3.73	3.4	2.26	3.48	2.62	1.07	0.98	0.71	>9	2.31	4.27	2.53	1.31
MAD*100	0	1.77	0.15	3.58	3.18	2	3.36	2.67	1.02	0.94	0.68	0.07	2.16	3.97	2.64	0.43

When it comes to out-of-sample performance (figure 5.4 and table 5.8), it turns out that unlike in the basic scenario, Factor Analysis models based on the Kaiser-criterion selection are often too sparse, so that even OLS accomplishes a better

Figure 5.4: Results dispersed loadings scenario (LVM): Out-of-sample.



performance. They leave the short-list of candidate methods for this reason. The PCA based counterpart neither performs particularly well nor badly. Amongst the best methods are again FarPfBar, PCovR, FarLasso, FarAdaLasso and the Elastic Net. Out the latter set, PCovR exhibits the most stable performance. LassoCV exhibits reasonable results in this sample, given the accurate fit in-sample owing to which one could expect overfitting tendencies as in the basic scenario. In consequence, this method dominates, for instance, the AdaLasso and the stepwise methods.

Table 5.8: Simulation results: Dispersed loadings scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.1	1.04	1.1	1.22	1.22	1.02	1.06	1.06	1.09	1.1	1.05	1.12	1.02	1.06	1.02	1.03
Median	1.09	1.04	1.1	1.21	1.21	1.02	1.05	1.07	1.09	1.09	1.05	1.09	1.01	1.05	1.02	1.02
STD*100	3.89	2.82	4	5.02	4.89	2.01	4.27	4.12	4.42	4.75	3.07	>9	1.99	3.79	2.19	2.32
MAD*100	4.13	3.06	4.62	4.38	4.61	2.47	4.47	4.03	3.86	3.93	3.65	4.22	1.04	4.15	2.17	2.04

Small sample size

The in-sample results for the small sample size scenario are presented in figure 5.5 and table 5.9. Sensitivity to small sample sizes is an important factor in the assessment of a method since some of the analyses in the empirical part of this thesis are characterized by it.

Concerning the training data fit of the methods which deviate substantially from

Figure 5.5: Results small sample scenario (LVM): In-sample.

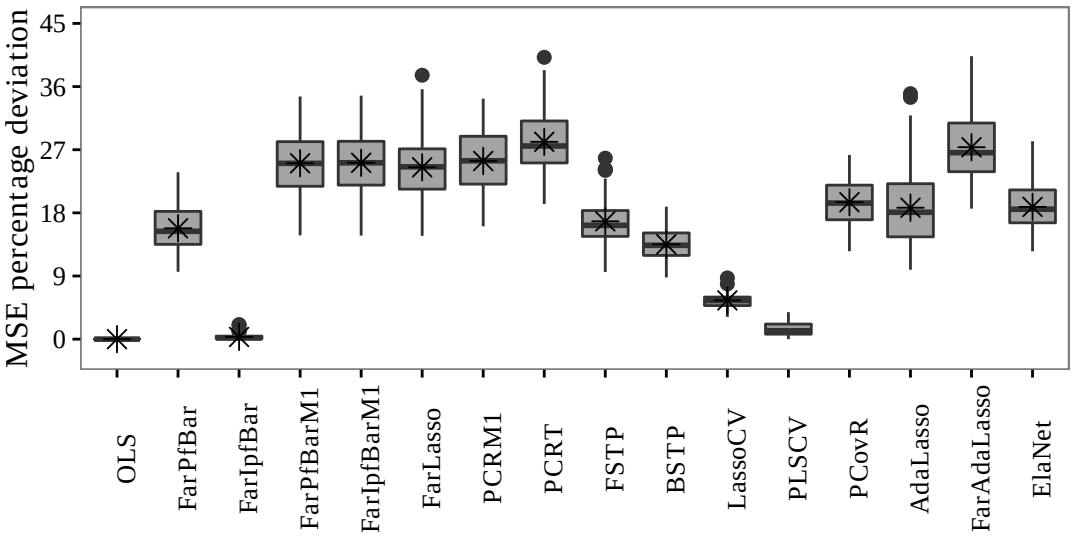


Table 5.9: Simulation results: Small sample scenario in-sample

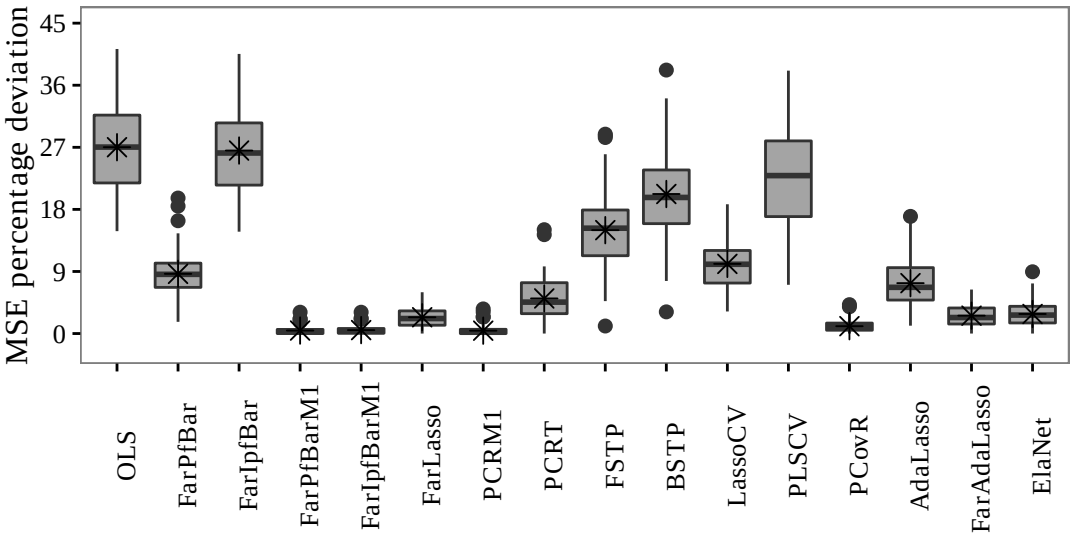
	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.16	1	1.25	1.25	1.24	1.25	1.28	1.17	1.14	1.06	>9	1.2	1.19	1.27	1.19
Median	1	1.15	1	1.25	1.25	1.25	1.25	1.28	1.16	1.13	1.05	1.01	1.19	1.18	1.27	1.19
STD*100	0	3.21	0.39	4.2	4.2	5	4.29	4.35	3.08	2.16	1.05	>9	3.11	5.36	4.53	3.11
MAD*100	0	3.22	0.22	4.68	4.69	4.48	5.13	3.93	2.49	2.32	0.94	0.98	3.58	5.77	4.8	3.52

the OLS performance, LassoCV exhibits the best performance. It outperforms the remaining methods in terms of fit and variance. Stepwise selection methods and FarPfBar follow thereafter. Among the remaining methods, certain performance differences exist, especially the FarAdaLasso and FarLasso are not able to explain as much variance as the Elastic Net, for instance. None of the methods performs outstandingly badly, however. Differences between methods based on Principal Component Analysis and Factor Analysis are also small.

Table 5.10: Simulation results: Small sample scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.27	1.09	1.27	1	1.01	1.02	1	1.05	1.15	1.2	1.1	>9	1.01	1.07	1.03	1.03
Median	1.27	1.09	1.26	1	1	1.02	1	1.05	1.15	1.2	1.1	1.23	1.01	1.07	1.02	1.03
STD*100	6.41	3.24	6.36	0.56	0.59	1.51	0.64	3.02	5.33	5.98	3.59	>9	0.85	3.61	1.5	1.83
MAD*100	7.11	2.7	6.82	0.34	0.43	1.61	0.29	3.29	4.66	5.76	3.38	8.17	0.77	3.78	1.62	1.77

Figure 5.6: Results small sample scenario (LVM): Out-of-sample.



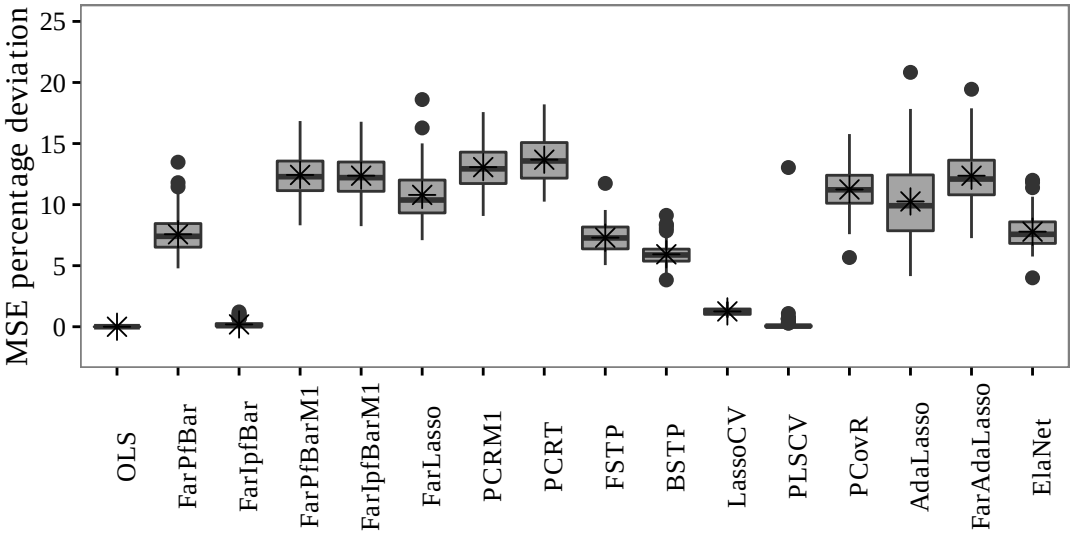
The results for the out-of-sample prediction are shown in figure 5.6 and table 5.10. The general pattern of a negative correlation between in-sample and out-of-sample performance is once more clearly observable. An exception is the Lasso, which is most able to produce a reasonable fit in both samples. Moreover, the results suggest that small sample sizes amplify the differences in predictive quality between the methods. OLS, for instance, performs much worse than in the basic scenario, where its mean percentage deviation was only about 16 % higher than the best performance, while it is about 27 % higher here.
Focusing on the out-of-sample fit only, the Kaiser-criterion methods show a remarkably good performance, irrespective of whether they are based on Factor Analysis or Principal Component Analysis.

Table 5.11: Simulation results: School factor scenario in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.08	1	1.12	1.12	1.11	1.13	1.14	1.07	1.06	1.01	>9	1.11	1.1	1.12	1.08
Median	1	1.07	1	1.12	1.12	1.1	1.13	1.14	1.07	1.06	1.01	1	1.11	1.1	1.12	1.08
STD*100	0	1.58	0.25	1.76	1.75	2.09	1.81	1.89	1.18	0.94	0.3	>9	1.77	3.09	2.26	1.41
MAD*100	0	1.34	0.11	1.79	1.75	1.91	1.85	2.17	1.37	0.74	0.32	0.07	1.72	3.23	2.05	1.22

From a performance point of view, these methods are followed by AdaFarLasso, FarLasso, the Elastic Net and PCovR. As remarked in the analysis of the basic scenario, the particular DGP may benefit Kaiser-criterion selection methods. It can hence be derived that the idea works out with a smaller sample size as well. From this point of view, the results could be interpreted as robustness of these

Figure 5.7: Results school factor scenario (LVM): In-sample.



procedures against small sample sizes. This judgement does not take account of the variance, though. It is considerably lower for methods that apply the Kaiser Criterion for selection. Since a subset of the factors used by FarLasso and FarAdaLasso also occurs in the methods that use the Kaiser-Criterion, the reason for higher variability must be due to the factors unique to the two shrinkage type methods. By definition, these are the factors which have rather low eigenvalues and might, therefore, be prone to instability. This becomes especially apparent in small samples. The reliability in terms of variance of PCoVR lies between methods based on the Kaiser-criterion and FarLasso.

School factor

If the DGP of the basic scenario is modified by having the school factor correlated with the latent factors, the results shown in figure 5.7 and table 5.11 emerge.
The performance in-sample shares the results of the basic scenario in many aspects. This broadly concerns the relative performance of the methods, although the variation is generally somewhat higher in this scenario.
Also the out-of-sample results (displayed in figure 5.8 and table 5.12) are almost the same as in the basic scenario. The only qualitative difference is that well-performing methods are even closer to each other in terms of fit.

Summing up the results for this scenario, it can be concluded that a factor which is non-influential for the observed regressors but correlated to their determinants does not induce any changes of practical relevance in the performance compared to the case of an independent school factor.

Figure 5.8: Results school factor scenario (LVM): Out-of-sample.

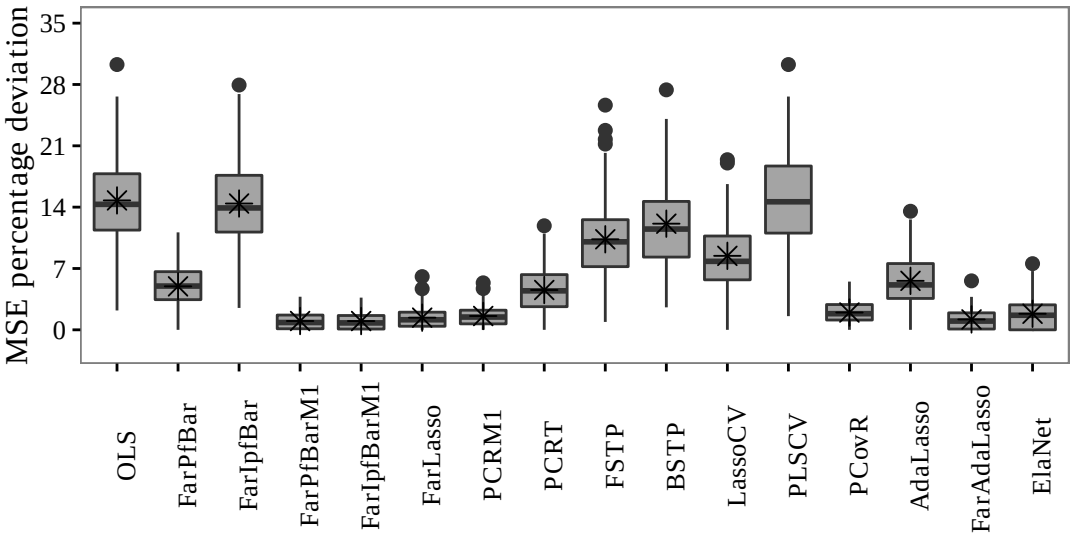


Table 5.12: Simulation results: School factor scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.15	1.05	1.14	1.01	1.01	1.01	1.02	1.05	1.1	1.12	1.08	>9	1.02	1.06	1.01	1.02
Median	1.14	1.05	1.14	1.01	1.01	1.01	1.01	1.04	1.1	1.11	1.08	1.15	1.02	1.05	1.01	1.02
STD*100	4.85	2.5	4.78	0.95	0.95	1.2	1.15	2.49	4.66	4.96	3.66	>9	1.2	3.07	1.12	1.82
MAD*100	4.89	2.43	5.21	1.07	1.12	1.13	1.16	2.71	4.17	4.72	3.88	6.04	1.36	2.6	1.37	2.19

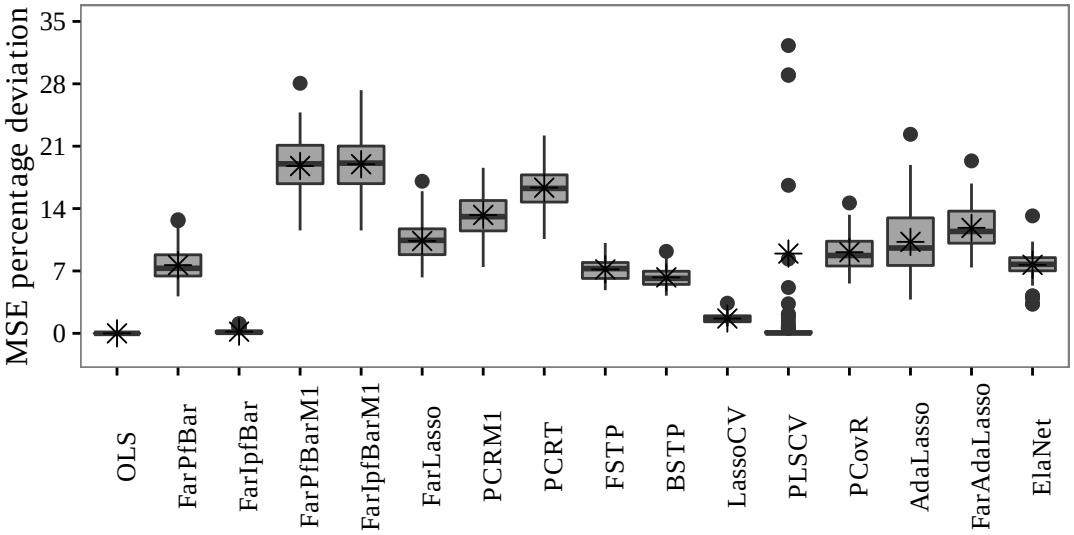
Noisy regressors

Table 5.13: Simulation results: Noise regressors in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.08	1	1.19	1.19	1.1	1.13	1.16	1.07	1.06	1.02	1.09	1.09	1.1	1.12	1.08
Median	1	1.07	1	1.19	1.19	1.1	1.13	1.16	1.07	1.06	1.02	1	1.09	1.1	1.11	1.08
STD*100	0	1.81	0.24	3.1	3.32	2.12	2.4	2.16	1.22	1.08	0.49	>9	2.05	3.55	2.33	1.43
MAD*100	0	1.67	0.15	3.25	3.31	2.06	2.39	2.27	1.27	1.03	0.46	0.05	1.93	3.72	2.32	1.11

The in-sample results for the noisy regressors scenario are presented in figure 5.9 and table 5.13. This scenario is special in that it contains regressors which are correlated to some of the other regressors but, at most, only mildly correlated to the outcome. When compared to the results of the basic scenario, differences in performance are somewhat more pronounced in this scenario. The relative ranking of methods remains for the most part unchanged, the few exceptions are PCRM1

Figure 5.9: Results noisy regressors scenario (LVM): In-sample.



and PCovR which show an improved performance. Again, the AdaLasso shows a significant amount of variance in performance, while LassoCV, for instance, is relatively stable in its performance gap to OLS.

The results for out-of-sample prediction are shown in figure 5.10 and table 5.14. Also here, similarities to the basic scenario are apparent, but the methods' results show less differentiation. In particular the performances of OLS and FarIpfBar are closer to the top methods. Further differences concern PCovR, which outperforms the other methods in terms of fit and variance and the Elastic Net, which performs worse than Factor Analysis Lasso type methods.

Also in this scenario, it is remarkable that LassoCV performs reasonably well despite its good fit in-sample, since there seems to be a notable trade-off for the other methods. Methods based on Factor Analysis and the Kaiser Criterion which performed best in the basic scenario keep a solid performance concerning the mean – their variance increases substantially, however. This suggests that in some of the replications noisy regressors must disturb the creation of factors so strongly that less important factors are driven to eigenvalues larger than 1.

Nevertheless, the results are similar to the basic scenario. There are two possible reasons, which can both contribute to this observation: Either noisy regressors do not confuse the methods, or the degree of disturbance was not sufficient to change the results significantly.

Figure 5.10: Results noisy regressors scenario (LVM): Out-of-sample.

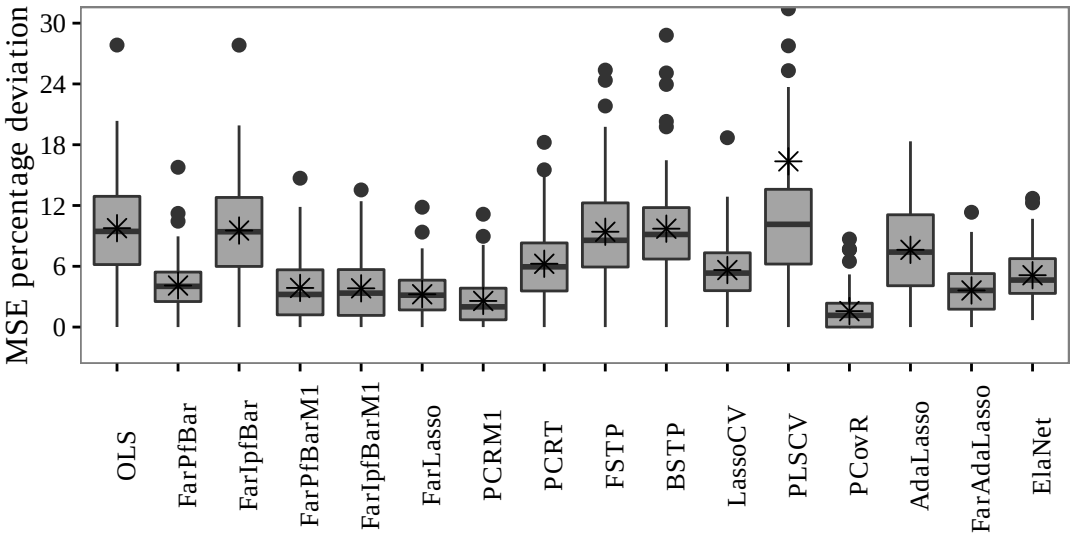


Table 5.14: Simulation results: Noise regressors out-of-sample

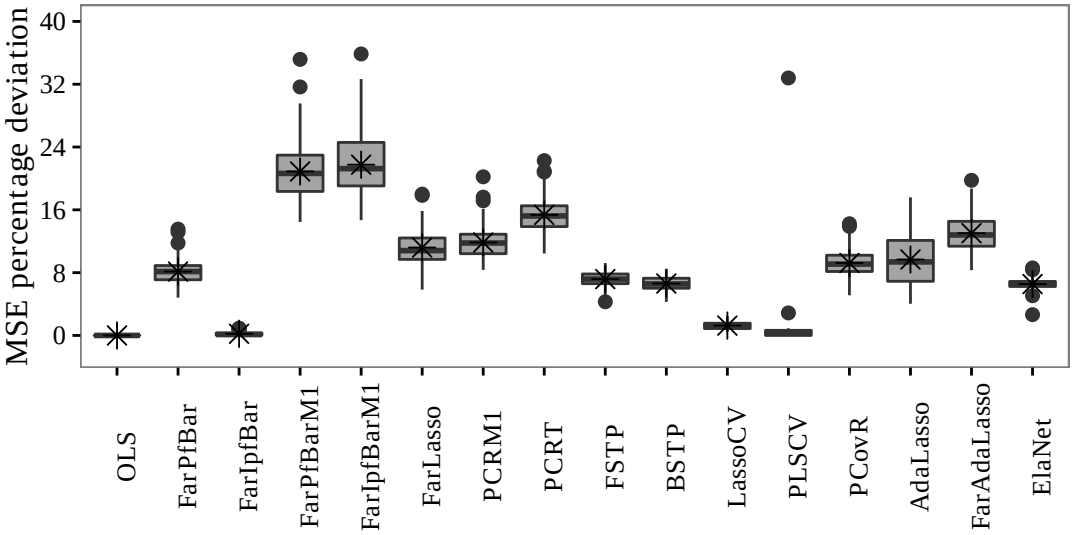
	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.1	1.04	1.1	1.04	1.04	1.03	1.03	1.06	1.09	1.1	1.06	1.16	1.02	1.08	1.04	1.05
Median	1.09	1.04	1.09	1.03	1.03	1.03	1.02	1.06	1.09	1.09	1.05	1.1	1.01	1.07	1.04	1.05
STD*100	4.8	2.72	4.79	3.07	3.16	2.21	2.4	4.02	4.84	4.85	3.23	>9	1.91	4.35	2.35	2.55
MAD*100	5.12	2.17	5.08	3.38	3.28	2.21	2.24	3.51	4.66	3.85	2.95	5.73	1.72	5.12	2.66	2.72

High Uniqueness

The results for the high uniqueness scenario are shown in figure 5.11 and table 5.15 for the training data set. The DGP leads to a decrease in the correlation of the observed variables, they become more specific and less driven by factors. With such a data structure at hand, one may be inclined to predict poor performances of methods which are based on Factor Analysis. The argument being that factors are based on common correlations which, in this scenario, are set low. On the other hand, low correlations may suffice to lead to factors. A performance prediction for this family of methods is thereby rendered difficult.

What can be observed in the results for in-sample data is that Factor Analysis methods with the Kaiser-criterion perform worse than methods based on PCA. As indicated by the large bars in the boxplot, there is also large variability. Decreasing the degree of correlation between variables seems to reduce the probability that the first k factors with eigenvalues larger than 1 cover the important factors. Principal Component Analysis, on the other hand, appears to work much better here. The

Figure 5.11: Results high uniqueness scenario (LVM): In-sample.



reason for this lies in the variance on which the spectral decomposition in both procedures is based. For Factor Analysis, in particular the Principal Factors extraction method, the variables' unique variances are removed prior to conducting the spectral decomposition, so that only the common variance for each variable remains. Owing to the decreased influences of factors, the communality is notably lower in this scenario. However, the communality is what is distributed over the factors as loadings. The consequence is that the loadings are much smaller than in the basic scenario. Since the sum of squared loadings on a factor determines its eigenvalue, fewer factors, ceteris paribus, exhibit an eigenvalue larger than 1. High uniqueness does not hurt Factor Analysis methods in general, however: This is demonstrated by the better performance of the other variants. Because they regard factors with smaller eigenvalues, the informative variance is still accessible.

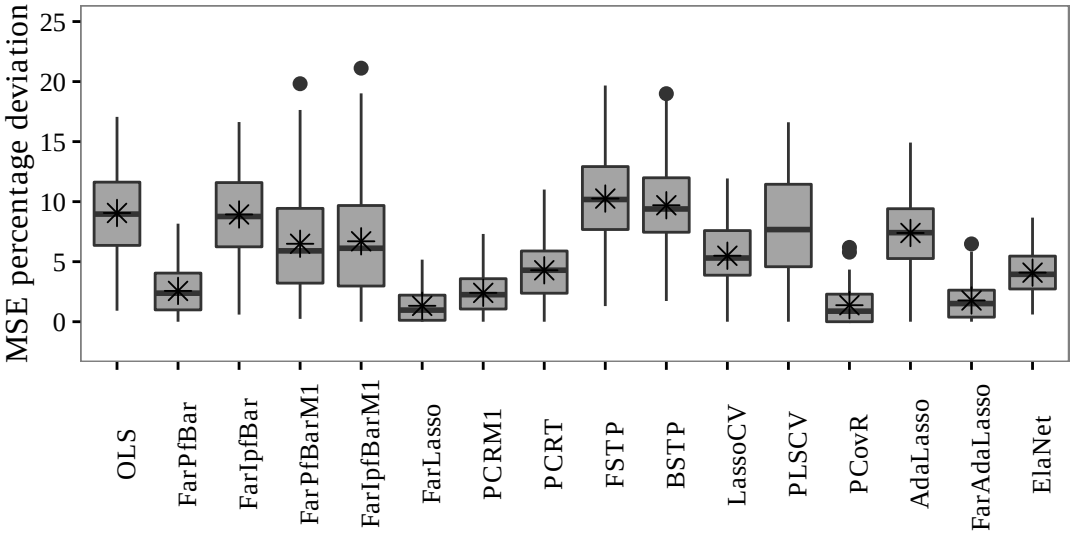
For the methods that deserve a closer look based on the insights of the previous scenarios, the FarLasso and FarAdaLasso deliver slightly worse performances, while the Elastic Net, LassoCV and FarPfBar deal well with the data.

The relatively good performance of the Elastic Net holds also out-of-sample, but it is beaten by the two Factor Analysis Lasso methods and PCovR in terms of fit. The results for the out-of-sample prediction are shown in figure 5.12 and table 5.16. The worst performance, even worse than OLS, is delivered by the stepwise methods. They do not seem to be the right choice for model selection in scenarios with clusters of lower correlations between variables.

Table 5.15: Simulation results: High uniqueness in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.08	1	1.21	1.22	1.11	1.12	1.15	1.07	1.07	1.01	>9	1.09	1.1	1.13	1.07
Median	1	1.08	1	1.21	1.21	1.11	1.12	1.15	1.07	1.07	1.01	1	1.09	1.09	1.13	1.07
STD*100	0	1.54	0.24	3.76	4.07	2.32	1.98	2.27	0.98	0.87	0.45	>9	1.59	3.31	2.31	0.7
MAD*100	0	1.36	0.16	3.49	3.78	2.01	1.92	2.06	0.91	0.94	0.5	0.38	1.56	3.92	2.23	0.5

Figure 5.12: Results high uniqueness scenario (LVM): Out-of-sample.



5.6.2 Regression model

In the regression model simulation, the dependent variable is not determined by the factors but by the observed variables.

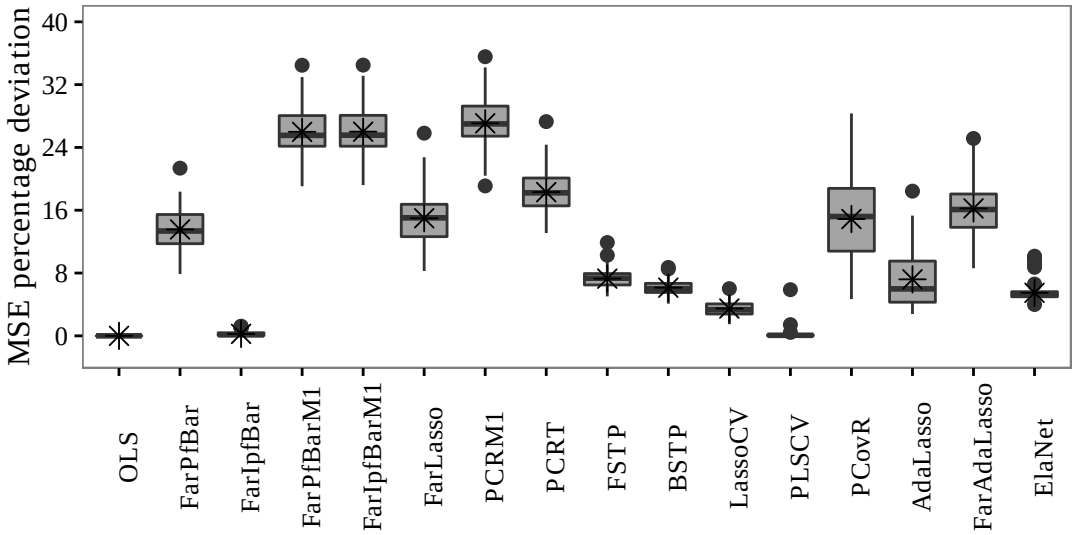
Basic Scenario

The results for the in-sample data performance of the methods in the basic scenario are shown in figure 5.13 and table 5.17. Comparing the results with the correspondence in the latent variable model framework, differences between the methods are more pronounced. Moreover, there is a general tendency of worse performance by methods using factors for regression. With the exception of FarIpfBar, which also performs like OLS in this specification, these models are outperformed by each of the variable based methods. Differences amongst the latter are generally rather small, although LassoCV performs somewhat better than the rest. In accordance with the latent variable model, AdaLasso exhibits a

Table 5.16: Simulation results: High uniqueness out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.09	1.03	1.09	1.06	1.07	1.01	1.02	1.04	1.1	1.1	1.05	>9	1.01	1.07	1.02	1.04
Median	1.09	1.02	1.09	1.06	1.06	1.01	1.02	1.04	1.1	1.09	1.05	1.08	1.01	1.07	1.02	1.04
STD*100	3.56	1.96	3.6	4.28	4.6	1.31	1.63	2.46	3.62	3.5	2.63	>9	1.49	3.06	1.53	1.97
MAD*100	3.97	2.42	4.08	4.48	4.94	1.44	1.91	2.6	4	3.48	2.69	5.2	1.3	3.02	1.69	2.16

Figure 5.13: Results basic scenario (RM): In-sample.



high variability in this specification. From this point of view, the performance is worse than BSTP.

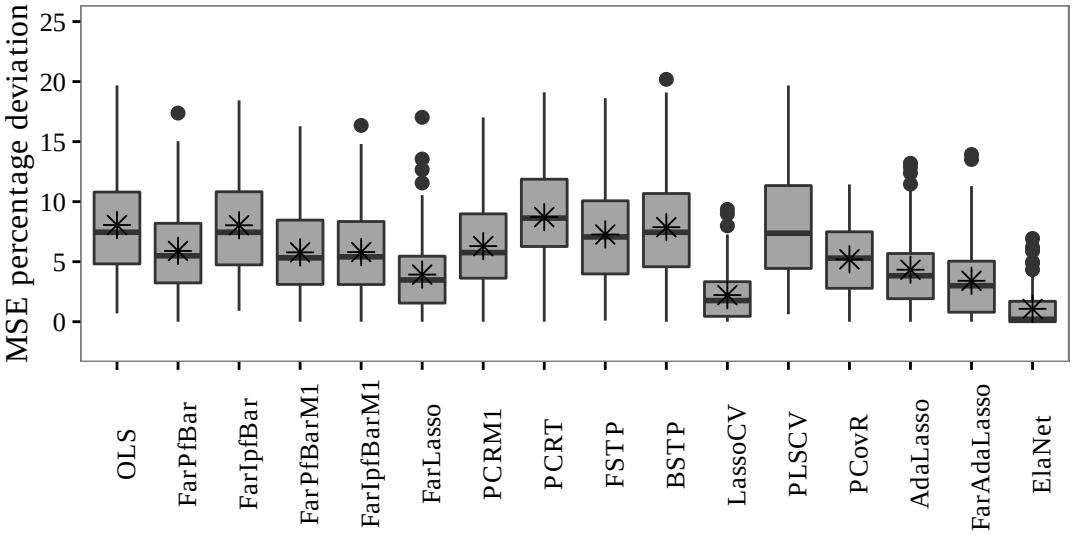
What also stands out in comparison to the results in the latent variable model framework, is the relatively high variability of PCovR. In the prior simulations, this method did not exhibit such a property. As can be observed in the following scenarios, the method's increased variability is a property for in-sample data in this simulation model. Among the factor regression models, the performance is worst when factor selection relies on the Kaiser-criterion. In particular, there is a significant gap in performance between PCRM1 and PCRT which, in the latent variable model specification, only appears in the in-sample results of the dispersed loadings scenario. With respect to the remaining factor regression models, the differences are small, merely FarPfBar performs somewhat better than the rest.

Figure 5.14 and table 5.18 summarize the results for the methods' performances on out-of-sample data in the basic scenario. They suggest a decrease of differences in prediction in comparison to in-sample results. Some observations justify an additional note, though. The Lasso and the Elastic Net perform best on these data,

Table 5.17: Simulation results: Basic scenario in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.14	1	1.26	1.26	1.15	1.27	1.18	1.07	1.06	1.03	>9	1.15	1.07	1.16	1.05
Median	1	1.13	1	1.26	1.26	1.15	1.27	1.18	1.07	1.06	1.03	1	1.15	1.06	1.16	1.05
STD*100	0	2.59	0.3	3.14	3.14	3.05	3.2	2.65	1.21	1	0.95	>9	5.45	3.53	3.22	1.17
MAD*100	0	2.75	0.17	3.3	3.24	3.08	3.17	2.68	1.11	0.91	0.92	0.09	6.3	3.32	3.33	0.48

Figure 5.14: Results basic scenario (RM): Out-of-sample.



despite their good performance in-sample. It leads to the observation that they outperform stepwise selection methods in both samples. The negative correlation between in-sample and out-of-sample fit hence seems especially weak for those methods. One reason for the particularly good performance of the Elastic Net could be found in the data structure, which contains clusters of moderately and strongly correlated variables.

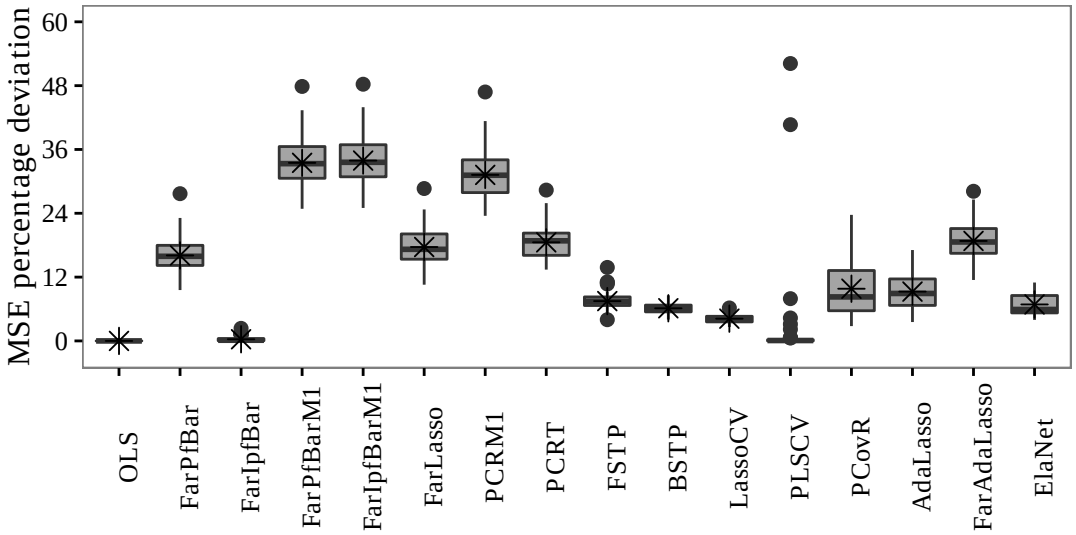
Factor Regression models, even when they are based on the Kaiser criterion, are surprisingly able to predict the outcome well. In fact, the two models using regularization techniques perform only mildly worse than the Lasso or the Elastic net.

Yet in sum, the model underlying this simulation design appears to grant advantages to variable-shrinkage methods. Index-building methods do not perform much worse on test data but they ceded parts of their success in contrast to the results from the latent variable model simulations by showing a significantly worse performance in-sample.

Table 5.18: Simulation results: Basic scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.08	1.06	1.08	1.06	1.06	1.04	1.06	1.09	1.07	1.08	1.02	>9	1.05	1.04	1.03	1.01
Median	1.07	1.05	1.07	1.05	1.05	1.03	1.06	1.09	1.07	1.07	1.02	1.07	1.05	1.04	1.03	1
STD*100	4.09	3.69	4.02	3.59	3.6	3.19	3.76	4.3	4.23	4.38	2.2	>9	3.03	3.22	2.97	1.61
MAD*100	4.36	3.81	4.3	3.72	3.89	2.95	4.27	4.36	4.58	4.4	2.17	4.71	3.32	2.84	3.23	0.3

Figure 5.15: Results dispersed loadings (RM): In-sample.



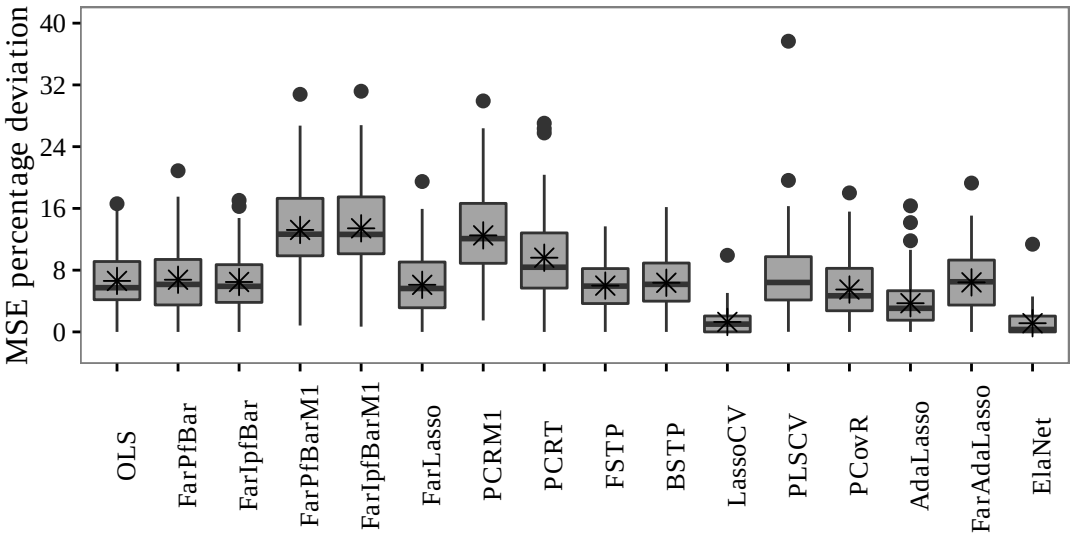
Dispersed loadings

Figure 5.15 and table 5.19 wrap up the in-sample results for the dispersed loadings scenario. Compared to the basic scenario, there are hardly any qualitative changes, apart from generally increased level variances and performance differences. This stands in contrast to the changes which emerged in the latent variable model framework and required additional inspection and interpretation.

Table 5.19: Simulation results: Dispersed loadings scenario in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.16	1	1.33	1.34	1.18	1.31	1.19	1.07	1.06	1.04	4.81	1.1	1.09	1.19	1.07
Median	1	1.16	1	1.33	1.34	1.17	1.31	1.19	1.07	1.06	1.04	1	1.08	1.09	1.19	1.06
STD*100	0	3.09	0.43	4.06	4.1	3.32	3.96	2.82	1.43	0.88	0.81	>9	5.26	3.2	3.42	2.02
MAD*100	0	2.91	0.17	4.39	4.24	3.01	4.43	3.18	1.19	0.95	0.83	0.11	4.8	3.6	3.62	1.37

Figure 5.16: Results dispersed loadings (RM): Out-of-sample.



One observation that deserves attention is the performance of PCovR. In spite of its high volatility, it performs much better than the remaining methods that form indices. Among the methods that do not use derived inputs as regressors, the method nevertheless ranks at the end in terms of the mean.

Table 5.20: Simulation results: Dispersed loadings scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.07	1.07	1.06	1.13	1.13	1.06	1.13	1.1	1.06	1.06	1.01	3.29	1.05	1.04	1.06	1.01
Median	1.06	1.06	1.06	1.13	1.13	1.06	1.12	1.08	1.06	1.06	1.01	1.06	1.05	1.03	1.06	1
STD*100	3.72	4.22	3.74	5.84	5.87	3.89	5.79	5.43	3.3	3.4	1.51	>9	3.95	3.2	3.85	1.7
MAD*100	3.53	4.31	3.77	6.09	6.19	4.07	5.83	4.6	3.37	3.98	1.5	4.38	3.54	2.94	4.29	0.47

The results out-of-sample (shown in figure 5.16 and table 5.20) confirm the ability of the Lasso and, in particular, the Elastic Net to find a suitable model under this specification. As in the basic scenario the performance differences have become smaller, but it meets the eye that the Kaiser Criterion for variable selection tends to generate inaccurate predictions. This inaccuracy is, however, not as extreme as it was in the latent variable model for this scenario.

In sum, the DGP leads to stronger differences between the methods' performances in the latent variable specification than under the regression model specification.

Figure 5.17: Results small sample scenario (RM): In-sample.

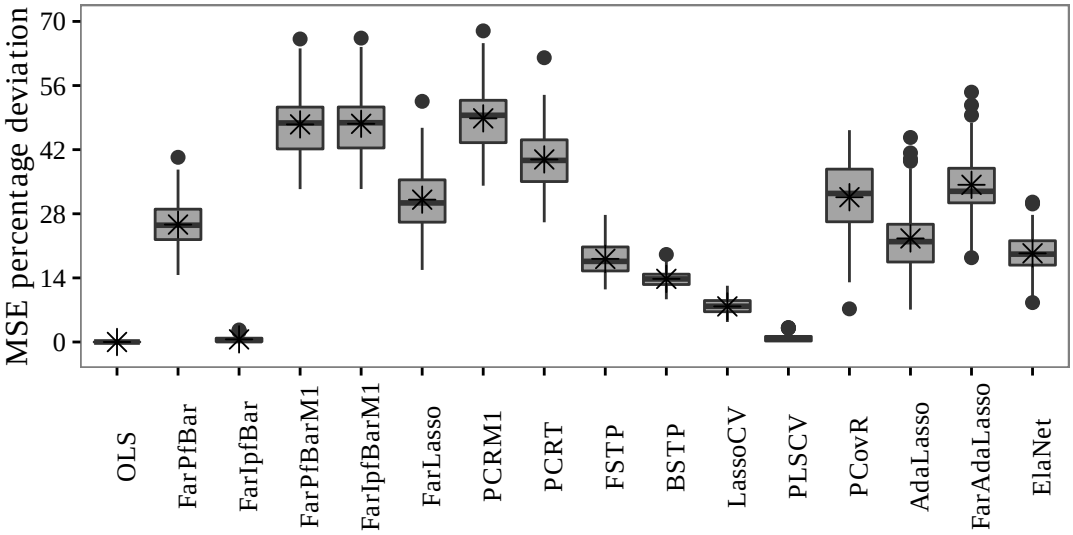


Table 5.21: Simulation results: Small sample size scenario in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.26	1.01	1.48	1.48	1.31	1.49	1.4	1.18	1.14	1.08	>9	1.32	1.23	1.34	1.19
Median	1	1.26	1	1.48	1.48	1.3	1.5	1.4	1.18	1.14	1.08	1.01	1.32	1.22	1.33	1.19
STD*100	0	5.12	0.63	6.82	6.84	6.71	6.89	6.71	3.66	1.95	1.67	>9	7.87	7.19	6.56	4.09
MAD*100	0	4.95	0.45	6.88	7.14	7.08	7.68	6.76	3.76	1.77	1.81	0.71	8.3	6.24	5.42	3.75

Small sample size

When the sample size is reduced substantially, the results shown in table 5.21 and figure 5.17 emerge. It can be observed that also in this scenario, the results exhibit strong similarities to the basic scenario. Although the differences between the methods are markedly more distinct and the variance on average higher, the relative ranking is almost unchanged.

When it comes to the results of the prediction of out-of-sample data(displayed in figure 5.18 and table 5.22), the performance differences are also more pronounced. Methods prone to overfitting, but also PCRT, are not able to keep the gap as small as in the basic scenario.
Despite their poor predictive power in-sample (compared to the Lasso and Elastic Net), methods such as the FarLasso, PCovR and the FarAdaLasso achieve a reasonable accuracy out-of-sample, and rank only slightly behind the variable shrinkage techniques. Their variance is, however, significantly larger.

Figure 5.18: Results small sample scenario (RM): Out-of-sample.

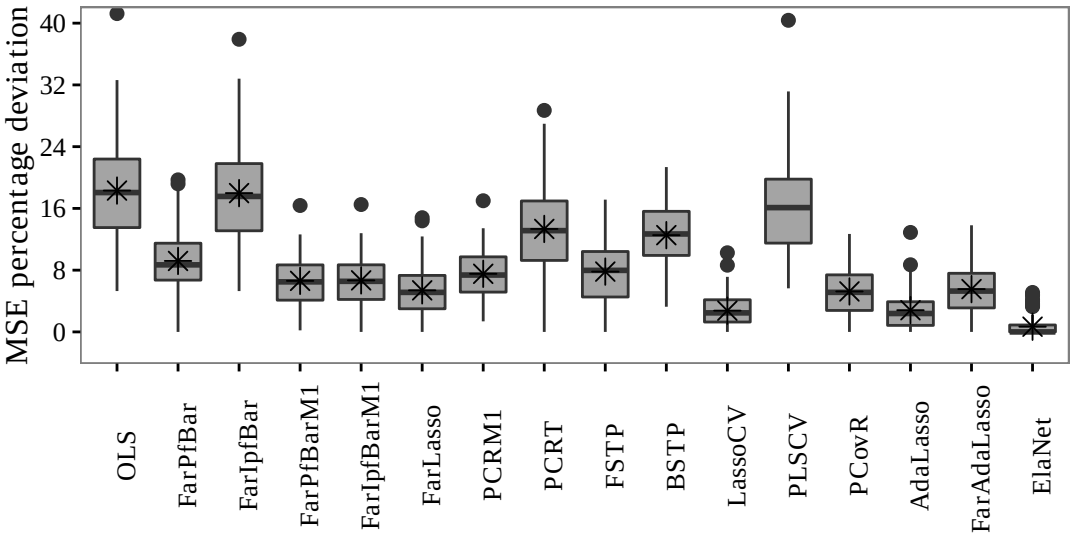


Table 5.22: Simulation results: Small sample size scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.18	1.09	1.18	1.07	1.07	1.05	1.08	1.13	1.08	1.13	1.03	>9	1.05	1.03	1.06	1.01
Median	1.18	1.09	1.18	1.06	1.07	1.05	1.07	1.13	1.08	1.13	1.02	1.16	1.05	1.02	1.05	1
STD*100	6.44	3.86	6.14	3.06	3.08	3.01	3.09	5.46	4.16	4.3	2.15	>9	3.07	2.4	3.02	1.24
MAD*100	6.72	3.32	6.64	3.4	3.32	3.17	3.44	5.77	4.51	4.26	2.26	6.42	3.52	2.32	3.42	0

School factor

The results emerging under the school factor scenario are almost identical to the ones under the basic scenario. Their presentation is hence omitted.

Noisy regressors

Figure 5.19 and table 5.23 show the results for the noisy regressors scenario. In terms of training data, they qualitatively display similar patterns as observed in the basic scenario. Noticeable differences to the basic scneario concern PCovR, whose performance is less variable in this scenario and more similar to the one of AdaLasso. A second interesting observation is found for FarIpfBar: While the method usually performs like OLS, the results indicate deviations from this behavior in some of the replications, leading to an unusually high variance. Also the Elastic Net is substantially more variant under this scenario, whereas the Lasso shows a robust performance.

Figure 5.19: Results noisy regressors scenario (RM): In-sample.

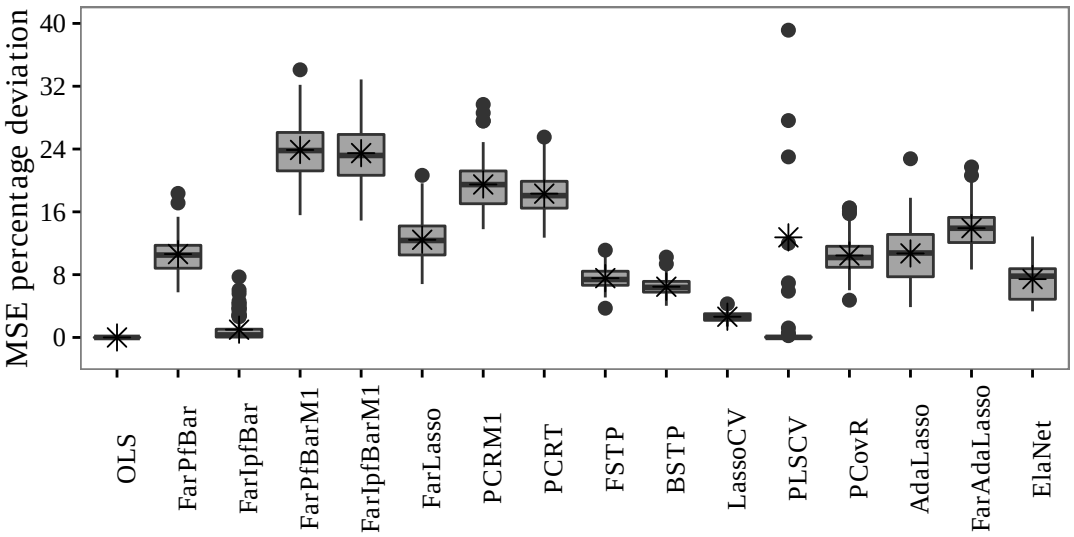


Table 5.23: Simulation results: Noisy regressors scenario in-sample

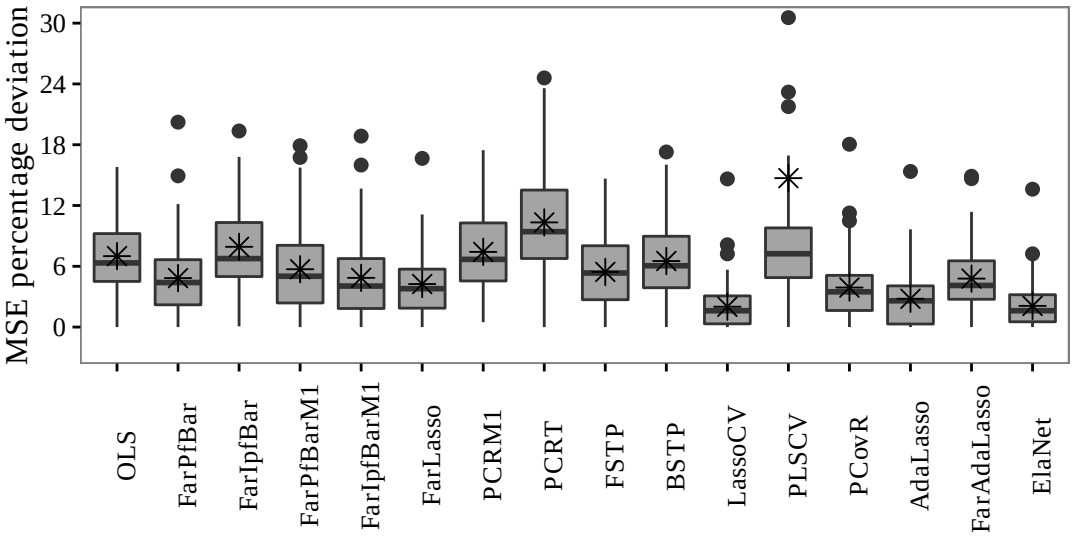
	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.11	1.01	1.24	1.23	1.12	1.19	1.18	1.08	1.06	1.03	1.13	1.1	1.11	1.14	1.07
Median	1	1.1	1	1.24	1.23	1.12	1.19	1.18	1.07	1.06	1.03	1	1.1	1.11	1.14	1.08
STD*100	0	2.31	1.54	4.01	3.9	2.72	3.3	2.56	1.4	1.02	0.63	>9	2.34	3.66	2.58	2.31
MAD*100	0	2.19	0.45	3.71	3.89	2.83	3.35	2.54	1.42	0.95	0.64	0.02	1.95	4.09	2.26	1.85

The results out-of-sample are displayed in figure 5.20 and table 5.24. The good predictive capability of shrinkage methods, especially the Lasso and Elastic Net, becomes markedly evident once more. The two methods perform best in all four performance measures. However, the performance improvement is small in comparison to the other methods. One exception is PCR with component selection by t-values, which exhibits a large variance and a relatively large performance gap. Since the median and the absolute deviation measures are more favorable, it seems as if this method is affected by bad models in single replications.

Table 5.24: Simulation results: Noisy regressors scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.07	1.05	1.08	1.06	1.05	1.04	1.07	1.1	1.05	1.07	1.02	1.15	1.04	1.03	1.05	1.02
Median	1.06	1.04	1.07	1.05	1.04	1.04	1.07	1.09	1.05	1.06	1.02	1.07	1.03	1.03	1.04	1.02
STD*100	4.33	3.43	4.69	4.15	4.09	3.15	3.86	5.54	3.38	3.58	2.17	>9	3.07	2.64	3.08	2.14
MAD*100	3.62	3.34	3.6	4.18	3.61	2.87	4.19	4.5	3.96	3.63	2.05	3.77	2.73	2.7	2.7	1.86

Figure 5.20: Results noisy regressors scenario (RM): Out-of-sample.



High uniqueness

Table 5.25: Simulation results: High uniqueness scenario in-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1	1.23	1	1.78	1.77	1.24	1.36	1.19	1.07	1.06	1.05	>9	1.06	1.07	1.25	1.05
Median	1	1.22	1	1.77	1.76	1.23	1.36	1.18	1.07	1.06	1.05	1	1.06	1.06	1.24	1.04
STD*100	0	4.38	0.44	6.83	6.77	4.66	5.61	3.04	1.17	1.07	0.89	>9	2.31	2.73	4.65	1.27
MAD*100	0	3.97	0.3	6.08	6.3	4.36	5.9	2.69	1.13	1.02	0.92	0.02	2.43	2.54	4.39	0.45

The results in the high uniqueness scenario are depicted for training data in figure 5.21 and table 5.25. Clearly, differences in performance are more pronounced than in the basic scenario. Particularly far off are the two Factor Regression methods which base the selection of factors on the Kaiser-Criterion. They even perform worse than the outlier-affected results of PLS. Arguments for the observed problems in such a data structure were given in the latent variable model section of this scenario and apply equally well here.

The results of the remaining methods are, on the contrary, less affected by low correlations between the variables since their results are similar to the ones observed in the basic scenario. One exception is PCovR which is among the best methods in this scenario and, from the viewpoint of the mean squared error, almost on the same level as the Elastic Net or the Lasso.

The results out-of-sample (shown in figure 5.22 and table 5.26) emphasize the

Figure 5.21: Results high uniqueness scenario (RM): In-sample.

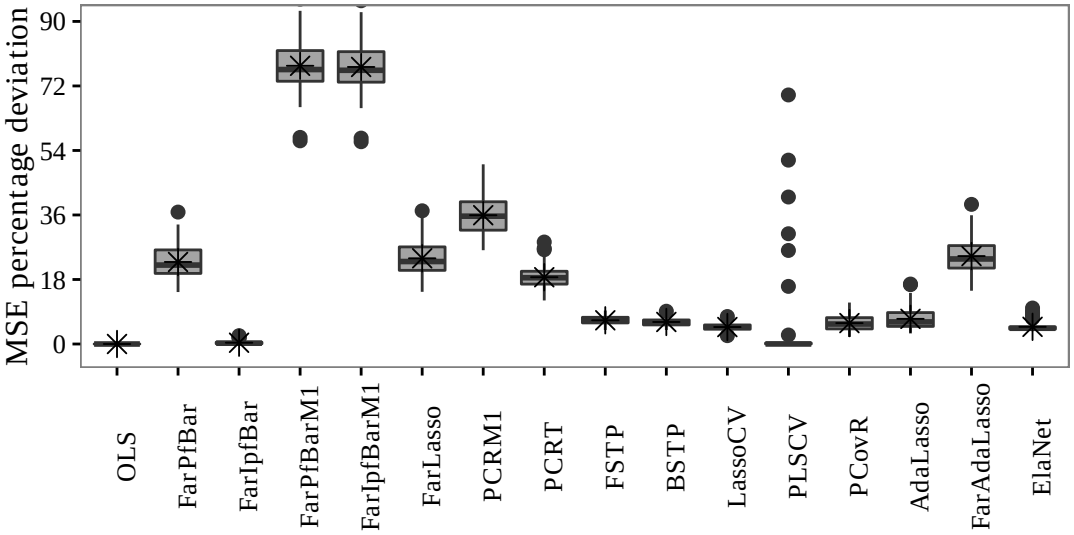
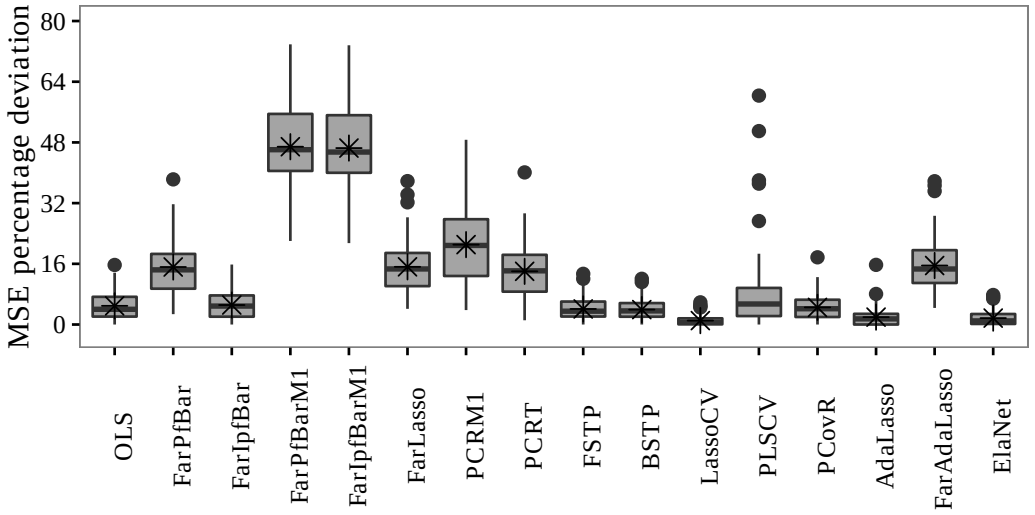


Figure 5.22: Results high uniqueness scenario (RM): Out-of-sample.



problems that can occur when using Factor Regression methods. Except for PCovR and FarIpfBar they even perform worse out-of-sample than OLS. One derivation is that a high share of uniqueness appears to affect Principal Factors as a factor extraction method. However, the PCR methods are also affected and unable to perform better than OLS on average. Of the three variable shrinkage methods, the Lasso performs best both in terms of mean squared error and variance. Also the Adaptive Lasso has a low error, but, similar to the other cohorts, the variation in performance is more volatile.

Table 5.26: Simulation results: High uniqueness scenario out-of-sample

	OLS	FarPfBar	FarIpfBar	FarPfBarM1	FarIpfBarM1	FarLasso	PCRM1	PCRT	FSTP	BSTP	LassoCV	PLSCV	PCovR	AdaLasso	FarAdaLasso	ElaNet
Mean	1.05	1.15	1.05	1.47	1.47	1.15	1.21	1.14	1.04	1.04	1.01	>9	1.05	1.02	1.16	1.02
Median	1.04	1.14	1.05	1.46	1.45	1.15	1.21	1.14	1.04	1.04	1.01	1.05	1.04	1.01	1.15	1.01
STD*100	3.57	6.86	3.57	>9	>9	6.93	>9	6.7	2.91	2.75	1.26	>9	3.35	2.38	7.04	1.88
MAD*100	3.96	6.25	4.18	>9	>9	6.38	>9	7.6	3.06	2.56	0.86	5.25	3.46	2.21	6.2	1.42

5.7 Conclusion

This section sums up the main insights of the simulation study, in which the candidate methods and their model selection features are compared to each other in different scenarios concerning their explanatory/predictive capabilities.

A central result is the performance dependence on the underlying specification. Index building models generally fare better under the latent variable model than under regression model specification. Although there are (scenario-)dependent exceptions and the performance is not substantially worse compared to regression methods, this result indicates that the assumption about the DGP can be crucial.

Results from the latent variable model:
OLS as the baseline approach shows the expected behavior of the best fit in-sample and tendencies of overfitting in out-of-sample data. The latter disadvantage is exacerbated in small training samples. Exceptions occur in the dispersed loadings scenario and the high uniqueness scenario, in which other methods, due to underfitting, perform worse out-of-sample. FarIpfBar always performs similar to OLS, implying that hardly any dimensional reduction takes place when Iterated Principal Factors is used as factor extraction method. Also, the model generated by PLS is close to OLS, when the outliers of this method are ignored, i.e. when the median and the mean absolute deviation instead of the mean and variance are considered. Generally, none of the three methods seem suitable for the pursued goal in this thesis.
FarPfBar shows a different behavior and can be best compared with FarLasso, or FarAdaLasso respectively, as these methods only differ in the additional shrinkage step. In consequence, FarLasso and FarAdaLasso yield a slightly worse in-sample fit but trump FarPfBar out-of-sample instead. Thus, they can be attributed a higher generalizing ability. In general, their performance is sufficiently good and stable across different scenarios, so the two procedures constitute a reasonable compromise between fit and interpretation. Their differences in performance are

usually small. This stands in contrast to the relationship between Lasso and AdaLasso when applied to the original variables: The latter method usually delivers less reliable predictions, while the Lasso is among the top performing methods most of the time. Similarly good results are obtained by PCovR.
Index building methods that select factors based on the Kaiser-criterion perform well if the data structure is one in which the factors or components with the largest eigenvalues are the most relevant ones. When loadings are dispersed and many factors become important, these methods fail. In such cases, selection methods based on the p-value in regression yield a better performance. Moreover, when the index creation is based on Factor Analysis, low correlation between the variables is harmful, as the poor performance in the high uniqueness scenario indicates.
Concerning the stepwise approaches, there are hardly any differences between backward or forward methods. Typically, backward stepwise selection has better fit in-sample, while forward stepwise selection performs better out-of-sample. Both methods perform mediocrely out-of-sample.
Among the regression methods based on Principal Component Analysis, it can be observed that one often fares better using the Kaiser-criterion to select components instead of t-/p-values. This holds for both training and test data. The dispersed loadings scenario is the exception where this observation does not apply.
Within the set of regularization methods, the Elastic net shows a robustly good performance. The Adaptive Lasso does not perform much worse, sometimes even better, but its variance is notably higher. The Lasso tends to perform better in-sample but steps behind the Elastic Net when it comes to prediction. In general, however, both methods deliver a robust and reliable performance.

Results from the regression model:

General performance tendencies in the latent variable model are confirmed in the regression model simulations. However, shrinkage methods, in particular the Lasso and the Elastic Net, are constantly best-performing in this simulation. With regards to out-of-sample performance the two methods share the top position, in spite of the fact that their performance in-sample is quite good. It appears as if the right degree of shrinkage is applied in these methods. Their performance is robust even to small sample conditions.
The general observation that methods based on variables instead of indices fare better in this simulation becomes clear when the out-of-sample performance of forward stepwise selection is compared to the one of FarLasso or FarAdaLasso. While the latter two dominate in the latent variable model specification, forward stepwise selection gives a better fit in a number of scenarios under the regression model. The second index-building method that delivered good results in the latent

variable model, PCovR, can also not maintain the degree of supremacy as it suffers from a highly unstable performance. The high uniqueness scenario constitutes the only exception.

In sum, the results from the regression model do not exhibit significant changes across the scenarios. The latent variable model specification shows more distinctly where the potential weak points of the used methods lie.

The decision for a method based solely on performance would arrive at the Elastic Net or the Lasso. Both methods yield stable and good results across specifications and scenarios. The drawback is the aggravated interpretability of single variables in the context of latent variables, which is why index-building methods, in particular those based on Factor Regression, have a head start. Balancing the benefits of a good model performance against interpretability is hence the key task in the decision. The set of index-building methods is quickly narrowed down, as only two methods within Factor Regression achieve reasonable results – the FarLasso and the FarAdaLasso. PCovR performs well in the latent variable model specification, too, but its high variance in the regression model simulations leads to its removal from the short-list. Moreover, its constructs are expected to be more difficult to interpret than the factors in FAR-methods. While FarLasso and the FarAdaLasso do not perform as well in the regression model as they do in the latent variable model, the latter model carries more importance for this work owing to its theoretical foundation. Based on the simulation results and the aspects of interpretability, both methods could be chosen since they combine reasonable predictive capability with interpretable factors.
Yet, the decision is made in favor of the FarAdaLasso as it tends to engender models that retain many factors with small coefficients lowering a model's interpretability. This behavior is believed to emerge as a byproduct of applying the same shrinkage factor onto all factors. Since Factor Analysis is unsupervised, the factors exhibit significant differences in their relevance for the outcome. However, the degree of shrinkage is usually small because the C_p-statistic depends on the fit of the model. This does not contradict the results of the simulation study, though. Despite being included in the model, factors with small coefficients only have a small influence. This observation is in line with the findings in the simulation, since the mean squared error as the evaluation criterion does not punish the absolute number of factors/variables in the model. One can, therefore, perhaps speak of visual parsimony that the Lasso cannot deliver on such data. Provided one were willing to ignore factors with small coefficients, similar results to the FarAdaLasso could be attained. Using the FarAdaLasso, however, yields more compact models with about the same degree of precision.

Chapter 6

Empirical analysis

This section presents the results of the empirical analysis. Starting with general information about the utilized data set and its structure, the steps to arrive at the dependent variables are presented in detail. This part is followed by an overview of the explanatory variables. The remainder of the chapter presents the results of the empirical analysis and gives a conclusion.

6.1 Data

The analysis is based on data from the German Socio Economic Panel Study (SOEP, v30) which is a representative panel survey of German households (Wagner et al., 2007). Fundamental information about children and their individual school success are gathered from the Youth Questionnaires of the years 2000-2013. These are completed once by children from SOEP households when they are about 17 and ready to enter the SOEP themselves. The (retrospective) questionnaires ask among others for information about the school career, particularly the school type attended/attending and the last grades obtained. In addition, the data set comprises unique identifiers that enable the researcher to link a child to its (social) mother and/or its (social) father, provided that they are known and surveyed in the SOEP. Given this information, parental data from the standard questionnaires can be linked to the outcome.
In order to measure cognitive skills, data on test scores can be obtained from the "COGDJ" data set. This is an addendum to the Youth Questionnaire and based on a questionnaire containing a modified I-S-T 2000-test (Schupp and Hermann, 2009, p.2). First introduced in 2006, it contains test scores of the adolescents that also answered the youth questionnaire from 2004 onward. Hence, the individual test scores can be linked to the youth questionnaire data set and thereby to the

parental data as well. Details on the procedures and tests underlying the COGDJ data set are described in Schupp and Hermann (2009).

While the school leaving degree is measured one-time by definition, also the cognitive skills test is conducted only once for an adolescent. Moreover, also the grades are solely surveyed for a single, the most recent, point in time. Any here proposed measure will hence be based on a cross section data set and thereby not allow the consideration of fixed effects.
The matched data contain roughly 200-300 observations per year in the case of school achievement indicators, the overall number of cases for test scores is lower on grounds of the later survey start. Further, there is also a higher non-response rate which reduces the number of observations on a yearly basis such that there remain about 150-250 observations per year.

6.1.1 Specifications

There are some caveats with the treatment of the explanatory parental data, because the outcome is measured only once. One relates to the question of how to link parental data to the child's outcome: Because the SOEP is an annual survey, parental data are typically available for multiple years before the child's school outcome is measured. It is not immediately clear how to prioritize the single bits of information from different years. Another point concerns the meaning of parental characteristics in the context of this framework: They are regarded as expressions of certain facets of family background or milieus. But as society changes, these facets and also their relations to the parental characteristics may change, too. Over a time span of 13 years, such developments cannot be excluded and may lead to misleading results. These two points are addressed in the following, starting with the first one.

In contrast to the survey about the school achievement or the test scores, parents receive a survey to answer every year. Despite this fact, not all characteristics are surveyed yearly since questionnaires change. There are some variables which are surveyed periodically but not necessarily annually, for instance every two years, and there are other variables which are irregularly surveyed. Table 6.1 shows an example of such a data structure. In this illustration, the census of Variable 1 is of annual nature, Variable 2 is periodically surveyed, but not annually, and Variable 3 is irregularly surveyed.

If a survey item was asked more than once, like Variable 1 and 2, there are data available for the same characteristics in different years. Hence, this information has

Table 6.1: Visual example for data structure

Year	t	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9
Outcome									X
Variable 1	X	X	X	X	X	X	X	X	X
Variable 2		X		X		X		X	
Variable 3					X				
X: Occurs in survey for the particular year									

to be subordinated or condensed in some way in order to relate to the outcome's single data point. How this is done optimally depends on the variable. Personality traits, for example, are considered stable and any year of measurement could be used, whereas income varies more strongly such that an average value could be deployed.

If a variable is collected irregularly, like Variable 3, a different issue emerges: Since the outcome's time of measurement varies depending on a child's birth year, it is not possible to set up a correspondence between the time of measurement of the parental variable and the age of all children in the sample. What would happen if such a correspondence were not required? The implication for a given parental characteristic could be that it is measured at the child's age of 7 for some children but at the age of 15 for others. The information would all be treated the same, although its meaning might depend on the child's age. Hence, it is sensible to demand that the timing of measurement should correspond to a certain age of a child or be in bounds of a certain age range at least.

An approach to solve this problem will be suggested after a second data issue has been highlighted. Up to now, the status is that one yields either no correspondence for irregularly surveyed variables or numerous missing values – unless they are removed from the set of explanatory variables. A related point concerns the meaning of character traits over time. As society changes, time spent on activities and attitudes are likely to change in social milieus. Characteristics that are an expression of an important facet in early survey years, might not be linked to this facet in later years. Moreover, also the facets themselves may change as society develops. Pooling the data over 13 years would assume constancy of these relations over time.

In an attempt to solve the data-related and theoretical issues, the data set is split into several "cohorts" which are analyzed separately. A cohort has a special meaning here, for it encompasses several adjacent birth years to ensure having a sufficient number of observations. These cohorts are created according to the demarcation scheme displayed in table 6.2.

Table 6.2: Cohort scheme

Survey Year	Cohort		Comprised Ages	
Dependent Variable	Grades	Test scores	Grades	Test scores
2000	1		16, 17, 18	
2001	1		17, 18, 19	
2002	1		17, 18, 19	
2003	2		16, 17, 18	
2004	2		16, 17	
2005	2		16, 17, 18	
2006	3	1	17	17, 18, 19
2007	3	1	17	17
2008	3	1	17	17
2009	3	2	17, 18	17
2010	4	2	17	17
2011	4	2	17	17
2012	4	2	17	17
2013	4	2	17	17

The selected time span of at most four years reduces the risk of evoking problems related to changing meanings of characteristics. Moreover, irregularly surveyed parental characteristics can now be linked to certain cohorts because the addressed time discrepancy has been reduced. In consequence, however, the sets of regarded variables differ across the cohorts which limits the comparability of cohort analyses.

What remains to be discussed is at which child's age the measurement of parental characteristics takes place. The literature emphasizes the importance of the first years of a child, comp. e.g. (Cunha and Heckman, 2007, p. 33), but there is a significant trade-off with the number of observations. On grounds of panel mortality, the sample size decreases the further one goes back in the past. In addition, the gathering of many interesting parental characteristics has only started recently. On the other side of this range, the Youth Questionnaire is filled in when the child is 17. This age, however, cannot be the reference age for every child since some have already graduated from school at the age of 15 or 16. Using parental characteristics that have been measured after the outcome measurement should be abstained from for potential endogeneity reasons.
Based on these considerations, there remain two options. Either the parental variables refer to the time of outcome measurement, i.e. when the child is 15, 16 or 17, or they refer to a specific age for all children in which case they would have to be measured before the child turns 15. There are two arguments in favor of the latter

option: One is that many school-related outcomes have been settled by the age of 15, some pupils have already graduated or will graduate in the upcoming year. Big leaps in both school type or grades rarely occur. Secondly, the parameters of the suggested latent variable model are unlikely to change substantially within a few years which offsets the potential advantage of the first option. Therefore, the measurement of parental characteristics usually takes place when their child's age is in the window of 12 and 15 years. In some cases, where the traits can be taken nearly constant, this window is opened somewhat further. In sum, variables gathered at this age range are to explain the variation for outcomes measured at the age of 15, 16 and 17 respectively.

The last point in this paragraph on data handling concerns the treatment of different family backgrounds. A thorough approach to the topic would require to exploit all available parent information. However, using data of both parents from children with separated, divorced or unknown parents is problematic. This problem and an attempt to deal with it is described in Boll and Hoffmann (2015) for a similar data structure. Their approach is adapted here in a reduced form and with a slight modification.
The fundamental idea is to analyze two kinds of samples: One in which partner data are neglected, even if they were available, and another one that takes partner information into account but thereby excludes all lone parents. A partner can both be a married partner or a life partner. The first sample hence contains all available parent-child pairs, while the second is a subset of the first and contains all parent-partner-child triples. The procedure to implement this in the SOEP data set is as follows:
The first step is to separately link the child's data to the parental data once via the mother's id (Identification number) and once via the father's id. This gives two different samples. The parent whose id was used to establish the link between the data is called "reference parent". For instance, when the child's data are linked via the mother's id, the sample is called "Sample with mother as reference parent". There is one additional requirement to be included the estimation sample which is called "Survey Restriction": The reference parent was surveyed in the SOEP latest from the child's age 12 onwards. This number stems from the above discussion on the time of data measurement. At the expense of not considering potential partner data, the resulting two samples can include children with lone parents.
In a second step, all children are regarded whose reference parents are in stable partnerships while the child is between 12 and 15 years old. The partner does not have to be the other biological parent but this is most often the case. The mentioned Survey Restriction is also imposed on the partner in these samples.
Doing all that, four samples emerge in total. Table 6.3 visualizes the previous

discussion. It shows all possible combinations of reference parents (row 1) and partner context (row 2). As an example, consider the setting in which the mother is the reference parent and partner data are disregarded. This sample consists of children whose mothers either have no partner (1) or have a partner who is not the child's father (2) or live with the child's father (3). When partner data are regarded, only the latter two (2+3) options remain. As the classic family forms the majority of families, there is a considerable overlap between the two samples with partner data. In general, the added value of reference-parent separated analyses is low, in particular on the father's side for which the number of cases is lower. The analysis in this thesis therefore focuses on the case in which the mother is defined as the reference parent and omit the father's samples.

Table 6.3: Overview of sample specifications

Mother			Father		
No Partner (1)	Partner (2)	Father (3)	Mother (4)	Partner (5)	No Partner (6)

Sample without partner data, reference parent mother:	(1+2+3)
Sample without partner data, reference parent father:	(4+5+6)
Sample with partner data, reference parent mother:	(2+3)
Sample with partner data, reference parent father:	(4+5)

6.1.2 Dependent variables

Referring to the discussion in the introduction on measuring human capital, it was noted that the choice between a school-related measure and one based on cognitive skills depends on four factors: What one's expectations are about the way human capital works, the context and aim of the analysis, the heterogeneity in the sample and data availability. Data availability has already been addressed in favor of a school-related measure because there are more observations available. The degree of heterogeneity is more difficult to judge. Since the analysis deals with pupils at German schools only, potential quality differences will not be as large as with international data. Nevertheless, school quality might still be heterogeneous enough (Anger et al., 2015) to justify using cognitive skills as an alternative measure of success. With regard to the effect channel of human capital, there are also some reasons for not ignoring them, provided that one does not believe Spence's signaling theory Spence (1973) to be the answer to everything. Cognitive skills can, for instance, be important in application processes which include intelligence tests. In sum, there are arguments for both measures such that a decision for one measure would restrict the analysis unnecessarily. A comparison between the two

measures could, by contrast, even yield some insights. Under the premise that test scores more strongly reflect innate ability, parental variables are expected to have a higher influence on the school-related measure. In line with the goals of this thesis, the focus should lie on measures of crystallized intelligence since they are supposed to be determined by environmental factors rather than by nature. Yet, owing to the larger number of cases and the single country analysis, the school-related measures are considered more important, while cognitive skills are deemed as a robustness check.

The endogenous variable for the school related measure is based on information about the latest grades in the child's main subjects, i.e. German, Mathematics and the first foreign language (mostly English). The adolescent states these when he or she turns 17, but some individuals left school at the age of 15 or 16. Hence, the age of grade measurement will differ by the graduation type. All grades are, however, measured while the individuals are attending a secondary school. Grades are typically integer values measured on a scale from 1 to 6, where 1 denotes the best and 6 the worst grade. Sometimes the grades come on a different integer scale ranging from the best grade 15 to the worst 0, which corresponds to the 1 to 6 scale, but with finer distinctions and in reversed order. Accordingly, the scales can be united by appropriate scaling (that is half point steps on a 1 to 6 scale).
Grades are still not readily comparable. This is because there are three different secondary school types in Germany which differ not only in their duration but also in their grade requirements. The highest standards are found in the upper track secondary school (Gymnasium, short: UTSS), the lowest in the lower track secondary school (Hauptschule, short: LTSS). In between the two is the middle track secondary school (Realschule, short: MTSS). To obtain a measure that can be compared across school types, differences due to valuation standards have to be taken account of. However, there is no universal transformation scheme to achieve this and only approximations exist. Leaving out finer distinctions in grades, the following table depicts the transformation scheme employed to obtain a comparable measure, denoted as "Unified scale".

Table 6.4: Transformation scheme for grades

		Grade								
School type:	UTSS	1	2	3	4	5	6			
	MTSS		1	2	3	4	5	6		
	LTSS				1	2	3	4	5	6
Unified scale:		1	2	3	4	5	6	7	8	9

This transformation scheme is derived from the Hamburg Stadtteilschule grade transformation scheme (according to Hamburger Senat, 2011) and is read as

follows: If a pupil achieved the best grade, that is a 1, in a subject at a LTSS, the corresponding grade at a UTSS in that subject would be defined as a 4. If the same pupil achieved a 5, the corresponding grade is worse than the worst grade on the UTSS, which is the reason for extending the scale up to the value of 9.
Having rescaled the grades, the next step is to invert the unified scale, such that favorable values are numerically larger. As the requirements to achieve a certain grade differ not only by school type but also across schools and classes, the grades should be standardized on classroom level. Through data limitations, however, a sufficient number of cases is first available on federal state level. Standardization on this level can still be justified by the fact that education policy differs between federal states which may result in different requirements. Observed differences in mean grade values strengthen this notion. Standardization at federal state level is done for each subject separately and data from all available survey years are used. The average of the three standardized grades defines the final measure for school achievement.

There are some assumptions to this generation process. A basic one, underlying the calculations, is that grades are metrically scaled implying equally large distances between the grades. However, grades are on an ordinal scale so treating them as metric is only an approximation. Yet, such a treatment is not uncommon when thinking of the calculation of mean grades. Moreover, this is probably only a minor imprecision in comparison to the subjective character of grades in general.

One might further argue that any transformation scheme, although required to compare grades between school types, is somewhat arbitrary. If one does not want to use such a scheme, only the type of graduation or the years of schooling remain as measures of school achievement. In contrast to grades, however, they have two weighty disadvantages: Substantial heterogeneity within a type of school cannot be taken account of. Furthermore, the measures exclude the possibility for the best pupils on a lower school type to be better than the worst ones on the next better school type. Graduation-based measures simply compress any variation to three categories. Signaling theory and the argument on formal institutional requirements speak in favor of the type of graduation. However, the grade point average measure contains the type of graduation implicitly as a level shifter. On account of these arguments, the grade-based measure is maintained and the transformation scheme's robustness is examined. One way to analyze the impact of assuming such a scheme is to use a different one and evaluate how strongly this affects the results. A robustness check is based on the following transformation scheme, where the difference between LTSS and MTSS has decreased by one grade:

The second robustness check is based on the test scores. The cognitive tests

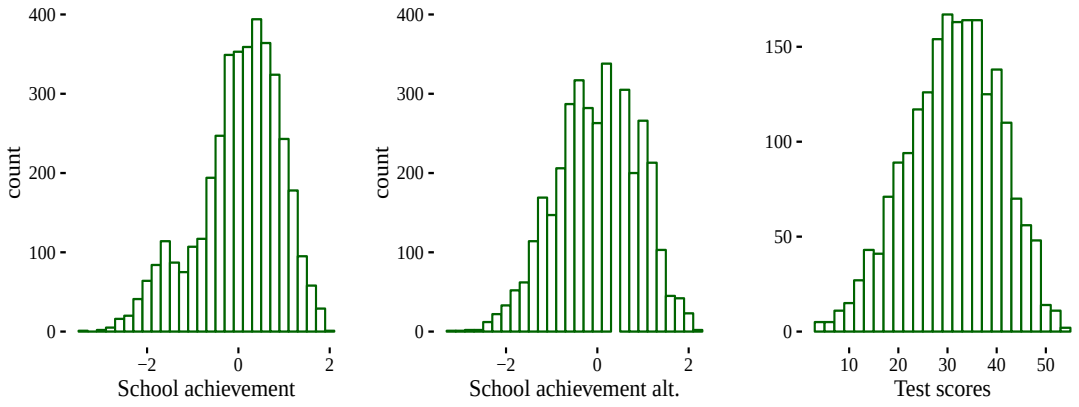
Table 6.5: Alternative transformation scheme for grades

		Grade							
School type:	UTSS	1	2	3	4	5	6		
	MTSS		1	2	3	4	5	6	
	LTSS			1	2	3	4	5	6
Unified scale:		1	2	3	4	5	6	7	8

encompass three categories with twenty tasks each. These categories are analogy tasks which are language related, insertion of correct arithmetic operators and finding figures which logically continue the displayed order. According to Dahmann (2015, p. 13) the first two tests record crystallized intelligence because they refer to learned competences. The third test, on the other hand, gives an indication of fluid intelligence. A sum index called "deductive thinking" is created as the number of right answers in the first two categories.

All three dependent variables exhibit characteristics of a bell-shaped curve, as can be observed in the following diagram:

Figure 6.1: Distributions of endogenous variables



6.1.3 Explanatory variables

This subsection is dedicated to a description of the explanatory variables. If not stated otherwise, they refer to a parent of a child. With many characteristics available, the description proceeds in correspondence to the grouping of parental characteristics in chapter 2. But not all variables should be used for each cohort analysis. For reasons like irregular questionnaires or a high non-response rate, missing values arise which decrease the available number of cases for an analysis. Therefore, there is often a trade-off between information gains from additional

samples and the information that an additional variable carries. In order to have a sensible balance, the available number of cases was thus checked (cohort-wise) prior to the analyses. If the inclusion of a variable led to a decline in observations of roughly more than 3-5 %, the variable was disregarded. This decision was made on the basis of two indicators. One was the count of available observations. However, missing values rarely occur truly randomly, instead the "missingness" has patterns across variables. Thus, a second check was done for each variable conditionally on the inclusion of the other variables.

A detailed tabular overview on all variables along with the information for which cohort a variable is used is provided in the appendix (Table 8.1 - Table 8.9).

Personality Traits

Starting with personality traits, the first challenge is to capture them, because these traits are latent. They can hence only be approached indirectly, e.g. through observed behavior or from answers to adequate questionnaires. For the two concepts discussed in the theory section, the Big-5 factors and the locus of control, the SOEP provides statements (items) in the questionnaires to which the respondent can agree (or disagree) on Likert scales. Since one item is usually not considered as a reliable enough indicator of a personality trait, several items per trait are required to be answered. In the case of the Big-5 indicators, for instance, there are fifteen statements in the SOEP, three for each trait. These fifteen items are already a significant reduction from the originally suggested 240 items. The concept of locus of control is covered by nine to ten questions relating to different facets of control orientation.
How this information is handled has to be discussed at this point. Generally, there are two options: The first is to preprocess this information in a way that obtains the "correct" number of personality dimensions and use these dimensions as input variables. For example, reducing the fifteen statements to five dimensions would be the way to go for the Big-5 indicators. The second option is to pass on any preprocessing and use the statements directly as explanatory variables.
The advantage of the latter option is that it allows for individual links to the set of other characteristics. That means each item can have an idiosyncratic relationship to the other input variables as well as the outcome. The less correlated the items are the stronger their idiosyncrasy would be in which case any preprocessing would run a risk of discarding possibly interesting information. On the other hand, a single statement is usually not conclusive enough about a personality trait. It will therefore be very difficult to grasp its meaning as an input variable, even more so if the statements belonging to a specific trait exhibit different associations to the

outcome. The first way, i.e. preprocessing, is for these reasons considered more closely related to the theoretical foundation of this work and therefore chosen. The validity of the interpretation in the empirical analyses depends, however, on the manner the preprocessing is conducted.

To obtain surrogate variables for the Big-5 traits based on the fifteen available items, two procedures have been suggested in the literature. Wichert and Pohlmeier (2010, p. 9ff.) suggest an equally weighted average of the statements belonging to a trait. Dehne and Schupp (2007) use Factor Analysis to determine the weights. The idea of using Factor Analysis instead of simple averages stems from the notion that personality dimensions are latent and can only be noisily measured. Categorical scaling of the indicators exacerbates the noise. According to Piatek and Pinger (2010, 20ff.), using Factor Analysis models can reduce measurement error in comparison to simple indexing with equal weights. In the analyses by Dehne and Schupp, items designed to belong to a certain factor (trait) receive the highest weight in the respective linear combination. Each Big-5 trait emerges as a separate factor and the individual's propensity to a factor is calculated as the weighted linear combination of all items.

Piatek and Pinger (2010) proceed similarly as Dehne and Schupp for items concerning the locus of control. In attempt to obtain a measure for the locus of control, Caliendo et al. (2015) also use Factor Analysis to determine which items are correlated, and compute average values based on that.

In this work variants of the factor-analysis weighted approach described by Dehne and Schupp are applied. Given the description in their paper, it was, however, not possible to reproduce the results using Stata 11.2. There might be some confusion about the terminology, but presumably there are also some inaccuracies. It is not obvious whether the authors applied a PCA or a Factor Analysis with Principal Component Factors as extraction method. Moreover, the coefficients for the linear combinations that are noted in the appendix did not lead to uncorrelated Big-5 dimensions in the data at hand. Uncorrelated dimensions by a linear combination can only be obtained from an unrotated PCA solution, but the stated solution is Varimax-rotated. Having tested several alternatives, the following procedure seemed to come closest to the results by Dehne and Schupp: Firstly, a Factor Analysis with Principal Component Factors is applied, then factors with a minimum eigenvalue of 1 are selected. Subsequently, the loading matrix undergoes a Varimax-rotation and the Big-5 traits are predicted using the regression scoring method. This approach is separately conducted on the overall sample, for each available survey year (2005, 2009, 2013). As the factor loadings are highly similar in each year of analysis, so are the resulting constructs. Their meaning corresponds to the postulated characteristics, i.e. high values on the openness variable indicate

an individual who is open to new experiences, those with high values on the neuroticism construct tend to be emotionally unstable.
The same approach was initially applied on the items for the locus of control. Although being a unidimensional concept, two factors emerged: One indicating a high external locus of control, the second a high internal locus of control. Piatek and Pinger (2010) also observed this and decided to use only the first factor. In this work, the solution to this problem is to extract the factors by the Principal Factors method instead of the Principal Component Factors method. This yields a single dimension in all survey years. The resulting construct correlates positively with variables indicating a high internal locus of control and is hence interpreted as that. Also for this measure, the procedure is applied for each survey year separately on the complete SOEP sample.

Attitudes

Attitudes comprise various topics. There are time preferences, risk preferences, attitudes towards reciprocity and trust to other people as well as importance indicators. Both the first three self-reports and subjective degree of importance are inquired on Likert scales.[24] Importance indicators, i.e. whether an individual considers a certain aspect important for satisfaction, are derived from the work by Kluckhohn and Strodtbeck (1961) and can be divided into attitudes towards materialism, family life and altruism. Several sub items are contained in each category, for example the category materialism contains the importance of being able to afford things, income and work. Since these variables have clear interpretations, no preprocessing is performed here. Attitudes also cover interests here, for instance the interest in politics, and memberships, such as in a labor union, an environmental organization or a professional/occupational association.
Four indicators which capture dimensions of subjective well-being related to helplessness, future confidence, and isolation, are also included. Here, too, a Likert scale with four items is used.
A topic that is related to education is the attitude towards further education. It can range from disinterest to interest, but does not necessarily imply (non-) participation. Nevertheless, this information may help to identify aspirations and valuing of education.

[24] Time preferences: How would you describe yourself: Are you generally an impatient person, or someone who always shows great patience?
Risk preferences: How do you see yourself: Are you generally a person who is fully prepared to take risks, or do you try to avoid taking risks?

Time use indicators

The SOEP is also rich in time use indicators. Details on the single variables used and their summary statistics can be found in the variable description in the appendix. One point concerns the measurement of them. Most of the indicators are inquired on Likert scales as frequencies of occurrence, ranging from 1 (never) to 4 (weekly) or 5 (daily). They are rescaled by applying the scheme of Büchel and Duncan (1998) who transform the statements in order to represent the number of times per year.[25] For some activities precise statements of the hours per day spent in them are provided. This concerns the variables housework, child care, leisure/hobbies and crafts/repairs/gardening. About 0.6 % of the individuals of the sample stated to spend more than 20 hours on such activities. Such values seem implausible as an average and cast doubt on the validity of those individual's statements, which is the reason they are removed from the sample.
As with the attitude towards further education, there are also time use indicators on this topic. One can differentiate (non-)participation by reason or goal and context. The context can be occupationally motivated which can differ between retraining or rehabilitation in order to keep a job or qualify for a new one, and advanced training, aimed at promotion. In a private context, further education is broadly defined and can therefore cover numerous topics.

Demographics

Characteristics in this group cover remaining aspects of the family background. In addition to general information such as parental age and education, operationalized as years of education, the data set contains information about work related variables, such as work hours and labor income. The actual number of weekly working hours measures the degree of labor market activity. In comparison to the employment status, they avoid creating too crude categories[26] and in comparison to contract hours includes self-employed parents. With regard to income, five measures are included here: On household basis, post-government household income, income from assets, income from private and public transfers. On individual basis, labor income is used. Two versions of income variables are created in this dissertation: One as an average over the child's age of 13 to 15, the other as an average over the age of 10 to 12. Not always are both timeframes used – in some samples the second variation leads to high loss in observations and is therefore excluded.

[25] That is: Daily=365, Weekly=52, Monthly=12, Less than monthly=4, Never=0

[26] For example, 18 and 34 weekly working hours would fall in the same category although they are qualitatively different.

Occupational prestige, measured by the Wegener scale[27], is not included because the variable is only gathered for employed individuals. While labor income and work hours can be set to zero in the case non-employment, a value of zero does not make sense for occupational prestige. Including it as a regular input variable would implicitly restrict the analyzed sample to children with employed parents. Information about the family background in general is also included, e.g. the number of children of a parent. Due to cultural differences, family backgrounds are likely to differ across geographical areas, hence the analysis includes high-level indicator variables for living in the north, west, east and south of Germany.

6.2 Results

The complete list of variables used for each cohort analysis split by category can be found in the appendix.
The following algorithm calculates the results: Using Factor Analysis in which Principal Factors as extraction method is applied, the complete set of factors is obtained. To facilitate the interpretation of the factors, a Promax oblique rotation with $1.0 \leq g \leq 1.5$ follows, where g is chosen to yield a good interpretability of the loading matrix. A value for g higher than 1.5 turned out to lead to factors that are difficult to interpret as they rely on a single or very few variables only. After the rotation, the factor scores based on the rotated factor model are estimated by the Bartlett method. The resulting scores are standardized and used as regressors. Their parameters are estimated by applying the Adaptive Lasso. Owing to the standardization, the estimated parameters coincide with the factors' economic significance.

To evaluate the performance of the overall model, the coefficient of determination, R^2, is attached to the tables. In order to compare the approach with a model that is based on parental education and income measures only, the Adaptive Lasso is applied on the set of those variables. The corresponding coefficient of determination, R^2_{EduInc}, is also denoted and the percentage change to R^2 calculated. The reason for using the same statistical approach is that shrinkage induces estimation bias, also on the factors left in the model, which reduces the coefficient of determination in-sample. Using a standard linear regression model without regularization as baseline would, therefore, render a comparison unfair.

[27] Occupations are ranked with regard to their social prestige on this scale and can be interpreted as metric owing to open scales. Unlike the SIOPS scale, for instance, the Wegener scale is adjusted for Germany (Boll, 2011, p. 71)

With regards to the presentation of the results, some simplifications are undertaken. To avoid a numeric inundation, neither the (rotated) loading matrices nor the coefficient matrices connected to estimation of factor scores are displayed. However, to find out about the exact contribution of a variable to a factor and, hence, its association with the outcome, this information is required. This issue is deepened on the basis of inspecting one factor more closely in the analysis of cohort 1. In this context, the main reason for interpreting a factor based on the variables driving it strongest is given. This information can be found in the loading matrix. Since a typical loading matrix for a cohort fills several pages with numbers, its content is practically reduced to the following information: When $g > 1$, i.e. an oblique rotation is conducted, only variables with a loading > 0.15 are deemed relevant for a factor. For $g = 1$, this absolute lower bound is set to 0.2. For the purpose of readability, the following thresholds indicate the strength of a loading: + for loadings bigger than +0.5, ∘ for loadings between +0.5 and +0.15 (+0.2), -∘ for loadings in the range of -0.15 (-0.2) and -0.5 and lastly − for loadings more negative than -0.5. The factors are numbered by their eigenvalue size from largest to smallest, which highlights the importance of considering factors independent of their eigenvalue size. Moreover, it facilitates the identification of factors with low eigenvalues that sometimes tend to yield unstable results.

6.2.1 Dependent variable: School achievement

This section presents the results for the standardized averaged grade over the subjects Math, German and the first foreign language.

Analysis of the first cohort without partner data:
Starting with the analysis of the first cohort, whose results are presented in table 6.6, a note concerns the low number of observations. Even though the choice of variables was made carefully, this cohort is concerned by the entry of many new households around the year 2000. Children of those households cannot be included in this analysis because there is not sufficient retrospective information on the parents available. A general approach to remedy the drawbacks of a low number of cases would be to impute values. With the statistical method of choice, however, the implementation of multiple imputation is problematic, which is for two reasons: First, the results that Factor Analysis produces change with every imputation. If the changes are severe enough, it becomes impossible to compare the factors and, thus, also the predicted factor scores. The second reason is that the decomposition of variance into actual variance and variance induced by imputation has not yet been developed in the framework of estimation by the Adaptive Lasso.

Table 6.6: Cohort 1 - Without partner data:

Variable	Coefficient	Description
Factor 33	.1832	∘: Years of living with both parents. -∘: Importance of having political influence, no interest in further education at all.
Factor 21	.1726	+: Education. ∘: Membership in environmental association.
Factor 13	.1471	+: Frequency of cultural activities. ∘: Interest in politics, participation in further education, importance of having political influence, education.
Factor 1	.1137	+: Income (Household labor, Post government). ∘: Asset income, education, individual labor income, frequency of cultural activities, age, exercise.
Factor 6	-.0743	+: Importance of income and flat. ∘: Importance of own mobility, health, freedom, job. -∘: Education
Factor 22	.0208	∘: Honorary post, religious activities.
Factor 32	-.0206	∘: Membership in labor union, membership in environmental association.
Factor 8	.0007	+: Living in city.
Intercept	- .1000	
N=318	R^2 =.2	Change: 67 % (R^2_{EduInc} =.12).

On the upside, by considering the number of factors, it becomes apparent that the chosen strategy to reduce dimensionality worked out for this data set. Another positive aspect is that the results of this cohort show similarities to the other cohorts. Thus, it appears that the consequences of a small number of observations are not too severe.

The factor with the largest coefficient size in absolute terms, Factor 33, relates, among others, to the years of living in a classic family, i.e. together with both parents, during childhood. The more years, the more positive the outcome. This factor appears in similar forms in other samples, too, but typically has a smaller coefficient. The large coefficient for this cohort might be due to the propensity to instability of factors with small eigenvalues. The low number of cases for this analysis may exacerbate the issue.

Another important factor is Factor 21, which is principally driven by formal education and less by holding a membership in an environmental association. Since this factor behaves more normal, it serves as an example on whose basis a technical interpretation of the parameter values is conducted. As the model is linear, the interpretation of an increase in the dependent variable of .1726 units, when the factor score increases by one unit, holds. Setting this into perspective, the dependent variable ranges from about -3.5 to 2, while the factor scores (for this particular factor) lie between -4 and 3.6. Given this information, the estimated

coefficient may appear small even though it is one of the largest in the model. One has to bear in mind, however, that the coefficient has been dragged towards zero during the model selection process.
When the relation between a factor and the outcome is inspected, the coefficient provides information about the association strength. To analyze the meaning of a unit increase in the factor score, however, the relation between the factor and the set of original variables becomes relevant. In this thesis, only the variables with the largest loadings are used for interpretation. However, as each factor depends on all variables in the original set, many possible circumstances could lead to a unit change. To validate the interpretation, it is necessary to quantify the impact of an isolated change in any predictor on the score. The exact calculation is involved since it requires going backwards through the previous computational steps, which are the standardization of factors, and the estimation of both factor scores and loading matrix. Doing so suggests that only the variables with the largest loadings on a factor are able to move the score substantially.
In this concrete example, one additional year of education raises the (standardized) factor score by .29 points. With this information, its implicit association with the outcome can be computed: It amounts to an increase of about .05 units. The second largest loading is found for the indicator of holding a membership in an environmental association. Holding one in comparison to not holding one leads to a .79 points higher factor score that corresponds to a .13 higher value in the dependent variable. On the other hand, a variable with an absolute loading below the considered threshold, such as the binary coded measure for "disinterest in further education" with a loading of -.11, leads to a decrease in the factor score of about .25. This corresponds to a comparably small impact of -.04 on the outcome. One could relate all predictors to the outcome conditional on a factor by continuing in this fashion. While this example shows on the one hand that every original variable influences a factor, it also shows that the magnitude depends on the loading size. For some applications, the total impact may be of interest. It is obtained when the sum of the predictor's influences over all retained factors is computed. Such an explicit interpretation is not conducted in this thesis, however. Owing to the premise of the underlying latent variable model, observed variables are expressions of factors, so that their direct relation to the outcome is taken to be uninformative for the purpose.

Factor 13 is associated with the frequency of undertaking cultural activities such as concerts, theaters, lectures and also with political interest, formal education and participation in further education. Cultural activities also appear in other samples when it comes to relevant facets which are not primarily driven by formal education or income. Factor 1 is also relevant and loads highly on household (labor) income.

There is a frequent appearance of this factor in analyses disregarding partner data. Although a potentially existing partner is not explicitly regarded in this analysis, this factor depicts the existence of (an employed) one. The reason for this is that the broad majority of partners in this sample is typically employed. A higher value in the two household income types can hence also reflect the presence of a partner in the household. In sum, the factor's large, positive coefficient emerges from a mixture of higher disposable income and a stable parental relationship.
The association for Factor 6 is slightly negative. It is driven by several importance indicators, among which the items considering income and housing important for satisfaction are most important. Since it is hard to argue why such an attitude could be harmful, the factor's negative link with education is likely to cause it.

In sum, the results for this cohort do not indicate important facets of family background that are simultaneously unrelated to formal education, income or family structure. Factor 13 comes closest to this by emphasizing societal participation, indicated by frequent cultural activities and interest in politics. Partly, these results may be ascribed to the small set of time use and attitude characteristics on which the analysis bases. While this cohort analysis includes variables of geographical location and city size, they turn out to be independent of relevant facets of family background. This independence is a stable observation across the remaining analyses. In this analysis, also the measurement variable for the internal locus of control is unrelated to any relevant factor.

The number of cases is already low for the analysis without partner data, so an analysis with partner data, which would lead to a further reduction, is omitted for this cohort.

Analysis of the second cohort without partner data:
For the analysis of the second cohort, more variables and observations are accessible. The results for the analysis without partner data are presented in table 6.7.
Factor 2 bundles correlating variation between labor and post-government income, asset income, education and participation in further education as well as the child living as few as possible years with a lone mother. Among the factors in the model, this factor has the strongest association with the outcome. As in the first cohort, an implicit driving force of this factor is the existence of an employed partner or husband. A weaker influence on this factor is the mother's propensity to acquire education, which is reflected in the participation in further education but also in the level of formal education.
Factor 14 is the factor with the second largest coefficient. It relates positively to education-related variables (negatively to disinterest in further education) and to variables relating to societal participation, such as interest in politics and

memberships. Moreover, the factor is linked to an open character in the Big-5 sense. This factor hence depicts a well-educated mother who is interested and engaged in society and open to learn something new. Another factor that hints at the positive association of further education is Factor 5. Although it also lacks independence of formal education, the factor principally describes educational alienation depicted by the disinterest in further education. In addition, it points at structural disadvantages of children whose parents immigrated.
A factor that is virtually independent of demographic characteristics is Factor 3 that loads highly on the five trust variables in the data set. This association hints at a positive relation between mothers who hold an optimistic view of other people and their child's school success. This relation may be interpreted in the framework of social capital theory. People who trust other people more/easier could have more acquaintances which impact positively on social capital. Higher social capital can be linked to a positive development of the child by the arguments given in the theory section.
Contradictory statements can be observed for the variable measuring a household's income from private transfers. A negative link to the child's school achievement is implied by Factor 13, which also relates to a child growing up with separated parents. The high loading on private transfers on this factor makes most sense, when it is interpreted as child alimony payments. On the other hand, Factor 25 is chiefly driven by private transfers, but is linked positively the outcome. There are two possible approaches to an explanation of these results: One is to relate Factor 13 rather to the family structure than to the amount of income, and Factor 25 to the income source alone which yields, ceteris paribus, positive effects. A second one is to interpret the association of private transfers as non-linear. While there is a baseline negative association that hints at a dearth of financial means, its marginal negativity is decreasing. Stated differently, an additional Euro of private transfers does not hurt as much as the Euro before.

Negative associations emerge through Factor 31 and Factor 9. The first mentioned relates positively to the time spent on manual activities like garden work, housework and repairs. From a perspective of a limited time budget, this could describe a parent who frequently engages in such activities has little time and maybe also little interest in education. In this case, the parent might not be not a role model who encourages investments in education. Factor 9 is associated with mothers of many children who specialized on housework and child-rearing instead of earning income. Thus, this factor is essentially the opposite of Factor 1, which displays a working mother with few children. Owing to a smaller coefficient Factor 1 is judged less important, though. There may be several reasons for the negative relation of Factor 9. One may refer to the division of time and monetary resources

Table 6.7: Cohort 2 - Without partner data:

Variable	Coefficient	Description
Factor 2	.2330	+: Post-government income, household labor income. ∘: Household asset income, education, labor income, general participation in further education. -∘: Years of living with mother alone.
Factor 14	.1700	∘: General participation in further education, formal education, frequency of cultural activities, interest in politics, openness, membership in environmental association. -∘: Disinterest in further education.
Factor 5	-.1564	+: Migration background of father and mother. ∘: No interest in further education at all. -∘: Education.
Factor 3	.1131	+: All five trust variables.
Factor 31	-.1054	∘: Time spent on garden work and repairs.
Factor 9	.1024	−: Household income from public transfers, number of children. -∘: Hours spent on child care/housework. +∘: Work hours.
Factor 29	-.0641	∘: Years of living together with other relatives.
Factor 22	.0541	∘: Frequency of cultural activities, cinema visits.
Factor 25	.0418	∘: Household income from private transfers.
Factor 13	-.0380	+: Years of living with the mother and her partner. ∘: Household income from private transfers. -∘: Years of living with both parents.
Factor 1	.0374	+: Labor income, hours of work. ∘: Education, household labor income, membership in professional association. -∘: Number of children, hours spent on child care/housework.
Factor 21	.0296	+: Years of living with the father alone.
Factor 10	.0262	−: Living in a middle-sized town.
Factor 28	-.0190	∘: Extraversion.
Intercept	-.0610	
N=966	R^2 =.14	Change: 75 % (R^2_{EduInc} =.08).

when there are many children in the household. Although public transfers increase with each additional child, the share of disposable income for each child typically decreases. From the perspective of parental time on dedicated child care, a lower share is inevitable. Another explanation attributes the observed relation to being a consequence of self-selection. Mothers with worse chances on the job market are more likely to specialize in child-rearing. Worse chances on the job market might be due to insufficient education which in turn is negatively correlated with the considered outcome. However, this factor does not load particularly highly on education in any direction, making the first explanation more likely.

The last factor that carries importance for the pursued theory in this thesis is Factor 22. It measures the frequency of both parents going to high cultural events and pop culture ones such as cinema, pop concerts, disco and sports events. The

coefficient size is comparably small, yet it confirms the tendency of Factor 14 that societal participation, expressed by cultural activities, is related to the child's school achievement.
Variables of residence do not matter for any relevant factor, the same holds for some leisure time use characteristics like exercising or socializing. A general drawback of the model for this cohort is that it is unable to explain as much of the variation in the dependent variable as the models for the other cohorts. Apart from chance or unsuitable predictors, this could be attributed to stronger coefficient shrinkage. On the other hand, the percentage gain in the explained variance compared to the baseline model is still high.

Analysis of the second cohort with partner data:
Some of the previous results also emerge when partner data are regarded. A difference in this analysis is the increased number of factors deemed as relevant. Some of which, however, hardly contribute to explaining the outcome since their coefficients are small. From the viewpoint of model sparsity, such a result indicates a bad performance of the statistical procedure. The table of results is shown in table 6.8.
Factor 1, the one with the highest importance, describes the expected relationship between primary disparities and the outcome. It is a conglomerate of the (working) partner's income and both parents' education. The importance of income is strengthened by Factor 36, which indicates marginally increasing improvements in the outcome induced by the partner's labor income, and despite a small coefficient also Factor 4. As in the previous analysis, there is a factor that relates to a parental migration background and to low education. It is negatively linked to the outcome, while Factor 7, related to trust variables, repeatedly exhibits a positive correlation. It is striking that the trust variables solely refer to maternal trust; a similar factor, relating to the partner's trust variables, does not maintain a non-zero coefficient.

Factor 11 is of interest for this study since it relates solely to the propensity to undertake further education. When further education is done of one's own accord, this can be interpreted as having aspirations and ambitions. The observation holds for both parents and this factor loads neither highly on formal education nor on income. Another factor, which is mostly independent of primary disparities, is Factor 12. It measures the frequency of both parents' church attendances or other religious events. Although it is hard to argue that these attendances affect the child directly, the activity can be viewed as a surrogate for a religious family environment. In this case, there may be beneficial, rather traditional, ethics or norms that correlate with having a belief. In other cohort analyses, this factor is also connected to the frequency of exerting honorary posts, which can be

Table 6.8: Cohort 2 - With partner data:

Variable	Coefficient	Description
Factor 1	.2092	+: Partner's and household labor income, post-government income. ∘: Education (both parents) and partner's hours work.
Factor 3	.1680	−: Migration background of father and mother. -∘: No interest in further education at all. ∘: Education.
Factor 36	.1374	∘: Partner's income squared.
Factor 20	.1349	+: Partner's interest in politics. ∘: Interest in politics, education (both parents), partner's time spent in local (political) initiatives .
Factor 7	.1083	+: All five trust variables.
Factor 34	.0984	∘: Frequency of cultural activities (both parents).
Factor 12	.0913	+: Frequency of religious activities.
Factor 11	.0776	+: General further education (both parents). -∘: No interest in further education at all (both parents).
Factor 49	.0748	∘: Education.
Factor 30	-.0710	∘: Time spent on garden work (both parents), partner's time spent on repairing cars.
Factor 9	-.0648	+: Household income from public transfers, number of children. -∘: Work hours.
Factor 27	-.0626	+: Death of father, years of living with other relatives.
Factor 4	.0460	+: Household asset income. ∘: Post-government income.
Factor 38	-.0449	+: Partner's frequency of exerting honorary posts. ∘: Frequency of exerting honorary posts.
Factor 57	-.0381	∘: Approval of "Caution When Dealing With Strangers" (both parents).
Factor 37	.0293	∘: Membership in a professional association (both parents).
Factor 8	.0163	+: Living in eastern Germany.
Factor 51	-.0155	No loading large enough.
Factor 53	-.0067	∘: Sociability, extraversion.
Factor 22	.0042	+: Frequency of cinema visits (both parents). ∘: Frequency of cultural activities.
Intercept	-.0200	
N=696	R^2 =.18	Change: 63 % (R^2_{EduInc} =.11).

interpreted in the environment of the religious community.

Factor 20 can be summarized as a facet of family background showing interest and engagement in political issues, in particular that of the partner's. This facet can be interpreted as a surrogate for parents who have ambitions to shape the environment according to their ideas. This ambitious attitude can also hold for the child's development and school success. Moreover, the facet could indicate knowledge of developments in society, such that the importance of human capital for success in life is well-known. Both points could result in an increased parental engagement

in the child's school matters, explaining the strong positive link. Factor 34 is another factor virtually unrelated to formal education. It depicts active societal participation and the possession of cultural capital by a high frequency of cultural activities. As in the cohort analysis without partner data, a factor (Factor 30) relating to manual activities in the household is identified. It has a negative link to the outcome.
One general observation that traverses through these analyses is that factors related to interests and activities most often load on both the mother's and the father's variable. This observation likely reflects the consequences of assortative mating, according to which similar characters are more likely to find together.
Summing up the results of this cohort, heterogeneity in family environments which is not sufficiently captured by education and income is indicated. Markedly, factors relating to maternal trust, participation in further education, but also religiousness and cultural activities exhibit notable links. While factors related to formal education and income continue to be most important, refinements are discernible.

Analysis of the third cohort without partner data:
The results for the third cohort show some recurring patterns. However, by virtue of a larger variable set, known factors change and additional factors emerge.

New to this cohort is that variables referring to the child's age 10 to 12 are included. This does not induce substantial changes as they typically constitute factors with their equivalents for the age span 13-15. Thus, if there is no reason for doing otherwise, the age indication for these variables is omitted.

Starting with the analysis that excludes partner data, the results shown in table 6.9 emerge. It is observable that the first factor carries the major importance. Maternal formal education and household income drive this factor. As in the previous analyses for samples without partner data, the latter variable hints at the existence of an occupied partner in the household. Factor 32, related solely to education, completes the influence of the benchmark variables. However, additional primary disparities are prominent in this analysis. The large coefficient of Factor 25, solely driven by maternal age, suggests a positive relation of age to the child's school achievement. This relation can point towards the positive connection between higher education and the age of giving birth on the one hand, but could also depict routine and experience in raising children.
Factor 3 depicts a family background with many children and a high income of public transfers. As in previous analyses, this facet has a strong negative link to a child's school achievement. A facet in which frequent attendance of cultural events, including cinema (pop concerts, disco, sports events) visits are the norm is

Table 6.9: Cohort 3 - Without partner data:

Variable	Coefficient	Description
Factor 1	.2117	+: Household labor income, post-government income (both time periods). ∘: Education, household asset income (both time periods).
Factor 36	.1440	∘: Mindful diet, frequency of exercise, non-smoking.
Factor 35	-.1228	∘: Years of living with father alone, years of living with other relatives.
Factor 25	.1210	∘: Mother's age.
Factor 3	-.1134	+: Number of children, income from public transfers in both time periods.
Factor 32	.1130	∘: Education.
Factor 26	.1125	∘: Frequency of cultural activities, cinema visits.
Factor 38	-.1057	∘: Extraversion.
Factor 47	.0931	∘: Household and post-government income when the child was between 10 and 12 years old. -∘: Household and post-government income when the child was between 13 and 15 years old.
Factor 8	.0825	+: Importance: Being politically engaged, interest in politics. ∘: Mindful diet.
Factor 31	.0743	∘: Importance: To possess own house, time spent on garden work and repairs.
Factor 18	-.0713	+: Years of living with mother and partner. -∘: Years of living with both parents.
Factor 34	-.0693	∘: Risk attitude, openness.
Factor 44	-.0685	-∘: Income squared.
Factor 23	.0666	+: Importance of a happy marriage and having kids. ∘: Importance of being there for others.
Factor 41	-.0543	∘: Neuroticism.
Factor 13	.0528	+: Living in the new part of town.
Factor 10	-.0525	+: Migration background of father and mother.
Factor 19	.0406	+: Household income from private transfers.
Factor 6	-.0185	+: Years of living with mother alone. −: Years of living with both parents. ∘: Death of father. -∘: Importance of a happy marriage.
Factor 16	.0129	+: Living in a city.
Intercept	.0270	
N=667	R^2 =.22	Change: 22 % (R^2_{EduInc} =.18).

expressed by Factor 26. Factor 10, describing a parental migration background, appears less relevant for this sample as a comparably smaller coefficient than in the samples before is observed. A reason for this observation could be that the factor does not depend as strongly on education as before. As a consequence, the strong negative association in the samples before could be driven by lower education rather than structural disadvantages related to the parental migration background. Different cohort compositions, however, render a comparison speculative. Factor 19 loads solely on private transfers and is positively associated with the outcome like it is in the second cohort. A contradiction to the results for the second cohort is found for Factor 31, which has a moderately large positive coefficient. Although unrelated to the importance indicator of possessing a house, the factor of spending time on manual work on the house in the analysis of the second cohort has a negative association.

Remaining factors with enhanced importance include Factor 36, which loads highly on characteristics related to a health-oriented lifestyle as indicated by frequent exercising and a considered diet. There being a large coefficient for this factor, it carries some relevance and should be looked at closer. While literature emphasizes the positive correlation of exercising and the internal locus of control (Cobb-Clark et al., 2014), this factor does not connect the two. A health-conscious lifestyle may still be connected to the desire of preserving physical capabilities and avoiding diseases. As noted, the frequency of exercising has not influenced any relevant factor in the older cohort analyses. The relevance gained in this cohort analysis may be explained by societal changes in time use and attitudes.

A strong negative association can be found for a factor related to extraversion. The negative association is unexpected since this Big-5 dimension could be associated with sociability and thereby social capital. Another surprising observation is the negative relation of Factor 34, which loads highly on the mother's attitude towards risk and her openness to new experiences. The last relevant factor that is not linked to income or education is Factor 23 depicting a mother who values and cares for a stable family life as expressed by deeming a happy marriage, having kids and being there for others important. This factor shows a positive correlation to the dependent variable. These results indicate that the outcome is positively linked to maternal attitudes that are family-centered and traditional.

Owing to the increase in available variables, the ability to compare the results with those of the older cohorts is limited. One observation, however, is that the amount of explained variance is comparably high for this cohort. With few exceptions, however, the results for factors that are based on variables which are also available for previous cohorts are similar. For example, factors related to political interest

and to cultural activity maintain their important role. Parental age and time spent on manual work on the house, by contrast, do not play such a prominent role or a different one in the older cohorts. New is a factor that depicts a healthy or health-oriented lifestyle. Although this parameter has no substantial link to education or income, it shows a positive connection to the outcome. A factor related to further education, in particular occupational, has a parameter estimate of zero and does, hence, not play any role for this cohort.

Analysis of the third cohort with partner data:
Table 6.10 presents the results of the analysis of the same cohort when partner data are included. A notably sparser model emerges here and some factors appearing in the older cohorts are either not retrieved by Factor Analysis or deemed unimportant by Adaptive Lasso. Among the recurring ones are Factor 1, 4 and 22. Also Factor 9, related to parental age, and Factor 45, the caring factor, are known from previous analyses. Factor 5 and 12 are notable, because they have a relatively high importance and are not linked to education or income. Factor 5, which depicts the degree of the partner's involvement in housework and child care activities, is negatively associated and Factor 12, interpreted as the social participation factor, loads additionally on the partner's exercising frequency as well as the attendance of pop-culture events (cinema, pop concert, disco, sports event) and is positively related to the outcome. Factor 23 features a facet of family background in which none of the parents smokes and that is related to an introverted mother. The non-smoking aspect could point at a health-conscious lifestyle. The factor representing the frequency of attending religious events, indicating a religious family background, has decreased importance for this cohort compared to cohort 2.

Analysis of the fourth cohort without partner data:
As the first cohort, also the analysis of the fourth one is limited by a low number of cases. Nevertheless, the results resemble the ones of the older cohorts in that similar factors are deemed as relevant. Changes occur with regard to income-related factors. While the results of this cohort are shaped by numerous factors relating to income, the otherwise present factor loading on a high amount of public transfers and many children is not selected for this model.

Factor 2 and Factor 47 are recurring factors related mainly to demographic influences such as income, education and family type. Also, Factor 4 and Factor 19 appeared negatively in previous cohort analyses. The negative association for Factor 4 is, however, less severe than in the previous analyses. Factor 3, which typically loads solely on maternal trust, covers more aspects in this model for it additionally refers positively to the internal locus of control and future optimism

Table 6.10: Cohort 3 - With partner data:

Variable	Coefficient	Description
Factor 22	.2086	+: Education (both parents).
Factor 1	.1984	+: Various income variables related to the partner's labor income (both time periods). ∘: Partner's hours of work (both time periods), education (both parents), partner's interest in politics.
Factor 9	.1250	+: Parental age.
Factor 5	-.1011	+: Partner's hours of housework and child care. −: Partner's hours of work.
Factor 12	.0807	+: Frequency of cultural activities (both parents). ∘: Cinema visits, partner's exercising frequency.
Factor 4	-.0651	+: Number of children, income from public transfers in both time periods
Factor 23	.0518	+: Non-smoking (both parents). -∘: Extraversion.
Factor 55	.0475	∘: Cinema visits (both parents), partner's time spent on honorary posts.
Factor 45	.0339	+: Importance of being there for others (both parents). ∘: Partner's agreeableness.
Factor 3	.0264	+: Household asset income.
Factor 62	.0030	+: Death of father.
Factor 14	.0024	+: Frequency of religious activities (both parents). ∘: Partner's time spent on honorary posts.
Intercept	.0690	
N=532	R^2 =.21	Change: 23 % (R^2_{EduInc} =.17).

indicators. Its independence of income and education is preserved, though. Factors displaying a politically interested background (Factor 24) and frequent cultural activities (Factor 30) maintain their prominent role, which is indicated by a large coefficient. However, both factors are also related to maternal education in this cohort.

Analysis of the fourth cohort with partner data:

The results of the fourth cohort considering partner data are displayed in table 6.12. The observation of many factors with small coefficients indicates that the AdaLasso's factor selection has not eventuated in producing a sparse model here. One of the causes may be the large number of available regressors for this cohort. Since the SOEP started to gather more interesting variables over the years, a variety of different aspects can be accessed in this analysis. Connected to the increase in the number of factors is the rise in explained variance which jumps from 16 % in the analysis without partner data to 25 % in this sample.

Starting the analysis with the most important factors, Factor 1 and 33 depict the correlation patterns between the standard demographic indicators education

Table 6.11: Cohort 4 - Without partner data:

Variable	Coefficient	Description
Factor 2	.2781	+: Household labor income, post-government income. ∘: Education, frequency of eating out, age, household asset income. -∘: Number of years of living with the mother alone.
Factor 47	.1358	+: Education.
Factor 30	.1026	+: Frequency of cultural activities. ∘: Frequency of excursions, education, frequency of religious activities.
Factor 24	.1013	+: Importance of political engagement, interest in politics. ∘: Education, parental age.
Factor 3	.0902	+: All five trust variables. ∘: Internal locus of control, future optimism.
Factor 17	.0777	+: Living in a city
Factor 43	.0741	∘: Household income between the child's age 10 to 12.
Factor 19	.0692	−: Migration background of father and mother. ∘: Education.
Factor 5	.0396	+: Household asset income. ∘: Post-government income.
Factor 1	.0125	+ Income, income squared, hours of work. ∘: Education, interest in politics, membership in professional association. -∘: Hours of childcare, hours of housework.
Factor 6	.0054	+: Years of living with both parents. ∘: Importance of owning a house, non-smoking. -∘: Years of living with the mother alone/mother with partner.
Factor 4	-.0033	+: Number of children, income from public transfers.
Factor 31	.0009	+: Non-smoking. -∘: Extraversion.
Intercept	.2470	
N=470	R^2 =.16	Change: 0 % (R^2_{EduInc} =.16).

and income and the outcome. The results reveal that Factor 11, indicating a religious family background, once more plays an important role. In comparison to the analysis without partner data in which it was merged in the activity factor, the facet of a religious background appears on a separate factor in this analysis. Also, Factor 4 reappears in a similar version as in the results for older cohorts. A negative relation between a positive risk attitude of the mother and the outcome is observed by considering the results for Factor 59. This observation corroborates the results for cohort 3. The role of a high degree of future optimism on this factor is more challenging to interpret, as it is contradictory to the positive association observed in the analysis without partner data. With regards to a parental migration background, the strong negative links in older cohorts cannot be confirmed in this analysis. Although the coefficient is negative, it is comparably tiny.

The facet of a politically interested family background maintains its robust positive association. It is found on two factors in this cohort analysis, Factor 12 and Factor 17. Considering both together, the economic relevance of this dimension is once

Table 6.12: Cohort 4 - With partner data:

Variable	Coefficient	Description
Factor 1	.2208	+: Household labor income, post-government income. ∘: Education (both parents).
Factor 33	.1970	+: Education (both parents).
Factor 11	.1269	+: Frequency of religious activities (both parents).
Factor 4	-.0901	+: Income from public transfers. ∘: Number of children.
Factor 59	-.0774	+: Risk attitude. ∘: Future optimism.
Factor 30	.0767	+: Frequency of cultural activities (both parents).
Factor 17	.0729	+: Interest in politics. ∘: Partner's interest in politics, importance of political engagement.
Factor 12	.0638	+: Importance of political engagement (partner only). ∘: Importance of political engagement, partner's engagement in local initiatives, partner's interest in politics.
Factor 32	.0576	+: Importance of possessing an own house (both parents).
Factor 26	-.0548	+: Frequency of neighborly help (both parents). ∘: Frequency of meeting friends, relatives or neighbors.
Factor 16	-.0538	+: Partner's joy of work, coping with the circumstances, non-solitude.
Factor 15	.0433	+: Partner's openness, partner's curiosity.
Factor 62	-.0393	+: Partner's conscientiousness.
Factor 27	.0319	+: Household income from private transfers.
Factor 29	.0293	+: Non-smoking (both parents).
Factor 46	.0285	∘: Neuroticism (both parents), partner's openness.
Factor 63	-.0271	+: Partner's frequency of exerting honorary posts.
Factor 3	.0222	+: Trust variables.
Factor 10	-.0159	+: Migration background of father and mother.
Factor 24	.0132	+: Living in small town.
Factor 18	.0086	+: Partner puts importance on happy marriage, having kids and being there for others.
Factor 5	.0074	+: Household asset income.
Factor 68	.0060	∘: Post-government income when the child was between 10 and 12 years old.
Factor 70	.0044	∘: Partner has a considered diet.
Factor 51	.0010	+: Membership in professional association. ∘: Future optimism.
Intercept	.3070	
N=382	R^2 =.25	Change: 47 % (R^2_{EduInc} =.17).

more attested. An interesting observation for this cohort is that neither education nor income influences the two factors substantially.
One note concerns Factor 3, related to maternal trust variables, which shows a substantially smaller coefficient than in the older cohorts. A similar decrease can be observed for Factor 29, which is interpreted as holding a health-conscious attitude. Special to the factor in this cohort is the restriction to non-smoking; the variable referring to minding the diet does, despite its availability, not influence this factor.
A factor in this analysis, which is not present in older cohorts, is Factor 26. It describes a facet in which neighborly help and sociability are cared about. It is negatively associated to the school achievement of the child. Such a factor can be interpreted as having access to social capital, which lets the negative association seem surprising. An attempt to interpret this is to differentiate the type of social capital, which for this case can be viewed as limited to the close neighborhood, relatives and friends. However, instead of people who are close to the family anyway, it may rather be acquaintances located outside the daily life who transfer educational stimuli. While this explanation excludes a positive association, it does not take account of the negative association. One way to explain the negativity is to assume that such activities have little educational value and a high frequency of them leaves little time for more valuable activities.
An association, which is hard to explain, can be observed for Factor 16. It relates the partner's positive attitudes towards job and general circumstances negatively to the school achievement. The partner's urge for knowledge and new experiences, depicted by Factor 15, on the other hand, exhibits a comparably small, but positive association. One explanation refers to intergenerational transmission: The partner may transmit this attitude to the child and trigger curiosity, which could lead to an improved school achievement.

Before turning to the results of the robustness checks, it is worthwhile to examine the properties of the method on this data set in more detail by analyzing the relation between the actual and predicted dependent variable.

Diagnostics

This insertion deals with regression diagnostics. In spite of different models, the diagnostics are highly similar for each cohort so they are merged for this particular inspection. It is also sufficient to consider the samples without partner data, for the changes induced by considering partner data are not discernible. Moreover, the analyses are based on in-sample results, tacitly assuming that the results of the simulation study in regard to a low risk of overfitting apply.

Figure 6.2: Predicted vs. Actual

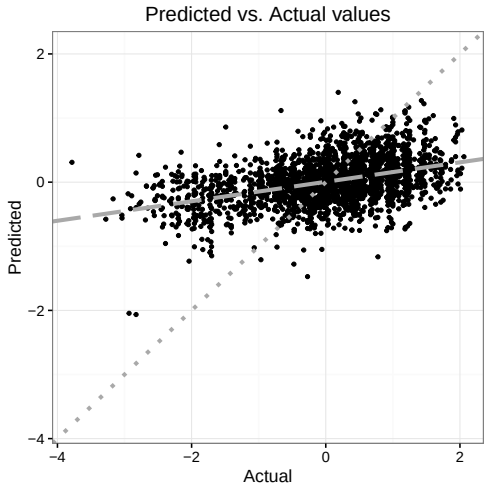
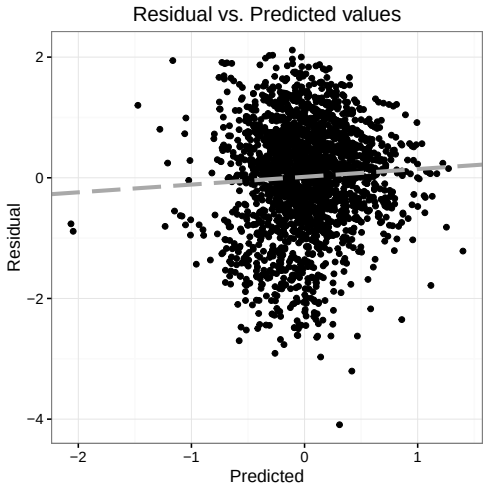


Figure 6.3: Residual vs. Predicted



In a first step, the match between the predicted dependent variable and the actual values is examined. Figure 6.2 displays the plot for which the cohort results have been merged. In addition to cloud of points, two lines are displayed: The dotted one can be considered the optimal line on which all points would be located if the model was perfectly accurate. The dashed line is the OLS-fitted line for a regression of the predicted on the actual values. The difference in slope and intercept between the two lines yields a quick impression of the performance. Since the underlying models were subject to regularization, the slope of the dashed line is likely smaller than it would be under OLS estimation.
By analyzing the plot, it is also visible that the model is unable to explain "extreme" values, but instead leads to a condensed point cloud around the center where most of the observations are located. The scaling of the axes has been chosen equally large, but while the actual values range over the complete interval, the predicted ones lie between 1 and -2, missing both the upper and lower end of the school achievement variable.

In this context, one can also consider the distribution of the residuals (defined as the actual value minus the predicted value) over the predicted values. This relation is plotted in figure 6.3, again including the line of best fit. Since it is difficult to display a point cloud, the graphical impression of increasingly negative residuals for larger predicted values is deceiving. Moreover, neither non-linearity nor heteroscedastic tendencies can be discerned in this relation.

6.2.2 Robustness check I: School achievement, alternative scheme

This section presents the results for the school achievement indicator using the alternative transformation scheme. To keep this section concise, the results of the robustness checks for Cohort 3 are shown as a representative example.

Analysis of the third cohort without partner data:
Under the modified transformation scheme for grades, the results in table 6.13 emerge. Changing the calculation of the dependent variable brings about only

Table 6.13: Robustness check: Cohort 3 - Without partner data:

Variable	Coefficient	Description
Factor 1	.2080	As before.
Factor 36	.1239	As before.
Factor 3	-.1209	As before.
Factor 32	.1193	As before.
Factor 26	.1186	As before.
Factor 25	.1023	As before.
Factor 35	-.0965	As before.
Factor 38	-.0925	As before.
Factor 8	.0889	As before.
Factor 47	.0854	As before.
Factor 34	-.0804	As before.
Factor 31	.0765	As before.
Factor 18	-.0749	As before.
Factor 41	-.0679	As before.
Factor 44	-.0642	As before.
Factor 23	.0583	As before.
Factor 10	-.0539	As before.
Factor 13	.0532	As before.
Factor 29	-.0226	+: Internal locus of control. ∘: Conscientiousness, agreeableness.
Factor 19	.0018	As before.
Intercept	.0000	
N=667	R^2 =.21	Change: 20 % (R^2_{EduInc} =.17).

slight changes. In comparison to the results for the original school achievement measure, Factor 6 and 16 are no longer deemed relevant enough. Instead Factor 29 appears, which describes a high internal locus of control, conscientiousness and agreeableness. Although small, the negative coefficient is unexpected because univariate regressions show a robust positive association between the maternal internal locus of control and the child's school achievement. The matter is different

for the two Big-5 Factors that also have a bearing on this factor. They typically have a weak negative relation which might drive the factor stronger than the internal locus of control.
The remaining results are close to the ones for the normal transformation scheme. If changes occur, they mostly concern the coefficient size, but the same tendencies predominate.

Analysis of the third cohort with partner data:
Using the alternatively generated endogenous variable in the analysis with partner data, the results presented in table 6.14 emerge. Compared to the model of the normal endogenous variable, there are five new factors in this model, whereof only two are of deeper interest, however. One is principally driven by the locus of control, which now loads on both parents' variables, but no longer large enough on conscientiousness and agreeableness as before. The other factor of interest relates to parental openness. Both factors are weakly negatively linked to the outcome. Except for Factor 41, the coefficients of the additional factors are comparably small and therefore less important. Since the statistical method is somewhat unstable regarding the choice of factors with small coefficients, this observation can be attributed to chance.

Table 6.14: Robustness check: Cohort 3 - With partner data:

Variable	Coefficient	Description
Factor 22	.2533	As before.
Factor 1	.2226	As before.
Factor 9	.1335	As before.
Factor 4	-.1039	As before.
Factor 12	.1028	As before.
Factor 5	-.0794	As before.
Factor 45	.0596	As before.
Factor 41	-.0559	+: Openness (both parents).
Factor 23	.0542	As before.
Factor 14	.0481	As before.
Factor 55	.0481	As before.
Factor 16	.0270	+: Living in eastern Germany.
Factor 15	-.0224	+: Internal locus of control (both parents).
Factor 3	.0218	As before.
Factor 62	.0121	As before.
Factor 53	-.0119	+: Labor income squared.
Factor 39	.0042	+: Living in northern Germany.
Intercept	.0330	
N=532	R^2 =.24	Change: 23 % (R^2_{EduInc} =.18).

When the relative importance of the central factors is changed, on the other hand, the coefficient sizes and changes of the retained factors are of interest. This is, for example the case for Factor 4 and 5, which swap positions. Factor 4, loading on the number of children and the amount of public transfers, has a more negative coefficient, while the one of Factor 5, loading on the partner's time spent on housework, decreases in absolute size. When the dependent variable under the standard transformation scheme is interpreted as punishing lower school achievements stronger, this result indicates that the time use of a partner could be a better indicator to describe the lower end of the dependent variable's distribution. Further noticeable is the increase in the model's coefficient of determination, which is .24 compared to .21. While this can be partly attributed to the five additional factors in the model, another reason may be the 0.5 units higher coefficient of Factor 22. Both arguments, as well as the extra factors in the model, point in the direction of a generally smaller shrinkage.

6.2.3 Robustness check II: Test scores

The results for the models explaining test scores are presented completely, which yields four different models in total. In general, using test scores as the dependent variable tends to engender smaller models. That means, fewer family background factors are deemed as relevant in terms of a non-zero coefficient. The set of selected factors bears, nevertheless, similarities to the most important factors for the school achievement.

Analysis of the first cohort without partner data:

The results for the first cohort, where partner data is excluded, are shown in table 6.15. A first look on the selected factors reveals a similar choice as in the analysis of school achievement measures. Given that this analysis excludes partner data, a notable difference is that the two most important factors do not contain the factor which relates to household income and thereby hint at the existence of an employed partner. This factor's relative importance is lower for this analysis. Factor 5, collecting social interest, engagement, activity and closeness to education plays a more important role here. This factor is based on many characteristics which also hold a beneficial association to the grade average. A negative sign is found for the factor relating to a parental migration background.

Further important associations originate from education-related factors, among which formal education, but also interest and participation in further education are found. Factor 1, in particular, appears similarly in the analysis of the second cohort's school achievement where the factor is interpreted as undertaking further

education of one's own accord and therefore linked to the appreciation of education and aspirations. Factor 22, on the other hand, displays a facet in which further education was not done out of an intrinsic motivation but rather to meet the needs of the job. Its association is strongly negative.

When it comes to other time use factors, Factor 16 is also reappearing. This factor indicates a religious family background; for this cohort, it is also connected to exerting honorary posts. Taken together the factor could be interpreted as community work in a religious environment. It could display various (societal) ambitions but, owing to the community component, also be linked to increased social capital. The facet of interest and participation in politics remains an important dimension of family background also in this cohort. Cultural activities show also positive links, as indicated by Factor 5 and Factor 19. The latter can be interpreted as social activity factor for it bundles indicators of activity and sociability.

A further interesting result is that two factors related to personality traits exhibit a relevant association in this analysis: Factor 32, loading mainly on maternal agreeableness, and Factor 20, mainly driven by a high internal locus of control and a low degree of neuroticism – both relate positively to the outcome.

Table 6.15: Robustness check: Cohort 1 Test Scores - Without partner data

Variable	Coefficient	Description
Factor 5	.1825	∘: Interest in politics, frequency of undertaking cultural activities/taking part in local initiatives/exerting honorary posts, education, participation in further education.
Factor 4	-.1801	+: Migration background of father and mother.
Factor 2	.1474	+: Household labor income, post-government income.
Factor 16	.1086	∘: Frequency of religious activities/exerting honorary posts.
Factor 32	.0983	∘: Agreeableness.
Factor 6	.0958	+: Years of living with both parents.
Factor 1	.0861	+: Further education for promotion. −: No interest in further education. ∘: General further education.
Factor 22	-.0829	∘: Further education in order to remain in the job.
Factor 13	-.0803	+: Living in a rural area. −: Living in a small town.
Factor 17	.0705	+: Parental age. -∘: Hours of child care.
Factor 8	-.0652	+: Income from public transfers, number of children.
Factor 20	.0525	+: Internal locus of control. -∘: Neuroticism.
Factor 19	.0466	∘: Frequency of exercising/undertaking cultural activities/visiting the cinema, sociability.
Factor 30	-.0097	∘: Sociability, extraversion.
Intercept	-.0200	
N=845	R^2 =.21	Change: 50 % (R^2_{EduInc} =.14).

Comparing the results to the ones for measures of school achievement, the model

is able to explain about the same level of variance. Moreover, many factors that have a relevant link to the school achievement also play a pronounced role here. For some, however, the relative importance differs.

Analysis of the first cohort with partner data:

Table 6.16 shows the results for same cohort when partner data are regarded. In comparison to the model calibrated on the sample without partner data, fewer dimensions drive these results. The main factors relate to a parental migration background, their income and education, and finally their interest and motivation in further education. Moreover, the factor describing participation in (high) cultural activities, in this cohort independent of education, exhibits a positive relation to the outcome once again.

While Factor 14, interpreted as appreciating education, also features a similar positive association as in the other cohorts, the facet indicating political interest and engagement plays a smaller role in the present model. A religious family background, which in the previous cohort is also connected to community work, does not play a role in this analysis. Slightly negative results can be found for Factor 16, which describes a partner who takes care of tasks at home but shows little activity on the labor market. The model is not able to improve the baseline model substantially. On the other hand, the coefficient of determination is relatively high.

Table 6.16: Robustness check: Cohort 1 Test Scores - With partner data

Variable	Coefficient	Description
Factor 7	-.1748	+: Migration background of father and mother.
Factor 5	.1622	+: Education (both parents). ∘: General further education.
Factor 1	.1313	+: Various income variables related to the partner's labor income. ∘: Father's hours of work, education (both parents).
Factor 14	.0891	+: Interest in further education.
Factor 8	.0885	+: Frequency of taking part in cultural activities (both parents).
Factor 4	.0882	+: Partner takes part in further education for job promotion. −: Partner has no interest in further education.
Factor 16	-.0272	+: Partner's hours of housework. −: Partner's hours of work. ∘: Partner's hours of child care.
Factor 20	.0217	∘: Interest in politics (both parents), partner's frequency of exerting honorary posts and taking part in local initiatives.
Intercept	.0000	
N=682	R^2 =.20	Change: 11 % (R^2_{EduInc} =.18).

Analysis of the second cohort without partner data:

Table 6.17 lists the results for the second cohort excluding partner data.

Table 6.17: Robustness check: Cohort 2 Test Scores - Without partner data

Variable	Coefficient	Description
Factor 2	.3268	+: Post-government income, household labor income. ∘: Interest in politics, household asset income, education, age, frequency of exercise, importance of political engagement.
Factor 39	.1646	∘: Post-government income while the child was between 10 and 12 years old.
Factor 24	.1544	+: Interest in politics, importance of political engagement. ∘: Education, age.
Factor 18	-.1178	+: Migration background of father and mother. -∘: Education.
Factor 29	.1146	∘: Parents are non-smokers. -∘: Extraversion.
Factor 1	.0787	+: Labor income. ∘: Hours of work, education, frequency of cultural activities, household labor income. -∘: Hours of housework/childcare.
Factor 47	.0527	No loading large enough.
Factor 16	.0463	+: Living in a city.
Factor 3	-.0457	+: Income from public transfers, number of children. ∘: Hours of housework.
Factor 27	-.0387	+: Internal locus of control. ∘: Future optimism, approval of "I have confidence in the future". -∘: Neuroticism.
Factor 5	.0328	+: Years of living with both parents. ∘: Importance of happy marriage.
Intercept	.0490	
N=558	R^2 =.15	Change: 15 % (R^2_{EduInc} =.13).

In this analysis Factor 2 stands out since it encompasses several characteristics which have exhibited beneficial associations in other cohorts. Primarily loading on a high income, which implicitly points to an employed partner in the household, this result does not confirm the tendency of the first cohort's results. Instead, these results are more in line with the ones of the analyses of school achievement indicators. Variables driving Factor 2 also appear in other factors, though. Factor 24, for example, depicts a household with higher educated and older parents who are politically interested and engaged. The stronger pecuniary influences in this cohort analysis are also described by Factor 39 which loads on an income variable, and Factor 1, which depicts a higher educated, working mother who frequently participates in cultural events.

Factor 29 is a factor, which describes introverted, and by non-smoking, possibly health-conscious parents. While the health aspect also appears in other analyses, the observed coefficient is larger for this analysis than usually. Another striking change in the coefficient size occurs for Factor 3, as its negative association is clearly smaller than in the other analyses. It is the recurring factor relating to income from public transfers and a high number of children in the household, sometimes

also including hours of housework and child care. A second contradiction to the results of the first cohort is observed for Factor 27, which loads mainly on the internal locus of control, but exhibits a weakly negative relation to the outcome. This stands in contrast to the positive association observed for Factor 20 in the analysis of the older cohort.

Analysis of the second cohort with partner data:
The results for this analysis are presented in table 6.18. They are special in that the selected model is comparably small. Furthermore, it performs worse than the baseline model from the perspective of explained variance. This indicates a case of underfitting and therefore makes a case against the Adaptive Lasso as a factor selection method, which otherwise performed satisfactorily. The relatively small sample in this analysis may be one reason for the observed problems. On the other hand, such a result could not be observed in other cohorts with similar or even smaller sample sizes. An unfortunate CV split can also be excluded, as similar results emerged for different (random) draws. When it comes to the factors selected, the ones that are most relevant in the other cohorts are chosen. In addition, the facet of non-smoking, which is interpreted as health-consciousness, exhibits increased relevance here. It is also the only dimension that refers to habits. The remaining factors are largely related to demographic aspects.

Table 6.18: Robustness check: Cohort 2 Test Scores - With partner data

Variable	Coefficient	Description
Factor 1	.1285	+: Partner's labor income, household labor income. ∘: Education (both parents).
Factor 12	-.0977	+: Parental migration background.
Factor 38	.0830	+: Both parents are non-smokers.
Factor 39	.0538	∘: Parental education.
Factor 5	-.0525	+: Number of kids, income from public transfers.
Intercept	.0760	
N=424	R^2 =.09	Change: -40 % (R^2_{EduInc} =.15).

6.3 Conclusion

The empirical approach presented in the previous section revolves around the discovery of family background dimensions and their associations to the child's school achievement and its cognitive skills. Fundamental to this investigation are family characteristics which describe the parental environment. Thus, the pool of characteristics does not contain individual determinants, such as the gender

of the child – its manifestation is independent of family background. Following the theoretical considerations about latent dimensions of the family background, Factor Analysis intended to disclose them is applied. The main drawback of such an unsupervised procedure is its independence of the outcome. In order to separate the important from the less important factors, it is necessary to evaluate each factor's association with the outcome. In this thesis, this is achieved by applying a regularization technique to the extracted factors: Using the Adaptive Lasso with 7-fold Cross-Validation selects factors by pulling the coefficients of the unimportant factors towards zero.

Comparing the theoretical expectations with the results of the empirical analysis, one note concerns factors with small eigenvalues. They often rely on only few variables, which means that the interpretation of these factors is based on one or two variables. The premise to yield facets of family background whose interpretation could be derived from several variables is partly undermined for this reason. Although such distinctness is within the bounds of possibility, there may also be other reasons leading to this observation. Oblique rotations which taper the factors to only their most correlated variables could be one. However, similar situations also occur for factors with small eigenvalues when the factors are Varimax rotated. With regard to using the Adaptive Lasso as a factor selection technique, the overall performance is satisfying, even though two drawbacks occurred in this analysis. Firstly, changes of the random seed on which Cross-Validation depends indicate somewhat unstable results concerning the selection of factors with small coefficients. However, since coefficient size is directly related to the importance of a factor, this instability pertains merely to the less interesting factors. Secondly, the method failed in one instance where it produced a too sparse model.

Despite these issues, the results provide some relevant insights. While they suggest that factors related to demographic indicators, like parental education and household labor income, have the highest relevance, there are other dimensions which contribute to a description of family background. Table 6.19 provides a compact overview of the core insights.

Since they play a weighty role in many cohort analyses, three factors stand out. Those apply often equally well to either the mother, her partner or both, as the data do not support a clear distinction at this point.

One family background dimension relates to frequent cultural activities that may be considered high culture, e.g. concerts, theaters and lectures. In some samples, the factor is also connected to cultural activities which could be considered pop culture, such as visits to cinemas, discos and sports events. Typically, the factor's

Table 6.19: Overview of associations

Outcome:	Grade average							Test Scores			
Cohort:	1wo	2wo	2w	3wo	3w	4wo	4w	1wo	1w	2wo	2w
Factor/Variables:											
Household income from labor	+	+	+	+	+	+	+	+	+	+	+
Receipt of public transfers, number of children	*o*	−	−	−	−	*o*	−	−	*o*	−	−
Parental education	+	+	+	+	+	+	+	+	+	+	+
Parental migration background	*o*	−	−	−	*o*	−	*o*	−	−	−	−
Years of living in classic family	+	+	+	+	*o*	+	*o*	+	*o*	+	*o*
Parental age	+	*o*	*o*	+	+	+	+	+	*o*	+	*o*
Political interest and derived characteristics	+	+	+	+	+	+	+	+	+	+	*o*
Frequent (high) culture activities	+	+	+	+	+	+	+	+	+	+	*o*
Religiousness of family	*o*	*o*	+	*o*	*o*	+	+	+	*o*	*o*	*o*
Frequent exercising activity	+	*o*	*o*	+	+	*o*	*o*	+	*o*	+	*o*
Time spent on garden work and repairs	*o*	−	−	+	*o*	*o*	*o*	*o*	*o*	*o*	*o*
Time spent on housework and childcare	na	−	*o*	*o*	−	*o*	*o*	*o*	−	−	*o*
Variables related to voluntary further education	+	+	+	*o*	*o*	na	na	+	+	na	na
Maternal trust in other people	na	+	+	na	na	+	*o*	na	na	*o*	*o*
Positive risk attitude and openness	na	na	na	−	*o*	*o*	−	na	na	*o*	*o*
Health-conscious lifestyle	na	na	na	+	+	*o*	+	na	na	+	+
Importance of income and job	−	na	na	*o*	*o*	*o*	*o*	*o*	*o*	*o*	*o*
Importance of a happy marriage and having kids	na	na	na	+	*o*	*o*	*o*	*o*	*o*	*o*	*o*
Extraversion and sociability	na	*o*	*o*	−	−	*o*	*o*	*o*	*o*	*o*	*o*
Internal locus of control	na	na	na	*o*	*o*	+	*o*	+	*o*	−	*o*
Neuroticism	na	*o*	*o*	−	*o*	*o*	+	−	*o*	+	*o*
Agreeableness	na	*o*	*o*	*o*	+	*o*	*o*	+	*o*	*o*	*o*

Notes: Factors do not always completely correspond across different cohorts, hence the reference is sometimes to the most prominent superordinate variable(s). Also, differences between maternal and paternal variables are omitted, if not especially noticeable. "wo" stands for analyses without, "w" for with partner data. +/− indicates a positive/negative association of a factor in a cohort analysis, o displays no or a negligibly small association, while na fills a cell, when the manifest variables needed for the factor are not included in a cohort analysis. To be concise, association strength is not noted at this stage and factors with small coefficients are left out.

importance is stronger the more it is connected to the first type of cultural activity. It carries high relevance in almost all cohort analyses, and is interpreted as active participation in society, more specifically, cultural life. From a sociological point of view, this might express the possession of cultural capital. Although one can argue that participation in cultural life is costly and therefore linked to income, Factor Analyses often separates the two. The implication is that parents with lower income also conduct such activities. On the other hand, the imprecise definition of cultural activity hampers an accurate interpretation. Also, linking this facet to a child's school success depends on the assumption of the cultural events' manifestation. If suitable, one could argue for a knowledge gain for the child that results from this activity, but this interpretation additionally assumes that the child takes part in these activities. Interpreting a culturally active family environment as yielding a general cultural stimulus, however, does not require making such assumptions.

Another important dimension is related to the interest and involvement in politics. Depending on the sample, a particular interest in politics, considering political influence important or, in some cases, active participation in local (political) initiatives characterizes this facet. It can be interpreted as active societal participation with the objective of enforcing certain political and societal ideas. In this sense, it can be linked to aspirations as well as the desire to influence and control the surrounding circumstances. On the contrary, there seems to be no connection to the internal locus of control, even though it would match well with this interpretation. While there are some exceptions, this facet is often also related to parental education. And although the influence of education on such a factor is often smaller than for politics variables, the factor cannot be considered completely independent of it. There are several possibilities to link this facet to a child's school success: The high degree of aspirations, expectations and involvement that parents show for society may reflect an ambitious attitude, which they also hold for the child's development and school achievement. Moreover, by exhibiting political interest, information about societal developments is possessed. This knowledge might improve decisions concerning the child's education. In addition, one could hypothesize that such parents also hold an above-average degree of information about school-related issues, for instance, by involvement in classroom matters such as parent-teacher meetings. This could benefit the child directly. Yet another way to interpret the results is to refer to the intergenerational transmission of interests. One can hypothesize that the child, like its parents, is interested in politics and aware of societal developments which could have a positive influence on the outcome. Connected to this point, political interest may be expressed in ways that influence the school outcome, for instance through political discussions

within the family.

Interest in further education, particularly the contrast between general disinterest in further education and participation in it, forms the content of the third important facet. Further education directed at occupational goals, such as continuance or promotion, does not exhibit a positive link, though. This is documented in the analysis of cohort 3 in which the corresponding factor is irrelevant. An interpretation, consistent with the observed results, is to consider the factor related to (general) further education as a proxy variable for the attitude towards education. If this attitude reflected a general view on educational achievements, it could influence the parental appreciation of a child's efforts and achievements in school. Also, it might affect the degree of encouragement of school matters. On the other hand, further education with occupational goals is not stringently caused by one's own initiative and may on this account not describe a positive attitude towards education clearly enough. From the perspective of the intergenerational transmission of attitudes, a further approach to explain the results can be derived. If a parent holds a positive view towards education, the child could take it up, leading to more efforts in school. Similarly, an actively learning parent could be an education-enhancing role model. The partner's thirst for knowledge and new experiences, which plays a role in the results for cohort 4, may be connected to this facet. Described by openness and curiosity, this factor may benefit the child's educational achievement if the characteristics are intergenerationally transmitted. However, the finding is partly ambiguous since the Big-5 factor openness is not always positively related to the outcome as the results for cohort 3 indicate. A comparison between these results is limited, however, since the variable, which indicates a curious person, is not available for cohort 3.

Another facet showing robust positive associations to the outcome is when the mother states having a high degree of trust to other people. In some cases, this facet is also driven by the attitude of future optimism and a high degree of an internal locus of control. In this analysis, the dimension is linked to social capital, which is easier to generate when one trusts people. This component could be complemented by trust in the future, a general optimism possibly ignited by the belief in one's own abilities. The potential benefits of social capital are described in the theory section. An explanation for the positive link of future optimism and trust in one's own abilities is given if one follows the theory of intergenerational transmission. Such characteristics may increase the child's resilience towards failures.

Negative factors in the results are often linked to family homes in which the parents separated at an early child's age. The lack of a mother's partner and their

respective income is associated with a significantly worse outcome for the child. Income by private transfers, which consists mostly of alimony payments, also shows a negative but convex relation to the outcome. Since this structural disadvantage is unlikely offset by high values on beneficial softer indicators, it is particularly important to be addressed by policy makers. A second negative dimension is described by a family environment with many children and a high receipt of public transfers. The associated coefficient is not always large in absolute terms, but the link is stable. This association is explained by interpreting public transfers as unemployment benefits, in which case either effects of parental unemployment or a latent variable influencing both the probability of unemployment and the child's school success causes this relation. Interpreting the transfers as mainly consisting of child benefits, the argument of less resources per child can be brought forward. Parental age, which has its own factor in some analyses, is positively connected to the outcome in a few cases. This could reflect experience in raising a child or, when connected to income variables, hint at a richer endowment or, when on a factor with education, emphasize the positive correlation between education and the age of giving birth.

If a child experiences a facet of family background in which parents frequently help their neighbors and meet with friends and relatives, a worse school outcome is observed in some of the samples. Two explanations are conceivable for this observation: Despite the presumed positive effects of social capital, it is conceivable that this kind of social capital is not valuable enough, as it is restricted to the close social environment. In addition, if the time spent on such activities has no positive impact, it reduces the disposable time for other activities that could have a beneficial impact.

Further negative links arise in some samples for time spent on housework and childcare, but not on work. In some cases, this factor emerges only for the partner. An interpretation of this relation is that parents who work transmit a specific set of values to their child.

For other facets, the results are more variable. A religious family background, identified by frequently attending religious events, is interpreted as a family background holding traditional norms and ethics. It plays a relevant positive role in many cohorts, but there are exceptions, such as the first and the third cohort. Similarly, a factor interpreted as pursuing a health-conscious lifestyle, marked by making conscious dietary choices and not smoking, has sample-dependent associations. In most cases, however, a positive link between this facet and the

outcome can be detected. Such results are also observed for facets explicitly relating to exercising frequency.

Parental personality traits play an ambiguous role in the results, as no trait shows a clear direction of association and the relevance fluctuates, too. This result is good news, since it gives an indication that the considered outcomes do not rely on presumably immutable parental traits.

Concerning the robustness checks, the results change little under a modified version of the standardized grade average. Although additional factors are retained, these are mainly ones with small coefficients. Their inclusion can hence also be caused by the statistical method which is somewhat unstable in the selection of less relevant factors. Other changes regard the relative ranking of the factors and their coefficient size. Neither aspect can be viewed as systematic, however. Concerning test scores, the models tend to be substantially sparser than the models for measures of school achievement. In spite of fewer retained factors, the share of explained variance is at about the same level with the exception of one cohort analysis. This observation is interpreted as fewer family factors playing a role for this outcome. One reason may be a higher objectivity of test scores, as they are less dependent on subjective assessments. For example, geographical indicators, which occasionally appear in the models for the school outcomes, play no role in this measure. In conclusion, the most important factors for school achievement are also the most important ones for test scores, while the less important ones play no role for the latter.

Chapter 7

Summary and conclusions

The starting point of this dissertation was to explore the associations of a child's family background with its educational achievement. This investigation is motivated by evidence that a substantial share of the variation in school success can be traced back to the family home. Since the term 'family home' could encompass many aspects, substantiating it has been of interest. In empirical studies, family background is often reduced to demographic indicators such as parental income and education, because they correlate substantially with the child's school achievement. As discussed in the introduction, however, it is questionable whether these indicators suffice to capture all important aspects of family background. Instead, one can argue that the parental mindset, particularly towards education, plays an important role. A mindset can find expression in various dimensions, for example certain attitudes and time allocation but also customs such as involvement in the child's school matters. Some of its expressions may have a decisive influence on a child's school achievement. Since there are reasons why to believe that this parental mindset could develop independently of the named demographic characteristics, this dissertation aimed to find such facets and examine them.

As it is unclear which dimensions of the parental mindset are of interest and how this could be operationalized, one approach is to consider the social environment, the milieu, of the family. Milieu theories state that there are groups in society which are characterized by similar life circumstances and attitudes. The observation that people who live in similar circumstances influence each other through social interaction and attempt to demarcate their group from other groups serves as a link to the mindset, since it could cover dimensions which are relevant for the educational achievement of the child, for instance ambitions and aspirations.

While the concept of milieus is appealing in theory, practical milieu concepts entail drawbacks for the purpose of this dissertation. The main reason is their inherent function of providing a model for the social structure of a society. Of interest in this context are, however, family environment characteristics which are connected to

the child's school achievement. There is no guarantee this information is included in existing milieu concepts. In fact, milieu concepts often use income as a classifier, which would contradict the premise of this investigation. In addition to other, less severe drawbacks, this motivates an extension of the milieu approach.
Its starting point was to interpret the milieu idea in a latent variable model. In this framework, the milieus act as the (predefined) latent variables, which influence the set of observed variables. These observed variables include all characteristics in which similar attitudes and life circumstances as well as the child's school success are reflected. In this dissertation, this model is generalized by detaching the latent variables from being fixed milieu indicators. Instead of a qualitative definition, the notion of similarities in social environments is exploited by deriving their content from data, i.e. through patterns in the set of observed variables. The latent variables are, for that reason, no longer called milieus, but are instead referred to as facets or dimensions of family background.
In order to derive their content, the set of observed variables must contain sufficient information. The data set at hand offers a large choice of different parental characteristics, but they are restricted to those that a child may, in some way, be exposed to. The selected variables are grouped into four categories: parental personality traits, parental attitudes, parental time use indicators and family demographic characteristics. While the number of personality traits and demographic characteristics is limited, data on attitudes and time uses can be plentiful, especially when gathered separately for both parents.

For these reasons, the empirical approach is based on a large number of variables, between which clusters of moderate to high correlations exist. Using these characteristics unfiltered in a regression is prone to lead to a model which is overly optimistic and too difficult to interpret on grounds of high dimensionality, expressed by a multitude of variables. In this context, the value of sparse models concerning interpretability, prediction and generalization ability is highlighted. In reaction to the data structure, several refined methods are suggested and described in detail. Their approach to dimensional reduction classifies these methods. Those forming indices, which are linear combinations of the original variables, is one group; the second one is regularization methods, which embark from the full linear model and drag estimated coefficients towards zero to reduce variance. Methods forming indices are additionally separated into supervised types, i.e. regarding the dependent variable in the construction of indices, and unsupervised types, i.e. creating indices independent of the outcome.
Based on theoretical considerations, predictions of the performance are given. While the evaluation of some methods can be backed up with findings from the literature, other proposed methods, especially those related to Factor Analysis,

have rarely attracted attention in the literature. Considering Factor Regression methods in this dissertation is justified, however, by its particular purpose: The latent variable model notion corresponds best to its underlying model. This is also the reason why a method combining Factor Regression with regularization techniques is proposed. In the decision for the applied method, the aspects of yielding an interpretable model and the performance in terms of fit have to be regarded. While the first property can be theoretically addressed, a simulation study is necessary to be able to assess the accuracy of models in different data environments.

The simulation study evaluates the proposed methods and their model selection criteria within two model DGPs, a latent variable model and a regression model. Their model parameters vary over six scenarios, which allow an examination of performance differences in varying data structures. While the results in the regression model are often stable, the scenario can be influential for the results in the latent variable model specification. There are, however, generally valid observations. Certain methods are prone to overfitting in about the same degree as OLS, which serves as a reference point in this simulation. This concerns Partial Least Squares and Factor Regression, when Iterated Principal Factors is used as an extraction and sole dimension reduction method. These methods perform exceptionally well on training data but tend to overfit the data, which is apparent when predictions on hold-out data are evaluated. Partial Least Squares occasionally suffers, in addition, from bad estimates which drive the mean squared error substantially upwards. The results of the simulation also point out the dangers of using heuristic criteria for factor selection, such as the Kaiser criterion, which selects all factors with a large enough eigenvalue. Such methods fail in the scenario where variable selection is aggravated by dispersed loadings of the underlying factor model. While this risk is unavoidable for heuristic methods in such a scenario, the bad performance in the low correlation scenario is connected to the inherent variance deduction in Factor Analysis. Principal Component Regression is not concerned in this case.
Moreover, the simulation evaluates the performance of popular stepwise selection methods. It is mediocre in the latent variable model specification, and only slightly better in the regression model. While one can argue that stepwise selection methods are particularly affected by multicollinear data structures, the approach also does not work out well when applied to PCR, where the components are pairwise orthogonal. In most cases, this algorithm performs worse than eigenvalue-based component selection. Considering the complete set of results in the latent variable model specification, the top performing methods are the Lasso, the Elastic Net, Principal Covariates Regression and two types of Factor Regression with regularization (Lasso, Adaptive Lasso). They yield relatively stable results with low

error across the different scenarios. They also perform well in the basic scenario, which is, owing to its data structure, the most important one for the decision. The Adaptive Lasso typically performs worse than both the Lasso and the Elastic Net, while the remaining methods either do not perform particularly well or have troubles with certain data structures.

In the Regression Model specification, shrinkage methods perform much better than methods which create indices. The Lasso and the Elastic Net score highest across the scenarios. Among index models, Factor Regression types with shrinkage perform best, but also their performance ranks behind variable shrinkage methods and sometimes even stepwise selection methods. The results for Principal Covariates Regression are substantially worse than in the latent variable model, since the quality of out-of-sample predictions varies strongly.

Based on the results of the simulation study and the theoretical considerations, Factor Regression methods using Lasso or AdaLasso regularization for factor selection are considered suitable. They have the superior interpretability of rotated factors, while the disadvantages of the unsupervised part are addressed by regularization. Although they sometimes perform worse than the Elastic Net, in particular, in the regression model specification, they constitute a good compromise. The performance differences between the two methods in the simulation study are small, but there is a practical difference: Factor selection by Lasso often yields larger models in terms of retained factors than selecting factors by Adaptive Lasso. These additional factors often have small coefficients. As the results of the simulation study indicate, one could work with either model since factors with small coefficients do not play a substantial role in explaining the outcome. However, the results of the Adaptive Lasso are more parsimonious when it comes to looking at and interpreting the results. For the analysis of empirical data, the SOEP data set has been chosen because it provides various rich data for this topic. For reasons of data availability and to allow for possible changes over time, the data are split into cohorts according to certain birth year ranges. The analyses are divided into those that disregard partner data, even if available, and those that regard partner data. While the first type maximizes the number of observations, the second contains more information. Due to this twofold analysis, the child data are only matched to maternal data – the added value of matching them also to the paternal data is low.

The main measure of educational success is defined by a grade average of the three most important subjects. Owing to different school types, this requires the definition of a transformation scheme. Such a scheme can be found in German Education Acts. Since it is to a certain degree arbitrary, the results under this scheme are compared to the ones using a modified scheme. The differences are,

however, of minor relevance.
In general, measuring human capital by school achievement is not undisputed, however. The main critique is that this measure disregards the quality of schooling. While differences in quality gain importance for cross-country comparisons, smaller qualitative differences can also be found within Germany. To examine the robustness of the results of the school achievement indicators, cognitive skills as measured by test scores are analyzed. The corresponding results are robust in the sense that the same factors appear and play the main role for the dependent variable based on grades. In general, however, the models for this outcome tend to be smaller in terms of retained factors, while the degree of explained variance is similar apart from one cohort. Owing to different sets of variables and varying sample sizes, the results between cohorts are not completely comparable. The bigger picture shows, however, that the existence of a working partner or husband and the respective income has a stable and often highly relevant, positive association. In most cases, a differentiation between the income types household asset income, household labor income and post-government income is not necessary as these variables usually all contribute to the same factor.

Concerning facets of family background which are not mainly driven by education and income, the results suggest that frequent parental visits to cultural events, high culture ones in particular, are robustly positively linked to the outcome. By interpreting this facet as societal participation in terms of going out and receiving cultural knowledge, it can be linked to a family background that provides cultural stimuli for the child. Another relevant dimension of family background relates to political interest; either expressed by declaring interest for it, considering political influence important, or by taking part in political initiatives. The facet's robust positive relation to the outcome is explained by interpreting the factor as a high parental motivation to influence the surrounding environment, i.e. having aspirations and ambitions. A similar attitude can also be adopted when it comes to involvement in the child's school matters. Moreover, high political interest points to knowledge of societal developments, which can include the importance of education for later success in life. Although this aspect shares a factor with education in many cases, a close relationship is not observable. These results, hence, indicate additional heterogeneity in levels of education which interest in politics takes account of. A third important dimension describes a family background open to further education. It is interpreted as indicating appreciation of education and the recognition of its value. Moreover, occupationally related further education may indicate ambitions and, for certain sub groups, also point toward the belief or even realization of social mobility. Apart from the mindset, role model arguments support the positive link, too. When parents engage in learning, the child may

take up this behavior. The positive relation of a facet described by the partner's high curiosity and openness to new experiences can be interpreted in this context as well. On the contrary, showing disinterest in further education may be related to educational alienation and be associated with taking no measures to improve a child's school achievement. Other facets that repeatedly play a role include a facet of religiousness, measured by frequent visits to religious events. It is interpreted as describing a family background with specific, rather traditional, norms and ethics. The importance of this facet varies by cohort analysis, but it is most often positively related. Occasionally, positive links of a dimension relating to a health-conscious lifestyle, indicated by a mindful diet and non-smoking, can be found. However, this factor does not always emerge which diminishes its importance. Indicators of maternal trust show a strong positive association with the school achievement, in particular in the older cohorts, but not with test scores. The factor is interpreted as the mother having a higher level of social capital, which can be linked to social interconnectedness and benefit the child in several ways. As it only concerns school achievement, one can hypothesize that social capital is more important for this measure. This becomes comprehensible when social capital increases, for instance knowledge about teachers and their grading, such that educational decisions for the child can be optimized.
The remaining results concern facets linked to demographic indicators, such as a family environment with many children and a high receipt of public transfers. Stable negative associations are observed for this factor, which are explained by interpreting the factor to depict unemployment or less time and pecuniary resources left per child. Occasionally connected to this factor are hours spent on child care and housework, in contrast to work. Sometimes this is an extra factor and its negative association is explained by the role model argument according to which work transmits values which are useful outside the home.
Parental age, in spite of not being prominent in each of the analyses, shows positive links. A parental migration background, by contrast, is often negatively associated. Since formal education is often negatively correlated to this factor, this can depict structural disadvantages.

Wrapping up the main insights, the selected method has indicated a potential to yield sparse models with interpretable factors. Remaining issues stem from two aspects: Firstly, factors with small eigenvalues load only on few variables. Since variables are treated as noisy approximations to the underlying factors, an interpretation is rendered difficult in such cases. The severity varies with the cohort analysis. Secondly, an overly sparse model, which performs worse than the baseline model, emerged in one case. From a model valuation point of a view, this would be considered underfitting introduced by a too strong shrinkage factor.

Policy recommendations based on the results can only be indicative, as the analysis finds associations, but is not capable of identifying causal relationships. When relating the theoretical considerations to the results, it can be observed that personality traits do not play a major role in the models. Although conscientiousness and extraversion are sometimes negatively related to the outcome, family background facets related to other attributes show a higher and more stable relevance. This is advantageous, for personality traits are considered the least malleable of the considered characteristics. Since the facet of frequent cultural activities has been found positively related, a policy recommendation would be to foster cultural events, so they become better known and more attractive. The assumption of this recommendation is that there is a direct transmission channel, e.g. knowledge gain or increased societal participation, which affects the child in a positive way. By contrast, if an unobserved third variable is reflected in this facet and drives the child's school success, such a policy will come to nothing. A similar argument can be brought forward for measures enhancing political participation or further education. Here, the intrinsic motivation of shaping the surroundings according to one's ideas may play the major role for both political interest and the school success of the child. Although it may generally be desirable to increase political interest, it is not guaranteed that this is of benefit to the child's educational achievement. If appropriate measures, however, manage to change the mindset in this respect, an effect might be observable. Using the role model argument, this could, for instance, hold for fostering further education.
All in all, these facets have shown relevant associations, but the most important ones remain related to family demographics – in particular the presence of both parents where the mother's partner has an income. This factor is linked to the biggest improvements in most of the considered outcome measures.

For the researcher aiming to describe the mindset of a family background more accurately, this work gives clear references. Useful indicators include the propensity to deliberately undertake further education, a measurement of curiosity, participation in cultural activities and political interest or engagement. Maternal trust, combined with future optimism, could serve as an indicator of social capital. An insightful extension of this work could arise with the advent of more precise data. Obtaining more detailed information about the type of cultural event or the kind of local political initiative, for instance, may lead to more informative structures in the factors. Moreover, it could be insightful to analyze the key family background dimensions for educational achievement measured at different points in time. With this information, one could examine whether the relevant dimensions differ between early and late school achievement and thereby address the limitation of this study concerning the time of measurement.

Chapter 8

Appendix

Variable description and summary statistics

This section provides an overview of the used variables. It starts with a short description of the variables whose descriptive statistics are shown in the tables below and whose abbreviated names may carry ambiguity.
Grade Average, *Grade Average alternative*, *Test Scores* are the three endogenous variables.
Income types are individual labor income (*Labor Income*), household labor income (*HH Labor Income*), household asset income (*HH Asset Income*), household post-government income (*HH PostGov Income*), household income from private and public (*HH PrivTrans Income*, *HH PubTrans Income*) transfers. Income variables are measured on yearly basis in thousand Euros. To include possible marginal changes in association with increasing income, individual labor income appears also as a squared term (*sq*) and is measured in million Euros. An *A* after a variable's name denotes the average value of this variable at the time when the child was between 13 and 15 years old; a *B* denotes the average value of a variable at the time when the child was between 10 and 12 years of age.

Hours work denotes the number of factual working hours per week. *Parental Age* is measured when the child is 15 years old. Education is measured as the years of obtained education. *Children total* counts the children that the reference parent has. *Living1* – *Living8* are variables which count the number of years a child lived in a specific family situation. *Living1* counts the years of the situation in which the child lived together with both parents, *Living2* is living with the mother alone, *Living3* is living with the mother and her partner, *Living4* is living with the father alone, *Living5* is living with the father and his partner, *Living6* is living with other relatives, *Living7* is living with foster parents, *Living8* is living in a children's home.

Eastern Germany, *Northern Germany*, *Western Germany* and *Southern Germany* indicate the geographical location in Germany of the child's household. *S.o.t.* (size of town) variables categorize the location where the child's household resides according to the number of inhabitants. Its values can be rural, small town, medium-large town and city. *Resid. area* variables indicate the residential area of the child's household. It can take the values old town, new town, mixed area, commercial area or industrial area.

Childcare and *Housework* denote the average number of daily hours spent on these activities. Also here *A* and *B* refer to the time span.

Anomie 1 – Anomie 4 refer to the extent of (dis-)approval of the following statements in numerical order: approval of "I have confidence in the future", disapproval of "I often feel lonely", disapproval of "My work is no fun" disapproval of "Everything is so complicated".

Trust 1 – Trust 5 refer to the (dis-)approval of the following statements in numerical order: approval of "On the whole one can trust people", disapproval of "Nowadays one cannot trust anyone", disapproval of "Show caution when dealing with strangers", approval of "Most People Are Fair", disapproval of "Most People act in own interest".

Reciprocity 1 – Reciprocity 5 refer to the (dis-)approval of the following statements in numerical order: approval of "I return favors", disapproval of "I get revenge for severe injustices", disapproval of "I cause similar problems to those who cause me problems", approval of "I help those who help me", disapproval of "I insult those who insult me".

Interest in Politics denotes the interest in politics in general, where larger values measure higher interest.

Time use activity is measured on Likert scales with regard to the frequency of attending religious events (*Church visits*), taking part in cultural activities (*Cultural activities*) such as concerts, theaters, lectures, investing time into honorary activities in clubs, organizations or social service (*Honorary Post*), visits to cinema, pop concerts, dances, discos and sports events (*Cinema visits*), exercising (*Exercising*), helping out friends, relatives, or neighbors (*Neighborly help*), participation in citizen initiatives, parties, community politics (*Local Initiatives*), socializing (*Sociability*), going to art exhibitions (*Art exhibitions*), watching TV (*Watching TV*), eating out (*Eating out*), going on excursions (*Excursions*), repairing cars (*Repairing cars*) and family and neighbor visits (*Visiting family*, *Visiting neighbors*). Weekly hours used for repairs on and around the house and garden work is captured by *Repairs* and hobbies by *Hobbies*.

Memberships such as in a professional association (*Prof. assoc.*), in a labor union (*Labor union*) or in an environmental association (*Envir. assoc.*) are regarded.

The degree of having a considered personal diet is denoted by *Minding diet* and non-smoking is indicated by *Non-smoking*.
In order to keep variable descriptions concise, importance indicators, showing the individuals' view of importance regarding the following subjects, are also numerically labeled: *Importance 1* "To be able to afford things", *Importance 2* "To be there for others", *Importance 3* "To develop oneself", *Importance 4* "Success in the job", *Importance 5* "To have an own house", *Importance 6* "Having a happy marriage", *Importance 7* "Having kids", *Importance 8* "Political/Social participation", *Importance 9* "Traveling", *Importance 10* "Environmental protection", *Importance 11* "Religion", *Importance 12* "Work", *Importance 13* "Family", *Importance 14* "Friends", *Importance 15* "Income", *Importance 16* "Housing", *Importance 17* "Health", *Importance 18* "Having political influence", *Importance 19* "Leisure time", *Importance 20* "Own mobility", *Importance 21* "Residential Area".

Further education variables include *Never interested in FE*, indicating that an individual has never taken part in further education since joining the SOEP, *FE for maintaining job*, indicating that further education was done to stay in the current job, *FE for promotion* indicating job promotion reasons for undertaking further education, *No interest in FE* showing no interest in further education in a particular survey year and finally *General FE* which measures whether a parent undertook any kind of further education.

The following tables show the descriptive statistics for the whole sample with the mother as reference parent. *C1* to *C4* denote cohort 1 to 4 for the dependent variables based on the standardized grades. *TC1* and *TC2* denote cohort 1 and 2 for the dependent variable based on test scores. An *X* in a column denotes whether a variable has been used for a particular cohort analysis. When a variable related to the child or household appears solely in samples without partner data it is marked *.X*.

Table 8.1: Summary Statistics: Endogenous variables

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
Grade Average	3,921	0.02	0.9	−3.4	2.0	X	X	X	X		
Grade Average alternative	3,921	0.02	0.9	−3.3	2.3			X			
Test Scores	2,030	31.7	9.3	3	55					X	X

Table 8.2: Summary Statistics Ia: Individual demographic characteristics

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
Reference parent											
Labor Income A	3,869	16.1	17.9	0.0	182.3	X	X	X	X	X	X
Labor Income Asq	3,869	0.6	1.4	0.0	33.2	X	X	X	X	X	X
Labor Income B	2,892	14.0	16.8	0.0	165.4			X	X		X
Labor Income Bsq	2,892	0.5	1.4	0.0	27.4			X	X		X
Education	4,325	12.1	2.6	7.0	18.0	X	X	X	X	X	X
Parental age	4,345	42.6	5.1	30	64	X	X	X	X	X	X
Partner											
Labor Income A	3,278	49.7	39.9	0.0	592.5		X	X		X	X
Labor Income Asq	3,278	4.1	11.7	0.0	351.1		X	X		X	X
Labor Income B	2,535	47.0	32.6	0.0	419.1			X			
Labor Income Bsq	2,535	3.3	6.8	0.0	175.7			X			
Education	3,830	12.4	2.8	7.0	18.0		X	X	X	X	X
Parental age	3,828	45.2	6.3	23	81		X	X	X	X	X

Table 8.3: Summary Statistics Ib: Household demographic characteristics

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
HH Labor Income A	3,869	62.2	44.9	0.0	592.5	X	X	X	X	X	X
HH Labor Income B	2,892	57.7	37.2	0.0	419.1			X	X		X
HH Asset Income A	3,869	4.5	13.4	0.0	615.2	X	X	X	X	X	X
HH Asset Income B	2,892	3.6	7.5	0.0	138.8			X	X		X
HH PostGov Income A	3,869	51.6	28.9	7.3	545.8	X	X	X	X	X	X
HH PostGov Income B	2,892	48.2	23.4	6.4	261.9			X	X		X
HH PrivTrans Income A	3,869	0.4	1.8	0.0	38.3	X	X	X	X	X	X
HH PrivTrans Income B	2,892	0.3	1.5	0.0	22.0			X	X		X
HH PubTrans Income A	3,869	7.1	5.4	0.0	51.0	X	X	X	X	X	X
HH PubTrans Income B	2,892	6.8	5.3	0.0	43.1			X	X		X
Father migrated	4,295	0.1	0.3	0	1	X	X	X	X	X	X
Mother migrated	4,326	0.1	0.3	0	1	X	X	X	X	X	X
Father died	4,357	0.02	0.2	0	1	X	X	X	X	X	X
Children total	4,299	2.5	1.2	0	12	X	X	X	X	X	X
Living1	4,347	12.7	4.6	0	15	X	X	X	X	X	X
Living2	4,168	1.3	3.4	0	15	X	X	X	X	X	X
Living3	4,159	0.8	2.6	0	15	X	X	X	X	X	X
Living4	4,129	0.1	0.7	0	15	X	X	X	X	X	X
Living5	4,131	0.04	0.5	0	13	X	X	X	X	X	.X
Living6	4,132	0.03	0.5	0	15	X	X	X	X	X	X
Living7	4,128	0.03	0.6	0	15	X	X	X	X	X	
Living8	4,355	0.01	0.3	0	11	X	X	X	X		

Table 8.4: Summary Statistics Ic: Household geographical characteristics

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
Eastern Germany	4,357	0.24	0.4	0	1	X	X	X	X	X	X
Northern Germany	4,357	0.14	0.3	0	1	X	X	X	X	X	X
Western Germany	4,357	0.33	0.5	0	1	X	X	X	X	X	X
Southern Germany	4,357	0.38	0.4	0	1	X	X	X	X	X	X
S.o.t.: City	4,328	0.2	0.4	0	1	X	X	X	X	X	X
S.o.t.: Mid	4,328	0.2	0.4	0	1	X	X	X	X	X	X
S.o.t.: Small	4,328	0.3	0.4	0	1	X	X	X	X	X	X
S.o.t.: Rural	4,328	0.3	0.5	0	1	X	X	X	X	X	X
Resid. area old	3,166	0.32	0.5	0	1			X			
Resid. area new	3,166	0.44	0.5	0	1			X			
Resid. area mix	3,166	0.21	0.4	0	1			X			
Resid. area com.	3,166	0.00	0.1	0	1			X			
Resid. area ind.	3,166	0.01	0.1	0	1			X			

Table 8.5: Summary Statistics II: Personality traits

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
Reference parent											
Openness	3,835	49.6	9.5	15.4	76.0		X	X	X	X	X
Conscientiousness	3,835	51.1	9.0	−2.2	68.2		X	X	X	X	X
Extraversion	3,835	51.8	9.9	16.3	76.0		X	X	X	X	X
Neuroticism	3,835	51.8	9.9	25.4	78.7		X	X	X	X	X
Agreeableness	3,835	52.2	9.1	18.4	75.3		X	X	X	X	X
Locus of control	2,859	50.0	11.6	8.4	76.7	X		X	X	X	X
Partner											
Openness	3,204	49.0	9.7	15.5	83.6		X	X	X	X	X
Conscientiousness	3,204	51.6	9.3	−12.0	71.6		X	X	X	X	X
Extraversion	3,204	49.3	9.8	14.0	74.6		X	X	X	X	X
Neuroticism	3,204	48.2	9.5	20.2	78.4		X	X	X	X	X
Agreeableness	3,204	47.0	10.5	8.3	73.0		X	X	X	X	X
Locus of control	2,526	51.5	11.9	2.9	79.4			X	X	X	X

Table 8.6: Summary Statistics IIIa: Time use characteristics of reference parent

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
Reference parent											
Hours work A	3,841	19.4	15.8	0.0	75.0	X	X	X	X	X	X
Hours work B	2,874	17.3	15.4	0.0	72.3			X	X		X
Childcare A	3,564	3.5	3.7	0.0	24.0		X	X	X	X	X
Childcare B	2,875	4.9	4.1	0.0	24.0			X	X		X
Housework A	3,559	3.0	1.6	0.0	12.0		X	X	X	X	X
Housework B	2,865	3.2	1.6	0.0	10.3			X	X		X
Interest in Politics	3,895	2.1	0.7	1	4	X	X	X	X	X	X
Church visits	3,898	9.4	19.2	0.0	365.0	X	X	X	X	X	X
Cultural activities	3,901	4.1	5.6	0	52	X	X	X	X	X	X
Honorary post	3,896	7.7	21.2	0.0	365.0	X	X	X	X	X	X
Cinema visits	3,897	5.5	8.0	0	52	X	X	X	X	X	X
Exercising	3,951	25.0	48.2	0.0	365.0	X	X	X	X	X	X
Neighborly help	3,572	11.4	14.1	0	52	X	X		X		X
Local initiatives	3,405	0.9	7.7	0	365		X	X	X	X	X
Repairs	3,894	0.7	0.8	0	10	X	X	X	X	X	X
Hobbies	3,895	1.6	1.4	0	11	X	X	X	X	X	X
Sociability	3,509	24.8	20.9	0	52		X		X	X	X
Art exhibitions	2,389	16.5	57.5	0	365				X		
Watching TV	2,389	306.0	126.2	0	365				X		
Eating out	2,389	12.6	23.2	0	365				X		
Excursions	2,383	6.5	7.7	0	52				X		
Minding diet	1,817	2.7	0.7	1	4			X	X		X
Non-smoking	2,186	0.6	0.5	0	1			X	X		X
Prof. assoc.	3,236	0.1	0.2	0	1	X	X		X		X
Labor union	3,251	0.1	0.3	0	1	X	X		X		X
Envir. assoc.	3,235	0.04	0.2	0	1	X	X		X		X
Repairing cars	2,401	4.8	14.9	0	365		X				
Visiting neighbors	2,384	35.8	59.0	0	365		X				
Visiting family	2,371	53.0	93.1	0	365		X				
FE for maintaining job	2,002	0.2	0.4	0	1			X		X	
FE for promotion	2,002	0.6	0.5	0	1			X		X	
General FE	3,548	0.4	0.5	0	1	X	X			X	

Table 8.7: Summary Statistics IIIb: Time use characteristics of partner

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
Partner											
Hours work A	3,259	39.7	15.5	0.0	80.0		X	X		X	X
Hours work B	2,525	40.3	14.1	0.0	80.0			X			
Childcare A	3,042	1.1	1.5	0.0	24.0		X	X			
Childcare B	2,534	1.5	1.6	0.0	24.0			X			
Housework A	3,039	0.6	0.7	0.0	6.0		X	X			
Housework B	2,534	0.5	0.7	0.0	6.0			X			
Interest in Politics	3,422	2.5	0.8	1	4		X	X	X	X	X
Church visits	3,396	8.1	17.4	0	365		X	X	X	X	X
Cultural activities	3,397	3.8	5.6	0	52		X	X	X	X	X
Honorary post	3,395	10.8	30.0	0.0	365.0		X	X	X	X	X
Cinema visits	3,396	6.0	9.7	0	52		X	X	X	X	X
Exercising	3,389	22.2	44.9	0.0	365.0		X	X	X	X	X
Neighborly help	3,024	11.6	13.7	0	52		X		X		X
Local initiatives	2,965	1.6	9.4	0	365		X	X	X	X	X
Repairs	3,398	1.0	1.0	0	10		X	X	X	X	X
Hobbies	3,422	1.6	1.6	0	14		X	X	X	X	X
Sociability	2,222	22.1	20.2	0	52						X
Art exhibitions	1,645	9.8	41.4	0	365				X		
Watching TV	1,644	304.8	127.3	0	365				X		
Eating out	1,647	18.9	46.8	0	365				X		
Excursions	1,641	7.1	15.0	0	365				X		
Minding diet	1,545	2.3	0.7	1	4			X	X		X
Non-smoking	1,970	0.6	0.5	0	1			X	X		X
Prof. assoc.	2,737	0.1	0.4	0	1		X				X
Labor union	2,769	0.2	0.4	0	1		X				X
Envir. assoc.	2,733	0.1	0.2	0	1		X				X
Repairing cars	2,147	20.8	42.0	0	365		X				
Visiting neighbors	2,012	28.3	43.8	0	365						
Visiting family	2,005	36.8	66.7	0	365						
FE for maintaining job	1,740	0.2	0.4	0	1						
FE for promotion	1,740	0.7	0.4	0	1						
General FE	3,090	0.5	0.5	0	1		X				

Table 8.8: Summary Statistics IVa: Attitudes of reference parent

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
Reference parent											
Anomie 1	1,552	2.6	0.7	1	4				X		X
Anomie 2	1,552	1.8	0.9	1	4				X		X
Anomie 3	1,519	1.7	0.8	1	4				X		X
Anomie 4	1,550	1.7	0.8	1	4				X		X
Trust 1	2,350	2.7	0.6	1	4		X		X		X
Trust 2	2,351	2.6	0.7	1	4		X		X		X
Trust 3	2,348	1.7	0.7	1	4		X		X		X
Trust 4	2,319	1.5	0.5	1	2		X		X		X
Trust 5	2,322	1.3	0.5	1	2		X		X		X
Reciprocity 1	1,399	6.5	0.8	1	7				X		X
Reciprocity 2	1,398	5.1	1.7	1	7				X		X
Reciprocity 3	1,396	5.5	1.5	1	7				X		X
Reciprocity 4	1,397	5.9	1.2	1	7				X		X
Risk attitude	1,809	4.2	2.2	0	10			X	X		X
Curiosity	785	5.3	1.3	1	7				X		
Future optimism	1,438	2.8	0.7	1	4				X	X	X
Importance 1	2,056	2.9	0.6	1	4			X	X	X	X
Importance 2	2,059	3.3	0.5	1	4			X	X	X	X
Importance 3	2,059	2.8	0.7	1	4			X	X	X	X
Importance 4	2,458	2.8	0.7	1	4	X		X	X	X	X
Importance 5	2,059	2.8	0.9	1	4			X	X	X	X
Importance 6	2,056	3.7	0.6	1	4			X	X	X	X
Importance 7	2,053	3.7	0.5	1	4			X	X	X	X
Importance 8	2,056	2.0	0.7	1	4			X	X	X	X
Importance 9	1,665	2.3	0.8	1	4			X		X	
Importance 10	598	3.2	0.6	1	4						
Importance 11	595	2.2	0.9	1	4						
Importance 12	591	3.2	0.7	1	4						
Importance 13	599	3.9	0.3	2	4	X					
Importance 14	599	3.2	0.6	2	4	X					
Importance 15	599	3.5	0.5	2	4	X					
Importance 16	597	3.5	0.5	2	4	X					
Importance 17	597	3.8	0.4	1	4	X					
Importance 18	597	2.1	0.7	1	4	X					
Importance 19	597	3.1	0.6	1	4	X					
Importance 20	598	3.2	0.7	1	4	X					
Importance 21	598	3.1	0.5	1	4	X					
Never interested in FE	3,916	0.2	0.4	0	1	X	X	X	X		
No Interest in FE	1,975	0.3	0.5	0	1					X	

Table 8.9: Summary Statistics IVb: Attitudes of partner

Statistic	N	Mean	St. Dev.	Min	Max	C1	C2	C3	C4	TC1	TC2
Partner											
Anomie 1	697	2.6	0.7	1	4						
Anomie 2	696	1.6	0.8	1	4						
Anomie 3	697	1.8	0.8	1	4						
Anomie 4	696	1.7	0.8	1	4						
Trust 1	1,944	2.6	0.7	1	4		X				X
Trust 2	1,949	2.6	0.8	1	4		X				X
Trust 3	1,944	1.8	0.7	1	4		X				X
Trust 4	1,932	1.5	0.5	1	2		X				X
Trust 5	1,928	1.3	0.5	1	2		X				X
Reciprocity 1	1,226	6.4	0.8	2	7						X
Reciprocity 2	1,224	4.7	1.7	1	7						X
Reciprocity 3	1,223	5.0	1.6	1	7						X
Reciprocity 4	1,225	5.8	1.1	1	7						X
Risk attitude	1,529	5.1	2.2	0	10			X			X
Curiosity	636	5.4	1.3	1	7				X		
Future optimism	1,225	2.8	0.8	1	4			X			
Importance 1	1,739	3.0	0.6	1	4			X	X		X
Importance 2	1,739	3.1	0.6	1	4			X	X		X
Importance 3	1,737	2.8	0.7	1	4			X	X		X
Importance 4	2,094	3.1	0.6	1	4			X	X		X
Importance 5	1,735	3.0	0.9	1	4			X	X		X
Importance 6	1,741	3.7	0.5	1	4			X	X		X
Importance 7	1,732	3.6	0.6	1	4			X	X		X
Importance 8	1,735	2.1	0.7	1	4			X	X		X
Importance 9	1,402	2.3	0.8	1	4			X			
Importance 10	524	3.1	0.6	1	4						
Importance 11	522	2.1	0.9	1	4						
Importance 12	523	3.5	0.6	1	4						
Importance 13	524	3.9	0.4	2	4						
Importance 14	524	3.1	0.6	1	4						
Importance 15	524	3.5	0.6	2	4						
Importance 16	523	3.4	0.5	2	4						
Importance 17	523	3.8	0.4	3	4						
Importance 18	524	2.2	0.7	1	4						
Importance 19	523	3.1	0.6	1	4						
Importance 20	523	3.2	0.6	1	4						
Importance 21	524	3.0	0.6	1	4						
Never interested in FE	3,085	0.1	0.3	0	1		X				
No Interest in FE	1,726	0.2	0.4	0	1						

Bibliography

Ajzen, I., 2005. Attitudes, personality, and behavior (2nd ed.). Open University Press (McGraw-Hill Education), Berkshire, Maidenhead, England.

Aldrin, M., 2006. Reduced-rank regression. In: El-Shaarawi, A., Piegorsch, W.W., C. (Eds.), Encyclopedia of environmetrics. Wiley Online Library (John Wiley & Sons, Ltd.), West Sussex, England, 1724–1728.

Allmendinger, J., 2009. Der Sozialstaat des 21. Jahrhunderts braucht zwei Beine. Aus Politik und Zeitgeschichte 45, 3–5.

Allmendinger, J., Giesecke, J., Oberschachtsiek, D., 2011. Unzureichende Bildung. Folgekosten für die öffentlichen Haushalte. Eine Studie des Wissenschaftszentrum Berlin für Sozialforschung im Auftrag der Bertelsmann Stiftung. Bertelsmann-Stiftung; Wissenschaftszentrum Berlin für Sozialforschung, Gütersloh.

Amirazodi, F., Amirazodi, M., 2011. Personality traits and self-esteem. Procedia - Social and Behavioral Sciences 29, 713–716.

Anderson, T. W., Rubin, H., 1956. Statistical inference in factor analysis. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry. University of California Press, Berkeley, 111–150.

Anger, C., Esselmann, I., Konegen-Gernier, C., Plünnecke, A., 2015. Bildungsmonitor 2015. Research study, IW Köln - Studie im Auftrag der Initiative Neue Soziale Marktwirtschaft.

Anger, S., 2011. The Intergenerational Transmission of Cognitive and Non-Cognitive Skills During Adolescence and Young Adulthood. IZA Discussion Papers 5749, Institute for the Study of Labor (IZA).

Anger, S., 2013. Personality and educational attainment. Working paper, IWAEE 2013.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics surveys 4, 40–79.

Ascheberg, C., 2006. Die SIGMA Milieus – das globale Zielgruppen-und Trend System. http://www.sigma-online.com/de/Articles_and_Reports/zielgruppenforschung.pdf, online; accessed 24.03.2015.

Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by Supervised Principal Components. Journal of the American Statistical Association 101 (473), 119–137.

Barth, B., Flaig, B. B., 2013. Was sind Sinus-Milieus? Eine Einführung in die sozialwissenschaftliche Fundierung und Praxisrelevanz eines Gesellschaftsmodells. In: Thomas, P. M., Calmbach, M. (Eds.), Jugendliche Lebenswelten: Perspektiven für Politik, Pädagogik und Gesellschaft. Springer Spektrum Akademischer Verlag, Berlin, 11–36.

Bartlett, M. S., 1950. Tests of significance in factor analysis. British Journal of statistical psychology 3 (2), 77–85.

Bauer, P., Riphahn, R. T., 2006. Timing of school tracking as a determinant of intergenerational transmission of education. Economics Letters 91 (1), 90–97.

Baumert, J., et al., 2001. PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich Verlag, Opladen.

Beck, U., 1986. Risikogesellschaft : auf dem Weg in eine andere Moderne. Edition Suhrkamp : Bd. 1365. Suhrkamp, Frankfurt am Main.

Becker, A., Deckers, T., Dohmen, T., Falk, A., Kosse, F., 2012. The Relationship Between Economic Preferences and Psychological Personality Measures. Annual Review of Economics 4 (1), 453–478.

Becker, G. S., 1964. Human capital: a theoretical analysis with special reference to education. National Bureau for Economic Research, Columbia University Press, New York and London.

Behrman, J. R., Rosenzweig, M. R., et al., 2002. Does increasing women's schooling raise the schooling of the next generation? American Economic Review 92 (1), 323–334.

Bekkers, R., 2007. Intergenerational transmission of volunteering. Acta Sociologica 50 (2), 99–114.

Björklund, A., Jäntti, M., 2012. How important is family background for labor-economic outcomes? Labour Economics 19 (4), 465–474.

Björklund, A., Lindahl, L., Lindquist, M. J., 2010. What more than parental income, education and occupation? an exploration of what swedish siblings get from their parents. The BE Journal of Economic Analysis & Policy 10 (1).

Black, S. E., Devereux, P. J., Salvanes, K. G., 2005. Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital. American Economic Review 95 (1), 437–449.

Bleakley, A., Jordan, A. B., Hennessy, M., 2013. The relationship between parents and childrens television viewing. Pediatrics 132 (2), e364–e371.

Boll, C., 2011. Lohneinbussen von Frauen durch geburtsbedingte Erwerbsunterbrechungen: der Schattenpreis von Kindern und dessen moegliche Auswirkungen auf weibliche Spezialisierungsentscheidungen im Haushaltszusammenhang ; eine quantitative Analyse auf Basis von SOEP-Daten. Zugl.: Kiel, Univ., Diss., 2010.

Boll, C., Hoffmann, M., 2015. Parents' employment and children's school success in germany. SOEPpapers on Multidisciplinary Panel Data Research 735, DIW Berlin, The German Socio-Economic Panel (SOEP).

Boudon, R., 1974. Education, opportunity, and social inequality; changing prospects in Western society. Wiley, New York.

Boulesteix, A.-L., Strimmer, K., 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 8 (1), 32–44.

Bourdieu, P., 1987. Die feinen Unterschiede: Kritik der gesellschaftlichen Urteilskraft. Suhrkamp, Frankfurt am Main.

Breen, R., Goldthorpe, J. H., 1997. Explaining educational differentials: Towards a formal rational action theory. Rationality and Society 9 (3), 275–305.

Bremer, H., 2007. Schicht, Klasse, Milieu. Die Zeitschrift 57 (4), 26–31.

Bremer, H., Kleemann-Göhring, M., 2012. Familienbildung, Grundschule und Milieu. Eine Expertise im Rahmen des Projekts: Familienbildung während der Grundschulzeit. Sorgsame Elternschaft fünf bis elf. Tech. rep., Landesarbeitsgemeinschaften der Familienbildung in NRW, Wuppertal.

Brown, S., Ortiz-Nuñez, A., Taylor, K., 2012. Parental Risk Attitudes and children's Academic Test Scores: Evidence from the US Panel Study of Income Dynamics. Scottish Journal of Political Economy 59 (1), 47–70.

Büchel, F., Duncan, G. J., 1998. Do parents' social activities promote children's school attainments? Evidence from the German socioeconomic panel. Journal of Marriage and the Family 60 (1), 95–108.

Busch, A., 2013. Die Geschlechtersegregation beim Berufseinstieg–Berufswerte und ihr Erklärungsbeitrag für die geschlechtstypische Berufswahl. Berliner Journal für Soziologie 23 (2), 145–179.

Caliendo, M., Cobb-Clark, D. A., Uhlendorff, A., 2015. Locus of control and job search strategies. Review of Economics and Statistics 97 (1), 88–103.

Checchi, D., Fiorio, C. V., Leonardi, M., 2014. Parents' risk aversion and children's educational attainment. Labour Economics 30, 164–175.

Cheng, R.C.H., 1998. Random variate generation. In: Banks, J. (Ed.), Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice. John Wiley & Sons, New York, 139–172.

Chevalier, A., Harmon, C., Walker, I., Zhu, Y., 2004. Does education raise productivity, or just reflect it?*. The Economic Journal 114 (499), F499–F517.

Clarke, B., Fokoue, E., Zhang, H. H., 2009. Principles and theory for data mining and machine learning. Springer, New York.

Cobb-Clark, D. A., Kassenboehmer, S. C., Schurer, S., 2014. Healthy habits: The connection between diet, exercise, and locus of control. Journal of Economic Behavior & Organization 98, 1–28.

Corak, M., 2013. Income inequality, equality of opportunity, and intergenerational mobility. The Journal of Economic Perspectives 27 (3), 79–102.

Costa, P. T., McCrae, R. R., 1985. The NEO personality inventory: Manual, form S and form R. Psychological Assessment Resources, Odessa, Florida.

Couch, K. A., Dunn, T. A., 1997. Intergenerational correlations in labor market status: A comparison of the united states and germany. Journal of Human Resources 32 (1), 210–232.

Cunha, F., Heckman, J., 2007. The technology of skill formation. American Economic Review 97 (2), 31–47.

Cunha, F., Heckman, J. J., Schennach, S. M., 2010. Estimating the technology of cognitive and noncognitive skill formation. Econometrica 78 (3), 883–931.

Cunningham, M., 2001. The influence of parental attitudes and behaviors on children's attitudes toward gender and household labor in early adulthood. Journal of Marriage and Family 63 (1), 111–122.

Currie, J., et al., 2009. Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development. Journal of Economic Literature 47 (1), 87–117.

Dahmann, S., 2015. How does education improve cognitive skills? Instructional time versus timing of instruction. SOEPpapers on Multidisciplinary Panel Data Research 769, DIW Berlin, The German Socio-Economic Panel (SOEP).

Dahrendorf, R., 1965. Bildung ist Bürgerrecht. Nannen, Hamburg.

Davis-Kean, P. E., 2005. The Influence of Parent Education and Family Income on Child Achievement: The Indirect Role of Parental Expectations and the Home Environment. Journal of Family Psychology 19 (2), 294–304.

De Jong, S., Kiers, H. A., 1992. Principal covariates regression: Part I. Theory. Chemometrics and Intelligent Laboratory Systems 14 (1), 155–164.

de Vries, J., de Graaf, P. M., 2008. Is the intergenerational transmission of high cultural activities biased by the retrospective measurement of parental high cultural activities? Social Indicators Research 85 (2), 311–327.

Dee, T. S., August 2004. Are there civic returns to education? Journal of Public Economics 88 (9-10), 1697–1720.

Dehne, M., Schupp, J., 2007. Persönlichkeitsmerkmale im Sozio-oekonomischen Panel (SOEP) – Konzept, Umsetzung und empirische Eigenschaften. Research notes, DIW Berlin.

DiStefano, C., Zhu, M., Mindrila, D., 2009. Understanding and using factor scores: Considerations for the applied researcher. Practical Assessment, Research & Evaluation 14 (20), 1–11.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., 2012. The intergenerational transmission of risk and trust attitudes. The Review of Economic Studies 79 (2), 645–677.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., et al., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46.

Duncan, G., Kalil, A., Mayer, S. E., Tepper, R., Payne, M. R., 2005. The apple does not fall far from the tree. In: Bowles, S., Gintis, H., Groves, M. O. (Eds.), Unequal Chances: Family Background and Economic Success. Russell Sage Foundation, New York, 23–79.

Duncan, G. J., Yeung, W. J., Brooks-Gunn, J., Smith, J. R., 1998. How much does childhood poverty affect the life chances of children? American sociological review 63 (3), 406–423.

Durkheim, E., 1988. Über soziale Arbeitsteilung. Vol. 2. Suhrkamp Verlag, Frankfurt am Main.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al., 2004. Least angle regression. The Annals of statistics 32 (2), 407–499.

Fan, J., Ke, Z. T., 2014. Discussion: "a significance test for the lasso". The Annals of Statistics 42 (2), 483–492.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96 (456), 1348–1360.

Farre, L., Vella, F., 2013. The Intergenerational Transmission of Gender Role Attitudes and its Implications for Female Labor Force Participation. Economica 80 (318), 219–247.

Flom, P. L., Cassell, D. L., 2007. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In: NorthEast SAS Users Group Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland, 1–7.

Fortin, N. M., 2005. Gender Role Attitudes and the Labour-market Outcomes of Women across OECD Countries. Oxford Review of Economic Policy 21 (3), 416–438.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33 (1), 1–22.

Funcke, A., Menne, S., 2010. Familie als Bildungsort stärken - Familienleben ermöglichen. Bertelsmann-Stiftung, Gütersloh.

Geißler, R., 2005. Die Metamorphose der Arbeitertochter zum Migrantensohn. Zum Wandel der Chancenstruktur im Bildungssystem nach Schicht, Geschlecht, Ethnie und deren Verknüpfungen. Juventa Verlag, Weinheim.

Goeman, J., Meijer, R., Chaturvedi, N., 2016. L1 and l2 penalized regression models. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf.

Greene, W., 2012. Econometric analysis. Pearson, Boston London.

Gregg, P., Macmillan, L., Nasim, B., 2012. The impact of fathers' job loss during the recession of the 1980s on their children's educational attainment and labour market outcomes. Fiscal Studies 33 (2), 237–264.

Guttman, L., 1954. Some necessary conditions for common-factor analysis. Psychometrika 19 (2), 149–161.

Hahn, U., 2003. Das Verborgene Wort (German Edition). Deutscher Taschenbuch Verlag GmbH & Co., München.

Hamburger Senat, 2011. Hamburgisches Gesetz- und Verordnungsblatt Teil I.

Hanushek, E. A., 2009. The economic value of education and cognitive skills. In: Sykes, G., Schneider, B., Plank, D. N. (Eds.), Handbook of education policy research. Routledge Chapman & Hall, New York, 39–56.

Hanushek, E. A., Lavy, V., Hitomi, K., 2008. Do students care about school quality? Determinants of dropout behavior in developing countries. Journal of Human Capital 2 (1), 69–105.

Hanushek, E. A., Woessmann, L., 2006. Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries. Economic Journal 116 (510), C63–C76.

Hanushek, E. A., Woessmann, L., 2012. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. Journal of Economic Growth 17 (4), 267–321.

Hanushek, E. A., Woessmann, L., 2015. The knowledge capital of nations: Education and the economics of growth. MIT Press, Cambridge.

Hastie, T., Taylor, J., Tibshirani, R., Walther, G., et al., 2007. Forward stagewise regression and the monotone lasso. Electronic Journal of Statistics 1, 1–29.

Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2009. The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Vol. 2. Springer, New York.

Haveman, R., Wolfe, B., 1995. The determinants of children's attainments: A review of methods and findings. Journal of Economic Literature 33 (4), 1829–1878.

Hayashi, F., 2000. Econometrics. Princeton University Press, Princeton.

Headey, B., Muffels, R., Wagner, G. G., 2013. Choices which change life satisfaction: Similar results for australia, britain and germany. Social Indicators Research 112 (3), 725–748.

Heij, C., Groenen, P. J., van Dijk, D., 2007. Forecast comparison of principal component regression and principal covariate regression. Computational statistics & Data analysis 51 (7), 3612–3625.

Hendrickson, A. E., White, P. O., 1964. Promax: A quick method for rotation to oblique simple structure. British journal of statistical psychology 17 (1), 65–70.

Hesterberg, T., Choi, N. H., Meier, L., Fraley, C., et al., 2008. Least angle and l1 penalized regression: A review. Statistics Surveys 2, 61–93.

Higham, N. J., 1986. Newton's method for the matrix square root. Mathematics of Computation 46 (174), 537–549.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Horn, J. L., 1965. A rationale and test for the number of factors in factor analysis. Psychometrika 30 (2), 179–185.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24 (6), 417.

Hradil, S., 2006. Soziale Milieus – eine praxisorientierte Forschungsperspektive. Aus Politik und Zeitgeschichte 44 (45), 3–10.

Huebener, M., 2015. The Role of Paternal Risk Attitudes in Long-Run Education Outcomes and Intergenerational Mobility. Economics of Education Review 47, 64–79.

Jackson, J. E., 1991. A user's guide to principal components. Vol. 587. John Wiley & Sons, New York.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer, New York.

Jolliffe, I. T., 2002. Principal Component Analysis. Springer, New York.

Judge, T. A., Erez, A., Bono, J. E., Thoresen, C. J., 2002. Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? Journal of Personality and Social Psychology 83 (3), 693–710.

Kaplan, D. S., Liu, X., Kaplan, H. B., 2001. Influence of parents' self-feelings and expectations on children's academic performance. The Journal of Educational Research 94 (6), 360–370.

Kiers, H. A., Smilde, A. K., 2007. A comparison of various methods for multivariate regression with highly collinear variables. Statistical Methods and Applications 16 (2), 193–228.

Kimko, D. D., Hanushek, E. A., 2000. Schooling, Labor-Force Quality, and the Growth of Nations. American Economic Review 90 (5), 1184–1208.

Kluckhohn, F., Strodtbeck, F. L., 1961. Variations in Value Orientation. Row and Peterson, Evanston, Illinois.

Komarraju, M., Karau, S. J., Schmeck, R. R., 2009. Role of the Big Five personality traits in predicting college students' academic motivation and achievement. Learning and Individual Differences 19 (1), 47–52.

Kordi, A., Baharudin, R., 2010. Parenting attitude and style and its effect on children's school achievements. International Journal of Psychological Studies 2 (2), 217–222.

Kroch, E. A., Sjoblom, K., 1994. Schooling as human capital or a signal: Some evidence. Journal of Human Resources, 156–180.

Kuhn, M., 2008. Caret package. Journal of Statistical Software 28 (5), 1–26.

Lampert, T., Kroll, L. E., Kuntz, B., Ziese, T., 2013. Gesundheitliche Ungleichheit. In: Dallinger, G., Haensel, K., Martin, R., Petter, M., Habich, R., Wettig, M. (Eds.), Datenreport 2013 - Ein Sozialbericht für die Bundesrepublik Deutschland. Statistisches Bundesamt (Destatis), Wissenschaftszentrum Berlin für Sozialforschung (WZB), Zentrales Datenmanagement, Bonn, 259–271.

Lang, K., Kropp, D., 1986. Human capital versus sorting: the effects of compulsory attendance laws. The Quarterly Journal of Economics 101 (3), 609–624.

Lange, D., Print, M., 2013. Civic Education and Competences for Engaging Citizens in Democracies. Vol. 1. Sense Publishers, Rotterdam.

Lebanon, G., 2010. Bias, Variance, and MSE of estimators. http://www.cc.gatech.edu/~lebanon/notes/estimators1.pdf, online; accessed 19.7.2014.

L'Ecuyer, P., 1998. Random number generation. In: Banks, J. (Ed.), Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice. John Wiley & Sons, New York, 93–137.

Lochner, L., 2004. Education, work, and crime: A human capital approach*. International Economic Review 45 (3), 811–843.

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., 2014. A significance test for the lasso. Annals of statistics 42 (2), 413.

Loureiro, M., de Galdeano, A. S., Vuri, D., 2009. Smoking Habits: Like Father, Like Son, Like Mother, Like Daughter. Working Papers 402, Barcelona Graduate School of Economics.

Machin, S., Marie, O., Vujic, S., 2011. The crime reducing effect of education. The Economic Journal 121 (552), 463–484.

Mander, A., 2014. LARS: Stata module to perform least angle regression. Statistical Software Components, Boston College Department of Economics. http://ideas.repec.org/p/bge/wpaper/402.html.

Mankiw, N. G., Romer, D., Weil, D. N., 1992. A Contribution to the Empirics of Economic Growth. The Quarterly Journal of Economics 107 (2), 407–437.

Mazumder, B., 2008. Sibling similarities and economic inequality in the US. Journal of Population Economics 21 (3), 685–701.

McCrae, R. R., John, O. P., 1992. An introduction to the five-factor model and its applications. Journal of Personality 60 (2), 175–215.

McDonald, R. P., 1981. Constrained least squares estimators of oblique common factors. Psychometrika 46 (3), 337–341.

McDonald, R. P., Burr, E., 1967. A comparison of four methods of constructing factor scores. Psychometrika 32 (4), 381–401.

McGinn, K. L., Ruiz Castro, M., Lingo, E. L., 2015. Mums the word! cross-national effects of maternal employment on gender inequalities at work and at home. Working paper, Havard Business School.

Mevik, B.-H., Wehrens, R., 2007. The pls Package: Principal Component and Partial Least Squares Regression in R. Journal of Statistical Software 18 (2), 1–23.

Milligan, K., Moretti, E., Oreopoulos, P., August 2004. Does education improve citizenship? Evidence from the United States and the United Kingdom. Journal of Public Economics 88 (9-10), 1667–1695.

Mincer, J. A., 1974. Schooling and earnings. NBER Chapters, 41–63.

Mittelhammer, R. C., 1996. Mathematical statistics for economics and business. Vol. 78. Springer, New York.

Morrill, M. S., Morrill, T., 2013. Intergenerational links in female labor force participation. Labour Economics 20, 38–47.

Nechyba, T. J., McEwan, P. J., Older-Aguilar, D., 1999. The impact of family and community resources on student outcomes: An assessment of the international literature with implications for New Zealand. Ministry of Education, Wellington.

Nelson, R. R., Phelps, E. S., 1966. Investment in Humans, Technological Diffusion and Economic Growth. The American Economic Review 56 (2), 69–75.

OECD, 2012a. Education at a Glance 2012. OECD Indicators, OECD Publishing.

OECD, 2012b. Let's Read Them a Story! The Parent Factor in Education. PISA, OECD Publishing.

Oreopoulos, P., Salvanes, K. G., 2011. Priceless: The nonpecuniary benefits of schooling. The Journal of Economic Perspectives 25 (1), 159–84.

Paulus, W., Blossfeld, H.-P., 2007. Schichtspezifische Präferenzen oder sozioökonomisches Entscheidungskalkül? Zur Rolle elterlicher Bildungsaspirationen im Entscheidungsprozess beim Übergang von der Grundschule in die Sekundarstufe. Zeitschrift für Pädagogik 53 (4), 491–508.

Pearson, K., 1901. Principal components analysis. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 6 (2), 559.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peter, F., Storck, J., 2015. Personality Traits Affect Young People's Intention to Study. DIW Economic Bulletin 12, 1–9.

Pfeffer, F. T., 2008. Persistent inequality in educational attainment and its institutional context. European Sociological Review 24 (5), 543–565.

Piatek, R., Pinger, P., 2010. Maintaining (locus of) control? Assessing the impact of locus of control on education decisions and wages. SOEPpapers on Multidisciplinary Panel Data Research 338, DIW Berlin, The German Socio-Economic Panel (SOEP).

Plug, E., 2004. Estimating the Effect of Mother's Schooling on Children's Schooling Using a Sample of Adoptees. American Economic Review 94 (1), 358–368.

Prinzie, P., Stams, G. J. J., Deković, M., Reijntjes, A. H., Belsky, J., 2009. The relations between parents' Big Five personality factors and parenting: A meta-analytic review. Journal of Personality and Social Psychology 97 (2), 351–362.

Putnam, R. D., 1995. Bowling alone: America's declining social capital. Journal of democracy 6 (1), 65–78.

Rencher, A. C., 2003. Methods of Multivariate Analysis. Vol. 492. John Wiley & Sons.

Rotter, J. B., 1966. Generalized expectancies for internal versus external control of reinforcement. Psychological monographs: General and applied 80 (1), 1.

Saucier, G., Goldberg, L. R., 1998. What is beyond the big five? Journal of Personality 66, 495–524.

Savin, I., Winker, P., 2013. Lasso–type and Heuristic Strategies in Model Selection and Forecasting. In: Towards Advanced Data Analysis by Combining Soft Computing and Statistics. Springer, Berlin, 165–176.

Schindler, S., Reimer, D., 2010. Primäre und sekundäre Effekte der sozialen Herkunft beim Übergang in die Hochschulbildung. KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie 62 (4), 623–653.

Schmitz, H., 2011. Why are the unemployed in worse health? The causal effect of unemployment on health. Labour Economics 18 (1), 71–78.

Schneeweiss, H., Mathes, H., 1995. Factor analysis and principal components. Journal of multivariate analysis 55 (1), 105–124.

Schnitzlein, D., 2014. How important is the family? Evidence from sibling correlations in permanent earnings in the USA, Germany, and Denmark. Journal of Population Economics 27 (1), 69–89.

Schräpler, J.-P., Schupp, J., Wagner, G. G., 2010. Individual and Neighborhood Determinants of Survey Nonresponse – An Analysis Based on a New Subsample of the German Socio-Economic Panel (SOEP), Microgeographic Characteristics

and Survey-Based Interviewer Characteristics. SOEPpapers on Multidisciplinary Panel Data Research 288, DIW Berlin, The German Socio-Economic Panel (SOEP).

Schultz, T. W., 1963. The economic value of education. Vol. 63. Columbia University Press, New York.

Schulze, G., 1992. Die Erlebnisgesellschaft: Kultursoziologie der Gegenwart. Campus, Frankfurt.

Schupp, J., Hermann, S., 2009. Kognitionspotenziale Jugendlicher: Ergänzung zum Jugendfragebogen der Längsschnittstudie Sozio-oekonomisches Panel (SOEP). Data Documentation 43, DIW Berlin, The German Socio-Economic Panel (SOEP).

Schwartz, S. H., 2009. Universals in the Content and Structure of Values. Theoretical Advances and Empirical Test in 20 Countries. Advances in experimental social psychology 25, 1 – 65.

Shani, D., 2009. On the origins of political interest. Princeton University, Ann Arbor.

Sinus Sociovision, 2005. Die Sinus-Milieus® in Deutschland 2005: Informationen zum Forschungsansatz und zu den Milieu-Zielgruppen. Sinus, Heidelberg.

Spearman, C., 1904. "General Intelligence", objectively determined and measured. The American Journal of Psychology 15 (2), 201–292.

Spence, M., 1973. Job market signaling. The Quarterly Journal of Economics 87 (3), 355–374.

Spencer, S. J., Steele, C. M., Quinn, D. M., 1999. Stereotype threat and women's math performance. Journal of Experimental Social Psychology 35 (1), 4–28.

Stamm, M., 2005. Bildungsaspiration, Begabung und Schullaufbahn: Eltern als Erfolgspromotoren? Schweizerische Zeitschrift für Bildungswissenschaften 27 (2), 277–297.

Teachman, J. D., 1987. Family Background, Educational Resources, and Educational Attainment. American Sociological Review 52 (4), 548–557.

Ten Berge, J. M., Krijnen, W. P., Wansbeek, T., Shapiro, A., 1999. Some new results on correlation-preserving factor scores prediction methods. Linear Algebra and its Applications 289 (1), 311–318.

Thomson, G., 1951. The factorial analysis of human ability, 4th Edition. University of London Press, London.

Thurstone, L. L., 1931. Multiple factor analysis. Psychological Review 38 (5), 406.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B, 267–288.

Tibshirani, R. J., 2014. A general framework for fast stagewise algorithms. arXiv preprint arXiv:1408.5801.

Timm, N., 2002. Applied Multivariate Analysis. Springer, New York.

Tu, Y., Lee, T.-H., 2012. Forecasting using Supervised Factor Models. Working paper, University of California, Riverside.

Überla, K., 1968. Faktorenanalyse. Springer, Berlin - Heidelberg - New York.

Vervloet, M., Deun, K. V., den Noortgate, W. V., Ceulemans, E., 2013. On the selection of the weighting parameter value in Principal Covariates Regression. Chemometrics and Intelligent Laboratory Systems 123, 36 – 43.

Vervloet, M., Kiers, H., den Noortgate, W. V., Ceulemans, E., 2015. PCovR: An R Package for Principal Covariates Regression. Journal of Statistical Software 65 (8), 1–14.

Vester, M., 2009. Milieuspezifische Lebensführung und Gesundheit. Jahrbuch Kritische Medizin und Gesundheitswissenschaften, Bd 45, 36–56.

Vester, M., von Oertzen, P., Geiling, H., Hermann, T., Müller, D., 2001. Soziale Milieus im gesellschaftlichen Strukturwandel. Zwischen Integration und Ausgrenzung. Suhrkamp Verlag, Frankfurt am Main.

Vigneau, E., Devaux, M., Qannari, E., Robert, P., 1997. Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. Journal of chemometrics 11 (3), 239–249.

Wagner, G. G., Frick, J. R., Schupp, J., 2007. The German Socio-Economic Panel Study (SOEP) – Evolution, Scope and Enhancements. SOEPpapers on Multidisciplinary Panel Data Research 1, DIW Berlin, The German Socio-Economic Panel (SOEP).

Wang, H., Liu, Q., Tu, Y., 2005. Interpretation of partial least-squares regression models with varimax rotation. Computational statistics & data analysis 48 (1), 207–219.

Wichert, L., Pohlmeier, W., 2010. Female labor force participation and the big five. Discussion Paper 10-003, ZEW-Centre for European Economic Research.

Wippermann, C., 1998. Religion, Identität und Lebensführung: typische Konfigurationen in der fortgeschrittenen Moderne; mit einer empirischen Analyse zu Jugendlichen und jungen Erwachsenen. Leske + Budrich, Opladen.

Wold, S., Ruhe, A., Wold, H., W. J. Dunn, I., 1984. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. SIAM Journal on Scientific and Statistical Computing 5 (3), 735–743.

Wölfel, O., Heineck, G., 2012. Parental risk attitudes and children's secondary school track choice. Economics of Education Review 31 (5), 727–743.

Wößmann, L., Piopiunik, M., 2009. Was unzureichende Bildung kostet. Eine Berechnung der Folgekosten durch entgangenes Wirtschaftswachstum. Studie im Auftrag der Bertelsmann-Stiftung, Gütersloh.

Zhan, M., 2006. Assets, parental expectations and involvement, and children's educational performance. Children and Youth Services Review 28 (8), 961–975.

Zhang, P., 1993. Model selection via multifold Cross Validation. The Annals of Statistics 21 (1), 299–313.

Zou, H., 2006. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101 (476), 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2), 301–320.

Zou, H., Hastie, T., Tibshirani, R., et al., 2007. On the "degrees of freedom" of the lasso. The Annals of Statistics 35 (5), 2173–2192.

Zumbühl, M., Dohmen, T., Pfann, G., 2013. Parental investment and the intergenerational transmission of economic preferences and attitudes. SOEPpapers on Multidisciplinary Panel Data Research 570, DIW Berlin, The German Socio-Economic Panel (SOEP).

Table 8.10: Overview of used user-written packages

Package Name (Method)	Version	Author
lars (Lasso)	1.05	Mander (2014)
PCovR (Principal Covariates Regression)	2.6	Vervloet et al. (2015)
pls (Partial Least Squares)	2.0	Mevik and Wehrens (2007)
parcor (Adaptive Lasso)	0.2-6	Kraemer, N., Schaefer, J. (2014)
glmnet (Elastic Net)	2.0-5	Friedman et al. (2010),
caret (Cross-Validation)	6.0-70	Kuhn (2008)

The calculations in this thesis were mainly executed with Stata 11.2, Stata 13.1, R 3.0.2, R 3.1.1 and Microsoft Excel 2010. Table 8.10 lists the (user-written) packages that were accessed. Some illustrations were using Python (x,y) and the library scikit-learn (Pedregosa et al., 2011).

Declaration

Erklärung zum selbständigen Verfassen der Arbeit:

Ich erkläre hiermit an Eides Statt, dass ich meine Doktorarbeit "Exploring family life circumstances and their relationship to a child's school achievement – an econometric analysis in large data contexts" selbständig und ohne fremde Hilfe angefertigt habe und dass ich alle von anderen Autoren wörtlich übernommenen Stellen, wie auch die sich an die Gedanken anderer Autoren eng anlehnenden Ausführungen meiner Arbeit, besonders gekennzeichnet und die Quellen nach den mir angegebenen Richtlinien zitiert habe. Die Arbeit hat bisher in gleicher oder ähnlicher Form oder auszugsweise noch keiner Prüfungsbehörde vorgelegen.

Malte Hoffmann