Graphical Chain Models and their Application

Iris Pigeot, Stephan Klasen and Ronja Foraita

Abstract Graphical models are a powerful tool to analyze multivariate data sets that allow to reveal direct and indirect relationships and to visualize the association structure in a graph. As with any statistical analysis, however, the obtained results partly reflect the uncertainty being inherent in any type of data and depend on the selected variables to be included in the analysis, the coding of these variables and the selection strategy used to fit the graphical models to the data. This paper suggests that these issues may be even more crucial for graphical models than for simple regression analyses due to the large number of variables considered which means that a fitted graphical model has to be interpreted with caution. Sensitivity analyses might be recommended to assess the stability of the obtained results. This will be illustrated using a data set on undernutrition in Benin.

1 Introduction

The selection of an adequate model is a crucial task when modeling complex association structures. The results of a particular analysis about direct and indirect effects of covariates on response variables and the corresponding substantive conclusions can be strongly affected by the choice of the underlying model. Each of the candidate models has advantages but also limitations that impact the most relevant questions to be answered by the analysis, namely how do the variables involved affect our

Iris Pigeot, Ronja Foraita
Bremen Institute for Prevention Research and Social Medicine (BIPS), University of Bremen, D-28357 Bremen, Germany, URL: www.bips.uni-bremen.de,
e-mail: pigeot@bips.uni-bremen.de

Stephan Klasen
Department of Economics, University of Göttingen, D-37073 Göttingen, Germany, URL: www.uni-goettingen.de/en/64094.html,
e-mail: sklasen@uni-goettingen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_13,

outcomes of interest and are there possibly interactions between them which also influence our response variable. In addition to the choice of the statistical model itself, the selection and the coding of the variables to be included in the analysis is another critical aspect in an empirical investigation.

In this paper we focus on the application of graphical models that are still a rather novel, though powerful statistical tool to analyze multivariate data sets. Graphical models are specifically suited for the analysis of complex association structures and provide a graphical representation of certain independence properties among the variables of interest. We restrict ourselves here to so-called graphical chain models that allow to reveal indirect associations and to identify hidden relationships. We demonstrate the challenges that are related to the interpretation of graphical chain models by especially investigating their robustness with respect to a change of the variables included in the analysis or a different coding.

The challenges that are illustrated by using a highly complex data example are of course not limited to these statistical techniques. The complexity of the model, however, adds to the difficulties which are inherent to any statistical modeling approach in an empirical investigation.

We will illustrate the above mentioned challenges in obtaining valid and meaningful results by considering the example of childhood undernutrition which is one of the most important health problems in developing countries. That is we are interested in modeling the determinants of undernutrition among children which is a complex undertaking. Although it seems as if the determinants of undernutrition are quite clear, namely inadequate dietary intake and incidence, severity, and duration of disease, these factors themselves are related to a large number of intermediate, underlying, and basic causes operating at the household, community, or national level (UNICEF 1998). Among the most important factors are probably the education, wealth, and income situation of the parents, household size, birth order, religion, and sex of the child, the availability of clean water, adequate sanitation, immunization, and primary health care services, and the level of disease prevalence in the surrounding community. Noteworthy, the association structure between these factors is assumed to be fairly complex. In fact, UNICEF has made a useful distinction between immediate, intermediate, and underlying causes of undernutrition. Any empirical strategy that attempts to identify the determinants of undernutrition must recognize the existence of such a dependence chain. This suggests that a simple multivariate regression model is not appropriate to capture the indirect associations and the overall complex association structure.

To determine whether an individual child suffers from undernutrition of the three forms, i.e. insufficient height for age (stunting) indicating chronic undernutrition, insufficient weight for height (wasting) indicating acute undernutrition, and insufficient weight for age (underweight) indicating acute and/or chronic undernutrition, the anthropometric indicator of the child is compared with a reference population by means of a Z-score:

$$Z_i = \frac{AI_i - MAI}{\sigma}$$

where AI_i refers to the individuals anthropometric indicator (weight at a certain height, height at a certain age, or weight at a certain age), MAI refers to the median of the reference population, and σ refers to the standard deviation of the reference population (Gorstein et al. 1994, WHO 1995). The Z-score thus measures the distance, expressed in standard deviations of the reference population, between the individuals anthropometry and the median of the reference population, where both populations are presumed to be normally distributed. While the average Z-score is likely to give an accurate picture of undernutrition at the population level, for an individual child it might be misleading as genetic influences of the parents are likely to affect it and thus bias the findings. There is also some on-going debate whether this might bias findings on undernutrition between different continents as there might be genetic differences particularly in the height potential of children (WHO 1995, Klasen 2003, Klasen 2008). In particular, there is a question whether the very high reported rates of undernutrition in South Asia, compared to Sub-Saharan Africa, are partly related to this question.

For the purpose of this paper we use data from the 1996 Demographic and Health Survey to fit a graphical chain model for undernutrition in Benin, West Africa. Our discussion of the challenges related to the modeling of the association structure will be based on two analyses conducted by the authors (Caputo et al. 2003, Foraita et al. 2008). The research work on the analysis of undernutrition with the help of graphical chain models was started within a subproject of the DFG-funded Collaborative Research Center 386 “Statistical Analysis of Discrete Structures: Modelling and Application in Biometrics and Econometrics" which was successfully coordinated by Ludwig Fahrmeir. In the first analysis we made full use of the data being available for Benin. In contrast, the second analysis had to be restricted to variables that were available for both Benin and Bangladesh for comparative reasons. The comparison of these two analyses sheds some light on the differences in the results explaining undernutrition in Benin. In addition to discussing the major differences between the two resulting models we will also describe the most important overlaps.

The paper is organized as follows. Section 2 gives an introduction to the theory of graphical chain models. In Section 3 we briefly describe the selection strategy we used for fitting such a model to our multivariate data set. We then present the data set in Section 4 where we also provide some descriptive statistics. Section 5 gives a detailed discussion of the results, while Section 6 concludes.

2 Graphical Chain Models

Graphical models are probability models for multivariate observations to analyze and visualize conditional relationships between random variables encoded by a conditional independence graph. In contrast to regression models, graphical modeling is concerned with identifying association structures for all study variables, including those which usually are regarded as explanatory. They are therefore appropriate in situations where complex associations have to be dealt with. Due to the visualization

in graphs, these models make it easier to display complex dependence structures. Furthermore, they can handle simultaneously categorical and continuous variables.

We denote an arbitrary graph by $G = (V, E)$ where $V = \{1, \ldots, K\}$ is a set of vertices representing the components of a multivariate random vector $X_V = (X_1, \ldots, X_K)$ and $E \subseteq V \times V$ is a set of edges. For $i, j \in V$, there is a symmetric association between two vertices i and j and a line in the graph (also called undirected edge) if $(i, j) \wedge (j, i) \in E$ whereas $(i, j) \in E \wedge (j, i) \notin E$ corresponds to an asymmetric association and an arrow in the graph (also known as directed edge), pointing from i to j. Semi-directed cycles are not allowed, i.e. sequences $a = i_0, \ldots, i_r = a$ with $(i_{k-1}, i_k) \in E \wedge (i_k, i_{k-1}) \notin E$ for at least one value of k.

The structure of the conditional relationships among random variables can be explored with the help of Markov properties (Lauritzen & Wermuth 1989, Frydenberg 1990). For instance, the pairwise Markov property claims

$$X_i \perp\!\!\!\perp X_j | X_{V^* \setminus \{i,j\}} \text{ whenever } (i, j), (j, i) \notin E$$

where V^* consists of all variables prior to or at the same level as i and j and the symbol $\perp\!\!\!\perp$ stands for conditional independence between X_i and X_j given $X_{V^* \setminus \{i,j\}}$. This implies that a missing edge can be interpreted as conditional independence. However this is only justified if the underlying multivariate statistical distribution fulfills the Markov properties since they lead to a factorization of the multivariate density and thus to a decomposition into smaller models and equivalently cliques, which are maximal complete subgraphs.

Graphical chain models are suitable to account for prior substantial knowledge of an underlying dependence structure by forming a dependence chain where all variables are partitioned into an ordered sequence of disjoint subsets $V_1 \cup \ldots \cup V_R$. The subsets are called blocks and all edges within V_r are undirected and all edges between V_r and V_s are directed from V_r to V_s for $r < s$. The blocks $V_1, \ldots, V_R$ are ordered due to subject-matter knowledge, so that the rightmost block contains the pure explanatory variables, the leftmost block the pure responses and the blocks between contain variables that are simultaneously responses to variables in previous blocks and potentially explanatory to variables in future blocks. Variables in these in-between blocks are intermediates and, in contrast to usual regression models, allow for modeling possibly indirect influences. Variables in the same block are assumed to be on equal footing, i.e. no sensible response-explanatory relationship can be assumed within this subset. The random vector X_V is divided into subvectors $X_{V_1}, \ldots, X_{V_R}$ such that the joint density $f(x_V)$ factorizes into a product of conditional densities as

$$f(x_V) = f(x_{V_1}) \prod_{r=2}^{R} f(x_{V_r}) | x_{V_1}, \ldots, x_{V_{r-1}}). \tag{1}$$

Each of the factors in (1) corresponds to the distribution of variables at one level conditional on variables at all lower levels. Thus, one may regard a graphical chain model as a sequence of regression models that describe these conditional distributions and the choice of the recursive structure reflects that one is specifically interested in the latter.

If mixed models are investigated, i.e. models including continuous as well as discrete variables, the distribution considered is the Conditional Gaussian distribution (CG-distribution), where the continuous variables are multivariate normal given the discrete. For further reading we refer to Lauritzen (1996), Cox & Wermuth (1996), Edwards (2000) and Green et al. (2003) and the references therein.

3 Model Selection

In our study we have to deal with mixed variables and also with a large number of variables. Thus, we are confronted with the problem to select those variables that are most influential for the response and to find the most appropriate association structure among them. A possible solution to this problem is the data-driven Cox-Wermuth selection strategy (Cox & Wermuth 1993, Cox & Wermuth 1994). This strategy exploits that each conditional density of the factorization is described by a system of multiple univariate regressions. The kind of regression used depends on the measurement scale of the involved univariate response. A problem of this strategy is that fitting multiple univariate regressions neglects the multivariate structure of the data and the validity of the equivalence of the Markovian properties is not ensured for the whole graph. Nevertheless, for large and complex graphs with mixed variables it is still the only feasible computer algorithm which is implemented in the software GraphFitI (Blauth et al. 2000).

The Cox-Wermuth selection strategy consists of roughly two steps: First, a screening for second-order interactions and non-linearities is performed (Cox & Wermuth 1994); second, a system of forward and backward regressions depending on the scale of the response variable is carried out.

In the screening procedure, the search for second-order interactions is based on the calculation of t-statistics derived from trivariate regressions, such as X_a on $X_b, X_c, X_b X_c$ with $X_a \in V_s$ and $X_b, X_c \in V_r, s \geq r$, where each X_a has to be regressed on all possible pairs of variables in the same block and in previous blocks as well as on their pairwise interaction. In case of large sample sizes and if there is no interaction, the t-statistics approximately follow a standard normal distribution. The ordered t-statistics are plotted against their expected values obtained from the standard normal distribution. If the assumption of no interactions is fulfilled, the points spread along the diagonal. Checking for non-linearities is performed similarly. All interactions and non-linearities with a $|t|$-value > 4 are considered in further steps.

To derive the graph a multivariate response model is needed for each V_r given $V_1 \cup \ldots \cup V_{r-1}$. The Cox-Wermuth strategy splits the problem of multivariate regressions into a system of univariate regressions for each variable X_a on the remaining variables in the same block and on all explanatories in the previous blocks. First, a forward selection investigates whether the detected interactions or non-linearities from the screening step have to be added into the set of covariates regarding X_a. This selection is based on statistical tests with $\alpha = 0.1$. The corresponding p-values have, however, to be interpreted in an exploratory sense since no adjustment for multiplicity takes

place. Then, a backward selection strategy for X_a is used on the preliminary set of covariates. In each step, the covariate with the smallest corresponding $|t|$-value is excluded until the remaining covariates all come up with a p-value smaller than 0.05. After that the remaining variables are checked again for interactions and non-linearities. All qualitative interaction terms and mixed interactions terms are included in the model equation. Again, a backward selection as described above is carried out. Finally, all quantitative interaction terms and non-linearities are introduced into the model. The final backward selection leads to the reduced model that should capture the underlying association structure.

4 Data Set

The data is part of the 1996 Demographic and Health Surveys (DHS, Macro 1996). These surveys are conducted regularly by the National Statistical Institutes in collaboration with Macro International, a US-based company that operates on behalf of the US Agency for International Development, in several countries of Africa, Asia, Latin America and the Near East. The DHS is based on a representative sample of women of reproductive age. These women are administered an extensive questionnaire covering a broad range of items regarding household structure, socioeconomic status, health access and behavior, fertility behavior, reproductive health, and HIV/AIDS. The questionnaire also contains items about the children including prenatal and postnatal care, nutrition, health, immunization, and care practices. Some parts or questions of the survey have been disregarded in some countries. In this study we focus on the DHS data set from Benin and involve only children between twelve and 35 months. We focus on these age group since by that age the children surely have been introduced to additional foods and water and therefore have already been through the weaning crisis associated with this transition. For older children the DHS survey does not collect data about their nutrition and health status. If the respondent has more than one child belonging to this age group we only select the younger one.

We compare the data sets from Caputo et al. (2003) and Foraita et al. (2008) which in this paper are abbreviated with *A* and *B* respectively. While Caputo et al. (2003) only focus on Benin, Foraita et al. (2008) compare the different patterns of malnutrition in Benin and Bangladesh. Although both papers are based on the DHS 1996 Benin data set, they vary in the variables that causes different total sample sizes ($N_A = 1076, N_B = 1122$) since both data sets are constrained to complete cases. Additionally, some variables have a different coding scheme or the reference category has changed (see Section 4.1 for more details).

4.1 Summary Measures

In order to capture the determinants of undernutrition and not to miss a relevant influence, a large number of variables are included in the model. In this section, we briefly introduce the variables, their scales and coding. Table 1 gives absolute and relative frequencies of binary and polytomous variables and Table 2 summarizes mean, median and the 25^{th} and 75^{th} percentile of continuous variables. This distinction between the various scales is not only convenient for their presentation, but also needed for choosing the adequate regression models in later analysis.

The response variables *stunting* (*St*) and *wasting* (*Wa*) are both continuous anthropometric indicators that measure malnutrition using the *Z*–score. Stunting reflects chronic malnutrition, whereas wasting stands for acute malnutrition (see Section 1). The different composition of the two analysis data sets has already an impact on the stunting and wasting. To be more specific, in data set *B* the children are slightly more stunted and less wasted than in data set *A*. The investigation of the impact of the child's nourishment focuses on the quality of food, measured by the number of meals containing *protein* (*P*) during one day. In data set *A*, *P* shows the absolute frequency of meals containing milk, meat, egg, fish or poultry whereas in data set *B* only the meals containing milk or meat are counted. The difference in this operationalization heavily affects the distribution of *P*: for data set *A*, 20% of all children had no protein in their meal compared to 65% in data set *B*. Data set *A* contains the further aspect of food quantity (*F*) which counts the number of meals a child has had during the day. Since this variable was not available for the Bangladesh data set, the number of meals was not included in data set *B*. Comparing both data sets to further nutritional variables, no essential difference can be seen with respect to the time when the children are *put to breast* or the duration of exclusively *breast–feeding* that is on average unusually long with about 19 months which has to be interpreted as an indicator of high poverty and lack of alternative food. The security of nutrition for the child is represented by the mother's body mass index *BMI*. A large BMI can be interpreted as sufficient nourishment of the whole family, whereas a low BMI indicates an uncertain nourishment.

Another important influence on the child's physical status is its current health situation. Therefore, the variable *ill* counts children who suffered from diarrhea or cough during the last two weeks before the interview. Due to the short observation time, one may presume an effect on *wasting*.

In both data sets half of the children have to be regarded as ill. In both data sets nearly 78% of the mothers have access to modern health care, measured by *prenatal and birth attendance* score (*BPA*). The variable *vaccination* (*V*) counts the number of vaccinations a child has already had. It may be considered as a substitute of health knowledge, but also of access to health care. Furthermore, the access to clean water and clean sanitation is important. These variables have been operationalized differently for both data sets. In data set *A* only piped water and flush toilet or all kinds of pit latrines has been categorized as high quality (around 24%) compared to data set *B* where piped as well as well water (50%) and only flush toilets but no open latrines (13%) has been regarded as high quality.

Table 1 Absolute and relative frequencies of binary and polytomous variables.

			A ($N = 1076$)		B ($N = 1122$)	
	Variable	Category	Freq	%	Freq	%
P	***protein* intakes yesterday**	0	216	20.1	734	65.4
		1	607	56.4	307	27.4
		2	207	19.2	81	7.2
		3	46	4.3	-	-
ILL	child was *ill* during the last 14 days	no (*A*)	539	50.1	560	49.9
		yes (*B*)	537	49.9	562	50.1
PB	when child *put to breast*	immediately	241	22.4	249	22.2
		within 6 hours	326	30.3	340	30.3
		first day	271	25.2	282	25.1
		2 days or more	238	22.1	251	22.4
BPA	*prenatal and birth attendance*	nothing	28	2.6	29	2.6
		other	133	12.4	136	12.1
		traditional	78	7.3	84	7.5
		modern	837	77.8	873	77.8
W	**source of drinking *water***	low quality	822	76.4	566	50.5
		high quality	254	23.6	556	49.6
T	**type of *toilet* facility**	low quality	876	81.4	972	86.6
		high quality	200	18.6	150	13.4
Rel	*religion*	Islam (*B*)	235	21.8	245	21.8
		Traditional	278	25.84	290	25.8
		Christianity	246	39.6	444	39.6
		no religion (*A*)	137	12.7	143	12.8
Sex	*sex* of child	male	551	51.2	575	51.3
		female	525	48.8	547	48.8
HH	relationship to *household head*	relative	174	16.2	183	16.3
		wife	842	78.3	871	77.6
		head (*B*)	40	3.7	43	3.8
		not related (*A*)	20	1.9	25	2.2
H	*house* quality	low quality	555	51.6	580	51.7
		high quality	521	48.4	542	48.3
Wo	current type of *employment*	paid employee	83	7.1	85	7.6
		self-employed	891	82.8	933	83.2
		unpaid worker (*A*)	44	4.1	45	4.0
		did not work (*B*)	58	5.4	59	5.3

A and *B* mark different reference categories in the respective data sets and bold written variables mark a different coding scheme in both data sets.

Additionally, various socioeconomic factors have been included into both data sets like the current type of *employment* of the mother (*Wo*), having a different reference category in the data sets, three proxies for the economic situation of the household (house quality (*H*), durable goods (*G*) and mother's height (*Ht*)) or the educational

Table 2 Summary measures of the continuous variables.

		A				*B*			
	Variable	Mean	Median	Q_1	Q_3	Mean	Median	Q_1	Q_3
St	*stunting* (*Z*-score·100)	**-147.3**	**-153.0**	**-232.5**	-62.0	**-148.2**	**-154.5**	**-233.0**	-62.0
Wa	*wasting* (*Z*-score·100)	**-93.6**	**-95.0**	**-165.5**	**-20.5**	**-92.7**	**-94.0**	**-164.0**	**-20.0**
F	*food*	**4.2**	**4.0**	**-3.0**	**5.0**	-	-	-	-
BF	duration of *breast-feeding* in months	18.6	18.0	15.0	22.0	18.6	18.0	15.0	22.0
BMI	*body mass index*	**21.3**	20.8	19.2	**22.5**	**21.4**	20.8	19.2	**22.6**
Ht	mother's *height*	158.2	**158.2**	154.0	162.2	158.2	**158.3**	154.0	162.2
V	*vaccination* score	6.2	8.0	5.0	8.0	6.2	8.0	5.0	8.0
Bo	*birth order number*	4.1	4.0	2.0	6.0	4.1	4.0	2.0	6.0
Age	*age* in months	22.5	22.0	16.0	28.0	22.5	22.0	16.0	28.0
HM	no. of *household members*	8.9	8.0	5.0	11.0	8.9	8.0	5.0	11.0
TC	*total children* ever born	**4.2**	**4.0**	**2.0**	**6.0**	-	-	-	-
CD	*deceased children* in %	12.5	0.0	0.0	25.0	12.5	0.0	0.0	25.0
1^{st}	age of mother at *first birth*	19.0	19.0	17.0	21.0	19.0	19.0	17.0	21.0
G	durable *goods* in %	23.8	28.6	14.2	28.6	23.9	28.6	14.3	28.6
EM	*mother's education* in years	0.9	0.0	0.0	0.0	0.9	0.0	0.0	0.0
EP	*partner's eduction* in years	2.3	0.0	0.0	4.0	2.3	0.0	0.0	4.0

Q_1: 25^{th} percentile; Q_3: 75^{th} percentile. Bold figures mark differences in data sets.

background in the family. The duration of education in years is very short for mothers with a 75^{th} percentile of 0 years and it is slightly better for their partners with around 4 years.

4.2 Dependence Chain

In line with UNICEF (1998), Caputo et al. (2003) and Foraita et al. (2008) postulated a dependence chain as follows (see Figure 1), from left to right: in the first block we put our pure response variables *wasting* and *stunting*.

The second block includes all variables that have an immediate influence on these *Z*-scores, where we include variables relating to nutritional intake and illness episodes (*ill*, *protein* and *food* for data set *A*).

The next block includes intermediate variables that reflect care practices, health knowledge, and access to water and sanitary services. It includes child care (*breast-feeding*, time *put to breast*, *vaccination*, *prenatal and birth attendance*, mother's *BMI*) and sanitary facilities (quality of drinking *water*, quality of *toilet* facilities).

The last block on the right includes basic variables affecting the ability of households to take care for their children, including demographic factors (*age*, *sex*, *birth order number* and additionally in data set *A* the variable *total children ever born*, *religion*, number of *household members*, number of *deceased children*, age of mother at *first birth*, relation between mother and *household head*), socioeconomic factors (*house* quality, fraction of durable *goods*, *mother's education*, *partner's education*,

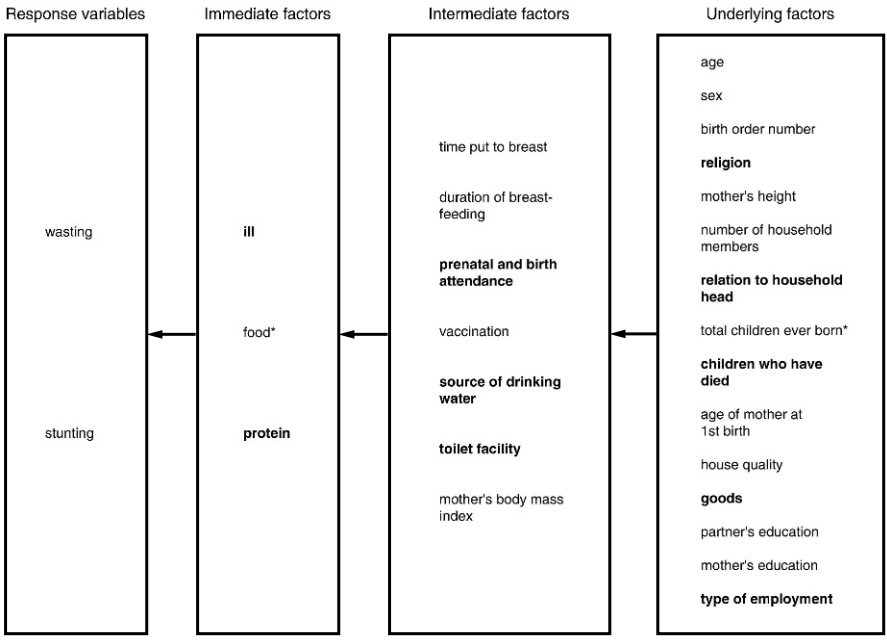

Fig. 1 The postulated chain. Pure responses are the undernutrition variables *wasting* and *stunting*. Immediate factors reflect food quality and the state of health, intermediate factors are variables of health care, health knowledge, food security, and sanitary facilities. Demographic and socioeconomic factors are put in the block of the underlying factors. Bold written variables indicate different coding schemes in both data sets and * indicates variables that are only included in data set *A*.

current type of *employment*) and mother's *height* as combination of socioeconomic and nutritional aspects that we call underlying factors. In the appendix a more detailed description of the variables is given.

5 Results

Figures 2–4 show the fitted graphical chain models for data set *A* and *B* and their common edges. The common edges in both analyses are shown in black in Figures 2 and 3 and are separately presented in Figure 4. The edges that are specific to analysis *A* and *B* are shown in grey in Figures 2 and 3, respectively.

While the figures may at first glance appear rather complicated, closer inspection reveals a number of interesting points.

Although both data sets differ only slightly, the graphs show several notable differences. This is surely on the one hand due to the omission of the variables *total children ever born* and in particular *food* that attracts many influences from the intermediate and underlying factors and on the other hand due to the change of some

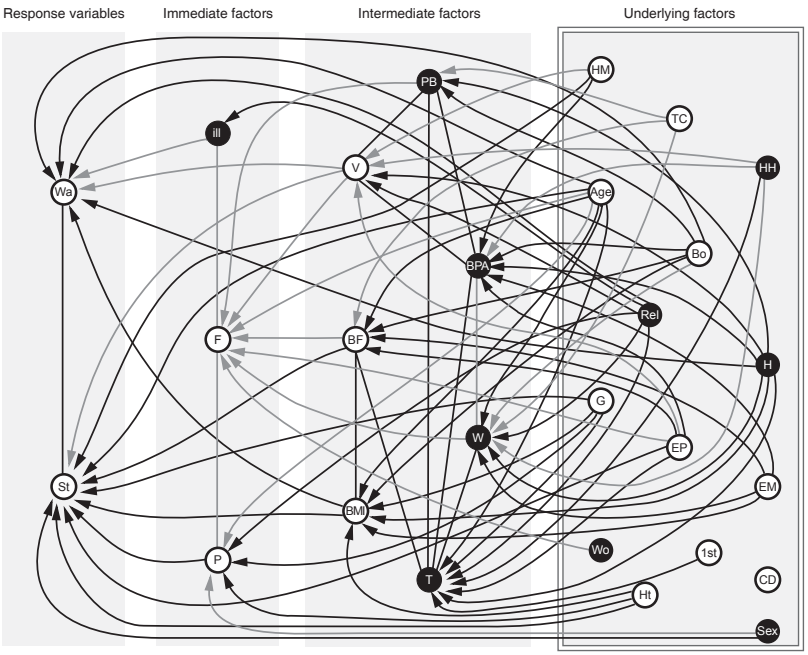

Fig. 2 Undernutrition in Benin – Fitted graphical chain model using data set *A*. Black dots represent discrete and white circles continuous variables. The double-lined box indicates that the associations among the variables within this box are not shown. Edges in black are common to both analyses; edges marked in grey are only specific to the analysis using data set *A*. For abbreviations of variables see Tables 1 and 2 or Figure 4.

categorical variables as well as the inclusion of further children which affects the variability between the data sets. Especially the variable *food* acts as some kind of hub in data set *A* that forwards the influences of many intermediate and underlying factor though its connections to *protein* and *ill*. In data set *B* it seems that *ill* has partly inherited the role since it works as endpoint for many underlying factors, with the difference that these influences are not carried forward to the response variables.

Although we have not substantially changed the data sets, we can see in Figures 2 and 3 that there is a certain amount of uncertainty in the data we have to be aware of. Edges that we detected in both analyses, seem to be more stable. Hence, our interpretation will only be based on those pathways.

First, in both data sets *protein* consumption has a direct influence on *stunting*, even though the variable has been recoded for data set *B*. Second, there are many direct and indirect influences from those variables reflecting the economic condition of the household, e.g. proxied by *partner's eduction*, *house quality* and *goods*. Third, mother's *BMI* and duration of *breast–feeding* are important players as intervening variable between underlying factors and the response variables *stunting* and *wasting*. However, the role of mother's *BMI* in the network is twofold. It can be regarded as

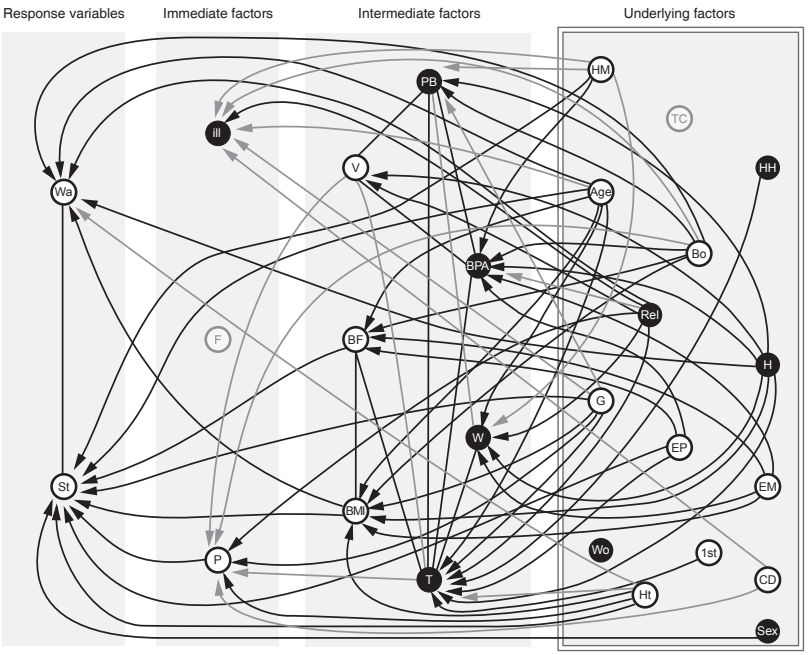

Fig. 3 Undernutrition in Benin – Fitted graphical chain model using data set *B*. The variables food (*F*) and total children ever born (*TC*) are not included in fitting the graph. Edges in black are common to both analyses; edges marked in grey are only specific to the analysis using data set *B*.

a variable that in some extent reflects the economic situation of the household in the sense that children of well-nourished mothers are also well-nourished. But *BMI* may also partly capture the genetic influences of the mother's anthropometry on her children. *Breast–feeding* is directly associated with *stunting* and has an indirect influence on *wasting* through *BMI*. The WHO (WHO 1995) recommends that after six months of exclusively breast–feeding, children need additional food. Hence extended periods of breast–feeding may be an indicator of the inability of the household to provide for such supplemental foods. Table 2 shows that on average the children are breast-fed for more than 18 months. The 25%-percentile equals 15 months. Forth, the eduction of the mother shows a rather indirect influence on the response via *BMI*, duration of *breast-feeding*, source of drinking *water* and *prenatal and birth attendance* which means that especially *stunting* is influenced in numerous ways by *mother's education*. The results suggest that more years of education is associated with a shorter duration of breast-feeding and better nourished mothers. It is more likely that these mothers make use of a modern health service and have access to clean water. Fifth, the *toilet* variable is noticeable since many influences point on it, but *toilet* itself is connected with *stunting* only via the *breast-feeding* link. Sixth, *religion* is an important factor in Benin. It is associated with many aspects in the dependence chain of undernutrition. It is directly linked with *wasting*, the current

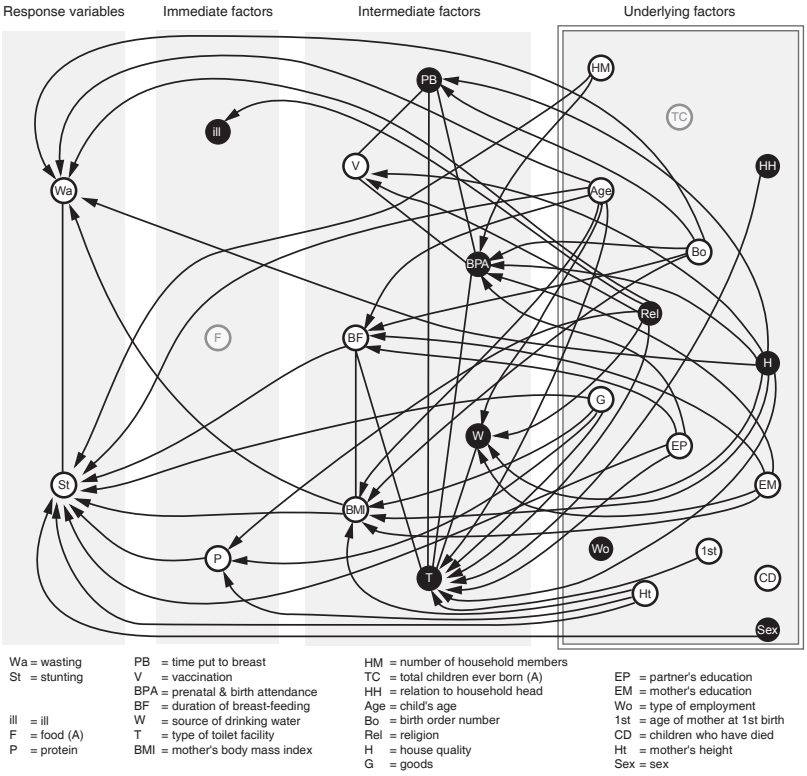

Fig. 4 Common edges in data sets *A* and *B*.

health situation of the child (*ill*), quality of food (*P*), access to health care (*V*) and to the access of clean water (*W*) and sanitation (*T*). Generally, belonging to any kind of religion seems to be favorable for children in Benin.

6 Discussion

As it became evident in the last section, the results that we obtain from a statistical analysis with graphical chain models should be interpreted with caution and critically reflected regarding their substantive message. Although the most important conclusions from the above exploratory data analysis remained stable regardless of the variables included and their coding, some results were rather different in both graphs. This problem is of course more severe in analyses with a huge number of variables involved that all may explain the response directly or indirectly. This is also

not so surprising as the omission of some variables will naturally lead to the effect being captured by closely correlated ones. A related problem is due to the fact that there is always more than one model which is consistent with the data, and typically different model selection strategies will lead to different results.

To get a better understanding of the mechanisms that led to these differences we again fitted a graphical chain model to the original data set *A* where we used the same coding but left out the variables *food* and *total children ever born*. The resulting Figure 5 (see Appendix) differs from Figure 2 mostly in those edges that are due to the above variables; only very few other edges are affected. Figure 3 that is based on data set *B* shows much more differences compared to Figure 5 which means that relatively small differences in coding and number of observations can have a substantial difference on the detected associations. Examples are that *ill* has a direct influence on *wasting*, or that *vaccination* directly influences *stunting* which is each shown in Figure 5, but does not appear in Figure 3. Thus, it is strongly recommended to not only think about the variables to be included but to also carefully think about the coding of variables and to carry out some sensitivity analysis based on various coding schemes.

Whereas a simple linear regression model can be assessed by the coefficient of determination R^2, no comparable measure exists to assess a graphical model. Thus, it is recommended to at least perform some kind of sensitivity analysis. For this purpose, often we suggest to estimate different reasonable models and compare their most important results. As an alternative, the bootstrap offers a valuable opportunity in two ways (Friedman et al. 1999b, Friedman et al. 1999a, Steck & Jaakkola 2004). On the one hand, it allows to generate repeated samples out of the original one which can be used to fit the model of interest repeatedly and to compare the variety of selected models. This gives an idea about the stability of the originally selected one. (For graphical models with only discrete data, the R-package gmvalid (Foraita & Sobotka 2008) provides functions that apply the bootstrap to investigate the uncertainty of graphical models.) The resulting models can, on the other hand, be exploited to derive measures of uncertainty which are especially appropriate to assess the validity of a selected graphical model. The development of such measures and their evaluation by means of real data examples and simulated data sets is currently under research by the authors.

References

Blauth, A., Pigeot, I. & Bry, F. (2000). Interactive analysis of high–dimensional association structures with graphical models, *Metrika* **51**: 53–65.

Caputo, A., Foraita, R., Klasen, S. & Pigeot, I. (2003). Undernutrition in Benin - An analysis based on graphical models, *Social Science & Medicine* **56**: 1677–1697.

Cox, D. R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion), *Statistical Science* **8**: 204–283.

Cox, D. R. & Wermuth, N. (1994). Tests of linearity, multivariate normality and the adequacy of linear scores, *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **43**: 347–355.

Cox, D. R. & Wermuth, N. (1996). *Multivariate Dependencies*, Chapman & Hall, London.

Edwards, D. (2000). *Introduction to Graphical Modelling*, 2 edn, Springer, New York.

Foraita, R., Klasen, S. & Pigeot, I. (2008). Using graphical chain models to analyze differences in structural correlates of undernutrition in Benin and Bangladesh, *Economics and Human Biology* **6**: 398—-419.

Foraita, R. & Sobotka, F. (2008). *gmvalid: Validation of graphical models*. R package version 1.2.

Friedman, N., Goldszmidt, M. & Wyner, A. (1999a). Data analysis with bayesian networks: A bootstrap approach.

Friedman, N., Goldszmidt, M. & Wyner, A. (1999b). On the application of the bootstrap for computing confidence measures on features of induced bayesian networks.

Frydenberg, M. (1990). The chain graph markov property, *Scandinavian Journal of Statistics* **17**: 333–353.

Gorstein, J., Sullivan, K., Yip, R., de Onis, M., Trowbridge, F., Fajans, P. & Clugston, G. (1994). Issues in the assessment of nutritional status using anthropometry., *Bull World Health Organ* **72**: 273–283.

Green, P. J., Hjort, N. L. & Richardson, S. (eds) (2003). *Highly Structured Stochastic Systems*, Oxford University Press, Oxford.

Klasen, S. (2003). Malnourished and surviving in South Asia, better nourished and dying young in Africa: what can explain this puzzle?, *Measurement and Assessment of Food Deprivation and Undernutrition*, FAO, Rome, pp. 283–287.

Klasen, S. (2008). Poverty, undernutrition, and child mortality: Some inter-regional puzzles and their implications for research and policy, *Journal of Economic Inequality* **6**: 89–115.

Lauritzen, S. L. (1996). *Graphical Models.*, Clarendon Press, Oxford.

Lauritzen, S. L. & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative, *Annals of Statistics* **17**: 31–57.

Macro (1996). MEASURE DHS datasets Bangladesh, Benin. `www.measuredhs.com`, last accessed: 10. Aug 2009.

Steck, H. & Jaakkola, T. S. (2004). Bias-corrected bootstrap and model uncertainty, *in* S. Thrun, L. Saul & B. Schölkopf (eds), *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA.

UNICEF (1998). *The State of the World's Children: Focus on Nutrition.*, UNICEF, New York.

WHO (1995). Physical status: The use and interpretation of anthropometry, *WHO Technical Report Series 854*, WHO, Geneva.

Appendix

Table 3 Further explanations of some variables. Abbreviations *A* or *B* indicate the respective data set.

Variable		Category	Comments
ill	child was *ill*	no yes	child suffered from diarrhoea or cough during the last 14 days
P	*Protein* intakes yesterday (*A*)	0-3	remembered number of meals containing milk, meat, egg, fish or poultry
P	*Protein* intakes yesterday (*B*)	0-2	remembered number of meals containing milk or meat
W	source of drinking *water* (*A*)	low quality high quality	unprotected well or surface water piped or well water
W	source of drinking *water* (*B*)	low quality high quality	unprotected well or surface water piped water
T	type of *toilet* facility (*A*)	low quality high quality	well water, open latrine, no facility or “other" toilets flush toilet, pit toilet latrine, open latrine
T	type of *toilet* facility (*B*)	low quality high quality	no facility or “other" toilets flush toilet or pit toilet latrine
G	durable *goods* in %	$[0, 1]$	averaged sum score out of if the house has electricity, radio, television, refrigerator, bicycle, motorcycle, car and telephone
H	*house* quality	low quality high quality	main floor material is natural all other materials (i.e. wood, cement ...)

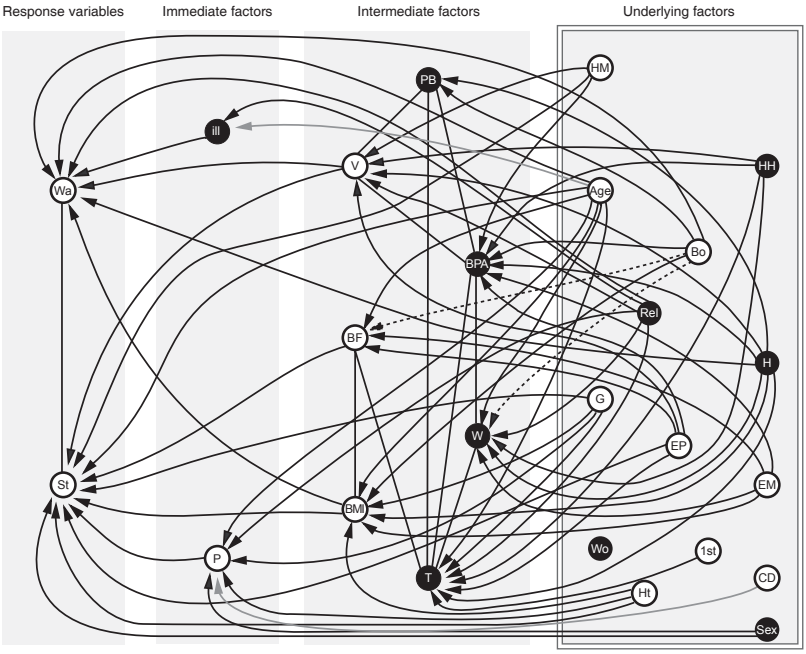

Fig. 5 Fitted graphical model using data set *A* without variables *food* and *total children ever born*. The associations $Age \rightarrow ill$ and $CD \rightarrow P$ are new (in grey), whereas the edges $Bo \rightarrow BF$ and $Bo \rightarrow W$ disappeared (marked as dotted lines).