BAYESIAN METHODS THAT OPTIMIZE CROSS-CULTURAL DATA ANALYSIS

Petri Nokelainen and Henry Tirri

January 15, 2004

HIIT
TECHNICAL
REPORT
2004–2

BAYESIAN METHODS THAT OPTIMIZE CROSS-CULTURAL DATA ANALYSIS

Petri Nokelainen and Henry Tirri

Helsinki Institute for Information Technology HIIT
Tammasaarenkatu 3, Helsinki, Finland
PO BOX 9800
FI-02015 TKK, Finland
http://www.hiit.fi

HIIT Technical Reports 2004–2
ISSN 1458-9478
URL: http://cosco.hiit.fi/Articles/hiit-2004-2.pdf

CHAPTER 8

BAYESIAN METHODS THAT OPTIMIZE CROSS-CULTURAL DATA ANALYSIS

Petri Nokelainen

University of Tampere

Henry Tirri

University of Helsinki

Introduction

In this chapter we first discuss about the typical problems of quantitative cross-cultural data analysis and describe the essential benefits of using Bayesian modeling. Next we describe the theoretical concepts of Bayesian modeling and illustrate their use in data analysis with excerpts from our preceding empirical studies. The last part of this chapter introduces B-Course, a free web-based online Bayesian classification and dependency modeling data analysis tool suitable for many data analysis needs rising from cross-cultural research.

In the social science researchers point of view the requirements that should be met in order to be able to conduct traditional frequentistic statistical analysis properly are very challenging. For example, the assumption of normality of both the phenomena under investigation and the data is prerequisite for traditional frequentistic calculations. Marini, Li and Fan (1996) state that in situations where a latent construct cannot be appropriately represented as a continuous variable, or where ordinal or discrete indicators do not reflect underlying continuous variables, or where the latent variables cannot be assumed to be normally distributed, traditional Gaussian modeling is clearly not appropriate. In addition, normal distribution analysis sets minimum requirements for the number of observations, and the measurement level of variables to be continuous. Bayesian modeling approach is a good alternative to traditional frequentistic statistics as it

In J. R. Campbell, K. Tirri, P. Ruohotie, & H. Walberg (Eds.) Cross-cultural Research: Basic Issues, Dilemmas, and Strategies. Research Centre for Vocational Education, University of Tampere, Finland.

is capable of handling both small discrete, non-normal samples and large scale continuous data sets.

Next we present the essential benefits of using Bayesian modeling with empirical samples in quantitative cross-cultural research.

1 Theoretical minimum sample size is zero. This is due to fact, that we model the phenomenon, not the data (the latter is the case in traditional Gaussian modeling approach). However, Bayesian modeling is also capable of scaling up to meet the requirements of large data modeling tasks.

2 Bayesian modeling is based on probabilities, thus allowing prediction with the model. For example, researcher is able to "fix" interesting values of variables in the Bayesian Network model, and further investigate the effect of her actions on conditional distributions of the other variables in the model.

3 Bayesian modeling is inductive, as the model is constructed from the data. In practice this means that we are not able to test hypotheses in a traditional way with the p-value.

4 Researcher is able to input a priori information to the model. The source of subjective information could be, for example, an interview with an expert of certain topic, or previously collected data. Nokelainen et al. (2002) have implemented an adaptive online questionnaire that profiles users online with Bayesian probabilistic modeling. A priori profile information is used to reduce the number of questions. Investigations with numerous empirical samples suggest that if a priori information (i.e., "learning data") is collected from the same or suchlike population, only approximately 35% of the questions are needed to achieve 99% accuracy in all the remaining (i.e. unasked) responses.

5 Cross-cultural researchers collect vast part of their comparative data with paper and pencil or web-based online surveys. The most typical question types in survey research are dichotomous and multiple-choice questions. In both cases the categories are discrete (i.e. have no overlap and are mutually exclusive) and exhaust the possible range of responses (Cohen, Manion & Morrison, 2000, 251). One of the major differences between traditional Gaussian and Bayesian models lies in the fact that the latter does not require multivariate normal distribution of

the indicators (i.e. observed variables) or underlying phenomena. This feature is especially useful for a social science researcher who collects her data with, for example, Likert –scale type questions as the response options from 1 to 7 produce data that is more qualitative than quantitative in nature. Measurement level of such items is ordinal and it is not advisable to model it with traditional statistical analysis that rely on the concept of normal distribution, and require calculation of mean and standard deviation.

6 Phenomena under cross-cultural investigation are seldom purely linear or continuous in nature. Unfortunately most commonly applied traditional linear Gaussian models (e.g. regression and factor analysis) are statistically inadequate for understanding non-linear dependencies between variables. However, Bayesian dependency models for discrete data allow also description of non-linearities. Bayesian theory gives a simple criterion, i.e., probability of the model, to select among such models (Nokelainen, Silander, Ruohotie & Tirri, 2003). Nokelainen, Tirri, Campbell and Walberg (2004) analyzed cross-cultural factors that account for adult productivity with three samples that came from U.S. (N=239), Germany (N=228) and Finland (N=157). We further investigated the number of non-linear and multi-modal relationships between variables in the three data sets in order to find how much they weaken the robustness of linear statistical methods. The results presented in Table 8.1 show that 63 percent of all dependencies are purely linear (linear mode, linear mean, unimodal). This is the best data for traditional linear analysis as no information is lost due to non-linearity. One should also observe that 37 percent of dependencies are to some extent non-linear. These dependencies are missed or poorly modeled using simple linear models.

Bayesian Modeling

Probability is a mathematical construct that behaves in accordance with certain rules (Berry, 1996) and can be used to represent uncertainty. The classical statistical inference is based on a frequency interpretation of probability, and the Bayesian inference is based on the "degree of belief" interpretation (Bernardo & Smith, 2000).

Table 8.1

Comparison of linear and non-linear dependencies in three empirical samples

Dataset	**U.S. N=239**	**Germany N=229**	**Finland N=159**	**Total**
Linear mode, linear mean, unimodal	59,4 %	71,4 %	59,1 %	*63,4 %*
Linear mode, linear mean, multimodal	3,1 %	0 %	4,5 %	*2,4 %*
Linear mode, non-linear mean, unimodal	21,9 %	21,4 %	22,7 %	*22,0 %*
Linear mode, non-linear mean, multimodal	9,4 %	7,1 %	4,5 %	*7,3 %*
Non-linear mode, linear mean, unimodal	0 %	0 %	0 %	*0 %*
Non-linear mode, linear mean, multimodal	0 %	0 %	0 %	*0 %*
Non-linear mode, non-linear mean, unimodal	3,1 %	0 %	9,1 %	*3,7 %*
Non-linear mode, non-linear mean, multimodal	3,1 %	0 %	0 %	*1,2 %*

Bayesian inference (Congdon, 2001) uses conditional probabilities to represent uncertainty. Therefore, we are interested in the probability P(M | D,I) — the probability of unknown things (M) given the data (D) and background information (I). The initial uncertainty about M is also represented as a conditional probability P(M | I). For example, we could have some initial belief that some answers are more likely than others. The essence of Bayesian inference is in the rule, known as Bayes' theorem (1763), that tells us how to update our initial probabilities P(M | I) if we see data D, in order to find out P(M | D,I).

Consequently Bayesian inference briefly comprises the following three principal steps:

1. Obtain the initial probabilities P(M | I) for the unknown things. These probabilities are called the *prior (distribution)*.
2. Calculate the probabilities of the data D given different values for the unknown things, i.e., P(D | M,I). This function of the unknowns is called the *likelihood*.
3. Finally the probability distribution of interest, P(M | D,I), is calculated using Bayes' theorem given above. This so called

posterior (distribution) will then express what is known about M after observing the data.

Bayes' theorem can be used sequentially. If we first receive some data D, and calculate the posterior P(M | D,I), and at some later point in time receive more data D', the calculated posterior can be used in the role of prior to calculate a new posterior P(M | D,D',I) and so on. The posterior P(M | D,I) expresses all the necessary information to perform predictions. The more data we get, the more certain we will become of the unknowns, until all but one value combination for the unknowns have probabilities so close to zero that they can be neglected.

The statistical procedures for analyzing cross-cultural data in this chapter include the following two stages: (1) variable selection based on Bayesian classification modeling and (2) inspection of probabilistic dependencies between the variables with Bayesian dependence modeling.

Bayesian Classification Modeling

The first stage is to conduct Bayesian classification modeling (Silander & Tirri, 1999) in order to find out which variables included in the study are the best predictors for different group memberships (in our example for example gender, productivity, level of giftedness).

In the classification process, the automatic search is looking for the best set of variables to predict the class variable for each data item. This procedure is akin to the stepwise selection procedure in the traditional linear discriminant analysis (Huberty, 1994, 118-126).

Nokelainen, Tirri and Campbell (2002) conducted the Bayesian classification analysis in order to find out which variables measuring computer literacy are the best predictors for the Mathematics Olympians country of origin. We derived the model for classifying data items according to the class variable "CON" ("U.S.", "Finland") with the 17 variables of computer literacy as predictors (Table 8.2). The estimated classification accuracy for the model was 82.64%.

Table 8.2

Finnish and U.S. Mathematics Olympians computer utilization

Variable		**Country**	
Code	**Description**	**Finland**	**U.S.**
V20a	Own computer (%)	79.0	65.0
US2	Work on computer daily (%)	92.0	82.5
V25	Hours per week on personal computer M (SD)	17.14 (14.16)	11.20 (15.11)
V26	Hours per week on main frame computer M (SD)	5.33 (8.38)	6.93 (10.82)
	Computer programs used (%)		
V27	Word processing	97.2	67.5
V28	Mathematics/Statistics	55.6	33.8
V29	Spreadsheet	61.1	26.3
V30	Internet	95.8	42.5
V31	Database	31.9	11.3
V32	Games	52.8	37.5
V33	Graphics	52.8	6.3
V34	Desktop publishing	43.1	18.8
E2	Other	44.4	16.3
V43	Have an e-mail address (%)	95.8	71.3
V44	Number of programming languages known M (SD)	4.24 (2.88)	4.16 (3.52)
COM1	Number of computer programs programmed M (SD)	10.24 (59.82)	0.64 (2.94)
V45	Self evaluated computer literacy (Scale: highest 5; lowest 1) M (SD)	4.04 (1.05)	4.08 (1.06)

Table 8.3 lists the variables ordered by their estimated classification performance in the model. The strongest variables, i.e. those that discriminate the two countries best, are listed first. The percentual value attached for each variable in the table indicates the predicted decrease in the classification performance if the variable is dropped from the model.

We learn from the table that variables in the model spread into three categories: Top (one variable), middle (three variables), and lower class (two variables). The most important variable is V30 ”I use the Internet”. Removal of that variable would weaken the performance of the whole model from 82.64% to 68.75%. In addition, middle group variables, variable V33 ”I use graphics software”, variable V27 ”I use word processing software”, and variable V45 ”Self-evaluated computer literacy”, have total effect of 17.37 percent. The weakest predictors of our model are variable E2 “I use other software”, and variable V31 ”I use database software”. Those variables are thus the most common computer literature variables among Finland and U.S. Mathematics Olympians. (Table 8.3.)

Table 8.3

Importance ranking of the variables in the Bayesian classification model

Variable name		**Decrease in predictive classification (%)**
V30	Internet	13.89
V33	Graphics software	7.64
V27	Word processing software	5.56
V45	Self evaluated computer literacy	4.17
E2	Other software	1.39
V31	Database software	0.69

As discussed above, in the classification process the automatic search tried to find the best set of variables that predict the country for each data item. The variables that were not selected for any subset are not good ones (under the classification model assumptions of multinomial distributions) to predict cross-cultural attitudes in our data. These variables are presented in Table 8.4.

Table 8.4

The variables excluded from the Bayesian discriminant analysis

V20a	Own computer (%)
US2	Work on computer daily
V25	Hours per week on personal computer
V26	Hours per week on main frame computer
	Computer programs used
V28	Mathematics/Statistics
V29	Spreadsheet
V32	Games
V34	Desktop publishing
V43	Have an e-mail address
V44	Number of programming languages known
COM1	Number of computer software programmed

The overall result of 82.64% is just an average performance rate of the classification model. Table 8.5 presents classification performance by groups. The second column of the table ("Success for different predictions") presents the estimated correctness of classification performance and its reliability by groups. The figure in this column shows the probability for correct classification for each country in percentages. Next to each estimate there is a figure indicating the percentage of the sample size used to calculate this estimate. The third column in the table ("Success in different classes") presents the group difficulty, i.e. how well the data items of different classes can be predicted. The fourth column of the table ("Predicted group membership") shows how many of the members of certain class were predicted to be members of certain other class. The entries denoting numbers of correct classifications are printed in bold face type setting. The Finland data was a slightly more coherent compared to U.S. yielding the predictive classification results with 10 misclassifications compared to 15 misclassifications of U.S. data. (Table 8.5.)

Table 8.5

Classification performance by groups

	Success for different predictions N (%)	**Success in different classes N (%)**	**Predicted group membership (N)**	
			U.S.	**Finland**
U.S.	68 (85)	73 (79)	58	15
Finland	76 (80)	71 (85)	10	61

Bayesian Dependence Modeling

The second stage of the analysis is to build a Bayesian network (Heckerman, Geiger & Chickering, 1995) to examine dependencies between variables by both their visual representation and probability ratio of each dependency.

A Bayesian network is a representation of a probability distribution over a set of random variables, consisting of an directed acyclic graph (DAG), where the nodes correspond to domain variables, and the arcs define a set of independence assumptions which allow the joint probability distribution for a data vector to be factorized as a product of simple conditional probabilities.

Graphical visualization of Bayesian network (Myllymäki, Silander, Tirri & Uronen, 2002) contains two components: (1) observed variables visualized as ellipses and (2) dependences visualized as lines between nodes. Solid lines indicate direct causal relations and dashed lines indicate dependency where it is not sure if there is a direct causal influence or latent cause. Variable is considered as independent of all other variables if there is no line attached to it. Previous research work has demonstrated that Bayesian networks are useful for explorative analysis of causal structures between observed variables (Ruohotie & Nokelainen, 2000; Nokelainen, Tirri, K., Nevgi, Silander & Tirri, H., 2001).

Nokelainen, Tirri and Campbell (2002) investigated probabilistic dependencies between all of the computer literacy variables (see Table 8.1 for variable description). Bayesian search algorithm (Myllymäki, Silander, Tirri & Uronen, 2001) evaluated three data sets, Finnish, U.S., and combined (Finnish and U.S.) in order to find the model with the highest probability. During the extensive search, great number of

models was evaluated: Finnish data, 3.657.122 models; U.S. data, 21.189.683 models; and combined data, 21.623.985 models.

Figure 8.1 presents causal model of the variables measuring computer literacy in Finnish, U.S., and combined data. Solid lines indicate direct causal relations and dashed lines indicate dependency where it is not sure if there is a direct causal influence or latent cause. In the Finnish data, core variables of the model measure extensive use (US2 "Work on computer daily") of basic computer software (V27 "Word processing", V33 "Graphics", V34 "Desktop publishing", V29 "Spreadsheet", and V28 "Mathematics/Statistics"). In the Finnish data there is only weak connection between the Internet (V30) and an e-mail address (V43). Working on computer daily is an important variable in the U.S data, too, but the strongest dependencies are found along two paths: First consisting of variables measuring use of mathematical software (V28) and programming (V44), and second measuring use of graphics (V33) and desktop publishing (V34) software. Analysis of the combined data reveals that the Internet (V30) is an important junction for two paths in the model: First path is publishing (V34, V33) oriented and second one is programming (V44, COM1, V28) oriented. The both U.S. and combined models show that working on computer daily (US2) is related to self-evaluated computing skills (V45).

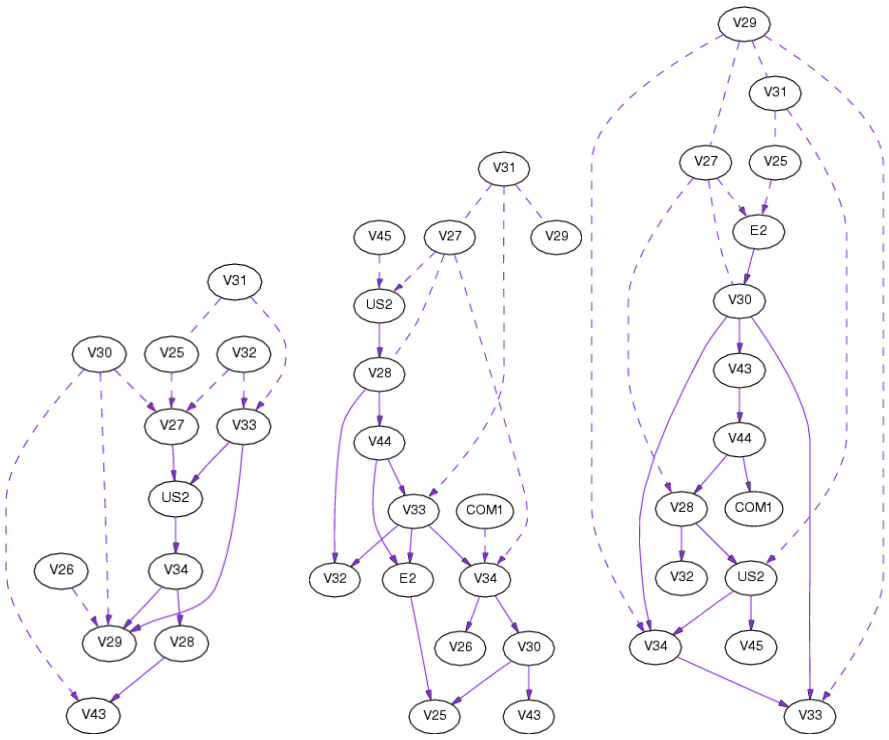

Figure 8.1
Causal model of the variables measuring computer literacy in Finnish (left), U.S. (middle) and combined (right) data

Table 8.6 presents the most dependent and independent variables of computer literacy in all three data sets (Finnish, U.S. and combined). The number of independent variables is highest in the Finnish model (4) while all the variables in the U.S. model seem to have statistical dependencies. The dependent variables list of both Finnish and U.S. data set show that country-specific structures do exist among variables measuring the computer literacy of Mathematics Olympians. The dependent variables list of combined data indicates that we are able to construct a cross-cultural structure of computer literacy variables.

Table 8.6

The most dependent and independent variables of computer literacy

Country	Dependent variables	Independent variables
Finland	V27, V33, US2, V34, V29, V28, V43	E2, V44, V45, COM1
U.S.	US2, V28, V44, V33, V32, E2, V34, V26, V30, V25, V43	–
Combined	E2, V30, V43, V44, V28, COM1, V32, US2, V34, V45, V33	V26

The B-Course: A Web-based Online Data Analysis Tool

Next we introduce a free web-based online data analysis tool, B-Course (Myllymäki, Silander, Tirri & Uronen, 2001; 2002) that is available at http://b-course.hiit.fi. We have conducted the preceding Bayesian classification and dependency modeling analysis in this chapter with this tool.

B-Course allows the users to analyze their data for multivariate probabilistic dependencies. These dependencies are represented as graphical models known as Bayesian networks. Although the analysis methods, modeling assumptions and restrictions are totally transparent to the user, this transparency is not achieved at the expense of analysis power. With the restrictions stated in the online material, B-Course is a powerful analysis tool exploiting several theoretically elaborated results developed recently in the fields of Bayesian and causal modeling.

B-Course can be used with most web-browsers, and the facilities include features such as automatic missing data handling and discretization, a flexible graphical interface for probabilistic inference on the constructed Bayesian network models, automatic pretty-printed layout for the networks, exportation of the constructed models, and analysis of the importance of the derived dependencies. (Figure 8.2.)

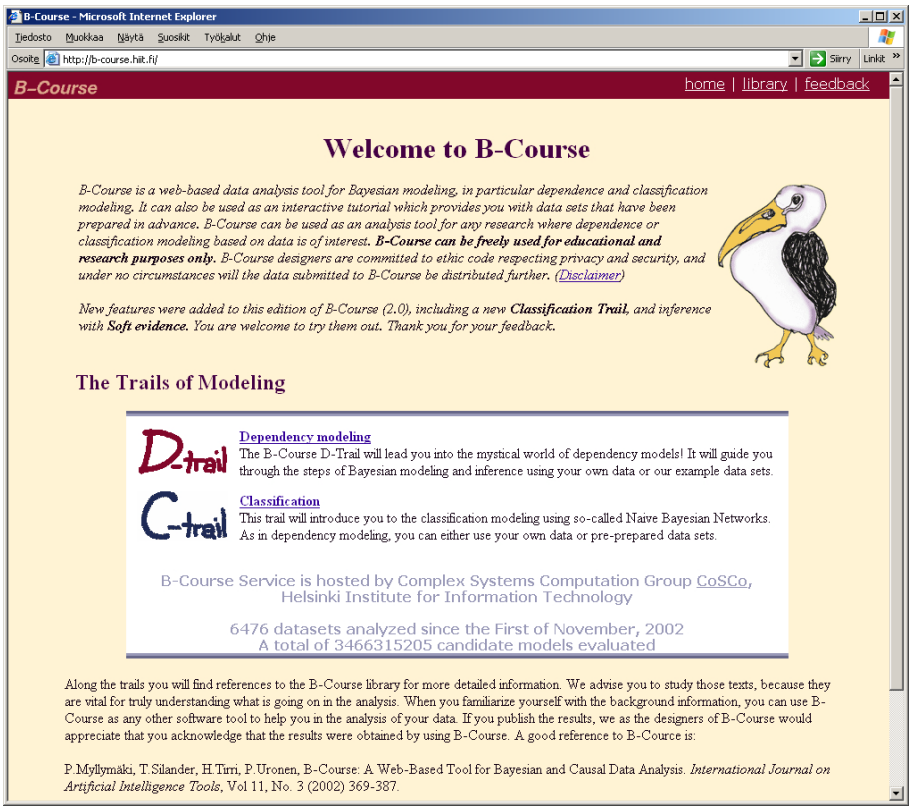

Figure 8.2
The B-Course web-based online data analysis tool

Tirri and Silander (2004) have discussed about the B-Course system user interface design aspects as follows:

1 No parameters. B-Course is meant to be used by social scientists and computer science students that either are taking (or have taken) an accompanying course in Bayesian modeling, or have some background in the topic. The user cannot be expected to be able to enter complex technical parameters or make decisions on selection of the mathematical methods used. Consequently, B-Course has no user definable technical parameters; all the data preprocessing (discretization, missing data handling etc.) and search related decisions (search criteria, search bias etc.) are handled automatically.

2 Ease of access. There are no problems of installation to various environments, as Application Service Provider (ASP)

allows a thin client at the user end for “non-power” users, and the computational load for searching models can be allocated to a server farm. B-Course can be used with most web-browsers and their early versions (even Lynx), and only requires the user data to be presented in tabular text format.

3 One resource — many trails. The B-Course is arranged around the notion of “trails”; it currently supports the “Dependency trail” and “Classification trail”. Also the ready-made examples are arranged in "trails" in order to simplify things.

4 Exporting results. In many cases a resource such as B-Course will be used for coursework, demonstrations or scientific work. For such purposes it is important that the results of the analysis can be easily exported in order to be used in reports, term papers etc. This does not only mean that systems like B-Course have to be able to store the results using some standard formats, it also forces the software to include additional features such as pretty-printing and sometimes verbose explanations about the results.

5 Interactivity. B-Course allows the user to study the inferred model interactively by providing an inference engine as an applet “Amazing Bayes-browser”. Implementation of the inference engine in B-Course has been the most time-consuming and error-prone interface task in the whole design. However, offering both model construction and inference with the model in the same service is quite essential for making the learning, teaching and research use easier. Integrated environment such as B-Course is not only simpler to use, it also eliminates many of the errors caused by switching between several tools.

Conclusions

In this chapter we have discussed about the typical problems of quantitative cross-cultural data analysis and described the essential benefits of using Bayesian modeling:

1) Robustness of statistical calculations with small sample sizes (theoretical minimum n=0),

2) Allowing prediction with the model as modeling is based on probabilities,

3) Inductive approach as the model is constructed from the data,

4) Possibility to input a priori (i.e. expert or subjective knowledge) information to the model,

5) Does not require multivariate normal distribution of the indicators (i.e. observed variables) or underlying phenomena, and finally,

6) Capability of understanding non-linear dependencies between observed variables.

We described the theoretical concepts of Bayesian modeling and illustrated their use in data analysis with excerpts from our preceding empirical studies in the research field of cross-cultural studies. The last part of this chapter introduced B-Course, a free web-based online Bayesian classification and dependency modeling data analysis tool.

Acknowledgements

This work was supported in part by the Academy of Finland, under the project Prose.

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosphical Transactions of the Royal Society, 53*, 370-418.

Bernardo, J., & Smith, A. (2000). *Bayesian Theory*. New York: John Wiley & Sons.

Berry, D. (1996). *Statistics - A Bayesian perspective*. Duxbury Press.

Cohen, L., Manion, L., & Morrison, K. (2000) Research Methods in Education. 5th edition. London: RoutledgeFalmer.

Congdon, P. (2001). *Bayesian Statistical Modelling*. Chichester: John Wiley & Sons.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*(3), 197-243.

Huberty, C. (1994). *Applied Discriminant Analysis*. New York: John Wiley & Sons.

Marini, M., Li, X., & Fan, P. (1996). Characterizing Latent Structure: Factor Analytic and Grade of Membership Models. *Sociological Methodology*, *1*, 133-164.

Myllymäki, P., Silander, T., Tirri, H., & Uronen, P. (2001). Bayesian Data Mining on the Web with B-Course. In N. Cercone, T. Lin and X. Wu (Eds.), *Proceedings of The 2001 IEEE International Conference on Data Mining,* (pp. 626-629). IEEE Computer Society Press.

Myllymäki, P., Silander, T., Tirri, H., & Uronen, P. (2002). B-Course: A Web-Based Tool for Bayesian and Causal Data Analysis. *International Journal on Artificial Intelligence Tools*, *11*(3), 369-387.

Nokelainen, P., Miettinen, M., Kurhila, J., Silander, T., & Tirri, H. (2002). Optimizing and profiling users online with Bayesian probabilistic modeling. In *Proceedings of the International Networked Learning Conference of Natural and Artifical Intelligence Systems Organization*, Berlin, Germany. Canada: ICSC-NAISO Academic Press.

Nokelainen, P., Silander, T., Ruohotie, P, & Tirri, H. (2003, August). Investigating Non-linearities with Bayesian Networks. *Paper presented at 111th Annual Convention of the American Psychological Association.* Toronto, Canada.

Nokelainen, P., Tirri, K., & Campbell, J.R. (2002, April). Cross-cultural findings of computer literacy among the Academic Olympians. *Paper presented at the Annual Meeting of the American Educational Research Association*, New Orleans, USA.

Nokelainen, P., Tirri, K., Campbell, J.R., & Walberg, H. (2004). Cross-cultural Factors that Account for Adult Productivity. In J. R. Campbell, K. Tirri, P. Ruohotie, & H. Walberg (Eds.), *Cross-cultural Research: Basic Issues, Dilemmas, and Strategies,* (pp. 119-139). Hämeenlinna, Finland: Research Centre for Vocational Education, University of Tampere.

Nokelainen, P., Tirri, K., Nevgi, A., Silander, T., & Tirri, H. (2001). Modeling Students' Views on the Advantages of Web-Based Learning with Bayesian Networks. In H. Ruokamo, O. Nykänen, S. Pohjolainen and P. Hietala (Eds.), *Proceedings of The 10th International Intelligent Computer and Communications Technology - Learning in On-Line Communities PEG2001 Conference,* (pp. 202-211).

Ruohotie, P., & Nokelainen, P. (2000). Modern Modeling of Student Motivation and Self-regulated Learning. In P. R. Pintrich and P. Ruohotie (Eds.), *Conative Constructs and Self-regulated Learning,* (pp. 141-193). Hämeenlinna, Finland: University of Tampere, Research Centre for Vocational Education.

Silander, T., & Tirri, H. (1999). Bayesian Classification. In P. Ruohotie, H. Tirri, P. Nokelainen and T. Silander, *Modern Modeling of Professional Growth,* (pp. 61-84). Hämeenlinna, Finland: Research Centre for Vocational Education, University of Tampere.

Tirri, H., & Silander, T. (2004). B-Course: Issues in designing a Web Service for Bayesian Data Analysis. In P. Ruohotie, P. Nokelainen, H. Tirri and T. Silander, *Modern Modeling of Professional Growth vol. 2*. Hämeenlinna, Finland: Research Centre for Vocational Education, University of Tampere.