Bayesian Model Selection for Human Development Indicators

Nitin Sawhney
MIT Media Laboratory
nitin@media.mit.edu

CS282: Probabilistic Reasoning
January 14, 2001

Abstract

In this paper, we explore the use of Bayesian estimation techniques towards modeling indicators related to Human Development, in particular Gender-related measures. The main emphasis and contribution of this work is demonstrating the use of Bayesian approaches towards modeling development indicators, rather than interpreting the validity of specific models obtained from the current limited dataset. Future work will seek to acquire and utilize more comprehensive data for a greater number of indicators. There is a need to explore methods that model indicators with *missing* data, which is quite often the case with a majority of development indicators today. As the topic of human development is an important concern for many, this paper has been written for a broader audience. Hence, the domain of human development indicators, concepts of Bayesian estimation and statistical approaches for model selection are clearly explained.

Introduction

Human Development Indicators

Statistics can provide quantitative information on trends in human development that can serve as inputs for the analysis of critical policy issues. A wide array of development indicators have been proposed in diverse areas such as economic growth, health, education, socio-political status, and even abstract concepts such as human freedoms [Amartya Sen 95]. There are many sources of such data, some are digitally-accessible such as that by the UNDP and World Bank, that put out surveys and collections of development statistics every year. However, many problems remain with coverage, consistency and comparability of data across countries and time. The indicators usually tend to be somewhat complex such as Literacy, which is reduced to simple reading/writing skills rather than one's capacity within a social context. Hence, the UNDP [2000] proposes composite measures e.g. using net enrollment data for literacy, but it is collected for very few countries. These indicators often demonstrate conflicting social effects, for example that many variables related to economic growth and modernization do not always lead to improved human development (see Figure 1). Overall, Human Development is much deeper and complex concept than what can be captured by statistical indicators. However, modeling causality between indicators may allow one to recognize broader patterns across countries or over time, and supplement existing socio-economic theories and fieldwork.

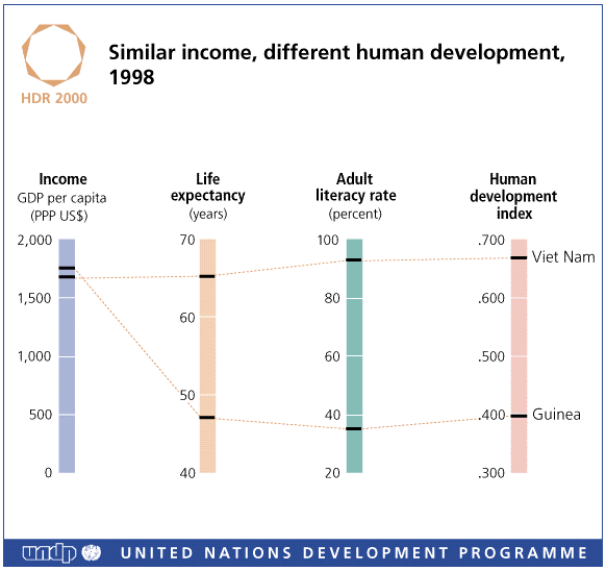


Figure 1: Contrasts in Human Development between 2 countries, despite having the same Income levels. (source: UNDP, 2000)

Overview of Bayesian Networks

The *Bayesian network formalism*, also referred to as probabilistic graphical models or belief networks, is a combination of probability theory and graph theory in which dependencies between random variables is expressed graphically. Hence a Bayesian network can be defined as a "graphical model for representing

conditional independencies between a set of random variables" [Ghahramani97]. Let us consider an example from a tutorial by Ghahramani. Figure 2 shows a graphical representation of the joint probability P(W,X,Y,Z) that can be factorized as a set of conditional independence relations, as follows:

$$P(W,X,Y,Z) = P(W)\,P(X)\,P(Y|W)\,P(Z|X,Y)$$

Given the values of X and Y, we can show that Z and W are independent.

$$P(Z,W|X,Y) = P(W|Y)\,P(Z|X,Y)$$

So the Bayesian network is a way of graphically representing a *particular factorization* of a joint distribution. This factorization implies a certain ordering of the random variables in a manner that defines a directed acyclic graph (DAG). Undirected graphical models are considered Markov networks, with a different set of semantics. In a DAG each node (variable) is conditionally independent from its non-descendents, given its parent nodes. For example, we can visually infer from the DAG that W is conditionally independent from X given the set {Y, Z}, but not necessarily from X given Z (cannot infer that from the graph). Here the set {Y, Z} *d-separates* the disjoint nodes W and X.

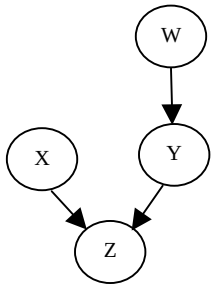


Figure2: A *directed acyclic graph* (DAG) representing the conditional independence between random variables in the joint probability distribution P(W,X,Y,Z).

The graph not only allows us to understand which variables affect others, but also serves as a means to efficiently compute marginal and conditional probabilities for inference and learning. For *singly connected networks*, in which the underlying undirected path has no more than one path between any two nodes (i.e. no loops), the general algorithm used is called *Belief Propagation.* For *multiply connected networks*, in which there can be more than one undirected path between any two nodes, a more general algorithm used is the *Junction Tree Algorithm*.

A Bayesian network can be constructed by combining *a priori* knowledge about conditional independencies between variables, either from an expert in a particular domain by asking questions about causality (as is often done in *Static Bayes nets*) or from observed temporal data (as modeled by *Dynamic Bayesian networks*).

Prior Work: Use of Bayesian Approaches in Social Research

Many sociological studies are observational and aim to infer causal relationships between a dependent variable and independent variables of interest. Linear regression and step-wise variable regression techniques are often used to select one model out of many proposed social theories. However the sampling properties of these model selection techniques are unknown in general, and choosing among a large number of models increases the possibility of finding 'significant' variables by chance alone.

Raftery [1994] was one of the first to apply Bayesian Model Selection in social research. In his paper he is critical of variable selection methods in sociology, such as P-values and T-tests. P-values reject plausible results, while many models may better explain results. It ignores uncertainty about model form, while finding "significant" variables by chance. Raftery proposes use of Bayesian hypothesis testing using BIC (Bayesian Inference Criteria) approximation. BIC tends to favor simpler models and null hypothesis, more so than P-values especially in large datasets.

Heckerman [1995] used a Bayesian approach to investigate factors that influence the intention of high school students to attend college. They used a number of socio-economic and demographic indicators to analyze data form over 10,000 high-school seniors. They assumed no hidden data, uniform priors and discrete variables. They used this data to compute the posterior probabilities of a number of pre-defined model structures. They found two most likely model structures after an exhaustive search. Their results were not surprising however they found that by introducing an additional hidden variable to the model structure, they were able to better explain the data. They interpreted this hidden variable to suggest "parental quality".

Approach: Bayesian Model Selection for GDI

Bayesian Estimation

For our domain of Gender-related Development Indictors (GDI), we assume the random variables X (indicators) have been *observed* (via surveys conducted in each country) and the data available is complete for each indicator (no unknown values). We wish to infer a set of plausible models to explain dependence relationships among variables represented by the data. We will generate a large number of *model structures*, i.e. graphical models with different patterns of connectivity, and assign values of *model parameters* θ, i.e. local conditional probabilities. We must assume that the parameters are mutually independent, allowing each to be updated independently. For each choice of parameter values, a different joint probability distribution $p(x \mid \theta)$ of the random variables will be obtained from the model, assuming a known prior probability $p(\theta)$ of the parameters (uniform or random priors). Using Bayes rule we can then estimate the posterior probability $p(\theta \mid x)$ of the parameters given the data, as follows:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

The conditional distribution $p(x \mid \theta)$ is referred to as the likelihood function, used to evaluate particular choices of the parameters to select ones that assign maximal probability for the data observed. Hence, this value of θ that maximizes the likelihood function is considered the *maximum likelihood estimate* of the true value of θ:

$$\hat{\theta}_{ML} = \text{argmax}_{\theta}\, p(x|\theta)$$

This maximum likelihood estimate can be computed by using the Expectation-Maximization (EM) algorithm. To find a local ML estimate, we assign a configuration to θ somehow (at random). We then compute the expected sufficient statistics for a complete data set, taken with respect to the joint distribution for X conditioned on the configuration of parameters θ and the known data. This computation is the *expectation* step of EM. Then the expected sufficient statistics are used to determine the configuration of θ that maximizes $p(x \mid \theta)$. This assignment is called the *maximization* step of EM. It has been shown that under certain regularity conditions, iteration of the EM steps will converge to a local maximum.

The models with the highest log likelihood are considered most plausible in explaining the data for any given year.

Implementation

For the purpose of our experiments, Kevin Murphy's Bayes Net Toolbox (BNT) for Matlab version 5 [Murphy2000] was utilized. The toolbox supports continuous (gaussian) probability distributions and a number inference engines for BNs (using popular algorithms such as Junction Tree and Variable Elimination) as well as batch EM parameter learning.

Pre-Processing GDI Data

At the time of the project, only data from the World Bank was available in electronic form (WDI CD-ROM 1999), and not the GDI measures from the UNDP [2000]. Raw data for all 6 indicators from 120 countries was extracted for two years (1990 and 1997) from specially formatted ASCII text files, generated by the World Bank CD-ROM. The indicators included: fertility rate (FR), female illiteracy rate (LR), female life expectancy (LE), female labor force participation (LF), infant mortality rate (IM) and the number of telephone-lines available per 1000 people (PP). This data was processed to convert the indicators to normalized values in accordance with the methodology developed by the UNDP for the Human Development Index (HDI). Each index is computed according to the general formula:

$$Index = \frac{Actual\ value - Min\ value}{Max\ value - Min\ value}$$

2 Gender-related development index

HDI rank	Gender-related development index (GDI) 1998 Rank	Gender-related development index (GDI) 1998 Value	Life expectancy at birth (years) 1998 Female	Life expectancy at birth (years) 1998 Male	Adult literacy rate (% age 15 and above) 1998 Female	Adult literacy rate (% age 15 and above) 1998 Male	Combined primary, secondary and tertiary gross enrolment ratio (%) 1997 Female	Combined primary, secondary and tertiary gross enrolment ratio (%) 1997 Male	GDP per capita (PPP US$) 1998 [a] Female	GDP per capita (PPP US$) 1998 [a] Male	HDI rank minus GDI rank [b]
101 Tunisia	86	0.688	71.0	68.6	57.9	79.4	68	74	2,772 [e]	7,982 [e]	-3
102 Moldova, Rep. of	81	0.697	71.7	63.8	97.9	99.5 [h]	71	69	1,548 [e]	2,381 [e]	3
103 South Africa	85	0.689	56.2	50.3	83.9	85.4	94	93	5,205 [e]	11,886 [e]	0
104 El Salvador	83	0.693	72.7	66.7	75.0	80.8	63	64	2,779 [f]	5,343 [f]	3
105 Cape Verde	88	0.675	71.6	65.8	64.6	83.7	76	79	1,931 [e]	4,731 [e]	-1
106 Uzbekistan	87	0.683	70.9	64.6	83.4	92.7	74	78	1,613 [e]	2,499 [e]	1
107 Algeria	91	0.661	70.6	67.7	54.3	76.5	64	71	2,051 [e]	7,467 [e]	-2
108 Viet Nam	89	0.668	70.0	65.3	90.6	95.3	59	64	1,395 [e]	1,991 [e]	1
109 Indonesia	90	0.664	67.5	63.7	80.5	91.1	61	68	1,780 [e]	3,526 [e]	1
110 Tajikistan	92	0.659	70.4	64.5	98.6	99.5 [h]	65	73	777 [e]	1,307 [e]	0

Figure 3: Data on normalized Gender-related Development Indicators, (Source: UNDP, 2000).

Female Literacy (LR) was computed from the illiteracy rates. The indicators LR and LF are available in percentages and normalized to values < 1.0. For IM and FR the indexes utilize the maximum and minimum values computed from the dataset. However, for Life Expectancy at birth (LE) the min and max values of 25 and 85 years, established by the UNDP are used. Finally, for the telephone density indicator (PP) the minimum value was set to 1 per 1000 and maximum to 500 per 1000 based on examining the data.

The data for all indictors was split evenly into training and test sets using a randomly ordered sequence. Hence, each dataset contained 60 observations for 5 of the GDI measures for 2 years each. However, for telephone-line density complete data for all countries was available for the year 1990 only.

Model Generation and Bayesian Inference

A model consists of a graph structure and its parameters. Initially several variations of plausible models for the GDI data were hand-constructed in the BNT tool, and evaluated to test the methods developed. However it became clear that one need to generate numerous Directed Acyclic Graphs (DAGs) to find the most likely ones. Hence the DAGs were automatically generated in Matlab, while ensuring their validity by eliminating DAGs that contained self-referencing links, bi-directional arcs to other nodes, and overall cyclic connections in the graph. From over 10,000 potential DAGs randomly generated, up to 1000 valid models were iteratively selected for computing log likelihood with each dataset while redundant duplicates were eliminated. The parameters in the model are represented by Gaussian conditional probability distributions (CPDs) of each node given its parents, stored as multidimensional arrays or tables (Tabular CPDs).

A variety of inference algorithms are provided in the BNT such as junction tree, variable elimination and loopy propagation, each having different tradeoffs between speed, accuracy, complexity and generality. The Junction Tree inference engine was used for these experiments as it provides exact inference for all topologies of static BNs with continuous-valued nodes, and handles any pattern of evidence. The junction tree algorithm runs reasonably fast in the BNT toolkit as it uses dynamic programming to compute all marginals in two passes when evidence is provided, making it more efficient during learning (than the variable elimination algorithm implemented here).

The initial parameters for the belief networks are set to random values, serving as the evidence for the inference engine. The log-likelihood for the model is computed and incorporated in a modified engine used for parameter learning. The maximum likelihood estimates of the parameters are now computed for up to 5 iterations, using batch EM learning. After learning the parameters for the model, the log likelihood for the training and test datasets is computed. For each model, the *log-likelihoods on the test data* is compared with that of the current top 10 scoring models, updating and sorting this list as needed.

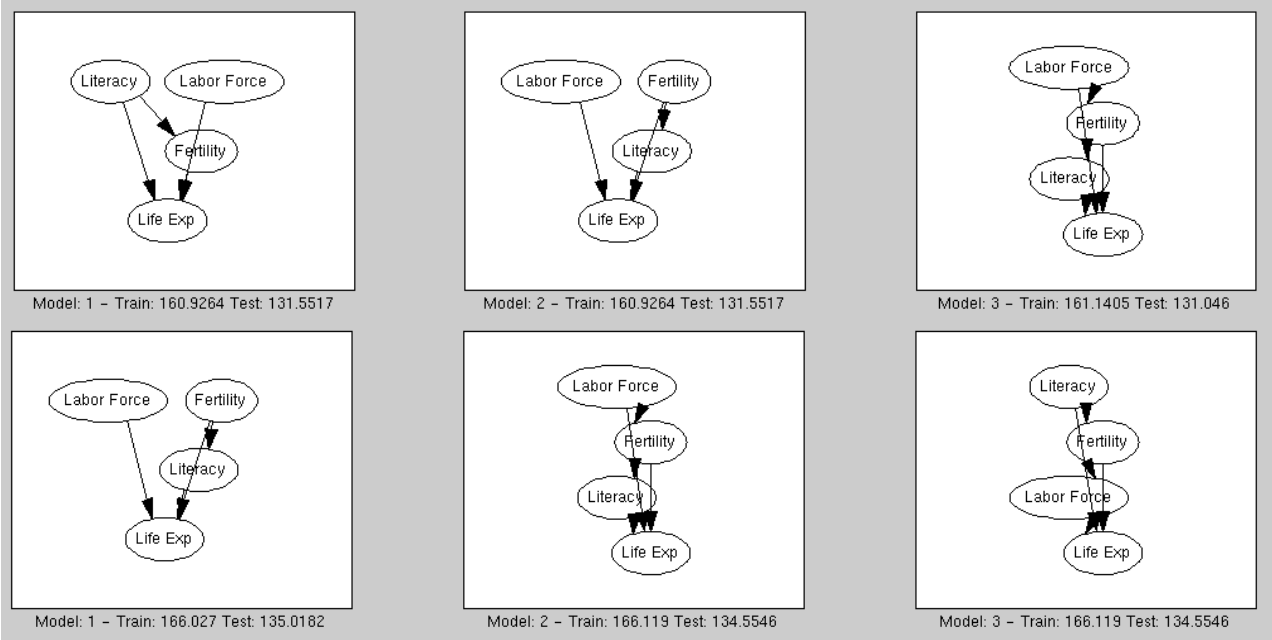
Interpreting Preliminary Results

Models for generated for three different sets of experiments conducted with a variety of datasets of development indictors. Here, three indicators are conceptually defined under the notion of Woman's Agency [Sen95], i.e. female Literacy, Life Expectancy, Labor force participation. We will discuss the preliminary results below and summarize them in the context of some previous theories and finding in developmental economics. For each dataset the top three maximally likely models are displayed along with the log-likelihood scores for the training and test data. The interpretations that follow summarize visually observed dependence relations among variables, in the most likely models. However, these interpretations are qualitative and must be subject to greater scrutiny by comparison with results from data in other years or comparison with models containing many more unrelated variables.

Experiment I: Basic GDI Indicators

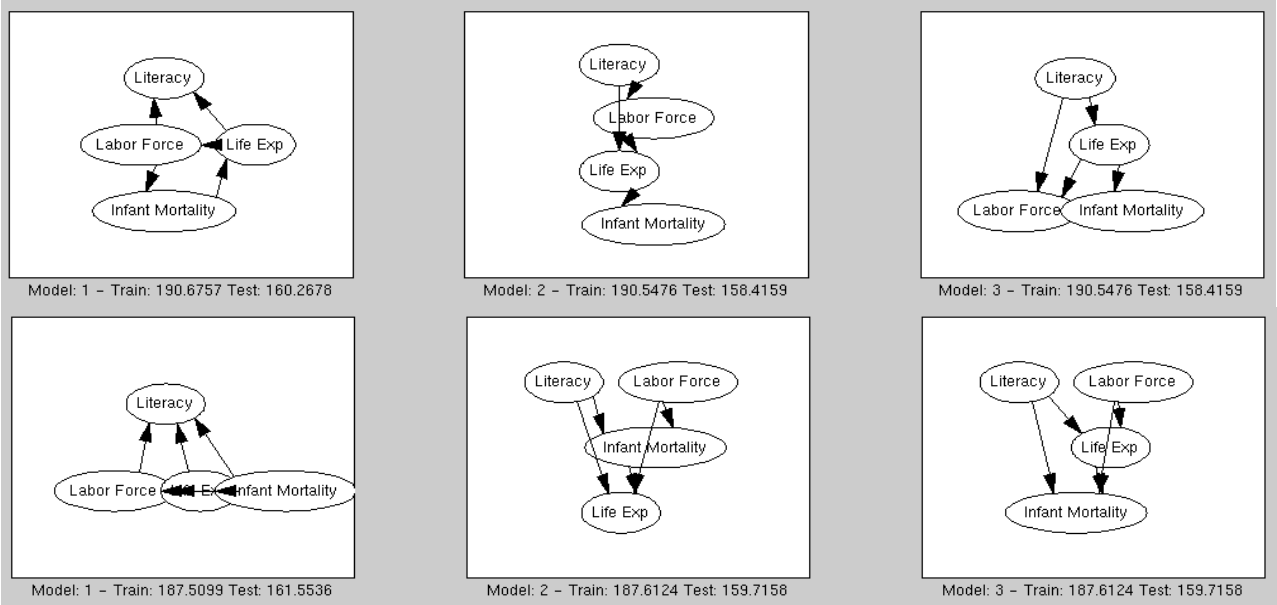
I.A. Woman's Agency (LE+LR+LF) and Fertility (FR) for 1990 and 1997

There seem to be two most likely models that have a similar structure for both years (Model 2, 1990 and Model 1, 1997). These indicate Life Expectancy being influenced by Literacy, Fertility and Labor Force Participation. A correlation between Fertility and Literacy is observed in all models. A positive dependence between Literacy and Fertility is seen in most models.



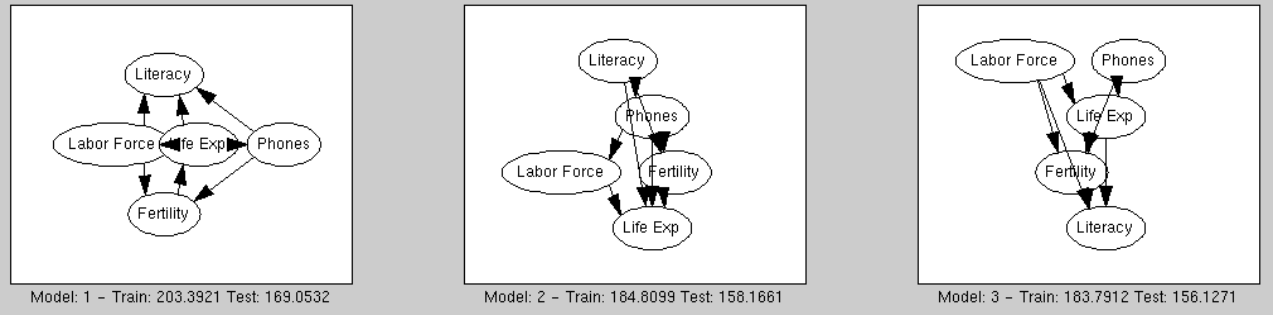
I.B. Woman's Agency (LE+LR+LF) and Infant Mortality (IM) for 1990 and 1997

The two most likely models 1 for 1990 and 1997 seem to share a similar structure. The overall models selected indicates general dependency between Literacy with Life Expectancy and Labor Force Participation, while Infant Mortality has a dependence relation among all 3 indicators of Life Expectancy, Literacy and Labor Force Participation. In particular many models, especially model 1 for 1990, show a dependence between infant mortality and female labor force participation, while models for 1997 show more evidence of dependence between infant mortality and literacy.

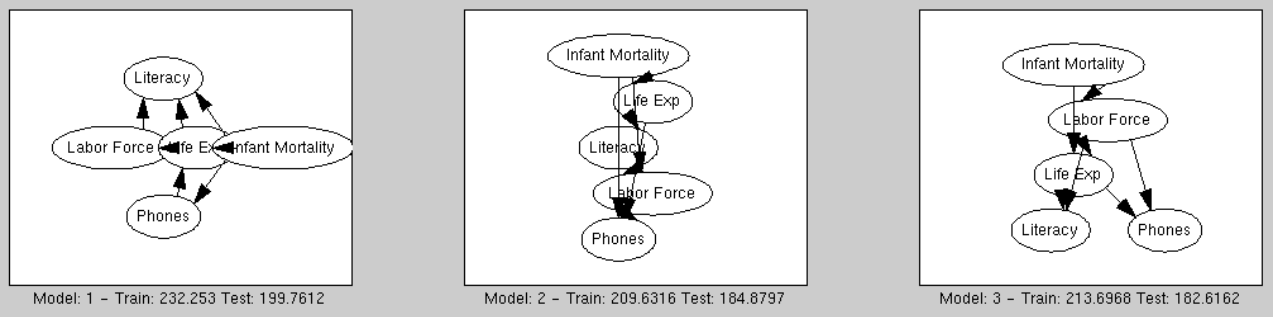


Experiment II: Influence of Telephone Density on GDI Indicators (Year 1990 only)

II.A. Telephone Density on Woman's Agency and Fertility (LE+LR+LF+FR)



II.B. Telephone Density on Woman's Agency and Infant Mortality(LE+LR+LF+IM)

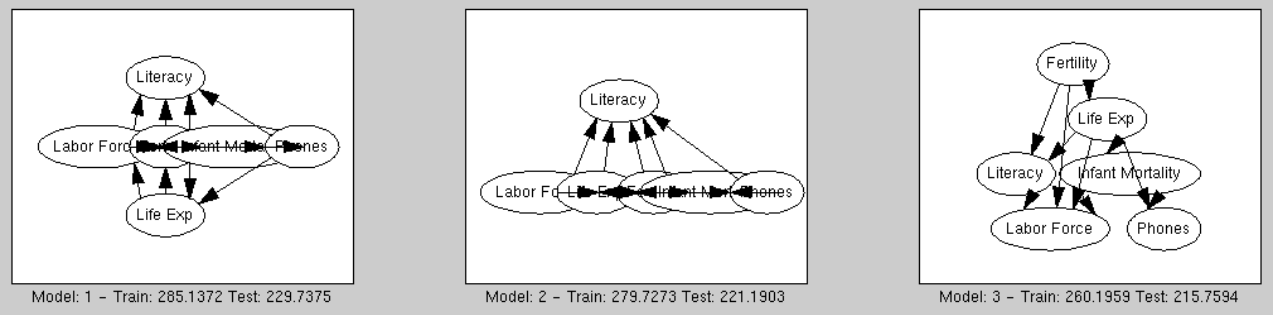


With the introduction of Telephone density, both these models are harder to interpret as they seem nearly fully connected. However the top two models for both datsets (models 1) seem to have much greater log-likelihood than the next best model, hence they suggest stronger models to explain the data. These models appear to show consistent dependence of Literacy with Life Expectancy and Labor Force Participation, consistent with results from dataset 1.B. However, any consistent dependence between telephone density and any other variables cannot be clearly observed.

Experiment III: Dependence among all six Indicators (Year 1990 only)

Telephone Density and GDI (PP+LE+LR+LF+FR+IM)

Modeling all 6 indicators, shows a preference for more fully connected DAGs, where most variables seem to influence literacy.



It is clear that all the indicators examined here are inter-related to a great extent and finding isolated dependence between specific variables is difficult, however some general trends may be considered for future experiments. The overall impression from modeling all datasets appears to show consistent

dependence relation between literacy and most other indictors. It is unclear whether literacy promotes greater Woman's agency or whether the presence of favorable conditions for Woman (labor force participation and life expectancy/health) influences greater literacy among female adults. Life Expectancy and Infant Mortality also seem to be dependent on a combination of such factors. In particular, models for 1997 verify the evidence of dependence between Literacy and Infant Mortality and, which is consistent with the prior work in the literature. The positive dependence between Literacy and Fertility as seen in most models confirms the widely observed link between the two in most countries.

Female Labor Force Participation usually tends to influence all other variables of Woman's agency in all models observed. This is consistent with prior hypothesis posed by economists such as Amryta Sen (1995). In particular the dependence between infant mortality and female labor force participation has been predicted in the past, yet it has not been established whether it is a positive or negative association. Finally the relationship of Telephone density (a sign of modernization in a region) remains ambiguous in the models generated. Many developmental economists like Dreze and Sen maintain that Gender inequity does not decline with economic growth and modernization.

Conclusions & Future Work

Overall, given enough data and experimentation one may begin to interpret the graphical models generated in our analysis of the GDI dataset to consider the importance of Woman's Agency (inter-related female indicators of education, employment and health) towards Fertility and Infant Mortality, as generally believed by many developmental economists, especially in comparison to weaker effects of variables relating to general economic progress (like telephone density and GNP). Future work will seek to acquire and utilize more comprehensive data for a greater number of indicators. There is a need to explore methods that model indicators with *missing* data, which is quite often the case with a majority of development indicators today.

Acknowledgements

Pierre Fallavier at MIT's Dept. of Urban Studies and Planning (DUSP) initially directed me towards UNDP's Human Development reports. Thanks to Prof. Avi Pfeffer and Kobi Gal in the CS dept. at Harvard University for engaging discussions on this topic. Tony Jebara at the MIT Media Lab was helpful in clarifying implementation issues for modeling the GDI dataset using Bayesian Networks.

References

[Ghahramani97] Ghahramani, Zoubin. 1997. Learning Dynamic Bayesian Networks. *Adaptive Processing of Temporal Information.* Lecture Notes in Artificial Intelligence. Springer-Verlag. See related tutorial paper here - http://www.cs.utoronto.ca/~zoubin/

[Heckerman96] Heckerman, David. 1996. A Tutorial on Learning with Bayesian Networks. Microsoft Research, Technical Report, MSR-TR-95-06.

[Murphy2000] Bayes Net Toolbox (BNT). http://www.cs.berkeley.edu/~murphyk/Bayes/bnt.html

[Raftery94] Raftery, Adrian E. Bayesian Model Selection in Social Research. Social Methodology. 1995.

[Sen95] A. Sen and Jean Dreze. INDIA: ECONOMIC DEVELOPMENT AND SOCIAL OPPORTUNITY. Oxford University Press. 1995.

[UNDP 2000] UNDP Human Development Report 2000. http://www.undp.org/hdr2000/home.html

[WDI 1999] World Development Indicators, World Bank. CD-ROM, 1999. http://www.worldbank.org/