Stony Brook University

Learning Mixed Sparse Factor Networks Structure: a Latent Variable Approach

A Dissertation presented

by

Ruofeng Wen

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

May 2015

Stony Brook University

The Graduate School

Ruofeng Wen

We, the dissertation committe for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Wei Zhu - Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Song Wu - Chairperson of Defense
Assistant Professor, Department of Applied Mathematics and Statistics

Xuefeng Wang - Faculty Committee Member
Assistant Professor, Department of Applied Mathematics and Statistics

Yuanyuan Yang - Outside Committee Member
Professor, Department of Electrical and Computer Engineering and Computer Science

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Learning Mixed Sparse Factor Networks Structure:

a Latent Variable Approach

by

Ruofeng Wen

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2015

Learning and visualizing complex causal or dependence structure among various variables is of great interest in applied science. Probabilistic Graphical Models, including Bayesian Networks and Markov Random Field, are well-developed tools for such problems, particularly when variables are either all categorical or continuous. In the first and major part of this text, we proposed a novel graphical structure learning approach, the MIxed Sparse FActor Network (MISFAN), to accommodate categorical and continuous variables seamlessly in one sparse Probit latent factor model. Such a network bridges the gap among latent variable models, traditional multivariate analysis and graphical models, can visualize the underlying interaction and clustering in a more extensive and succinct way, and simultaneously presents certain causal hypothesis as in a Bayesian Network, along with local conditional dependence structure as in a Markov Random Field.

Another independent application of latent variable models is from the Error-in-Variable (EIV) perspective. EIV considers the intrinsic and mostly inevitable measurement error affecting the latent true predictors of a regression model. Although proved to be inconsistent and biased on data with measurement error, Ordinary Least Square still dominates for its simple computation and interpretation, while the EIV models seem to be daunting and confusing

for its diverse formulations. We intend to give a clear, systematic and unified description with novel geometric insight for the common EIV models in the second part of the text. Additionally, model caveats, parameter specification and alternative estimation approaches are discussed for practical interests.

Contents

List of Figures

List of Tables

Preface

This dissertation is an original, unpublished and independent work by the author, Ruofeng Wen, and consists of two self-contained papers with their own abstracts, introductions and main texts. The numbered lists of Figures, Tables and Bibliographies are in a combined format.

The thesis includes selected results regarding to the theoretical aspect of Statistics, accomplished during the Ph.D. study of the author at Stony Brook University, New York. It mainly discusses the proposed derivative of latent variable models on learning the network structure of a Probabilistic Graphical Model with both continuous and discrete variables. It also summarizes a unified geometric perspective of another latent variable setting: Error-in-Variable models. The two parts can be read individually.

Part I
Learning Mixed Sparse Factor Network Structure

Learning Mixed Sparse Factor Networks Structure: a Latent Variable Approach

Ruofeng Wen

05/05/2015

Abstract

Learning and visualizing complex causal or dependence structure among various variables is of great interest in applied science. Probabilistic Graphical Models, including Bayesian Networks and Markov Random Field, are well-developed tools for such problems, particularly when variables are either all categorical or continuous. We proposed a novel graphical structure learning approach, the MIxed Sparse FActor Network (MISFAN), to accommodate categorical and continuous variables seamlessly in one sparse Probit latent factor model. Such a network bridges the gap between traditional multivariate analysis and graphical models, can visualize the underlying interaction and clustering in a more extensive and succinct way, and simultaneously presents certain causal hypothesis as in a Bayesian Network, along with local conditional dependence structure as in a Markov Random Field.

Keywords. Graphical Model, Bayesian Network, Latent Variable, Mixed Network, Probit Factor Model, Conditional Dependence

1 Introduction

In the field of economics, engineering, finance, social sciences and bioinfomatics, processing high dimensional data with various types of variables is routine, and a computationally efficient and comprehensive model is needed to discover and present significant association of all the possible factors at once. Probabilistic Graphical Model (PGM), including Bayesian Networks (BN), Markov Random Field (MRF), and some traditional approaches, e.g.

Structure Equation Modeling (SEM), have been thriving in the past decade, for their ability to model complex interaction among numerous variables while retaining the interpretability.

Bayesian Networks, essentially a directed acyclic graph (DAG) with causal variables as the starting nodes, and outcome variables as the ending ones, were originally exclusively designed for categorical variables and the network is built on the assumption of multinomial joint distribution, estimated by the co-occurrence frequency of the random variables. Subsequently, similar model structure has been investigated on the so-called Gaussian Bayesian Network (GBN), a graph where all nodes are continuous and normally distributed. The undirected version of such networks is referred to as Markov Random Field (MRF), or sometimes Markov Networks, where an edge indicates certain mutual probabilistic association. Both BN and MRF aim to present the *conditional independence* structure of the variables through directed and undirected graphs, respectively. These PGMs give a parsimonious description of the underlying complex association among many variables, visualize it with a graph and thus enable qualitative and quantitative inference. Numerous statistical models fall in such categories, with either all Gaussian variables or discrete variables: Principle Component Analysis, Factor Analysis and Structural Equation Modeling as GBN; Naive Bayes Classifier, Hidden Markov Model and State Space Model as BN; Ising Model and Gaussian Graphical Model as MRF.

A network model for *both* continuous (numeric) and categorical (discrete) variables, commonly referred to as Mixed Graphical Model or Hybrid Networks, is of practical interest since mixed variables co-exist in many sophisticated system. One direct approach to accommodate both types of variables is to simply categorize the continuous variable into ordinal levels, see Neil et al. (2008) and Monti and Cooper (1998) for examples. This makes full use of the mature discrete BN framework, but distribution assumptions are generally subjective and hard to justify. Another popular approach is to use the conditional Gaussian distribution to model the directed relationship between a pair of continuous and categorical variables, e.g. Lauritzen (1996), Murphy (1998), and Bø ttcher (2001). The downside is that the parameter space is prohibitively large, and the causal direction can only be from the categorical to the continuous variables due to its innate limitation of distribution assumption. By simply using the Generalized Linear Model, Skrondal and Rabe-hesketh (2005) showed that SEM can accommodate both discrete and continuous factors at ease, however SEM is usually used as a confirmatory model to learn linear parameters among a small group of variables, so learning a sparse and consistent network structure from data is necessarily a prerequisite. As for

the undirected graph, Friedman, Hastie, and Tibshirani (2008) proposed a $l1$-penalized precision matrix estimation, the Graphical Lasso, for learning partial correlation structures for a multivariate normal distribution and thus building a GMRF. A different but equivalent perspective is on learning a precision matrix through the estimation of a set of joint sparse linear regression models, as described in Peng et al. (2008), from which Cheng, Levina, and Zhu (2013) further modeled the mixed network case using a set of joint logistic/linear regression models with Group Lasso regulation, called the Mixed Graphical Model (MGM). Lee and Hastie (2013) independently proposed an equivalent formulation of MGM, but estimated as a convex optimization problem.

Causal networks is of great interest as a fundamental model to describe the interaction among variables. BN does not uniquely learns the *true* causal structure, but answers part of the question by giving hypothetical causal process with appropriate conditional independence structure, though the latter is NP-hard for the discrete case, as discussed in Chickering, Heckerman, and Meek (2004). MRF, on the other hand, shows the local conditional dependence instead of causality. The different possible mapping and tedious translation of causal, directed or undirected graphs makes the interpretation restricted, and bird-viewing all possible patterns at once difficult. Another type of graph named Factor Graph adds additional 'factors' into the node list which abstractly represents a factored part of the joint density function, as elaborated in Loeliger (2004) and Kschischang, Frey, and Loeliger (2001). Although Factor Graph delivers the general merged representation of BN and MRF, the ambiguous definition of 'factors' and the mixing of nodes and edges is bewildering and in fact limited because it shows neither causal relation nor conditional independence on the graph. A more succinct and comprehensive visualization is desirable, to unify both the causal implication and conditional dependence patterns.

In this paper, we pursued the merits of the mentioned missing properties and proposed a novel MIxed Sparse FActor Networks (MISFAN) model to learn and present the conditional dependence graphical structure. It assumes latent Gaussian variables with a multinomial Probit link to model the categorical variables, builds $l1$-penalized sparse latent factors as the origin of the low dimensional causal information source, and detects sparse conditional dependence as the network structure, yielding both stronger visual interpretability and more coherent formulation. With the mature framework of traditional factor analysis, it also enables plenty of extensions for graphical models. In Section 2 we introduced some background concepts of PGM network analysis that is relevant to our work. We presented the MISFAN model, its strong interpretive power supported by a so-called Factor Network, and some of its

features in Section 3, followed by its estimation via a Monte Carlo Expectation Conditional Maximization Either (MC-ECME) algorithm. In Section 4 we conducted an intense simulation experiment to assess the behaviors and properties of our fitting algorithm, and designed another series of simulated Bayesian Networks to demonstrate the superior performance of MISFAN compared to its competitors in different settings. Then in Section 5, an analysis of the World Development and Value data serves as a practical example for the application of MISFAN. Conclusion and possible future work were discussed in Section 6.

2 Background

In this section we introduce some background theory related to our proposed method.

2.1 Probabilistic Graphical Model

The fundamental structure among a series of random variables is depicted by their joint probability distribution. Probabilistic Graphical Models (PGM) is used to describe the *conditional independence or dependence* structure implied by the joint distribution with a graph-induced decomposition of the joint density function. In such a graph or network, a node is a random variable, and an edge between two nodes indicates certain stochastic association. Given a set of random variables $\mathcal{S}$ and their joint distribution $\mathcal{P}$, the set of its conditional independence structure is denoted as $\mathcal{I}(\mathcal{P}) = \{(x, y, \mathcal{C})|x, y \in \mathcal{S}, \mathcal{C} \subseteq \mathcal{S} \setminus \{x, y\}, x \perp\!\!\!\perp y|\mathcal{C}\}$ and a PGM graph $\mathcal{G}$ aims to model such structure. If $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{P})$, $\mathcal{G}$ is said to be an *I-map* of $\mathcal{P}$. If deleting any edge makes $\mathcal{G}$ no longer an I-map, $\mathcal{G}$ is called a *minimal I-map*. If $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{P})$, $\mathcal{G}$ is then a *perfect map* of $\mathcal{P}$. Here we briefly introduce the basic theory of Bayesian Networks, Markov Random Field and Factor Graph, while all details could be found in Bishop (2006) and Koller and Friedman (2009).

A Bayesian Network (BN) is defined to be a PGM represented by a directed acyclic graph (DAG), where each directed edge starts from a *cause* random variables and ends with a *result* random variable. See Figure 1(A) for an example of BN. The joint density function, according to the example, can be decomposed to a series of conditional density function of result given cause:

$$f(x, y, z, w, u) = f(u|w, z)f(w|z)f(z|x, y)f(y)f(x) \tag{2.1}$$

This decomposition describes conditional independence structure among

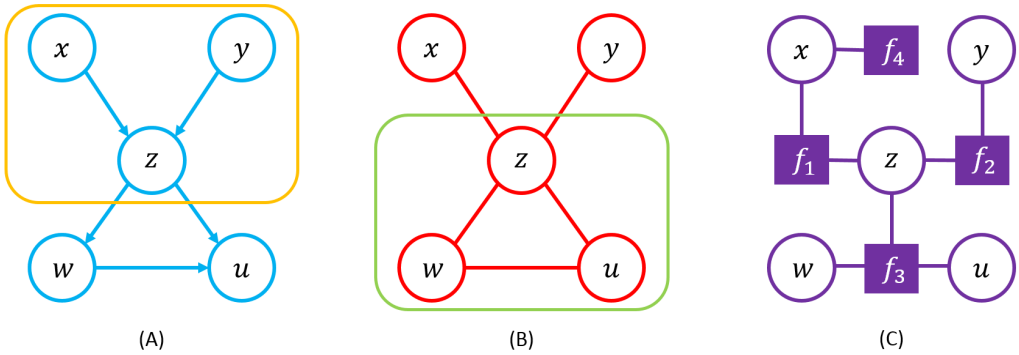

Figure 1: (A): an example of Bayesian Network. The outlined orange part is a V-structure. (B): an example of Markov Random Field. The outlined green part is a maximal clique. (C): an example of a Factor Graph. The solid squares are the abstract representation of factored potential functions.

the variables, e.g. $x \perp u|\{w, z\}$ thus the first term on the right hand side of (2.1) does not include x. It also specifies a conjectured generative causal process among the variables. The node of cause is called a *parent* while the result is called a *child*. Formally, a BN $\mathcal{G}$ specifies the joint distribution $\mathcal{P}$ of a random vector $X = (X_1, \ldots, X_p)^\intercal$ by assuming the density function can be decomposed according to the graph:

$$f(x) = \prod_{i=1}^{p} f(x_i|\text{pa}_i) \tag{2.2}$$

where pa_i is the set of all parents of x_i. If $\mathcal{P}$ can factor according to $\mathcal{G}$, then $\mathcal{G}$ is proved to be a minimal I-map, namely all its conditional independence claims are true for $\mathcal{P}$. BN commonly assumes all variables are discrete, thus each follows a multinomial distribution and forms an expanded grid of a joint contingency table, whose parameter size grows exponentially with the number of variables. Probability of falling into certain slots in such grid can be decomposed according to (2.2) to reduce parameter size. Another class of BN assumes all variables are normally distributed and their relationship is conditionally linear:

$$X_i|\text{pa}_i \sim N(\alpha_i + \sum_{X_j \in \text{pa}_i} \beta_{ij} X_j, \sigma_i^2) \tag{2.3}$$

which is often referred to as a Gaussian Bayesian Network (GBN). It has been proved that for every GBN $\mathcal{G}$ there exists an equivalent multivariate normal

distribution $\mathcal{P}$ that $\mathcal{I}(\mathcal{P}) = \mathcal{I}(\mathcal{G})$, and for every such $\mathcal{P}$ we can find a minimal I-map GBN as well.

A Markov Random Field (MRF) is a PGM with undirected graph, where each edge directly implies conditional dependence given all other nodes. In an undirected graph, a *clique* is defined as a subset of nodes that are pairwisely connected, and a *maximal clique* is a clique that will no longer be a clique if we add any nodes into the subset. A maximal clique in a graph is equivalent to a non-factorable function in the joint density. See Figure 1(B) for an example of MRF, where the subset bounded by green outline is a maximal clique. The joint density function can be factored according to the maximal cliques:

$$f(x, y, z, w, u) = K\psi_1(x, z)\psi_2(y, z)\psi_3(z, w, u) \tag{2.4}$$

where ψ_i is a function of the members of a clique, commonly referred to as the *potential function*, and K is a normalization constant to make the right hand side a valid density function. Generally, a MRF specifies the joint distribution $\mathcal{P}$ of a random vector $X = (X_1, \ldots, X_p)^\intercal$ by decomposing the density function:

$$f(x) = K \prod_{c \in C_m} \psi_c(x_c) \tag{2.5}$$

where c is a clique, X_c and ψ_c are the corresponding nodes and potential functions, and C_m is the set of all maximal cliques. Similarly, if $\mathcal{P}$ can factor according to such MRF $\mathcal{G}$, then $\mathcal{G}$ is proved to be a minimal I-map for $\mathcal{P}$.

The most appealing feature of a PGM is that the clear and sparse conditional dependence relationship can be visually inferred from the graph, for further quantitative and qualitative analysis. For a MRF:

- Consider all nodes that have been conditioned on becoming 'blocks' on the network paths, then two nodes are conditionally dependent, if and only if there is a clear path between them.

In Figure 1(B), if we condition on z, then $x \perp\!\!\!\perp y$ because there is no longer a path between them, but $w \not\perp\!\!\!\perp u$ since they are still connected. Simply setting z to a constant in (2.4) can be an alternative algebraic justification. Conditional dependence in BN is trickier. In Figure 1(A), if we condition on z, then x and y are dependent. This is due to the fact that the joint distribution of these three variables is $f(z|x, y)f(y)f(x)$, and by setting z to a constant, the first term is still non-factorable regarding to x and y. In fact, the shape $x \rightarrow z \leftarrow y$, called a *V-structure*, is the only exception we need to consider in a BN.

- Consider all nodes that have been conditioned on becoming 'blocks' on the network paths. And for every such V-structure, if the center node

and all of its descendants are not conditioned on, we block it; otherwise we add an edge between the parents. We then have that two nodes are conditionally dependent, if and only if there is a clear path between them.

This property is called the *d-separation* of a BN (*d* for *directed*). Note that BN and MRF are mutually convertible, with possible loss of structure. In fact, the conditional independence that BN and MRF can handle are two intersecting subsets of the universal set.

Factor Graph was proposed to have a comparatively more general decomposition, as shown in Figure 1(C):

$$f(x, y, z, w, u) = f_1(x, z) f_2(y, z) f_3(z, w, u) f_4(x) \tag{2.6}$$

The four functions are the so-called *factors* and are explicitly shown as nodes in the undirected graph. Factor Graph is said to be *bipartite* because it has two kinds of nodes: variables and factors. The edges can only be drawn between a variable and a factor, according to the decomposition (2.6). The factors are abstract nodes to mediate the effects among variable nodes, and such representation is not unique, e.g. $f_4(x)$ can be merged into $f_1(x, z)$ and thus omitted. Generally, a Factor Graph has the form

$$f(x) = \prod_s f_s(x_s) \tag{2.7}$$

where s is the index for a subset of the variables. Note both (2.2) and (2.5) can be seen as special cases of (2.7), therefore converting a BN or MRF to a Factor Graph is trivial. Factor Graph is flexible, can deal with directed and undirected models in the same framework and also can be constructed based on the possible prior information of the application. However the decomposition is subjective and does *not* indicate conditional independence or causal suggestion. The Factor graph is mainly used as an abstract tool to facilitate the inference and estimation of a BN or MRF, e.g. via sum-product and max-sum algorithm. More details could be found in Kschischang, Frey, and Loeliger (2001). In Section 3, we proposed a more straight-forward variation of the Factor Graph, that has better interpretability and retains the informative structure of both a Bayesian Network and a Markov Random Field.

2.2 Learning Sparse Gaussian Networks

The graphical structure of the networks, or equivalently the decomposition of joint distribution, is assumed to be known in the previous subsection, which

is not practical when we start to explore in a new territory. Learning an optimal BN for a set of discrete random variables is NP-hard and is thus usually performed via certain greedy searching algorithms, which we do not discuss further in this text. Interested readers are referred to Koller and Friedman (2009). GBN and GMRF (a.k.a. Gaussian Graphical Model), on the other hand, could be learned by computing the partial correlation coefficients of the random vector, since lack of partial correlation is equivalent to lack of conditional dependence given all others, for normal random variables.

For the data matrix $\mathbf{X}_{n\times p} = (x_1, \ldots, x_n)^\intercal$ containing i.i.d. samples of a p-dimensional multivariate normal random vector $X = (X_1, \ldots, X_p)^\intercal$ with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, the partial correlation coefficient between random variables X_i and X_j given all other variables is

$$\rho^\star_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \tag{2.8}$$

where $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = \{\omega_{ij}\}$ is the precision matrix. Since zero entries in $\mathbf{\Omega}$ indicate conditional independence between corresponding variable pairs, given all other variables, we can maximize the following log-likelihood with sparsity restriction as described in Friedman, Hastie, and Tibshirani (2008):

$$\log\det(\mathbf{\Omega}) - \mathrm{tr}(\mathbf{S\Omega}) - \lambda||\mathbf{\Omega}||_1 \tag{2.9}$$

where $\mathbf{S}$ is the sample covariance matrix and here $||\cdot||_1$ of a matrix is defined as the $l1$-norm of its vectorized form. Such model is named Graphical Lasso, since it directly estimates a sparse GMRF, i.e. an undirected graph, by putting an edge between X_i and X_j if $\omega_{ij} \neq 0$.

Another equivalent form was proposed by Peng et al. (2008), which is simply a set of linear regressions of one variable on all others:

$$X_i = \sum_{j\neq i} \beta_{ij} X_j + \epsilon_i \tag{2.10}$$

given $cor(\epsilon_i, X_{-i}) = 0$. And it can be shown that $\beta_{ij} = \rho^\star_{ij}\sqrt{\omega_{jj}/\omega_{ii}}$ and thus $\rho^\star_{ij} = sign(\beta_{ij})\sqrt{\beta_{ij}\beta_{ji}}$, therefore testing whether partial correlation coefficients equal zero, testing conditional independence, and variable selection of a set of regression models are all equivalent. The following loss function is used to estimate the joint regression models:

$$L_n(\rho, \theta, \mathbf{X}) = \sum_{k=1}^{n}\sum_{i=1}^{p}(x_{ik} - \sum_{j\neq i}\rho^\star_{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}x_{jk})^2 + \lambda||\rho^\star||_1 \tag{2.11}$$

where $\rho^\star = \{\rho^\star_{ij} | i < j\}$,$\omega = (\omega_{11}, \ldots, \omega_{pp})$. The structure of GMRF is determined by examining whether $\rho^\star_{ij} = 0$ likewise, and the directed version could be obtained by the conversion from GMRF to GBN as mentioned in the previous subsection.

2.3 Learning Mixed Networks

In the case of both continuous and categorical variables, apart from the heuristic discretization methods which we do not discuss in this text, Lauritzen (1996) described the first sound framework called Conditional Gaussian Bayesian Network. For a continuous random vector X and a categorical one Y, the joint probability distribution is factored into

$$f(x, y) = \prod_{j \in \Delta} P(y_j | \text{dpa}_j) \prod_{i \in \Gamma} f(x_i | \text{dpa}_i, \text{cpa}_i) \tag{2.12}$$

where Δ and Γ are the index sets of discrete and continuous variables, dpa_i and cpa_i are the sets of discrete and continuous parents of the variable of index i. Apparently, the model does not allow continuous parents for a discrete variable, since the mixed part of joint distribution can only be Gaussian, conditioning on the values of the discrete variables:

$$X_i | \text{dpa}_i, \text{cpa}_i \sim N(\mu_i | \text{dpa}_i + (\beta_i^\intercal | \text{dpa}_i) \cdot X_{\text{cpa}_i}, \sigma^2 | \text{dpa}_i) \tag{2.13}$$

where all the parameters are conditioned on the discrete parents. The discrete part of joint distribution is simply the right hand side first term of (2.12). The joint distribution can be rewritten in a clear form in the exponential family

$$f(x, y) = \exp(g_y + h_y^\intercal x - \frac{1}{2} x^\intercal \mathbf{K}_y x) \tag{2.14}$$

and here g_y, h_y, $\mathbf{K}_y$ are the canonical parameters conditioning on the discrete Y. It is straight forward to see that, the number of parameters grows exponentially as the number of possible combinations of levels within the categorical variables explodes. Bø ttcher (2001) presented a Bayesian method to learn the structure of such networks with Dirichlet and Inverse Gaussian priors respectively for discrete and continuous distribution parameters. The conditional Gaussian network was then simplified and applied to learning a mixed MRF, or a Mixed Graphical Model, by Cheng, Levina, and Zhu (2013). The exponential part in (2.14) was written in linear and pairwise quadratic forms of the variables, and it was shown that the coefficients can be obtained by solving a series of conditional likelihood problems, i.e. a set of generalized linear regression models, similar to Peng et al. (2008). The conditional

distribution of continuous variables are still Gaussian as in (2.13), while the category probability of a binary discrete variables is depicted by a Logistic regression on others variables:

$$\log \frac{P(Y_j = 1|X, Y_{-j})}{P(Y_j = 0|X, Y_{-j})} = g(Y_{-j}, X, XX^\intercal|\theta) \tag{2.15}$$

where $g(\cdot)$ is a linear equation with coefficients θ, and $XX^\intercal$ denotes all the cross-product terms of X. A weighted Lasso penalty is then imposed to assure the identification and sparsity of the model. Lee and Hastie (2013) independently developed an equivalent setting with minor difference, e.g. the direct use of the general form of multinomial Logistic regression, Group Lasso penalty and convex optimization form. MRF does not have the concern of continuous-parent-discrete-child problem because it has no implication on a generative causal model. The regression sets, each with its own penalty, are solved separately but still assuming one common tuning parameter for computational feasibility. The regression coefficients are at least estimated twice during the process due to parameter overlapping, but only the largest is suggested to be taken, and the grouped coefficients are coarse to model the complex inter-category interaction among discrete variables. These complications cost the model to lose the simplicity of processing a Gaussian Network as in the previous subsection.

2.4 Multinomial Probit Latent Factor Model

Here we briefly introduce the Probit latent factor model that inspired our work. Consider a categorical random variable Y with possible values $\{0, 1, \ldots, K\}$ to indicate the $K + 1$ levels (categories), we assume

$$Y = \begin{cases} 0 & \text{if } \max(U) \leq 0 \\ k & \text{if } \max(U) = U_k > 0, k = 1, \ldots, K \end{cases} \tag{2.16}$$

where $U = (U_1, \ldots, U_K)$ is the latent continuous normal random variables, sometimes called the *utilities* in decision theory, that models the underlying intention or priority for Y to fall in a certain category. Further more, we build the following multivariate linear model

$$U = \mathbf{B}X + \mathbf{L}F + \epsilon \tag{2.17}$$

where X is the $p \times 1$ covariates vector, F the $d \times 1$ unobserved latent factor, $\mathbf{B}$ and $\mathbf{L}$ the corresponding linear coefficients and loading matrices and ϵ is the zero-mean error term. We usually assume uncorrelated latent factors

$F \sim N(0, \mathbf{I})$ and error $\epsilon \sim N(0, \mathbf{\Psi})$ where $\mathbf{\Psi}$ is a diagonal matrix, then the model is identifiable up to a rotation matrix multiplying $\mathbf{L}$, and can be estimated via Maximum Likelihood. An example can be found in Zhou and Liu (2007). Particularly, the distribution of a binary variable $Y \in \{0, 1\}$ is actually modeled by:

$$P(Y = 1) = \Phi(\mathbf{B}X + \mathbf{L}F) \tag{2.18}$$

where Φ is the standard normal cumulative distribution function. For a multiclass Y, the probability structure can be similarly interpreted on each of its binary dummy variables.

Without the factor term F, (2.16) and (2.17) is simply a Generalized Linear Model with a multinomial Probit link, a not-so-popular alternative of the multinomial Logistic regression, also known as the Max-Entropy or Softmax regression, to describe the relationship between a categorical response and a series of predictors. Probit model is less used mainly because the estimation involves computing the integral of a multivariate truncated normal density function, which is often intractable and needs Monte Carlo integration. Classical Logistic model is by contrast easy to fit using analytical iterations, but often accused for its unjustified assumption of *independence of irrelevant alternatives* (IIA). IIA states that the probabilistic preference of one category over another does not change in the presence or absence of other possible categories. This is often undesirable because categories of a discrete random variable could have more complex association, such as hierarchy. For instance, given the choice of taking a blue bus, a red bus and a car to work, the probability of choosing either is the same. IIA leads to an unrealistic argument that, if we remove the red bus from the alternatives, the probability of choosing the blue bus is still equal to that of choosing the car. Consequently, Probit model can accommodate much more complex categorical variables at a cost of time efficiency. Another advantage of Probit link is that the latent utilities are Gaussian, which greatly facilitates its natural incorporation into Gaussian Networks as shown in the next section.

3 Main Results

We here propose the MIxed Sparse FActor Network (MISFAN) model, and unveil its interpretation and estimation. Unlike the above methods, MISFAN starts from another perspective by transforming the mixed network into a latent Gaussian Network with Probit links, and the association among variables is modeled by some sparse latent factors as the fundamental source of a

generative process.

3.1 Model

Suppose we have p continuous variables, and m nominal categorical variables each with $n_j + 1$ levels, $j = 1, \ldots, m$, the MISFAN of such a set of variables is defined as:

$$W \coloneqq \begin{pmatrix} X \\ Z \end{pmatrix} = \mu + \mathbf{L}F + \epsilon \tag{3.1}$$

$$Y_j = \begin{cases} 0 & \text{if } \max(Z_j) \leq 0 \\ k & \text{if } \max(Z_j) = Z_j^k > 0, k = 1, \ldots, n_j \end{cases}, j = 1, \ldots, m \tag{3.2}$$

where X is a p-dimensional random vector denoting the observed continuous variables, $Z = (Z_1^1, \ldots, Z_1^{n_1}, \ldots, Z_m^1, \ldots Z_m^{n_m})^\intercal$ is a random q-vector denoting the latent utilities, Y is a random m-vector with each entry $Y_j \in \{0, \ldots, n_j\}, j = 1, \ldots, m$ as a categorical variable indicating among the $n_j + 1$ levels which it takes. Each such Y_j is determined by n_j latent utilities, namely $(Z_j^1, \ldots, Z_j^{n_j})^\intercal$,thus $q = \sum_{j=1}^m n_j$. X and Z are combined into W for notation convenience, and assumed to have a low-dimensional linear structure represented by the d independent Gaussian latent factors $F \sim N_d(\mathbf{0}, \mathbf{I})$ $(d \leq p + q)$ with a loading matrix $\mathbf{L}$, and corrupted by some Gaussian error $\epsilon \sim N_{p+q}(\mathbf{0}, \mathbf{\Psi})$, where $\mathbf{\Psi} = diag\{\psi_1, \ldots, \psi_{p+q}\}$ allows different error variance for each random variable. The error variance ψ_j is sometimes called the *uniqueness* since it is the part of individual variation of a variable other than the common part shared with other variables. The formulation terms seem complex for its rigorousness, and the Bayesian Network representation of MISFAN in Figure 2 might be more intuitive.

From Figure 2 we can observe the following properties, from left to right:

- Each latent factor F_j is an independent root of the Bayesian Network, and portrays the low-dimensional *true* structure within the multivariate normal distribution. The number of factors d can be interpreted as the degree of freedom.

- The underlying dependence structure relies on the linear transformation between the first two columns, namely, on the loading matrix $\mathbf{L}$. A nonzero term l_{ij} indicates an edge from F_j to $\bar{W}_i$, and F_j can mediate this effect to $\bar{W}_k$ if $l_{kj} \neq 0$. Two variables that share the same factor are correlated.

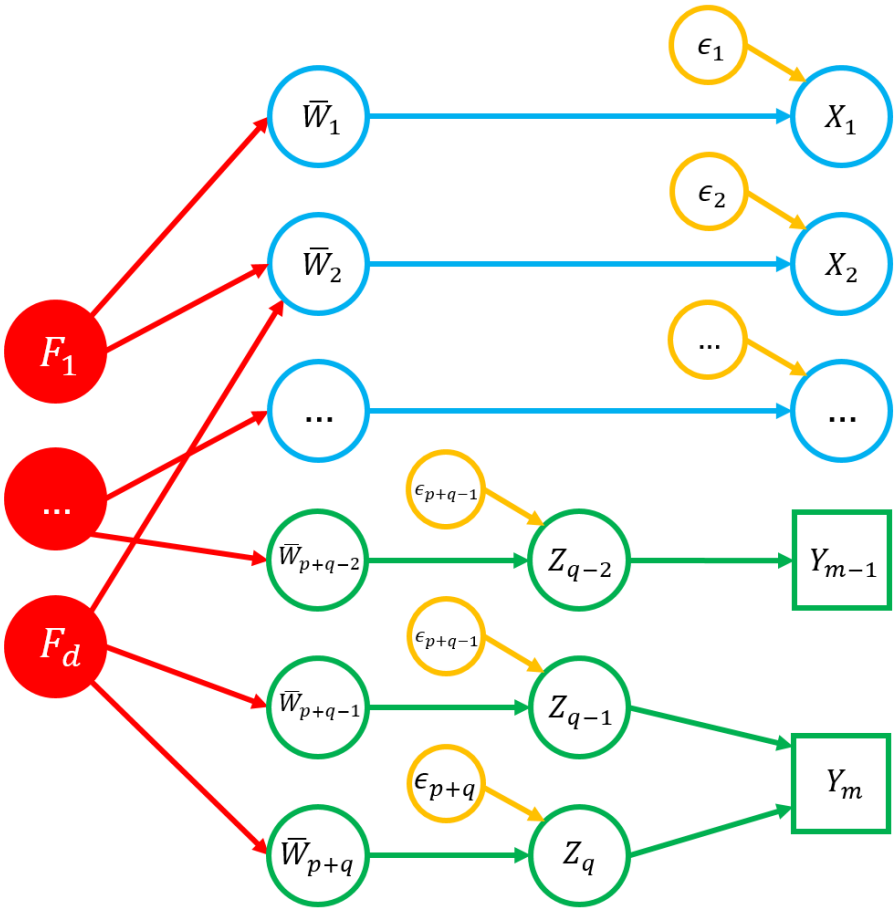

Figure 2: MISFAN illustrated as a Bayesian Network. The first column consists of the latent factors, and the second column is the $p + q$ *true* linear combinations (denoted by $\bar{W} := \mu + \mathbf{L}F$ instead of W) of such factors with a set of links. Independent Gaussian errors (yellow) are then added to the variables. For the q latent utilities (green), there is an additional step to convert them into observable categorical variables (shown as squares) via the Probit link. In this example, Y_{m-1} is a binary variable while Y_m has 3 unordered levels. As for the p continuous variables (blue), the error corrupted result is directly observable. For more details see text.

- The model can deal with heterogeneous error corruption, which almost always exists in application. Such error could be interpreted as either model error or measurement error.

- Both continuous and categorical variables are treated indistinguishably as latent normal random variables, and thus the dependence structure of a mixed network is simplified into the form of a latent multivariate normal distribution, i.e. a latent Gaussian Network. In this sense MISFAN directly applies to the all-discrete or all-continuous cases.

- A categorical variable is manifested by its latent utilities with the powerful Probit link, so its association to other variables is decomposed into the more sophisticated correlations between each of its utility component and other variables. The ambiguous relationship between a multi-level categorical variable with another variable can only be thoroughly investigated through the connection to its individual utilities.

The joint distribution of a MISFAN is

$$f(x,y) \quad = \quad \int_{z\in\mathcal{Z}(y)} f(x,z)dz \tag{3.3}$$

where $\mathcal{Z}(y)$ is the set of z that satisfies a given y in the sense of (3.2). It is clear that the factorization of (3.3) can be conducted directly through the multivariate normal density $f(w) = f(x,z)$, i.e. the latent Gaussian Network, and the conditional independence structure then depends on the precision matrix $\mathbf{\Omega} = (\mathbf{L}\mathbf{L}^\intercal + \mathbf{\Psi})^{-1}$. Therefore, we do not distinguish the structure of the latent Gaussian Network and MISFAN here after. Classically speaking, MISFAN is a variation of an exploratory multivariate multinomial Probit plus multivariate Gaussian mixed response latent factor model, and to the best our knowledge, this is also the first literature on details of the subject. For inference, the conditional distribution of a continuous variable is a normal distribution and can be represented in a linear regression model; the conditional distribution of a categorical variable is given by a multinomial Probit regression. Notably, MISFAN has a different model coverage from the Conditional Gaussian Network for mixed variables as in Lauritzen (1996), specifically on how variables interact. As a result, it gains certain better explanatory advantage. In the next subsection we build a better visualized graph and discuss how to interpret our results and compare it with other PGMs.

3.2 Interpretation

With a model as in (3.1), we might wonder how the parameters are associated with the underlying parsimonious graphical structure.

The number of factors d indicates a means of dimension reduction to filter out the noise and reveal a cleaner linear pattern. It also serves as the degree of freedom, i.e. the factors can be treated as independent roots of cause and all variables are the consequences. This interpretation does not limit its application when actually one *variable* is the true root, since switch the direction of a single edge between a factor and the variable does not change the BN's conditional independence structure, namely the resulting BNs are *I-equivalent* to the original one. Therefore d can be determined through domain knowledge regarding to the possible latent causal factors, for the sake of interpretation.

The loading matrix $\mathbf{L}$ is the connection between the factors and the variables and indicates the range of effect of a certain factor. In real life application with numerous variables, it is sensible and desirable that a particular latent cause only affects the system locally, instead of directly covering the whole realm. So sparsity in $\mathbf{L}$ is helpful to obtain an interpretable and also *unique* model, since $\mathbf{L}$ is identifiable up to a rotation matrix in the original factor model but becomes unique with a specific sparsity penalty added. As mentioned in the previous paragraph, within the group of variables connected to a certain factor, any of them could actually be the root cause. These sibling variables are undoubtedly correlated but with unknown conditional dependence structure, at least for the moment. Nodes that have more than one parent factor are at the intersection of the causal flow, i.e. they are affected by more than one cause. This observation naturally assigns hierarchical labels to the variables: nodes with only one parent are *close* to the hypothetical cause (the latent factor or one of its siblings), while multi-parent nodes are at the family boundary shared with others factors, if we assume factor influence tends to be local. In this paper we consider only exploratory factor analysis and estimate the whole loading matrix, but confirmatory factor model with prior restrictions and preference on $\mathbf{L}$ can be incorporated with ease.

The diagonal error covariance matrix $\mathbf{\Psi}$ allows heteroskedasticity across variables. It can be considered as the variation of measurement error and thus incorporate an Error-in-Variable model framework, or the unique information the variable carries regardless of others. With such a general diagonal covariance matrix, MISFAN can model any multivariate normal distribution and thus equivalently any induced Bayesian Networks. If the *true* variables are on similar scales with assumedly equal measurement error, than the value of ψ_j would contain the equation error. Note in a Bayesian Network variables

are sequentially generated from roots to leaves via linear equations and such a process will make error accumulates. Thus with proper assumption, the relative scale of ψ_j could be used to infer causal sequence as well.

In summary, the Bayesian Network with factors implies possible causal structure and plenty of other information like node clustering by family boundaries. See the first column in Figure 3 for some examples. The factors serve as the hubs among a set of nodes, so we put no care on a single isolated factor or variable. The plots resemble the Factor Graph, especially like the first row of Figure 3, where each factor and its neighbors actually indicate a clique in the Markov Random Field version (elaborated later). However the *factor* here is for a latent Gaussian variable, instead of a part of a joint distribution function, and the resemblance is not general because the underlying causal and dependence structure could be much more complex, yet we only have at most $p + q$ factors for such description as in a Factor Graph.

The only disadvantage, in fact a deal breaker, of such BN with factors is that it does *not* indicate conditional independence in the graph among the variables, just like Factor Graph, due to the existence of non-conditionable factors. As shown in the second column of Figure 3, the underlying dependence structure can be substantially different for seemly similar factor connections. Since MISFAN assumes an underlying Gaussian Network, obtaining conditional independence structure simply depends on computing the precision matrix and build a GMRF.

Counter-intuitively, the sparsity in the factor model does not induce sparsity in the conditional dependence. This can be algebraically justified: the off-diagonal elements of $(\mathbf{L}\mathbf{L}^\intercal + \mathbf{\Psi})^{-1}$ do not necessarily vanish for a sparse $\mathbf{L}$, even though a sparse $\mathbf{L}$ leads to a somewhat sparse $\mathbf{\Sigma}$. This is because the inverse of a sparse matrix is not generally sparse. Therefore, another constraint is needed, and the best candidate is, of course, on $\mathbf{\Omega}$ itself. As mentioned, the sparsity of $\mathbf{L}$ controls the number of nodes that are connected in a subgraph, and the complexity of this connection is then controlled by the sparsity of $\mathbf{\Omega}$. For instance, Figure 3 (A) (B) and (C) rows are for the pairs of (sparse $\mathbf{L}$, dense $\mathbf{\Omega}$), (sparse $\mathbf{L}$, sparse $\mathbf{\Omega}$) and (dense $\mathbf{L}$, sparse $\mathbf{\Omega}$) respectively.

To deal with the seemly separated generative process and conditional independence network structure, we design a Factor Network as the graphical part of MISFAN, to visualize the possible causal process as a Bayesian Network, a straight-forward conditional dependence structure as a Markov Random Field, and additionally the effect of variables clustering. Factor Network examples are shown in the third column of Figure 3. It basically combines the notation of the previous two plots. The graph edges comes from the MRF, and the

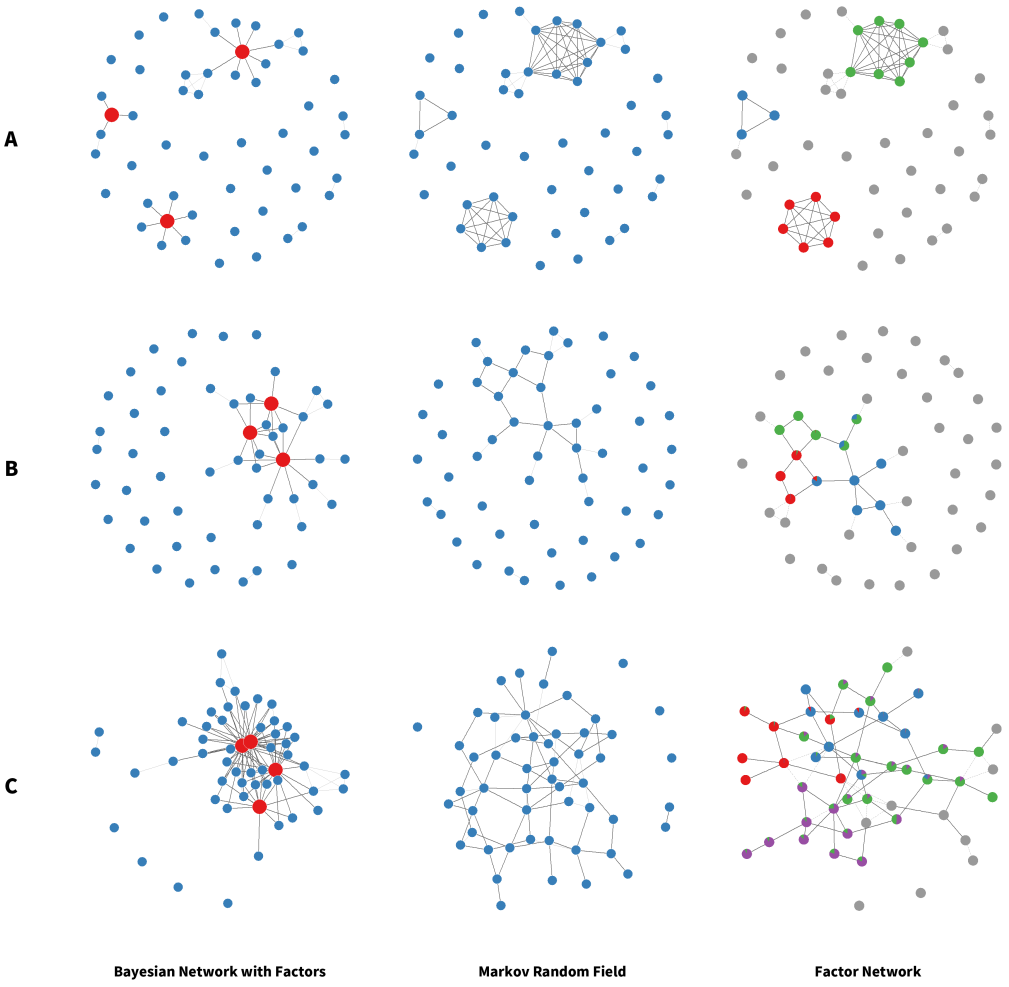

Figure 3: Two aspects of MISFAN and the combined Factor Network, with three typical examples. The left column consists of Bayesian Networks where an edge always goes from a latent factor (red) to a variable (blue). The middle column displays the Markov Random Field determined by the underlying multivariate normal distribution. The Factor Network combines both features and is listed in the right column. Nodes Color denotes different proportion of factor influence and can either determine variable clusters or provide suggestion for possible causal chain: nodes with a purer color combination are closer to the source of cause. Edges are the same as in the middle column. The three examples show some typical structures of interest, i.e. (A) isolated node groups with dense inner connection (e.g. cliques), (B) locally connected groups with simple structure (e.g. chains), and (C) globally connected graph with a more complex structure.

nodes are colored pie charts by the proportion of influence (absolute weights in $\mathbf{L}$) from the factors according to the first column. The clustering by factors coverage, causal flow by color impurity and conditional independence structure by edges are self-explanatory in the graph. Note since the two sparsity representation are partially independent, the factor coverage and path connection might not exactly agree with each other. Originally un-connected grouped latent utilities are here connected by a dashed grey edge, and originally connected ones a solid grey edge, just for visual convenience to show a categorical variable coherently. We next turn to the algorithm of learning such Factor Network structure from data, namely estimating the MISFAN.

3.3 Estimation

Let the parameter set $\theta = (\mu, \mathbf{L}, \mathbf{\Psi})$, and precision matrix $\mathbf{\Omega} = (\mathbf{L}\mathbf{L}^\intercal + \mathbf{\Psi})^{-1}$. An observed i.i.d. sample of size n is denoted by $\mathbf{X}_{n\times p} = (x_1, \ldots, x_n)^\intercal$, $\mathbf{Y}_{n\times m} = (y_1, \ldots, y_n)^\intercal$, corresponding unobserved latent variables and factors are denoted by $\mathbf{Z}_{n\times q} = (z_1, \ldots, z_n)^\intercal$ and $\mathbf{F}_{n\times d} = (f_1, \ldots, f_n)^\intercal$. Likewise we have $\mathbf{W}_{n\times(p+q)} = (w_1, \ldots, w_n)^\intercal = (\mathbf{X}, \mathbf{Z})$. Ideally, as the discussion of sparsity in the previous subsection indicates, the actual log-likelihood should be maximized with the dual sparsity penalty:

$$\max_{\theta} \ell(\theta|\mathbf{X}, \mathbf{Y}) - \lambda P(\mathbf{L}) - \rho P(\mathbf{\Omega}) \tag{3.4}$$

where $P(\mathbf{L})$ and $P(\mathbf{\Omega})$ are the penalty functions to ensure sparse factor loadings and precision matrix respectively. Inspired by Hirose and Yamamoto (2014), Zhou and Liu (2007), Liu and Rubin (1998) and Zhao, Yu, and Jiang (2008), we use a Monte Carlo Expectation Conditional Maximization Either (MC-ECME) approach to solve (3.4).

Assuming latent variables are observed, the complete penalized log likelihood function is

$$\begin{aligned}
\tilde{\ell}(\theta|\mathbf{W}, \mathbf{F}) &= \ell(\theta|\mathbf{W}, \mathbf{F}) - \lambda P(\mathbf{L}) - \rho P(\mathbf{\Omega}) \\
&= \ell(\theta, \mathbf{F}|\mathbf{W}) + \ell(\theta|\mathbf{F}) - \lambda P(\mathbf{L}) - \rho P(\mathbf{\Omega}) \\
&= -\frac{(p+q)n}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Psi}| \\
&\quad -\frac{1}{2}\sum_i (w_i - \mu - \mathbf{L}f_i)^\intercal \mathbf{\Psi}^{-1}(w_i - \mu - \mathbf{L}f_i) \\
&\quad -\frac{dn}{2}\log(2\pi) - \frac{1}{2}\sum_i f_i^\intercal f_i - \lambda P(\mathbf{L}) - \rho P(\mathbf{\Omega})
\end{aligned} \tag{3.5}$$

We derive the ECME steps as below.

E-step

$$\begin{aligned}\mathrm{E}[\tilde{\ell}(\theta|\mathbf{W},\mathbf{F})|\mathbf{X},\mathbf{Y},\hat{\theta}] &= \text{const} + \frac{n}{2}\log\det(\mathbf{\Psi}^{-1}) - \lambda P(\mathbf{L}) - \rho P(\mathbf{\Omega}) \\ &\quad -\frac{1}{2}\sum_i \mathrm{tr}(\mathbf{\Psi}^{-1}\mathrm{E}[(w_i-\mu-\mathbf{L}f_i)(w_i-\mu-\mathbf{L}f_i)^\intercal|\mathbf{X},\mathbf{Y},\hat{\theta}])\end{aligned} \tag{3.6}$$

For convenience, we later on use a *star* notation for the conditional expectation: $w^\star = \mathrm{E}[w|\mathbf{X},\mathbf{Y},\hat{\theta}]$. We have

$$\begin{aligned}[(w_i-\mu-\mathbf{L}f_i)(w_i-\mu-\mathbf{L}f_i)^\intercal]^\star &= \mathbf{L}(f_i f_i^\intercal)^\star\mathbf{L}^\intercal - \mathbf{L}((f_i w_i^\intercal)^\star - f_i^\star\mu^\intercal) \\ &\quad -((f_i w_i^\intercal)^\star - f_i^\star\mu^\intercal)^\intercal\mathbf{L}^\intercal \\ &\quad +((w_i-\mu)(w_i-\mu)^\intercal)^\star\end{aligned} \tag{3.7}$$

Analytical expression of each star term above on the right hand side is necessary. Note

$$\begin{pmatrix} W \\ F \end{pmatrix} \sim N_{p+q+d}(\begin{pmatrix} \mu \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma} & \mathbf{L} \\ \mathbf{L}^\intercal & \mathbf{I} \end{pmatrix}) \tag{3.8}$$

where $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\intercal + \mathbf{\Psi}$. Then

$$\begin{aligned} f_i^\star &= \mathrm{E}[f_i|\mathbf{X},\mathbf{Y},\hat{\theta}] = \mathrm{E}[\mathrm{E}[f_i|\mathbf{W}]|\mathbf{X},\mathbf{Y},\hat{\theta}] \\ &= \mathrm{E}[\hat{\mathbf{L}}^\intercal\hat{\mathbf{\Sigma}}^{-1}(w_i-\hat{\mu})|\mathbf{X},\mathbf{Y},\hat{\theta}] \\ &= \hat{\mathbf{L}}^\intercal\hat{\mathbf{\Sigma}}^{-1}(w_i^\star-\hat{\mu}) \\ (f_i f_i^\intercal)^\star &= \mathrm{E}[\mathrm{E}[f_i f_i^\intercal|\mathbf{W}]|\mathbf{X},\mathbf{Y},\hat{\theta}] \\ &= \mathbf{I} - \hat{\mathbf{L}}^\intercal\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{L}} + \hat{\mathbf{L}}^\intercal\hat{\mathbf{\Sigma}}^{-1}[(w_i-\hat{\mu})(w_i-\hat{\mu})^\intercal]^\star\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{L}} \\ (f_i w_i^\intercal)^\star &= \mathrm{E}[\mathrm{E}[f_i|\mathbf{W}]w_i^\intercal|\mathbf{X},\mathbf{Y},\hat{\theta}] \\ &= \hat{\mathbf{L}}^\intercal\hat{\mathbf{\Sigma}}^{-1}((w_i w_i^\intercal)^\star - \hat{\mu}w_i^{\star\intercal}) \end{aligned} \tag{3.9}$$

Also note $\hat{\theta}$ in (3.9) is the estimate from the previous iteration. From above we know that the E-step depends on finding $w_i^\star$ and $(w_i w_i^\intercal)^\star$. Since $w_i^\star = (x_i^\intercal, z_i^{\star\intercal})^\intercal$, essentially we need to compute the sufficient statistics $z_i^\star$ and $(z_i z_i^\intercal)^\star$. Getting the conditional distribution of Z involves computing integral over a rectangle region of a multivariate normal density function, and the estimation is usually obtained by a Monte Carlo simulation:

$$\begin{aligned} z_i^\star &= \frac{1}{K}\sum_{k=1}^K z_i^{(k)} \\ (z_i z_i^\intercal)^\star &= \frac{1}{K}\sum_{k=1}^K z_i^{(k)} z_i^{(k)\intercal} \end{aligned} \tag{3.10}$$

where $z_i^{(k)}$ is sampled from $f(z|x_i, y_i, \theta)$, a multivariate truncated normal distribution with mean $\mu_z + \mathbf{\Sigma}_{zx}\mathbf{\Sigma}_{xx}^{-1}(x_i - \mu_x)$ and covariance $\mathbf{\Sigma}_{zz} - \mathbf{\Sigma}_{zx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xz}$, where we decompose $\mu = (\mu_x^\intercal, \mu_z^\intercal)^\intercal$ and $\mathbf{\Sigma}$ to the corresponding block matrices likewise. The normal distribution is truncated so that $z_i^{(k)} \in \mathcal{Z}(y_i)$. It's worth mentioning that we here avoid computing $\mathbf{\Sigma}_{xx}^{-1}$ directly but apply Woodbury formula:

$$\begin{aligned} \mathbf{\Sigma}_{zx}\mathbf{\Sigma}_{xx}^{-1} &= \mathbf{L}_z\mathbf{L}_x^\intercal(\mathbf{L}_x\mathbf{L}_x^\intercal + \mathbf{\Psi}_{xx})^{-1} \\ &= \mathbf{L}_z(\mathbf{I} + \mathbf{L}_x^\intercal\mathbf{\Psi}_{xx}^{-1}\mathbf{L}_x)^{-1}\mathbf{L}_x^\intercal\mathbf{\Psi}_{xx}^{-1} \end{aligned} \tag{3.11}$$

where all block matrices are from the decomposition similarly as before. This trick makes the estimation feasible even if $\mathbf{\Sigma}_{xx}$ is singular, i.e. the $p > n$ case.

CME-steps:

The optimization problem is

$$\arg\max_\theta \mathrm{E}[\tilde{\ell}(\theta|\mathbf{W}, \mathbf{F})|\mathbf{X}, \mathbf{Y}, \hat{\theta}] \tag{3.12}$$

and from (3.6) (3.7) (3.9):

$$\begin{aligned} \mathrm{E}[\tilde{\ell}(\theta|\mathbf{W}, \mathbf{F})|\mathbf{X}, \mathbf{Y}, \hat{\theta}] &= \mathrm{const} + \frac{n}{2}\log\det(\mathbf{\Psi}^{-1}) - \lambda P(\mathbf{L}) - \rho P(\mathbf{\Omega}) \\ &\quad -\frac{n}{2}\mathrm{tr}(\mathbf{\Psi}^{-1}(\mathbf{L}\mathbf{A}\mathbf{L}^\intercal - \mathbf{L}\mathbf{B} - \mathbf{B}^\intercal\mathbf{L}^\intercal + \mathbf{C})) \end{aligned} \tag{3.13}$$

with

$$\begin{aligned} \mathbf{A} &= \frac{1}{n}\sum_i (f_i f_i^\intercal)^\star \\ &= \mathbf{I} - \hat{\mathbf{L}}^\intercal\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{L}} + \hat{\mathbf{L}}^\intercal\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{L}} \\ &= \mathbf{M}^{-1} + \mathbf{M}^{-1}\hat{\mathbf{L}}^\intercal\hat{\mathbf{\Psi}}^{-1}\mathbf{C}\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}}\mathbf{M}^{-1} \\ \mathbf{B} &= \frac{1}{n}\sum_i [(f_i w_i^\intercal)^\star - f_i^\star\mu^\intercal] \\ &= \hat{\mathbf{L}}^\intercal\hat{\mathbf{\Sigma}}^{-1}\mathbf{C} \\ &= \mathbf{M}^{-1}\hat{\mathbf{L}}^\intercal\hat{\mathbf{\Psi}}^{-1}\mathbf{C} \\ \mathbf{C} &= \frac{1}{n}\sum_i [(w_i - \mu)(w_i - \mu)^\intercal]^\star \end{aligned} \tag{3.14}$$

where we define $\mathbf{M} = \mathbf{I} + \hat{\mathbf{L}}^\intercal\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}}$. Woodbury formula is used again as in (3.11). Next we sequentially obtain estimate of each component of θ by Conditional Maximization. Let the partial derivative of (3.13) regarding to μ equals 0, and simple calculus gives the **CM step 1**:

$$\hat{\mu}_{new} = \frac{1}{n}\sum_i (w_i^\star - \hat{\mathbf{L}}\ f_i^\star) \tag{3.15}$$

and $\hat{\mathbf{L}}$ is estimated from the previous iteration. Then we estimate $\mathbf{L}$ with the updated $\mu \leftarrow \hat{\mu}_{new}$ plugged into (3.14) by solving the following optimization problem:

$$\max_{\mathbf{L}} -\frac{n}{2}\text{tr}(\mathbf{\Psi}^{-1}(\mathbf{LAL}^\intercal - \mathbf{LB} - \mathbf{B}^\intercal\mathbf{L}^\intercal + \mathbf{C})) - \lambda P(\mathbf{L}) \tag{3.16}$$

and here we use $P(\mathbf{L}) = \frac{n}{2}||\mathbf{L}||_1$ as an example, namely a Lasso penalty for a columnwisely vectorized matrix, with a constant factor for notation convenience. Alternatives exist, e.g. Adaptive Lasso in Zou (2006) or SCAD in Fan and Li (2001). Note the trace function in (3.16) can be decomposed into $p+q$ independent components for each diagonal entry:

$$\max_{\mathbf{L}} -\frac{n}{2}\sum_{j=1}^{p+q}[\frac{1}{\psi_j}(l_j^\intercal\mathbf{A}l_j - 2l_j^\intercal b_j + c_{jj}) - \lambda||l_j||_1] \tag{3.17}$$

where l_j, b_j and c_{jj} are respectively the jth row of $\mathbf{L}$, the jth column of $\mathbf{B}$ and the (j,j) element of $\mathbf{C}$. Now the problem has been reduced to a series of individual standard Lasso optimization problems. Let $\mathbf{A} = \mathbf{R}^\intercal\mathbf{R}$ be the Cholesky decomposition and $\tilde{b}_j = \mathbf{R}^{\intercal-1}b_j$ then we have standard Lasso problem-(j), $j = 1, \ldots, (p+q)$ in the form of:

$$\min_{l_j} ||\mathbf{R}l_j - \tilde{b}_j||_2^2 + \lambda\psi_j||l_j||_1 \tag{3.18}$$

We then use the standard generalized coordinate descent method to solve (3.17). The solutions $\hat{l_j}$ $(j = 1, \ldots, p+q)$ of problem-(1) to problem-$(p+q)$ can be merged to complete the **CM step 2**:

$$\hat{\mathbf{L}}_{new} = (\hat{l_1}^\intercal, \ldots, \hat{l_j}^\intercal, \ldots, \hat{l_{p+q}}^\intercal)^\intercal \tag{3.19}$$

because the problems are independent from each other . For the final update of $\mathbf{\Psi}$, we no longer use same expected conditional maximization but use a variation of Liu and Rubin (1998): maximizing the penalized likelihood marginalized on $\mathbf{F}$

$$\begin{aligned}\tilde{\ell}(\theta|\mathbf{W}) &= \ell(\theta|\mathbf{W}) - \rho P(\mathbf{\Omega}) \\ &= -\frac{(p+q)n}{2}\log(2\pi) + \frac{n}{2}\log|\mathbf{\Omega}| \\ &\quad -\frac{1}{2}\sum_i (w_i - \mu)^\intercal\mathbf{\Omega}(w_i - \mu) - \rho P(\mathbf{\Omega})\end{aligned} \tag{3.20}$$

and $P(\mathbf{\Omega}) = \frac{n}{2}||\mathbf{\Omega}||_1$ as well. Take expectation then the optimization problem is

$$\max_{\mathbf{\Psi}} \log|\mathbf{\Omega}| - \text{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) - \rho||\mathbf{\Omega}||_1 \tag{3.21}$$

where $\hat{\mathbf{\Sigma}}$ is the estimated covariance matrix by $\hat{\mathbf{L}}_{new}$ and previous $\hat{\mathbf{\Psi}}$. Note this is exactly the same as the Graphical Lasso problem described in the previous section, except we need to estimate $\mathbf{\Psi}$ based on $\hat{\mathbf{L}}_{new}$. After obtaining the estimated sparse precision matrix $\hat{\mathbf{\Omega}}_{new}$ and simultaneously $\hat{\mathbf{\Sigma}}_{new}$, we reach the final **CME step 3**:

$$\hat{\mathbf{\Psi}}_{new} = \text{diag}\{\hat{\mathbf{\Sigma}}_{new} - \mathbf{L}\mathbf{L}^\intercal\}_+ \tag{3.22}$$

with the previous update $\mathbf{L} \leftarrow \hat{\mathbf{L}}_{new}$. Here $\text{diag}\{\mathbf{M}\}_+$ is a diagonal matrix whose entries are those of $\mathbf{M}$ thresholded to be non-negative. The individual optimization problems are all convex thus the ECME algorithm should converge to global maximum. In practice, R package *tmvtnorm* by Wilhelm and G (2014), *glmnet* by Friedman, Hastie, and Tibshirani (2010) and *glasso* by Friedman, Hastie, and Tibshirani (2008) are used to solve corresponding subproblems in this algorithm.

Some additional details are worth noting: the model is not identifiable, because for each categorical variable, its latent utilities have their own mean and variance, which are both connected to the choice of levels, and multiplying any constant to the mean and standard deviation does not affect the likelihood of the data. Therefore we restrict the diagonal elements of the covariance matrices of all groups of latent utilities to be 1, by adjusting a constant factor within each group of latent variables. Another quick fact about identification is, the originally estimate of $\mathbf{L}$ up to a rotation matrix has become uniquely identifiable after adding the Lasso penalty, because $\ell 1$-norm is not invariant to rotation. Therefore a MISFAN is altogether uniquely identifiable given three parameters: d the degree of freedom or information sources, λ the locality of the factor influence, and ρ the sparsity of conditional dependence or neighbor density, which are actually features of the graph from different perspectives. These parameters could be determined empirically, by Akaike and Bayesian information criteria, or by using Cross Validation if sample size is moderate. We will describe some strategies to substantially reduce the parameter grid in the next subsection.

3.4 Miscellaneous

For the initial values of parameters for ECME, we simply use the scaled 0/1 dummy variables as the latent utilities and fit the multivariate normal distribution accordingly. The following example can justify this decision, yet also points to some interesting connection with other statistical models.

Let Y and X be two nominal variables with corresponding possible values $\{0, \ldots, K\}$and $\{0, \ldots, L\}$. We specify two multivariate normal latent random vectors $U = (U_1, \ldots, U_K)^\intercal$and $V = (V_1, \ldots, V_L)^\intercal$ as the utilities of Y and X respectively. The linear association between Y and X is thus solely determined by the cross-covariance matrix of U and V. Assuming such association is indeed linear and nontrivial, we have

$$\begin{aligned} U|F &\sim N_K(\mu_1 + \mathbf{L}_1 F, \mathbf{\Psi}_1) \\ V|F &\sim N_L(\mu_2 + \mathbf{L}_2 F, \mathbf{\Psi}_2) \\ F &\sim N_d(0, \mathbf{I}) \end{aligned} \tag{3.23}$$

where the Gaussian factor F is used to model the underlying linear relationship between the two random vectors. This is one of the simplest examples of a MISFAN, namely a multinomial Probit latent factor model to represent the correlation between two categorical variables. Another perspective of interpretation for this model is from Canonical Correlation Analysis (CCA). Bach and Jordan (2006) placed CCA in a probabilistic formulation and derived the Maximum Likelihood Estimation, in terms of Canonical direction and correlation coefficients between U and V, for all the parameters in (3.23), therefore it is equivalently modeling the canonical correlation between two sets of variables. They also proved equivalence between Linear Discriminant Analysis (LDA) and CCA, simply by replacing U with an indexing dummy vector T, such that $T = (0_{(1)}, \ldots, 1_{(k)}, \ldots, 0_{(K)})^\intercal \iff Y = k$, therefore CCA between (T, V) is equivalent to LDA with Y as the response and V as the predictor. This observation implies that using CCA directly between dummy variables, instead of latent utilities, can achieve comparable performance of LDA in classification. We employ this observation to start the iteration in estimating MISFAN by setting the scaled dummy variables as the initial values for the latent utilities.

The focus of this text is on learning the network structure, from the exponentially growing possibilities. The Markov Blanket, i.e. the neighbors in a MRF, of a variable is a naturally selected set of regressors and more accurate predictive models could be fitted accordingly. The network itself could be validated with further data and confirmatory analysis, e.g. Structural Equation Modeling. If cross validation is desirable for tuning parameters, the standard inference method of GMRF could be applied to give crude prediction:

$$\hat{W}_i - \mu_i = \frac{1}{\omega_{ii}} \sum_{j \in \mathcal{MB}(i)} \omega_{ji}(W_j - \mu_j) \tag{3.24}$$

where $\mathcal{MB}(i)$ is the set of Markov Blanket of W_j. This form is equivalent to (2.10). For latent variables on the right hand side, their Monte Carlo simulated conditional expectation $w^\star$ can be used, just as in the estimation algorithm. The obtained $\hat{w}$ can then be translated into $(\hat{x}, \hat{y})$.

Finally, we discuss about parameter tuning. In the unfortunate case when there is no prior information about the system, tuning a fine grid of (d, λ, ρ) by cross-validation could be computationally prohibitive. We later show in the simulation section that MISFAN is robust against over-specified d, mainly because sparsity in $\mathbf{L}$ can automatically reduce the number of actual effective factors by giving all-zero columns. As mentioned in previous text, λ controls the number of nodes that are connected, while the structure of each group could be from a clique to a chain and ρ governs this complexity. Empirically, we also found that when λ is fixed, the variation in ρ hardly changes the factor coverage structure, but edges between colored nodes can vanish. Therefore, we heuristically suggest choosing a ρ that filters the conditional dependence structure as simple as possible, while majority (say 90%) of the colored nodes still have at least one edge to other nodes, i.e. connected in both sense of $\mathbf{L}$ and $\mathbf{\Omega}$. This agreement of connection in the sense of BN with factors and MRF is defined as the *consistency*. Computing the whole consistency grid by parameters still requires fitting the some models but can substantially lower the burden in the case of cross-validation. Additionally we can further lower the computation by stepwisely increasing the parameter value and stop when inconsistency is first detected. This strategy will be validated in the first example of Section 4.

4 Simulation

In this section we first conduct a numerical experiment to evaluate the estimation algorithm on a simulated dataset generated from a MISFAN model and analyze its behavior. Then we compare performance with Mixed Bayesian Network (MBN) by Bø ttcher (2001) and Mixed Graphical Model (MGM) by Lee and Hastie (2013) on another simulated dataset generated from a toy Bayesian Network with either a latent Gaussian Network with Probit link or a conditional Gaussian Network as the underlying model, in order to put the models on the same page regardless of their distinct assumptions. MISFAN, data simulation, analysis and visualization are implemented in R 3.1.2. Com-

peting models are fitted using R package *deal* for MBG, and the officially provided MATLAB scripts for MGM.

The assessment is mainly on the reconstruction of conditional independence structure of the network, and, for MISFAN model only, the root mean square error (RMSE) between estimated and true parameters. The reconstruction goodness of fit is measured by True Positive Rate (TPR) and False Positive Rate (FPR) on detection of an edge. The two quantities can be calculated by:

$$\begin{aligned} \text{TPR} &= \frac{\#\text{ of estimated edges that agrees with true edges}}{\#\text{ of true edges between pairs}} \\ \text{FPR} &= \frac{\#\text{ of estimated edges that does not agree with true edges}}{\#\text{ of pairs that have no edge between them}} \end{aligned} \tag{4.1}$$

4.1 MISFAN Model

Generating an arbitrarily sparse $\mathbf{\Omega}$ and $\mathbf{L}$ simultaneous has no straightforward analytical process, so we use the following two-step simulation:

1. Given $\mathbf{\Psi}$ and a sparse $\mathbf{L}$, compute $\mathbf{\Omega}$. Randomly assign zeros to the non-zero off-diagonal entries in $\mathbf{\Omega}$ to make it sparse, and rescale the diagonal elements accordingly. Discard almost-zero entries.

2. Fit an Exploratory Factor Analysis on $\mathbf{\Omega}^{-1}$, get new $\mathbf{\Psi}$ and $\mathbf{L}$. Discard almost-zero entries in $\mathbf{L}$. Along with a pre-specified μ, we can then generate data from MISFAN as in (3.1) and (3.2).

The above procedure guarantees dual sparsity, and have two parameters to tweak, i.e. original sparsity in $\mathbf{L}$ and probability to assign non-zero off-diagonal entries of $\mathbf{\Omega}$ to zero, in order to get networks with different densities.

The dataset used in this section consists of $p = 20$ continuous variables and $m = 18$ categorical variables, and the degree of freedom $d = 4$. The categorical variables includes 11 binary, 5 three-level and 2 four-level variables, so there are $q = 27$ latent utilities in total. μ is uniformly distributed on $[-1, 1]$. For the first-step $\mathbf{\Psi}$ and $\mathbf{L}$, we sample ψ_j uniformly from $[0, 1]$, and $\mathbf{L}$ with 85% of zeros and 15% of ± 2. Then 85% of non-zero entries in $\mathbf{\Omega}$ are randomly selected and set to zero with certain rescaling. After the Factor Analysis, we rescale the data so that the covariance matrix of latent utilities is standardized, as discussed in Section 3.3. The final parameter value distributions are shown in Figure 4 (A). Note the partial correlation coefficients $(-\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}})$ are mostly within $[-0.5, 0.5]$ and some are close to 0, namely moderate to small

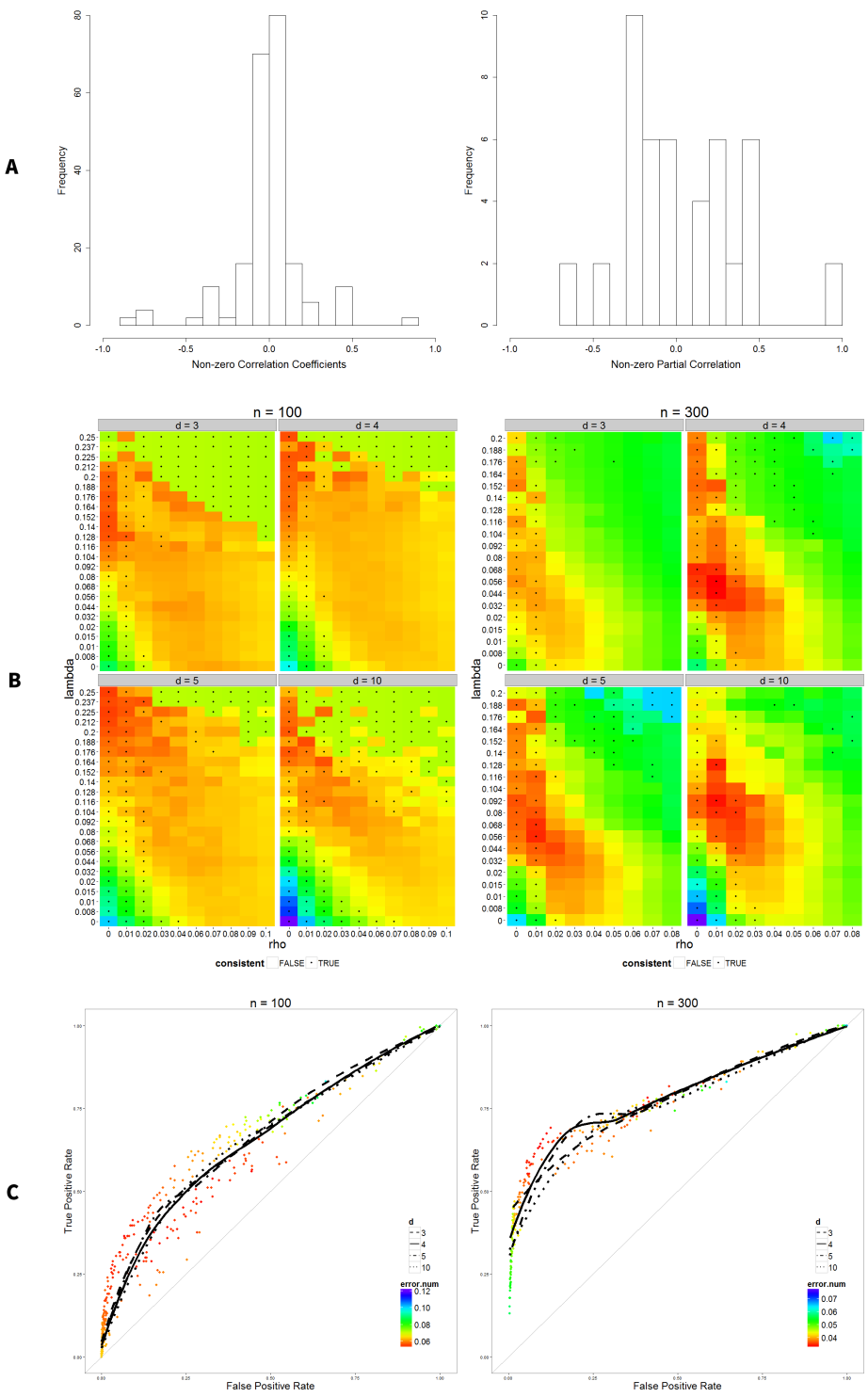

Figure 4: (Caption continues on the next page).

Figure 4: (continued from the previous page) Setting and results of the MISFAN dataset. (A) Histograms of the true non-zero correlation and partial correlation coefficients. Percentage of zeros is 87.9% and 95.8%, respectively. (B) Error heatmap of RMSE of $\mathbf{\Sigma}$, stratified by n and d (set in estimation) with grid of λ vs ρ. Black dot indicates the consistency of 90% of nodes. (C) ROC curve for detecting edges in $\mathbf{\Omega}$. Each point is a result from a consistent (d, λ, ρ) pair, averaged on 10 replicates, and the LOWESS smoothing curve for each d is drawn separately. Note (B) and (C) of the same n share the same color scale as shown in the two distinct colorbars.

correlations, so this dataset is by design pretty noisy. The network structure of this dataset is similar to row (B) of Figure 3.

We generated two datasets, with $n = 100$ and $n = 300$ respectively. MISFAN is fitted with the grid of $d \otimes \lambda \otimes \rho$, with $d \in \{3, 4, 5, 10\}$, $\lambda \in [0, 0.25]$ and $\rho \in [0, 0.1]$. Empirically we found that setting maximum iteration number as 20 is sufficient in most cases, and Monte Carlo sampling of size 50 is more than enough to reach a stable estimate of the truncated normal expectation. After fitting all the models, we test if the proportion of consistently connected nodes, namely it has an edge in sense of both $\mathbf{L}$ and $\mathbf{\Omega}$, is over 90%. Since the rotation-free feature of $\mathbf{L}$ will cause instability in the assessment of parameter fit, we only look at the root mean squared error (RMSE) of the covariance matrix $\mathbf{\Sigma}$ (estimation result of μ is more than satisfactory, not shown here). The error heatmap of $\mathbf{\Sigma}$ and Receiver Operation Characteristics (ROC) curves of detecting non-zero entries in $\mathbf{\Omega}$ are shown in Figure 4 (B) and (C). Note that here we have two free parameters which have overlapping influence on the estimation, so monotonicity of false positive versus true positive in common ROC plot does not hold for MISFAN, and we use Locally Weighted Scatterplot Smoothing (LOWESS) curves for visual resemblance.

From Figure 4 (B), it can be seen that the result is robust against different choice of d, since sparser $\mathbf{L}$ sometimes leads to all 0 in one column and automatically reduces d. This can also be observed in Figure 4 (C), where different curves of d basically overlap. The consistent combination, shown with a black dot, can always cover the optimal cases and thus using the consistency criterion to filter the grid is beneficial and can greatly reduce the computation intensity. The well shaped minimum valley of RMSE, especially with the large sample size case, indicates the stable convex behavior of parameter error of MISFAN. Asymptotic consistency is suggested by the reducing error (see the scale of the two colorbars) and higher ROC curves. The color correspondence between parameter error and ROC curve position gives some heuristics on choice of

FPR and TPR combination, to lower parameter error. This result of MISFAN on the noisy dataset is promising, and we then carry out the comparison with other methods in different model settings.

4.2 Comparison on Bayesian Networks

Here we compare the performance of MISFAN, MBN and MGM. The model assumptions for the three methods are in fact different, so we designed a series of simulated toy Bayesian networks for both cases, as a fair competition. MISFAN assumes a latent Gaussian Network with Probit link, while MBN and MGM assume a conditional Gaussian Network. We build 4 continuous, 2 binary and 1 three-level variables generated by 6 Bayesian Networks shown in Figure 5. The latent Gaussian Network in (A) is generated step-by-step: first generate roots from normal distributions, and simply plug them into the linear equations indicated by edges to get other nodes. Then categorical variables are computed by the Probit link. The conditional Gaussian Network in (B) works similarly, with categorical roots sampled from a multinomial distribution. If parents and child are both discrete, child is sampled from a joint multinomial distribution given the level taken by parents; if child is continuous, then it follows a conditional Gaussian model with linear transformation of its continuous parents while coefficients are determined by each level-combination of its discrete parents. For each model, we design 3 structures to represent the typical cases that practitioners might encounter in real life, as already shown in Figure 3 with some larger networks. The 3 structures are coded by different combination of solid, dashed or dotted edges, as described in the caption. For convenience, we later call these true models from (A1), (A2), ... , (B3), where $1 \sim 3$ denotes (1) only solid edges exist, (2) solid and dashed edges exist, and (3) all edges exist *except* those connected with Z_6, Z_7, X_{67} or X_8. For model parameters, we set:

- Continuous nodes have mean 0; latent utilities have mean ± 0.5; root nodes have standard deviation 10; linear equations have coefficients ± 1 or ± 0.5, and are corrupted by Gaussian noise with standard deviation of 1 or 2.

- Discrete root nodes in conditional Gaussian Network take levels with probability randomly set at $p = 0.3$, 0.4 or 0.5, and the corresponding $1 - p$; the mean of each Gaussian linear equation are ± 10 or 0, and standard deviation 1 or 2, depending on the value of discrete parents; two discrete parents have no interaction on their child.

The three models being compared here do not give the same kind of output, and we only demonstrate their ability of learning conditional independence structure. The simulated Bayesian Networks are translated into undirected graphs, so is the result from MBN, and we evaluate the performance by plotting the ROC curve of edge detection. Unlike the ROC in the previous subsection, here the edge is between variables, and we do not consider latent utility any more since it's not comparable. Therefore utility nodes of a categorical variable in a MISFAN are combined into one group with an edge denoting at least one edge to this nodes group. Sample size $n = 100$ with 10 replicates is used in this numeric experiment.

In MISFAN we set $d = 3$ ($d = 4$ or 5 gives indistinguishable results, not shown), $\rho \in [0, 0.5]$ with 8 cutting points, and λ is set as $\{0, 0.01\}$ for network (A1) (A2) (B1) (B2) and $\{0.05, 0.1\}$ for (A3) (B3). Fewer λ improves the monotonicity of ROC curve, and as a result a smoothing line is no longer necessary. For MGM, we set tuning parameter $\lambda' \in [0, 1.5]$ with 14 cutting points. This interval is computed as suggested in their paper. MBN does not have a simple parameter to control the network density, so for simplicity and fairness we do not get in to the trouble of setting all possible prior distributions and initial network structures, but use the default non-informative setting as presented in the example section of their paper. The search algorithm of MBN is heuristic and only guarantees local minima, so we perturb the initial null network by switching the existence and direction of edges for $k \in \{30, 100, 1000\}$ times and for each k we have 10 replicates. The results are then averaged and presented by only one point per setting as the default, and connected with $(0, 0)$ and $(1, 1)$ only to show visual resemblance with the other two methods for intuitive reference. See results in Figure 6.

From the plot we can infer that: for (1) the simple isolated structure case, MISFAN ($\text{AUC}_{A1} = 0.88$, $\text{AUC}_{B1} = 0.87$) greatly outperforms the other two if it gets the right model, and only slightly lags behind on their home court; MGM ($\text{AUC}_{A1} = 0.69$, $\text{AUC}_{B1} = 0.95$) and MBN ($\text{AUC}_{A1} = 0.67$, $\text{AUC}_{B1} = 0.83$, note that this is an under-estimate with only one point) have comparable results. For (2) the more complex connected structure covering all points, MGM ($\text{AUC}_{A2} = 0.69$, $\text{AUC}_{B2} = 0.87$) stands out by breaking even with MISFAN ($\text{AUC}_{A2} = 0.69$, $\text{AUC}_{B2} = 0.72$) on latent Gaussian model, and gets even better result for the conditional Gaussian; MBN ($\text{AUC}_{A2} = 0.58$, $\text{AUC}_{B2} = 0.56$) has poorest performance for this setting. Finally, for (3) the densely connected but isolated subgraphs, MISFAN ($\text{AUC}_{A3} = 0.96$, $\text{AUC}_{B3} = 0.96$) gets nearly perfect result, MBN ($\text{AUC}_{A3} = 0.72$, $\text{AUC}_{B3} = 0.94$) has similar performance only for its own model, while MGM ($\text{AUC}_{A3} = 0.74$, $\text{AUC}_{B3} = 0.69$) fails on both comparatively. From these observation we can

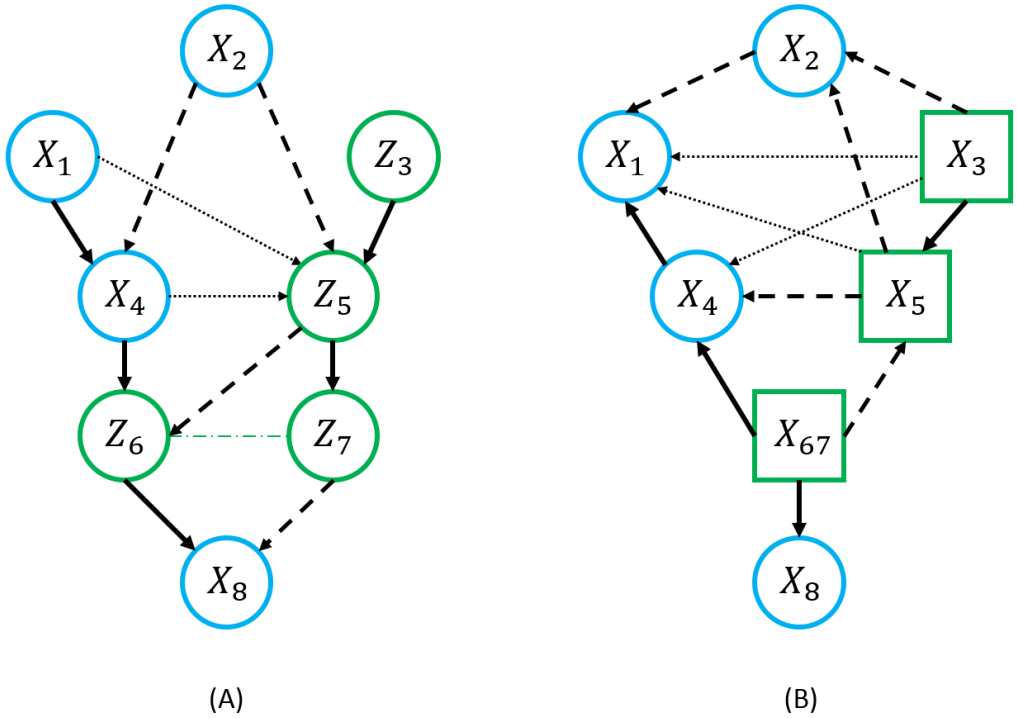

Figure 5: The Second Simulation Setting. (A) Latent Gaussian Network with 4 continuous variables (X_i in blue) and 4 latent utilities (Z_i in green). Z_6 and Z_7 form a three-level variable, others binary. Each edge denotes a Gaussian linear equation. (B) Conditional Gaussian Network with 4 continuous variables and 3 categorical variables (as squares). Edges are changed so that discrete nodes do not have continuous parents, and each edge denotes either a conditional Gaussian linear equation or a joint multinomial distribution, depending on nodes continuous or discrete. Each graph implies 3 networks: (1) only solid edges exist, (2) solid and dashed edges exist, and (3) all edges exist *except* those connected with Z_6, Z_7, X_{67} or X_8. These are 3 typical examples corresponding to those in Figure 3 B, C and A respectively, to provide representability for general applications. Notably, the 3 networks in (A) is designed to share the same conditional dependence structure with the corrsponding ones in (B), namely they are I-equivalent.

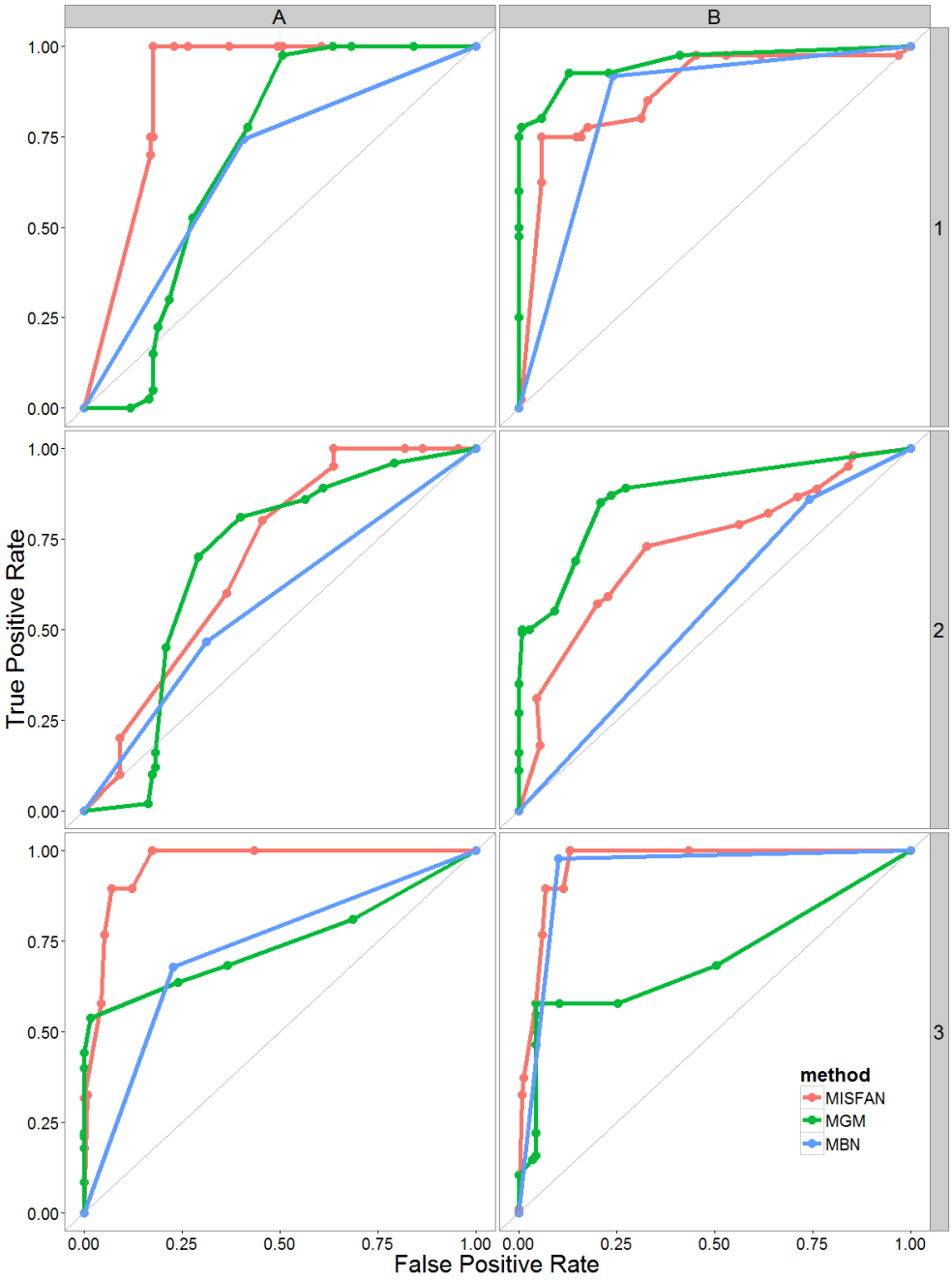

Figure 6: ROC curve for edge detection, stratified by latent Gaussian Network (A, assumed by MISFAN) or conditional Gaussian Network (B, assumed by MGM and MBN), and the three structures described in Figure 5. Each point on the curves is an average of 10 replicates using the same parameter setting, except that MBN has only one averaged point plus $(0, 0)$ and $(1, 1)$ and the broken line is just for reference.

Method	Model Assumption	Free Parameters	Estimation	Resulting Network
MISFAN	Latent Gaussian with Probit link	Effectively 2	MC-ECME	Factor Network
MGM	Conditional Gaussian	1	Proximal Gradient	Markov Random Field
MBN	Conditional Gaussian	Prior/Structure	Bayesian	Bayesian Network

Table 1: Model Comparison of MISFAN, MGM and MBN. The Factor Network, as described in Section 3, contains a full Markov Random Field and partial causal indication of a Bayesian Network.

summarize:

- MISFAN can learn with high accuracy on isolated subgraphs wherever locally simple or complex, thanks to its flexible two-parameter setting, and also it is robust on model violations in these cases
- MGM excels on connected graph with many edges evenly spread, instead of densely clustered
- MBN has good performance only if model assumption is satisfied and the network has not too many edges

Besides the numeric performance, their model difference is also summarized in Table 1. With different domain knowledge and expertise, these network models can be chosen accordingly.

5 World Development and Value Data

Every year The World Bank collects and compiles a series of World Development Indicators across different nations and regions. The dataset describes the global development from various perspectives, including education, environment, economy, health, infrastructure etc. In 2014, 112 indicators across 214 countries and regions are available. Another comprehensive dataset is from the World Values Survey Association, an international network of social scientists, who periodically gather data for over 300 survey questions and have

received over 400,000 of response from subjects across nations since 1981. The questions are mainly related to the belief, value and motivation of the subjects, including topics as political view, religion, gender values, life satisfaction etc. From 2010 to 2014, about one thousand responses were received from each of the 60 country and regions and aggregated into the so-called Wave 6 data. Both datasets contain numerous variables, continuous or discrete, and limited sample size on the country level.

The World Development Indicators data have been well-analyzed mainly for the interest of interaction among political, educational and social-economical development, while the World Value Survey is often cited in cross-cultural organization and comparison studies. Combining the two datasets, it is intriguing to investigate whether people's belief and opinion, often subjective and abstract, is directly correlated with the quantifiable concrete development status of their region of residence, controlling all other observed confounding factors. For instance, can the dominance of a certain religion/political ideology/convention statistically explain (or be explained by) the current business environment of that nation? Here we apply the MISFAN model on the data to explore the underlying association.

The 2014 World Development Indicators and 2010-2014 World Value Survey dataset can be downloaded from their websites (*data.worldbank.org* and *www.worldvaluessurvey.org*). Assuming life values were stable during the 4-year interval, the two datasets are then combined, and samples with missing values are either removed or imputed, depending on the proportion of missing entries. The survey data is summarized into country-level, and a couple of selected questions are included in the analysis for their representability. Finally we have a cleaned dataset, consisting of 55 countries and 48 variables. 42 of the variables are numeric (including ordinal discrete variables), and 6 of them are categorical with in total 25 levels, then for MISFAN there are altogether 61 continuous or latent/dummy variables thus a $p > n$ case. By empirically setting $d = 9, \lambda = 0.001, \rho = 0.33$, the resulting conditional dependence network structure is shown in Figure 7.

The network successfully illustrates the complex association structure among the variables, some of them can be validated by common sense, and some are worth noting. The cluster on the right mainly includes natural and geographical factors. The human population directly affects the number of endangered mammal, bird and fish species, but not (or less) the plants. Labor tax is negatively correlated with the area of the region, and this might be due to the confounding effect of overall population availability of workforce. Tax rate then connects to the industrial administrative efficiency, indicated by the time and procedures to register a business. The *hub* contains two important factors,

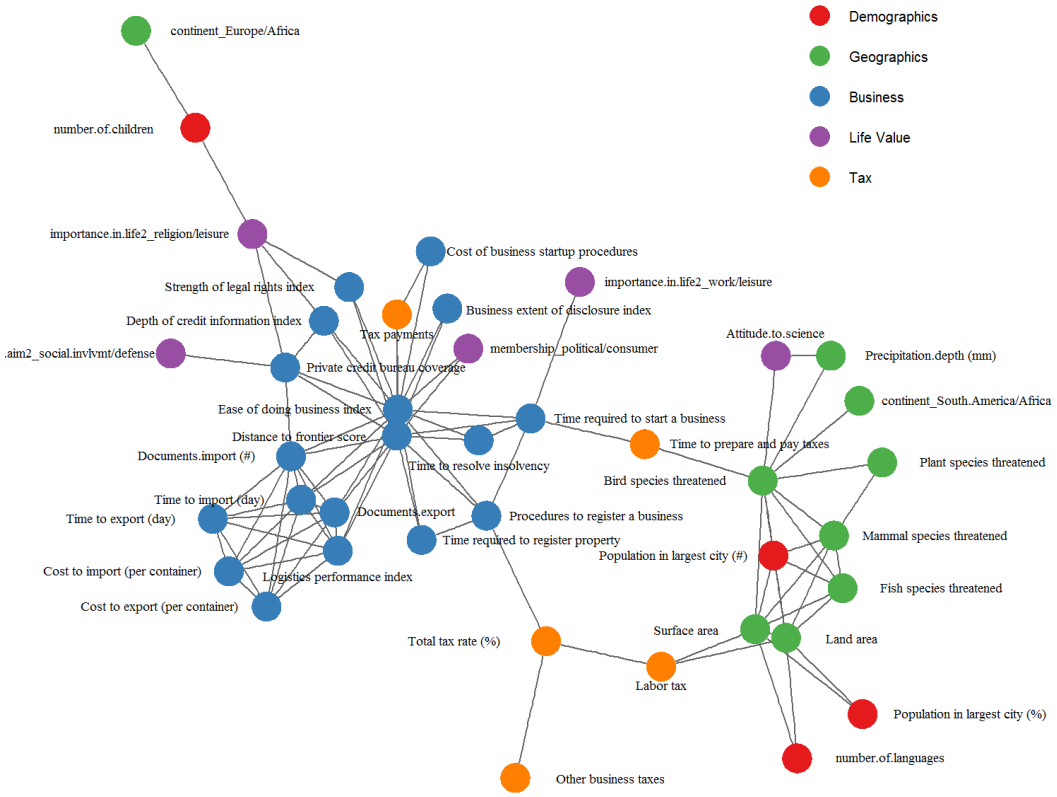

Figure 7: Conditional dependence network of the variables. Nodes are colored by artificial categories to show the variable clustering effect. Each node is either a continuous variable or a latent utility, and an edge indicates conditional dependence given all other variables. Single isolated nodes are not shown. See text for variable meanings and analysis.

one is the overall ease of doing business, which is basically the summary of most business factors. And similarly, the other one is the distance to the best (frontier) performance of business related indicators. The densely connected cluster, actually a clique, next to the hub is related to import, export and logistics. On the far left, the number of children is connected with the location of the region at the level of continent. The Europe/Africa notation means this node is the latent utility of the *Europe* level of the categorical variable *continent*, with *Africa* as the reference level. Apparently the difference of birthrate on these two continents is the most outstanding.

The interaction of beliefs and values (in purple) with other development factors are of particular interest. Positive attitude to Science leads to less threatened bird species. The *importance.in.life.2* indicator is about the second popular 'most importance thing in life' that subjects chose (all regions have *family* as the first choice, except *religion* in Libya). *Work* over *leisure* leads to reduction of time to prepare for a start-up company, and *religion* over *leisure* gives more births and correlates with higher prevalence of credit system and better legal rights. The *membership* variable gives the most prevalent kind of communities that subjects actively join in, and *political* over *consumer* is negatively correlated with ease of doing business and positively associated with the distance to best business performance, and thus have large impact on the whole business indicators group. Finally the second most important aim of a country or state (the first choice is unanimously to develop better *economy*) to be promoting people to feel *socially involved* in the community, over better military *defense* ability, is negatively correlated with percentage of adults that has *credit* history record in a private institution, and it might be due to the confounding of preference of institutionalization for certain cultural groups.

This exploratory analysis of the joint dataset demonstrated the power of MISFAN model, on detecting conditional dependence for a large group of mixed variables with limited sample size. The network is built mainly for illustration and the parameters are set with preference on readability. Smaller values of ρ could reveal the weaker association, and larger λ gives disconnected subgraphs. For this dataset, a slightly larger λ would isolate the tax group, the right geographic cluster and left major cluster, which eliminates some possible spurious edges, e.g. the edge between the number of days to prepare tax report and the number of endangered bird species.

6 Discussion

We presented a novel approach of learning the sparse structure of a Probabilistic Graphical Model with a latent factor network, to accommodate both continuous and categorical variables, named MIxed Sparse FActor Network (MISFAN). Its ability to take advantage of both Bayesian Networks and Markov Random Field is demonstrated in both theory and visualization. An ECME algorithm to fit the model is proposed and validated by simulation. Another simulation experiment compared the performance to two state-of-the-art methods, and MISFAN shows superior or comparable result in different cases. Real life data with small sample size and high dimension, from The World Bank and World Value Survey, is analyzed as an example to illustrate the use of MISFAN in practice.

Future work of this framework could be extensive. MISFAN has a comparatively low computational performance, mainly due to the Monte Carlo sampling from a multivariate truncated normal distribution in each EM step. Bayesian Estimation has been thriving in the area of Probit models, with examples like Imai and Dyk (2005) and Talhouk, Doucet, and Murphy (2012), and usually considered to have faster computation with appropriate chosen priors. Ordinal categorical variable in this paper is treated as a continuous one, as shown in Section 5, but could be extend to one binned latent Gaussian variable with ease and achieve possibly better performance. Another immediate direction is the extension to non-linear cases. Factor analysis with the kernel trick, as discussed in Guo-En and Pei-Ji (2009) and the incorporation of the similar Independent Component Analysis could be some quick candidates. In plenty of application, including the datasets used in Section 5, the network could change with time. Dynamic Bayesian Network, e.g. Hidden Markov Model, has been well developed to describe such a system. In the work of Song, Kolar, and Xing (2009) and Tucker and Liu (2008), network structure could be changing smoothly and the model is essentially a linear dynamic model, or a Vector Autoregression (VAR) with time-varying coefficients. At the same time, dynamic factor models have long been popular in finance and economics, and MISFAN is compatible with such frameworks and could be extended to model multivariate time series.

Part II
A Unified Geometry of Error-in-Variable Models

A Unified Geometric Perspective of Error-in-Variabless Models

Ruofeng Wen

04/06/2015

Abstract

Error-in-Variabless (EIV) models consider the intrinsic error noise in the independent variables of a regression model. Measurement errors exist in virtually all observed variables - be it independent or dependent. However, although proven to be inconsistent and inefficient, simple linear model estimated by the ordinary least square still dominates for its simple computation and interpretation, while the EIV models seem to be daunting and confusing due to its diverse formulations. In this text, we intend to give a clear, systematic and unified description with novel geometric insight for the common EIV models in a non-parametric way: all these models are equivalent in terms of minimizing the sum of a general squared metric between the sample points and the proposed model. This perspective is then naturally generalized to higher dimensions for multiple regression. In light of this geometric perspective, model caveats, parameter specification and alternative estimation approaches are discussed for both theoretical and practical interests. Finally, the performance of EIV models are numerically compared with simulation datasets.

Keywords. Deming Regression, Error-in-Variables, Measurement error, Total Least Square

1 Introduction

In real-life applications, most measured variables are inevitably subject to a certain degree of error. Dealing with such error has became a fundamental

component of nearly every applied research areas. However, due to the lack of communications across disciplines, we often have multiple formulations of the same problem developed in different areas, causing confusions. Discussion can be found in Carroll and Ruppert (1996) and Ludbrook (2010). Meanwhile, linear regression with ordinary least square (OLS) estimation is still widely adapted when we have errors in independent variables, even if it would yield biased and inefficient estimates of the model parameters, as validated and compared in Linnet (1993).

Error-in-Variables (EIV) models can date back to the 1800's, when Adcock (1878) developed the Orthogonal Regression (OR) by minimizing the sum of the squared orthogonal distances from points to the regression line. The more general univariate EIV model has since then been well discussed:

$$\begin{aligned} y_i &= \eta_i + \epsilon_i \\ x_i &= \xi_i + \delta_i \\ \eta_i &= \alpha + \beta\xi_i \end{aligned} \tag{1.1}$$

where (x_i, y_i), $i = 1, \cdots, n$, are the ovserved independent sample points, with (ξ_i, η_i) being their corresponding true values, and (δ_i, ϵ_i) the measurement errors. Furthermore if (ξ_i, η_i) are treated as random variables, we have the structure EIV models, otherwise the functional EIV models.

It is well known that the slope parameter β estimated in (1.1) is between the OLS of y on x and the OLS of x on y. Based on this observation, another model, commonly known as Geometric Mean Regression (GMR), simply takes the geometric mean of the two OLS estimates of the slope parameter.

As OR and GMR were criticized for their implicit, and often unjustified assumptions, the majority of the literature on EIV has taken on the maximum likelihood approach (MLE), by assuming that δ and ϵ in (1.1) are normally distributed. The model is unfortunately not identifiable for this 2-dimensional case, and needs further information supplied. Despite several alternatives, Lindley (1947) argued that assuming the error ratio $\lambda = \sigma_\epsilon^2/\sigma_\delta^2$ as known would be the most convenient. This formulation is also known as Deming's Regression, since Deming (1931) mentioned the same specification of λ, without assuming normal distribution.

It turned out that by setting λ to different values, we could obtain estimates of OR, GMR, OLS(y on x) and OLS(x on y) as special cases, when $\lambda = 1, S_{yy}/S_{xx}, \infty, 0$ respectively, where $S_{ab} = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})$. Although a seemly good property, it as well reveals the suffer of choosing and justifying λ from all the possible values, if no reliable prior information can be provided.

An independent development to deal with the same problem emerged in the field of computational mathematics, known as the Total Lease Square (TLS) (Markovsky and Van Huffel (2007)). TLS is in nature a low rank matrix approximation problem, and has later been proved to be equivalent to OR, and also a special case of Principal Component Analysis (PCA). Unlike the previous examples, TLS is intrinsically multidimensional, and can accommodate heteroscedasticity and autocorrelation. If λ (or the weight matrix in multivariate case, elaborated in later sections) is known, the dataset could be normalized accordingly so that TLS or PCA can be applied directly.

The statistical multivariate form of (1.1) is in fact the well-known Exploratory Factor Analysis (EFA), if we restrict the model to be structural with normal distributions. With some restriction on the number of variables, the model becomes identifiable and thus the suffering of choosing λ is no longer necessary in certain multivariate cases.

In this text, we illustrated that all the methods above, as diverse and irrelevant as it seems in their original formulation, can all be considered as a form of minimizing the sum of area of a series of ellipses/ellipsoids, or a certain squared geometric weighted distance between the data points and the regression line/hyperplane. This systematic non-parametric perspective is by nature multidimensional, can easily accommodate non-linear relationships, bridging the existing gap of Deming's regression (i.e. univariate EIV MLE), TLS, EFA, design with replicates and other general models, and thus revealing the underlying essence of EIV models for better comprehension. We also discuss the means of choosing or by-passing λ or weight matrix, the popular alternative Method of Moments estimation with no requirement for λ, and other guidelines for practical use.

In Section 2, we describe the details of the mentioned methods. Their association is revealed in Section 3 by first investigating two intuitive approaches that are similar, but not equivalent, to Deming's regression, and then we propose our geometric framework inspired by these two step stones. Some practical model specification and extension are also discussed and summarized. The paper ends with a simulation study in Section 4 and a conclusion in Section 5.

2 Background

In this section we elaborate some common EIV models. These methods, either commonly recognized as EIV models or not, are organized in a way that their intrinsic equivalence could be conveniently unveiled.

2.1 Orthogonal Regression, Total Least Square and Geometric Mean Regression

We first introduce methods that were not originally designed or treated as EIV models, but later turned out to be a special case. The well-known model of OLS of y on x is:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{2.1}$$

and we minimize a loss function to obtain estimates of parameters:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2 \tag{2.2}$$

We will focus on slope β, since estimating the intercept is usually trivial and avoidable if data are centered. The minimal loss in (2.2) can also be interpreted as minimizing the sum of the squared vertical distance from the data point (x_i, y_i) to the estimated point on the model line $(x_i, \alpha + \beta x_i)$. The estimated slope is simply:

$$\hat{\beta}_{ls(x)} = \frac{S_{xy}}{S_{xx}} \tag{2.3}$$

where $S_{ab} = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})$.

Orthogonal regression instead minimizes the sum of the squared perpendicular distance from the data point to the line:

$$\min_{\beta} \frac{1}{1+\beta^2} \sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2 \tag{2.4}$$

Note the use of the equation to calculate distance between a point (x_i, y_i) and a line $\alpha + \beta x - y = 0$ in elementary geometry. Take derivative of (2.4) regarding to β, the minima is given by:

$$\hat{\beta}_{or} = \frac{S_{yy} - S_{xx} + \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}}}{2S_{xy}} \tag{2.5}$$

Total Least Square gives exactly the same solution from another aspect. With the centered data matrix (or vector, if the number of independent variables $p = 1$) $\mathbf{X}_{n\times p} = (x_1, \ldots, x_n)^\intercal$ where x_i is now a p-vector and $\mathbf{y}_{n\times 1} = (y_1, \ldots, y_n)^\intercal$, Markovsky, Kukush, and Huffel (2006) described multiple TLS problem as:

$$\min_{\beta,\hat{\mathbf{X}},\hat{\mathbf{y}}} |||[\mathbf{X}-\hat{\mathbf{X}}, \mathbf{y}-\hat{\mathbf{y}}]||_F \quad s.t. \quad \hat{\mathbf{y}} = \hat{\mathbf{X}}\beta \tag{2.6}$$

where $||\cdot||_F$ is the Frobenius norm of a matrix. Suppose $[\mathbf{X}, \mathbf{y}]$ is of full rank $k = \min(p+1, n)$, then (2.6) is to find a rank-deficient matrix $[\hat{\mathbf{X}}, \hat{\mathbf{y}}]$, i.e. a low rank matrix, to best approximate $[\mathbf{X}, \mathbf{y}]$ in the sense of Frobenius norm. According to Eckart-Young-Mirsky theorem, the best rank $k-1$ matrix of such approximation is by setting the smallest singular value to zero, in the sense of the following singular value decomposition (SVD):

$$[\mathbf{X}, \mathbf{y}] = \mathbf{U}\mathbf{S}\mathbf{V}^\intercal \tag{2.7}$$

Here $\mathbf{S} = diag(s_1, \ldots, s_k)$ is the matrix with singular values on the diagonal, in descending order. The low rank matrix estimate could be $[\hat{\mathbf{X}}, \hat{\mathbf{y}}] = \mathbf{U}diag(s_1, \ldots, s_{k-1}, 0)\mathbf{V}^\intercal$, and the corresponding slope estimate is

$$\hat{\beta}_{tls} = -\frac{1}{v_{kk}}(v_{1k}, \ldots, v_{(k-1)k})^\intercal \tag{2.8}$$

where $v_k = (v_{1k}, \ldots, v_{kk})^\intercal$ is the right singular vector with the least singular value. In fact, with some simple algebra, (2.8) would degenerate to (2.5) when $p = 1$, which proves the equivalence between TLS and OR.

One important observation is that TLS, like OR, is symmetric to all variables, which means the linear equation do not distinguish predictors and responses, and the use of $\mathbf{y}$ and $\mathbf{X}$ here is just to keep the notation consistent.

Another quick catch is that (2.7) (2.8) is actually extracting the kth principal component (PC) of the augmented data matrix, and the low rank approximation can be seen as dimension reduction. In the 2-dimensional case, the second PC is the normal vector, i.e. the coefficients of the line equation, of the first PC direction, which is apparently the *natural* line to represent the dataset. Another argument is, Principal Component Analysis (PCA) manages to get rid of dimensions of the least variation that are also orthogonal to the remaining, and thus is doing the same work as the geometric definition of OR. Thus TLS and OR are both special cases of PCA. See Figure 8 for illustration.

As Figure 8 shows, the regression line of OR is always across the center of the data, and between the lines of OLS on y and OLS on x. Geometric Mean Regression (GMR) is proposed based on such observation as well, and literally takes the geometric mean of the one slope and the reciprocal of the other. By using (2.3) and $\hat{\beta}^{-1}_{ls(y)} = \frac{S_{xy}}{S_{yy}}$, we have

$$\hat{\beta}_{gm} = sign(S_{xy})\sqrt{\frac{S_{yy}}{S_{xx}}} \tag{2.9}$$

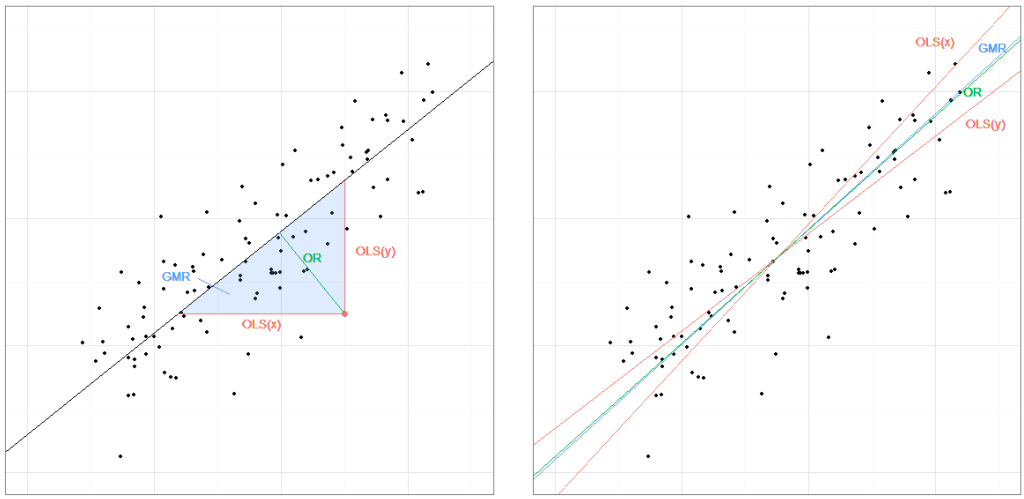

Figure 8: Left: the geometric squared distance/area to be minimized. OLS: vertical or horizontal distance, OR: orthogonal distance, GMR: area of the triangle. Right: estimated regression lines for each approach.

Like OR, the estimate of GMR is also symmetric of both variables. However two things are different: first it does not naturally generalize to the high dimensional case; but good thing is that it's scale-free, i.e. invariant under linear transformation of original data, while OR is not. Thus GMR is popular in certain applied areas where statistical analysis is mainly about univariate regression models. Barker, Soh, and Evans (1988) derived another view of GMR: it is actually a *least-triangle* approach, i.e. minimizing the sum of area of the right triangles surrounded by the regression line, a vertical line across a data point and a horizontal line across the data point.

As we will see in the next subsection, GMR, along with direct use of OR and TLS, has some implicit assumption which is usually unnoticed while also hard to justify.

2.2 Error-in-Variables Models and Maximum Likelihood Estimation

In this subsection we describe the formal Error-in-Variables model for both univariate and multivariate cases.

2.2.1 Univariate: Deming's Regression

We write down the parametric EIV model again as the same form in (1.1)

$$\begin{aligned} y_i &= \eta_i + \epsilon_i \\ x_i &= \xi_i + \delta_i \\ \eta_i &= \alpha + \beta\xi_i \end{aligned} \tag{2.10}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2), \delta \sim N(0, \sigma_\delta^2)$, with i.i.d. samples, $\xi = \{\xi_1, \ldots, \xi_n\}$ in a function model, or $\xi \sim N(\mu, \sigma^2)$ in a structural model. We take the functional model for convenience, but note the MLE of α and β are the same in both cases. (2.10) is not identifiable, therefore different assumptions or prior knowledge is necessary for learning the parameters. Here we follow Deming's Regression approach, by assuming $\lambda = \sigma_\epsilon^2/\sigma_\delta^2$ is known. Then we can estimate the parameters by solving $\max L(\alpha, \beta, \sigma_\delta^2, \xi|\lambda)$, or

$$\max \frac{1}{(2\pi)^n} \frac{\lambda^{-n/2}}{\sigma_\delta^{2n}} \exp[-\sum_{i=1}^{n} \frac{(x_i - \xi_i)^2 + (y_i - \alpha - \beta\xi_i)^2/\lambda}{2\sigma_\delta^2}] \tag{2.11}$$

Take log-transform and we can find MLE of ξ_i first: $\hat{\xi}_i = \frac{\lambda x_i + \beta(y_i - \alpha)}{\lambda + \beta^2}$. Plug it back and we have

$$\max \frac{1}{(2\pi)^n} \frac{\lambda^{-n/2}}{\sigma_\delta^{2n}} \exp\{-\frac{1}{2\sigma_\delta^2}[\frac{1}{\lambda + \beta^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2]\} \tag{2.12}$$

It can be shown that MLE of α is simply $\bar{y} - \hat{\beta}\bar{x}$. We then pursue β in the following minimization of the exponential term

$$\min_\beta \frac{1}{\lambda + \beta^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{2.13}$$

And in fact we can reach a closed form solution

$$\hat{\beta}_{mle} = \frac{S_{yy} - \lambda S_{xx} + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}} \tag{2.14}$$

Immediately we observed resemblance between (2.13) and (2.4). The solution of OR can be achieved by setting the error variance ratio to 1. If we define $\lambda^{-1} = 0$ when $\sigma_\delta^2 = 0$, from (2.14) we can also obtain the solution of OLS on x or OLS on y, by plugging in $\lambda^{-1} = 0$ or $\lambda = 0$ respectively. These

two extreme cases are for $\sigma_\delta^2 = 0$ or $\sigma_\epsilon^2 = 0$, when (2.10) degenerate to ordinary linear model, with no error in the independent variable.

For GMR, it is not that straight forward but still achievable. Note in (2.9), the GMR estimator should be consistent so that

$$\hat{\beta}_{gm} = sign(S_{xy})\sqrt{\frac{S_{yy}}{S_{xx}}} \to sign(\rho_{xy})\sqrt{\frac{\mathrm{Var}(y)}{\mathrm{Var}(x)}} = sign(\rho_{xy})\sqrt{\frac{\beta_{gm}^2\mathrm{Var}(\xi)+\sigma_\epsilon^2}{\mathrm{Var}(\xi)+\sigma_\delta^2}} \tag{2.15}$$

when $n \to \infty$. To make (2.15) an identity, $\lambda = \sigma_\epsilon^2/\sigma_\delta^2 = \beta_{gm}^2$ must be true, namely the error variance ratio is the same as the observation variance ratio. For validation, simply plug $\lambda = \beta^2$ into (2.13) and then we get the solution of GMR (2.9) once again.

It has been shown that all the univariate models described in the previous section are special cases of Deming's Regression, namely the EIV Gaussian model with preset error variance ratio λ. Given the geometric significance depicted in Figure 8, we managed to reveal a universal geometric framework with arbitrary λ, multivariate cases, non-linear equations and other general and complicated models.

2.2.2 Multivariate: Factor Analysis

Although usually not recognized as an EIV model, Factor Analysis, also known as Exploratory Factor Analysis (EFA, in contrast to Confirmatory Factor Analysis), is indeed the multivariate extension of structural Deming's Regression.

Consider a centered random p-vector X, EFA assume the following model:

$$X = \mathbf{L}F + \epsilon \tag{2.16}$$

where $\mathbf{L}$ is a $p \times q$ matrix with $q \leq p$, also known as the *loading matrix*, F is a latent random q-vector known as the *factor* and indicates the low dimensional linear structure of the manifest variable X. We further assume that $F \sim N_q(\mathbf{0}, \mathbf{I}_q)$, $\epsilon \sim N_p(\mathbf{0}, \mathbf{\Phi})$ and $\mathbf{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_p)$.

Set $p = 2$, $q = 1$ then (2.16) degenerates to (2.10) with $X = (y_i, x_i)^\intercal$, $\epsilon = (\epsilon_i, \delta_i)^\intercal$, $\mathbf{\Phi} = \mathrm{diag}\{\sigma_\epsilon^2, \sigma_\delta^2\}$, $\mathbf{L} = (1, \beta)^\intercal$,$F = \xi$. Apparently the λ-equivalent parameter in EFA should be $\mathbf{\Lambda} = \mathrm{diag}(\frac{1}{\phi_1}(\phi_2, \ldots, \phi_p))$. p and q can be interpreted as the number of variables and the number of linear relationships for them.

Tipping and Bishop (1999) pointed out the equivalence between PCA and EFA when $\mathbf{\Phi} = \sigma\mathbf{I}_p$, and thus built up a parametric model for PCA called

probabilistic PCA (also known as maximum likelihood PCA in Wentzell and Andrews (1997), another independent work). In this case, the singular value decomposition (SVD) gives the MLE of EFA. Note $\mathbf{\Phi} = \sigma\mathbf{I}_p$ or $\mathbf{\Lambda} = \mathbf{I}_{p-1}$ gives an extended condition of $\lambda = 1$ in OR, and once again the equivalence among PCA, OR and TLS in the multivariate case is shown.

For the general situation when $\mathbf{\Phi} \neq \sigma\mathbf{I}_p$, MLE of (2.16) no longer has an analytical solution as all the methods previously described. Also note $\mathbf{L}$ is identifiable up to a rotation matrix, i.e. $\mathbf{LR}$ is also an MLE for any q-by-q orthogonal matrix $\mathbf{R}$. This is due to the fact that the low dimensional factor F has no practical meaning and only important for its q degrees of freedom, and thus can be arbitrarily rotated. In practice, a rotation is usually used to assign meanings to the factors, by either making them sparse, or satisfy domain-specific indicators. Estimation is usually carried out by an iterative algorithm like Expectation Maximization (EM).

Although (2.16) is the high dimension generalization of (2.10), there are two features essentially different:

1. Unlike Deming's regression, EFA can accommodate distributions other than Gaussian.

2. We usually do not need prior knowledge about $\mathbf{\Lambda}$ if there aren't too many linear restrictions, because $\mathbf{\Phi}$ can be estimated if certain condition is satisfied.

To explain the second point, we first assume $\mathrm{Var}(F) = \mathbf{\Psi}$ is an arbitrary covariance matrix that needs to be estimated. Note $\mathbf{L}$ has pq parameters, $\mathbf{\Psi}$ has $\frac{1}{2}q(q+1)$ free parameters (for its symmetry) and $\mathbf{\Phi}$ has p parameters, then we totally have $pq + \frac{1}{2}q(q+1) + p$ covariance parameters to estimate. But note that we can construct an arbitrary q-by-q matrix $\mathbf{G}$ such that $\tilde{F} = \mathrm{Var}(\mathbf{G}^{-1}F)$ and set the new $\tilde{\mathbf{L}} = \mathbf{L}\mathbf{G}^{-1}$, and this new estimate also fits the data as well as before. This observation means the real number of free covariance parameters we need to estimate is $pq + \frac{1}{2}q(q+1) + p - q^2$, because q^2 parameters cannot be uniquely estimated due to the presence of the arbitrary $\mathbf{G}$. In practice we usually let $\mathrm{Var}(\mathbf{G}^{-1}F) = \mathbf{I}_q$ as specified in (2.16). Next we note the sample covariance matrix is the only data we have, and it provides $\frac{1}{2}p(p+1)$ estimation equations. The number of equations should be no less than the number of free parameters, so the condition for identifiability is

$$(p-q)^2 \geq p + q \tag{2.17}$$

Therefore, for $p = 2$, $q = 1$, the inequality does not hold, so Deming's regression is not identifiable and addition information about λ is needed.

2.2.3 Replicates

Perhaps the most direct solution for the existence of unknown measurement error is to simply estimate it by repeatedly measuring the same (group of) samples, i.e. to have replicates. With this information, we can naturally estimate the error variance first and then remove some free parameters. Instead of this naive two-step estimation, a direct maximum likelihood estimation of functional Deming's regression with replicates was well-discussed in Barnett (1970).

$$\begin{aligned} y_{ij} &= \eta_i + \epsilon_{ij} \\ x_{ij} &= \xi_i + \delta_{ij} \\ \eta_i &= \alpha + \beta\xi_i \end{aligned} \tag{2.18}$$

where $i = 1, \ldots, m$ indicates different observations grouped by underlying true values, and $j = 1, \ldots, n_i$ is the number of replicates in each group. $\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon i})$ and $\delta \sim N(0, \sigma^2_{\delta i})$ allow heterogeneous variance of the error for each group. $\lambda = \sigma^2_{\epsilon i}/\sigma^2_{\delta i}$ is still assumed fixed across all groups, but unknown and estimatable. The log-likelihood function is:

$$\begin{aligned} l(\alpha, \beta, \sigma^2_\delta, \xi, \lambda) &= \text{const} - \sum_{i=1}^{m} n_i \log(\sigma^2_{\delta i}\sqrt{\lambda}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \xi_i)^2\sigma^{-2}_{\delta i} \\ &\quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \alpha - \beta\xi_i)^2\lambda^{-1}\sigma^{-2}_{\delta i} \end{aligned} \tag{2.19}$$

Maximizing (2.19) does not have a close-form solution, but some implicit estimation equations can be obtained for iterative computation.

2.3 Method of Moments

We have showed that all the methods above can be derived from the MLE of Deming's regression or EFA, assuming normal distributions. The Deming's regression and certain cases in EFA are not identifiable and thus extra information, usually in the form of assumption on λ or replicated samples, is needed. Apart from adding replicates, another popular approach is to use Method of

Moments to obtain the distribution-free estimation by making use of higher-order moments of the data. As discussed in Gillard and Iles (2005),we consider using the first and second moments of the structural model (2.10):

$$\begin{aligned} \bar{x} &= \mu \\ \bar{y} &= \alpha + \beta\mu \\ S_{xx} &= \sigma^2 + \sigma_\delta^2 \\ S_{yy} &= \beta^2\sigma^2 + \sigma_\epsilon^2 \\ S_{xy} &= \beta\sigma^2 \end{aligned} \tag{2.20}$$

where S_{ab} is the sample covariance of a and b. In (2.20) there are 6 free parameters and five equations, making it unidentifiable. However it will be a different picture if we utilize higher moments, e.g. the third moments:

$$\begin{aligned} S_{xxx} &= \mu^{(3)} + \mu_\delta^{(3)} \\ S_{xxy} &= \beta\mu^{(3)} \\ S_{xyy} &= \beta^2\mu^{(3)} \\ S_{yyy} &= \beta^3\mu^{(3)} + \mu_\epsilon^{(3)} \end{aligned} \tag{2.21}$$

where $S_{abc} = \sum(a_i - \bar{a})(b_i - \bar{b})(c_i - \bar{c})/n$ and $\mu_a^{(3)}$ is the population third moment of a, namely the non-standardized skewness. Together (2.18) and (2.19) have 9 parameters and 9 equations. To estimate β, we need these equations to be nontrivial, so the assumption $\mu^{(3)} \neq 0$ should hold. Asking for enough skewness for the underlying true variable or factor is not always reasonable, since any symmetric including normal distribution has zero skewness. However, using high moments for estimation remains a good alternative to solve Deming's regression without assumption of λ, when data is not normal, strongly skewed and has large sample size. Note the higher order moments need more samples to accurately estimate, fourth order moments and above should be used with care. Some further discussion can be found in Dagenais and Dagenais (1997).

Sometimes we might have extra instrumental variables at hand, then Generalized Method of Moments (GMM) can be used to achieve an asymptotically efficient estimator. The topic is out of the scope of this paper, and we refer readers to Hall (2005), Erickson and Whited (2002) and Lewbel (1997) for details.

3 Geometric Perspective of the General EIV Model

We have described some common EIV models in the previous section, for both univariate and multivariate cases. Figure 8 gives some intuition about the geometry of EIV models. OLS and OR are minimizing the sum of squared vertical/horizontal and orthogonal Euclidean distance, while GMR is minimizing the sum of area of the triangle. Inspired by the relationship between these regression, and λ in Deming's regression, we might suspect that as λ slides from 0 to ∞, the squared distance that the corresponding regression is minimizing is from the data point to an arbitrary point on the hypotenuse part of that triangle.

This naive speculation turns out to be incorrect, but nevertheless leads to another kind of regression, which will be immediately be introduced. After that we can then uncover the actual general distance that is being minimized and extend it to higher dimensions.

3.1 Slant Regression

We refer to this naive approach as Slant Regression, with respect to Orthogonal Regression. And we show that such model is close to Deming's regression, but not equivalent.

See Figure 9 left. Instead of minimizing any special distance, we draw a general line from a data point to the hypotenuse, with a parameter $\theta = |AD|/|AB|$. Some elementary geometry can help here: $|\vec{CA}| = |y_i - \alpha - \beta x_i|, |\vec{CB}| = |x_i - (y_i - \alpha)/\beta|$, then $\vec{CD} = \theta\vec{CB} + (1-\theta)\vec{CA}$, $|\vec{CD}|^2 = \theta^2|\vec{CB}|^2 + (1-\theta)^2|\vec{CA}|^2$.

Therefore, slant regression is to solve

$$\min_{\beta} \theta^2 \sum_i [x_i - (y_i - \alpha)/\beta]^2 + (1-\theta)^2 \sum_i [y_i - \alpha - \beta x_i]^2 \tag{3.1}$$

Now how do we choose $\theta \in [0, 1]$, or equivalently, how do we choose the position of *true* point D for each sample point? In the functional EIV model, when a true point $(\xi_i, \alpha + \beta\xi_i)$ is given, the sample follows a bivariate normal distribution

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N_2(\begin{pmatrix} \xi_i \\ \alpha + \beta\xi_i \end{pmatrix}, \begin{pmatrix} \sigma_\delta^2 & 0 \\ 0 & \lambda\sigma_\delta^2 \end{pmatrix}) \tag{3.2}$$

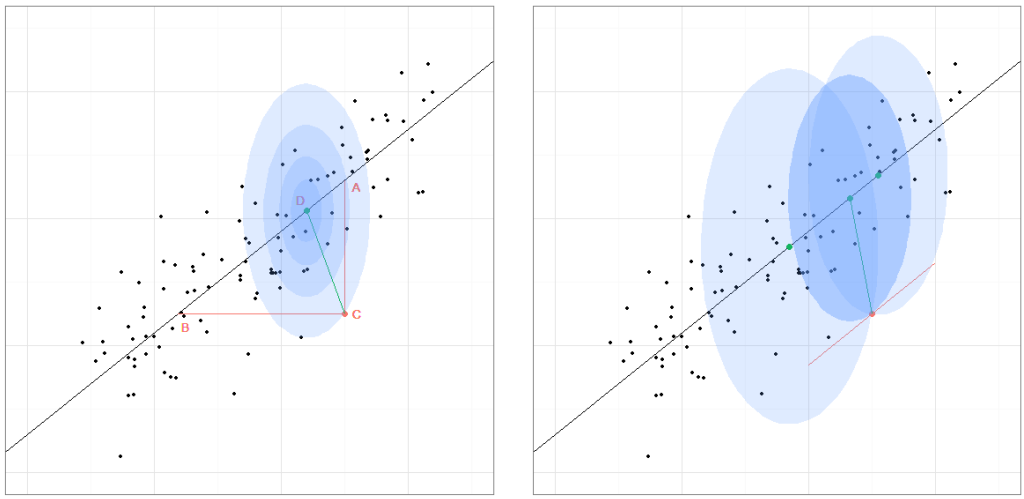

Figure 9: Left: the arbitrary oblique distance $|CD|$ with density contours of error distribution around the true point. Right: the choice of the correct, i.e. smallest, ellipse and thus infer the true point of a corresponding sample point.

So sample points around the true point could be considered as the realizations, with confidence quantified by the joint probability density function, of the bivariate normal.

See Figure 9 left, the probability density contours of the bivariate normal are shown in blue. Note that the contour of a bivariate normal is actually an ellipse $(x-\xi)^2/\sigma_\delta^2 + (y-\alpha-\beta\xi)^2/\lambda\sigma_\delta^2 = c_0$, or

$$(x-\xi)^2 + \frac{(y-\alpha-\beta\xi)^2}{\lambda} = c \tag{3.3}$$

Different values of c generate larger or smaller ellipse contours, while contours with larger c indicate less probability coverage (shown with lighter color). In order to find an appropriate true value for the sample, we use the idea of maximum likelihood, i.e. we try to find the smallest ellipse contour (or the largest probability density value) that covers the sample. See Figure 9 right. It is straight forward to show that the smallest ellipse is the one which has the red line (parallel to the true line) across the sample as its tangent line. Then the center of the ellipse $(\xi_i, \alpha+\beta\xi_i)$ is the true value of the sample.

Therefore, we see $c(\xi) = (x_i-\xi)^2 + (y_i-\alpha-\beta\xi)^2/\lambda$ is minimized when $\xi = \xi_i$. Take $dc(\xi)/d\xi = 0$, we have

$$\lambda(x_i-\xi_i) + \beta(y_i-\alpha-\beta\xi_i) = 0 \tag{3.4}$$

With some simple triangle geometry, we can find the relationship between θ, β and λ

$$\theta = \frac{|AD|}{|AB|} = \frac{|x_i - \xi_i|}{|x_i - \frac{y_i - \alpha}{\beta}|} = \beta \frac{|x_i - \xi_i|}{|\beta(x_i - \xi_i) + \beta\xi_i - y_i - \alpha|} = \frac{\beta^2}{\lambda + \beta^2} \tag{3.5}$$

The last equality comes from (3.4). Substitute back to (3.1), then Slant Regression is equivalent to the following problem:

$$\min_{\beta}[\frac{\lambda^2 + \beta^2}{(\lambda + \beta^2)^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2] \tag{3.6}$$

Note this is different from the MLE (2.13). We can also get estimates of ξ_i from (3.4), and they agree with the MLE solutions, which is natural, since the idea to find true point is exactly an intuitive version of maximum likelihood. A 4-degree polynomial equation can be found by let the derivative of (3.6) to be 0 and $\hat{\beta}_{sr}$ is its real root with the same sign of S_{xy}:

$$\begin{aligned} S_{xy}\beta^4 + ((2\lambda - \lambda^2)S_{xx} - S_{yy})\beta^3 + 3(\lambda^2 - \lambda)S_{xy}\beta^2 & \\ +((\lambda - 2\lambda^2)S_{yy} + \lambda^3 S_{xx})\beta - \lambda^3 S_{xy} &= 0 \end{aligned} \tag{3.7}$$

OLS and OR are still special cases of Slant Regression, by assuming $\lambda = \infty, 0, 1$. However this is not the case for an arbitrary λ. By substituting $\beta_{GMR} = sign(S_{xy})\sqrt{S_{yy}/S_{xx}}$ to the above 4-degree polynomial equation of β and solve the resulting 4-degree polynomial equation of λ, we can find the extreme case when GMR and Slant Regression gives the same answer.

Slant Regression is an alternative to EIV MLE and can easily extend to high dimensional case. In later sections we argue why it is comparatively not desirable.

3.2 Ellipse Area and Mahalanobis Distance

Based on the observation from the previous two subsections, especially the derivation of Slant Regression, we obtained some insights about the geometry of the EIV model.

3.2.1 Minimizing Ellipse Area

Counter-intuitively, Slant Regression is actually not equivalent to MLE, for its main idea is quite straight-forward in the sense of maximum likelihood.

Under the same framework of Slant Regression as shown in Figure 9, we still first locate the true point in the ellipse fashion, but this time instead of minimizing the sum of squared distance between the true and sample points, we minimize the area of that ellipse:

$$\hat{\beta}_{elli} = \arg\min_{\beta} \sum_i S_i \tag{3.8}$$

where S_i is the area of the smallest ellipse (3.3) covering the ith sample point.

The area can be calculated

$$\begin{aligned} S_i &= \pi \times (MajorSemiAxis) \times (MinorSemiAxis) \\ &= \pi \times c_i \times \sqrt{\lambda} c_i \\ &= \sqrt{\lambda}\pi[(x_i - \xi_i)^2 + \frac{(y_i - \alpha - \beta\xi_i)^2}{\lambda}] \\ &= \frac{\sqrt{\lambda}\pi}{\lambda + \beta^2}(y_i - \alpha - \beta x_i)^2 \end{aligned} \tag{3.9}$$

and therefore

$$\hat{\beta}_{elli} = \arg\min_{\beta}[\frac{\sqrt{\lambda}\pi}{\lambda + \beta^2} \sum_i (y_i - \alpha - \beta x_i)^2] \tag{3.10}$$

Note there's only a constant difference between (3.11) and the MLE solution (2.13), so equivalence is shown. Thus the MLE of EIV functional model can be interpreted as minimizing the sum of area of the smallest possible ellipse centered at true values and covering the sample points, with the shape of the ellipse given by λ. This ellipse represents the *error cloud* covering the true underlying data point.

This geometric interpretation is universal to Deming's regression with all possible values of λ. For OLS on x or y, the area of the ellipse is degenerated to the square of Major Semi-axis or Minor Semi-axis of the ellipse. For OR, since $\lambda = 1$, the ellipse degenerates to a circle and the area of the circle is just proportional to the square of its radius. For GMR, we have $\lambda = \beta^2$, $\theta = \beta^2/(\lambda + \beta^2) = 0.5$, then D is the midpoint of AB, also A and B are actually on the ellipse. Simple calculation shows that the area of ellipse is in fact $S_i = \pi S_{\triangle ABC}$, which agree with the triangle minimization interpretation of GMR. See Figure 10 upper row for all these special cases in 2-dimension.

This geometric view can be extended to high dimensions, i.e. EFA, with ease. One change is from error ellipse to error (hyper-)ellipsoid, the shape of

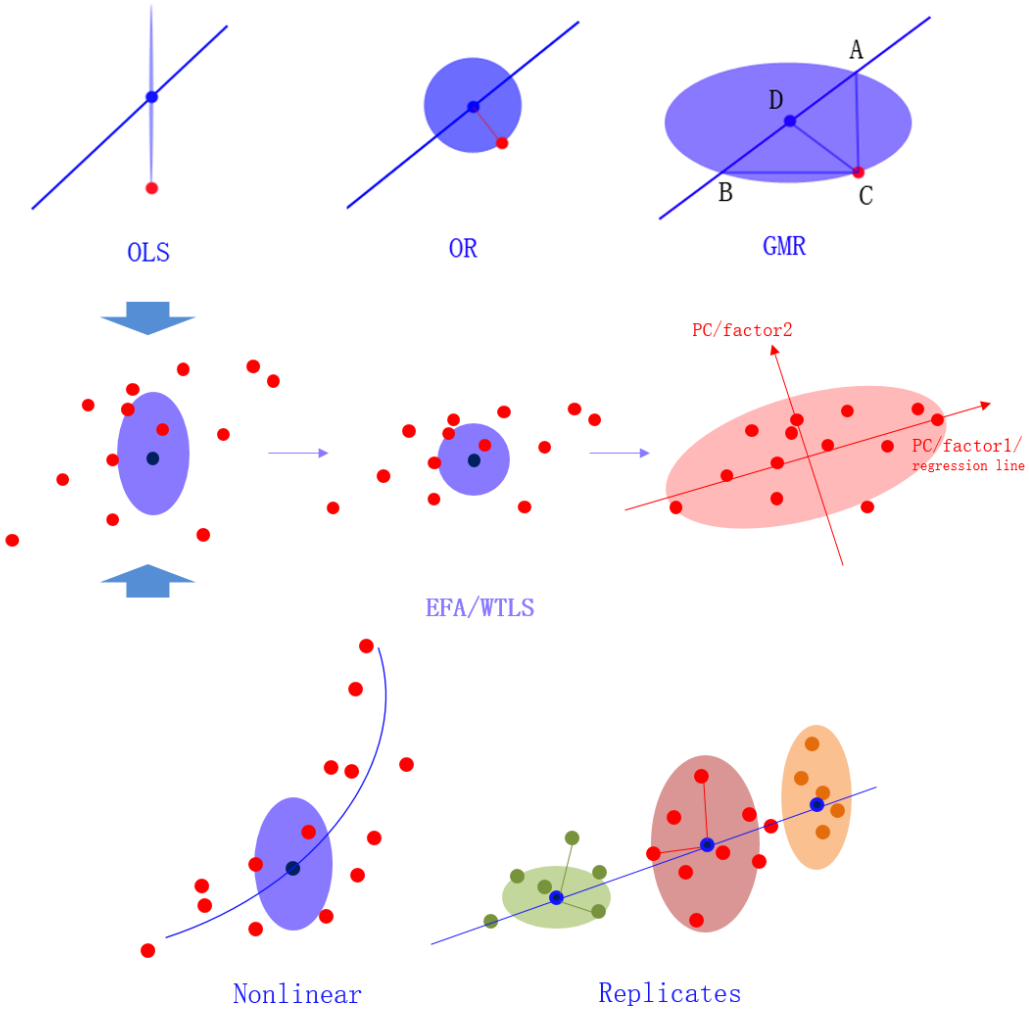

Figure 10: Illustration of ellipse view. Upper: special cases; Middle: for EFA and WTLS, based on the shape of the error ellipse, data points are considered as transformed to a space with even error in each axis, and then the regression line/low dimension structure is extracted, similar to PCA. Lower: nonlinear and replicated model could be explained in the same fashion.

which given by the usually estimatable $\mathbf{\Lambda}$ or Φ. Note the original 1-dimensional line is now extended to a q-dimensional hyperplane, spanned by the factor F. The ellipsoid is the contour of a multivariate normal distribution, but can also be seen as determined by the second moment of the arbitrary error joint distribution, and indicate the *signal-to-noise density* of each variable axis in a non-parametric way. Accordingly, if second moments is not sufficient to describe the error cloud, other complex shapes could be used, corresponding to higher order method of moments. Another observation is that this is also applicable to non-linear EIV, since the non-linear part is only in the low dimensional manifold as opposed to hyperplane, but the measurement error cloud covering it is usually still assumed linear.

The geometric perspective also works for replicated samples. Each group of samples are covered by some concentric small ellipses, shaped by λ and centered at $(\xi_i, \alpha + \beta\xi_i)$. We maximize the overall likelihood, i.e. making the sum of the area of these ellipses as small as possible. See Figure 10 middle and lower for illustration.

Note if $\mathbf{\Lambda}$ is known, we can simply re-normalize the data so that the ellipse becomes a circle, i.e. $\mathbf{\Lambda} \to \mathbf{I}$, and then use OR, TLS or PCA for their simpler computation. When $\mathbf{\Lambda}$ is not known with an identifiable model, some iterative algorithms can be used to reach the optimal solution.

3.2.2 Minimizing Squared Mahalanobis Distance

This ellipse view might at first seem confusing: why area of ellipse? In fact, the area of the ellipse is simply a special squared weighted distance. From (3.1) and (3.8) we have a new formulation of the minimization problem: decomposing the loss function into multiple parts (two in this case) and assigning different weight to each of them. Actually Deming's regression can also be written in such a format, as to minimize:

$$SS_{residual} = \sum_{i=1}^{n}(\frac{\epsilon_i^2}{\sigma_\epsilon^2} + \frac{\delta_i^2}{\sigma_\delta^2}) = \frac{1}{\sigma_\delta^2}\sum_{i=1}^{n}[(x_i - \xi_i)^2 + (y_i - \alpha - \beta\xi_i)^2/\lambda] \qquad (3.11)$$

which agrees with the exponential term of likelihood function (2.11). The right hand side is also proportional to the third line of (3.10), which means the area of the ellipse can be seen as the squared weighted residual/distance multiplied by a constant. To reveal the connection between the algebraic relationship and ellipsoid interpretation, we further investigate the form of multivariate TLS.

TLS has some more general formulation, commonly known as Weighted TLS and Generalized TLS as reviewed in Markovsky and Van Huffel (2007).

Method	Dim	To minimize sum of (squared)	Assumptions
OLS(y)	1	vertical distance	no error in x
OLS(x)	1	horizontal distance	no error in y
OR/TLS	$1/n$	orthogonal distance/Frobenius norm	$\lambda = 1$ or $\mathbf{\Lambda} = \mathbf{I}$
GMR	1	area of the triangle	$\lambda = S_{yy}/S_{xx}$
Slant	$1/n$	Euclidean distance to ellipse center	λ known
Deming	1	area of the ellipse/Mahalanobis dist.	λ known
Replicates	1	area of the concentric ellipses	$\lambda = \sigma^2_{\epsilon i}/\sigma^2_{\delta i}\ \forall i$
Moments	$1/n$	none	Skewed distribu.
WTLS	n	weighted matrix norm/Mahalanobis	$\mathbf{W}$ known
EFA	n	area of the ellipsoid/Mahalanobis	$(p-q)^2 \geq p+q$

Table 2: Summary of EIV models. For all except Slant Regression and Method of Moments, assuming normal distribution leads to MLE.

Consider the multivariate case of WTLS: the data matrices are $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^{n\times q}$, with n sample points as rows. Linear coefficient $\mathbf{B} \in \mathbb{R}^{p\times q}$ is unknown. The WTLS problem is

$$\min_{\mathbf{B},\hat{\mathbf{X}},\hat{\mathbf{Y}}} ||[\mathbf{X}-\hat{\mathbf{X}}, \mathbf{Y}-\hat{\mathbf{Y}}]||_W \quad s.t. \quad \hat{\mathbf{Y}} = \hat{\mathbf{X}}\mathbf{B} \tag{3.12}$$

If $\|\cdot\|_W = \|\cdot\|_F$, Frobenius norm treat every entry in the matrix equally, thus give the natural result of OR/PCA/TLS. WTLS instead applies a weighted matrix norm

$$\|\mathbf{A}\|_W = \sqrt{vec(\mathbf{A})^\intercal \mathbf{W} vec(\mathbf{A})} \tag{3.13}$$

where $vec(\mathbf{A})$ is the vector induced by the matrix column-wisely, and $\mathbf{W} \in \mathbb{R}^{n(p+q)\times n(p+q)}$ is the weight matrix. It is straight-forward to show that $\mathbf{W} = \mathbf{I}$ gives Frobenius norm and thus OR/PCA/TLS, $\mathbf{W} = diag\{\lambda\mathbf{I}_n, \mathbf{I}_n\}$ for $p = q = 1$ gives Deming Regression, $\mathbf{W} = diag\{\mathbf{I}_n, \lambda_1^{-1}\mathbf{I}_n, \ldots \lambda_{p+q-1}^{-1}\mathbf{I}_n\}$ gives multivariate EIV with known Λ. General diagonal $\mathbf{W}$ gives solution for heteroscedasticity EIV, i.e. errors do not have the same variance, and non-diagonal $\mathbf{W}$ assume correlation among errors, e.g. autocorrelation. Estimation of (3.14) needs special care and further details can be found in Markovsky and Van Huffel (2007) and Huffel and Vandewalle (1989). In certain aspect, WTLS is

more general than EFA since it can deal with non-i.i.d. dataset. However, it still assumes the weights are known while EFA can estimate weight if condition satisfied. To make it worse, the computation of the more general cases, like non-diagonal $\mathbf{W}$, is usually extremely ill-conditioned and not computationally practical.

The above notation is based on data matrix, $[\mathbf{X} - \hat{\mathbf{X}}, \mathbf{Y} - \hat{\mathbf{Y}}]$, the residual matrix, represents all the distance from the sample to the true model and $||\cdot||_W$ sums them together with weights. By writing it down sample-wise, we can see WTLS is actually minimizing the sum of distances from a particular sample point to its true value. This distance is known as the Mahalanobis Distance. The Mahalanobis distance between a vector x and a random vector Y with mean μ and covariance matrix $\mathbf{\Sigma}$ is defined as

$$d_M(x, Y) = \sqrt{(x - \mu)^\intercal \mathbf{\Sigma}^{-1} (x - \mu)} \tag{3.14}$$

Simple derivation from (3.13) to (3.14) shows that WTLS is equivalent to minimizing the sum of squared Mahalanobis distance from the samples to the underlying model, by setting $\mathbf{W}$ accordingly regarding to the assumedly known precision matrix $\mathbf{\Sigma}^{-1}$ of the error ellipsoid cloud:

$$\min \sum_i [(x_i, y_i) - (\hat{x}_i, \mathbf{B}\hat{x}_i)]^\intercal \mathbf{\Sigma}^{-1} [(x_i, y_i) - (\hat{x}_i, \mathbf{B}\hat{x}_i)] \tag{3.15}$$

To connect with the geometric view, it is easy to show that squared Mahalanobis distance is proportional to the area/volume of an ellipse/ellipsoid, and can be the exact area if we rescale the covariance matrix. Consider the simple case when $\mathbf{\Sigma}$ is diagonal in (3.15), then we have

$$d_M(x, Y) = \sqrt{\sum_{i=1}^{d} \frac{(x_i - \mu_i)^2}{\sigma_i^2}} \tag{3.16}$$

when $d = 2$. Note the similarity to (3.12). If we rescale $\{\sigma_1^2, \sigma_2^2, \dots\}$ to $\{1, \lambda_1, \dots\}$, they will be the same. When $\mathbf{\Sigma}$ is not diagonal, the ellipse/ellipsoid is oblique, namely the major/minor axis is not parallel to the coordinate system, indicating correlated error in the model.

Mahalanobis distance is well developed in the clustering and classification models. Training points in each cluster can be used to calculate the mean and covariance matrix of the cluster, and distance between unclassified points and each cluster is calculated in the sense of Mahalanobis distance. Fisher's Discriminant Analysis is one classic example.

We come to the conclusion that, in general, all above EIV models could be interpreted as minimizing ellipsoid area or squared Mahalanobis distance. See Table 2 for summary.

4 Simulation

In this section we conduct a numerical experiment on the simple $y = \alpha + \beta x$ model to validate and analyze the superior performance of EIV MLE or Deming's regression over all other univariate methods, when measurement error exists. The simulation settings are listed as below:

- ξ_i is generated from a χ^2_{10} distribution as the underlying factor of the system
- ϵ_i and δ_i are sampled from $N(0, \lambda\sigma_\delta^2)$ and $N(0, \sigma_\delta^2)$ respectively as the noise terms
- $(\alpha, \beta) = (1, 0.8)$
- (y_i, x_i) are then computed according to (1.1)
- $n \in \{100, 1000\}$ for analysis of small and large sample size effect
- $\sigma_\delta \in [1, 5]$ with 8 equal-distance points; note $\mathrm{Var}(\xi_i) = 20$, so noise variance ranges from $5\% \sim 125\%$ of signal variance
- $\lambda \in [1/3, 3]$ with 7 equal-distance points, more extreme λ values along with large σ_δ will have trivially huge estimation error

We build the parameter grid $(n, \sigma_\delta, \lambda)$, and randomly simulate 100 datasets for each combination. Then we apply the univariate regression models: OLS(x), OLS(y), GMR, OR/TLS, Slant Regression, Method of Moments, and MLE (Deming regression). For Slant regression and MLE, we assume λ is known. The 100 regression models give the mean and standard deviation of $\hat{\beta}$, and we summarize the results in Figure 11.

Firstly, as σ_δ increase, the signal-noise ratio decrease and worse estimation is expected. OLS on x is constantly biased and inconsistent for its underestimated slope, while OLS on y is the opposite: inconsistent overestimation. Increasing n does not help, only to reduce variance of the inconsistent estimator. GMR has the most robust and stable behavior to error increase, but nevertheless biased and inconsistent when λ moves to extreme values. OR/TLS has

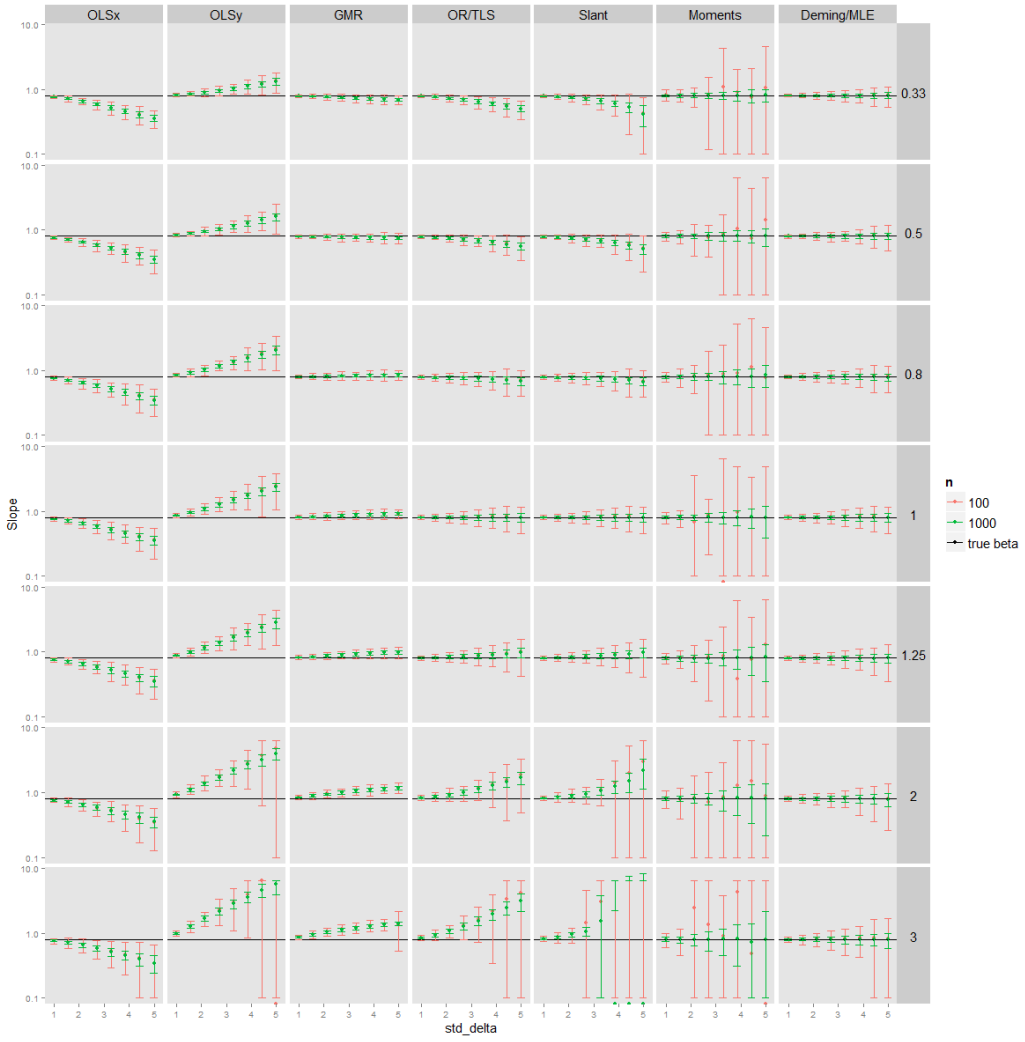

Figure 11: Error-bar plot for the simulation results on the parameter grid. For each subplot, y-axis is for $\hat{\beta}$ in log-scale, with true β as a black horizontal line, x-axis is σ_δ. A point denotes the mean $\hat{\beta}$, with an error-bar showing 2 standard deviation upper and lower bound. Error bar extended to the boundary of the subplot indicates further error but ignored for readability. $n = 100$ and $n = 1000$ are distinguished by red and green color. For the whole grid plot, each column is a result from a specific regression method, and each row is a particular value of λ. See text for analysis.

good performance near $\lambda = 1$ as expected, but will overestimate or underestimate the slope as λ gets larger or smaller, respectively. Slant regression has a similar behavior as OR/TLS but with more variance in the estimator. Method of Moments utilizes third moments to estimate β and achieve an unbiased and consistent estimate, but higher moments consume much more samples in order to be accurately estimated, so method of moment estimator has the largest variance and practically not desirable unless sample size is huge. Finally EIV MLE/Deming's Regression gives an unbiased, consistent and low-variance estimator at a cost of the necessity of a known λ. In practice, such λ can be determined by replicates, regression inefficiencies or cross-validation, if no prior information is present.

5 Summary

We have reviewed a family of Error-in-Variables regression models with seemly diverse formulation, and unified the underlying concepts via a novel and intuitive geometric interpretation. We show both algebraically and geometrically that this family of EIV models is equivalent and extensive to multivariate or nonlinear cases, in order to reduce the confusion and hesitation that practitioners might experience when facing real life data with unignorable measurement error. The superiority of EIV models is then validated in a simple simulation experiment, with helpful notes for application.

References

Adcock, Robert James (1878). "A problem in least squares". In: *The Analyst*, pp. 53–54.

Bach, Francis R and Michael I Jordan (2006). "A Probabilistic Interpretation of Canonical Correlation Analysis Review : probabilistic interpretation of PCA Probabilistic interpretation of CCA Definition of CCA". In: *Dept Statist Univ California Berkeley CA Tech Rep* 688, pp. 1–11. ISSN: 15264998. DOI: 10.1002/ps.2016. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.4942\&rep=rep1\&type=pdf.

Barker, F, YC Soh, and RJ Evans (1988). "Properties of the geometric mean functional relationship". In: *Biometrics*, pp. 279–281.

Barnett, VD (1970). "Fitting straight lines-the linear functional relationship with replicated observations". In: *Applied Statistics* 19.2, pp. 135–144. URL: http://www.jstor.org/stable/2346543.

Bishop, Christopher M (2006). *Pattern Recognition and Machine Learning*. Vol. 4, p. 738. ISBN: 9780387310732. DOI: 10.1117/1.2819119. arXiv: 0-387-31073-8. URL: http://www.library.wisc.edu/selectedtocs/bg0137.pdf.

Bø ttcher, Susanne G (2001). "Learning Bayesian Networks with Mixed Variables". In: *Proceedings of the Eight International Workshop in Artificial Intelligence and Statistics*, pp. 149–156. URL: http://phase.hpcc.jp/mirrors/stat/R/CRAN/doc/packages/deal.pdf.

Carroll, RJ and D Ruppert (1996). "The use and misuse of orthogonal regression in linear errors-in-variables models". In: *The American Statistician*. URL: http://www.tandfonline.com/doi/abs/10.1080/00031305.1996.10473533.

Cheng, Jie, Elizaveta Levina, and Ji Zhu (2013). "High-dimensional Mixed Graphical Models". In: *arXiv preprint arXiv:1304.2810*, p. 21. arXiv: 1304.2810. URL: http://arxiv.org/abs/1304.2810.

Chickering, David Maxwell, David Heckerman, and Christopher Meek (2004). "Large-Sample Learning of Bayesian Networks is NP-Hard". In: *Journal of Machine Learning Research* 5, pp. 1287–1330. ISSN: 1532-4435.

Dagenais, MG and DL Dagenais (1997). "Higher moment estimators for linear regression models with errors in the variables". In: *Journal of Econometrics* 76, pp. 193–221. URL: http://www.sciencedirect.com/science/article/pii/0304407695017895.

Deming, W Edwards (1931). "XI. The application of least squares". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 11.68, pp. 146–158.

Erickson, T and TM Whited (2002). "Two-step GMM estimation of the errors-in-variables model using high-order moments". In: *Econometric Theory*, pp. 776–799. URL: http://journals.cambridge.org/abstract_S0266466602183101.

Fan, Jianqing and Runze Li (2001). "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". In: *Journal of the American Statistical Association* 96, pp. 1348–1360. ISSN: 0162-1459. DOI: 10.1198/016214501753382273.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." In: *Journal of statistical software* 33, pp. 1–22. ISSN: 1548-7660. DOI: 10.1359/JBMR.0301229. arXiv: NIHMS201118. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2929880\&tool=pmcentrez\&rendertype=abstract.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2008). "Sparse inverse covariance estimation with the graphical lasso." In: *Biostatistics (Oxford, England)* 9.3, pp. 432–41. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxm045. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3019769\&tool=pmcentrez\&rendertype=abstract.

Gillard, JW and TC Iles (2005). "Method of moments estimation in linear regression with errors in both variables". In: *Cardiff University School of Mathematics Technical ...* October. URL: http://cf.ac.uk/maths/resources/Iles_Gillard_Tech_Report.pdf.

Guo-En, Xia and Shao Pei-Ji (2009). "Factor analysis algorithm with Mercer Kernel". In: *Proceedings - 2nd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2009*, pp. 202–205. ISBN: 9780769535791. DOI: 10.1109/IITSI.2009.55.

Hall, Alastair R (2005). *Generalized method of moments*. Oxford University Press Oxford.

Hirose, Kei and Michio Yamamoto (2014). "Estimation of an oblique structure via penalized likelihood factor analysis". In: *Computational Statistics and Data Analysis* 79.Kaiser, pp. 120–132. ISSN: 01679473. DOI: 10.1016/j.csda.2014.05.011. arXiv: arXiv:1302.5475v1.

Huffel, SV and J Vandewalle (1989). "Analysis and properties of the generalized total least squares problem AXB when some or all columns in A are subject to error". In: *SIAM Journal on Matrix Analysis and Applications* 10.3, pp. 294–315. URL: http://epubs.siam.org/doi/abs/10.1137/0610023.

Imai, Kosuke and David a. van Dyk (2005). "A Bayesian analysis of the multinomial probit model using marginal data augmentation". In: *Journal of Econometrics* 124.2, pp. 311–334. ISSN: 03044076. DOI: 10.1016/j.jeconom.2004.02.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S0304407604000351.

Koller, Daphne and Nir Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. Vol. 2009, p. 1231. ISBN: 0262013193. DOI: 10.1016/j.ccl.2010.07.006. URL: http://mitpress.mit.edu/catalog/item/default.asp?ttype=2\&tid=11886.

Kschischang, F. R., B. J. Frey, and H. a. Loeliger (2001). "Factor graphs and the sum-product algorithm". In: *IEEE Transactions on Information Theory* 47.2, pp. 498–519. ISSN: 00189448. DOI: 10.1109/18.910572.

Lauritzen, S.L. (1996). *Graphical Models*. Clarendon Press. ISBN: 9780191591228. URL: http://books.google.com/books?id=mGQWkx4guhAC.

Lee, J and T Hastie (2013). "Structure Learning of Mixed Graphical Models". In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. Vol. 31, pp. 388–396. DOI: 10.1080/10618600.2014.900500. arXiv: 1205.5012. URL: http://jmlr.org/proceedings/papers/v31/lee13a.html.

Lewbel, A (1997). "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D". In: *Econometrica: Journal of the Econometric Society* 65.5, pp. 1201–1213. URL: http://www.jstor.org/stable/2171884.

Lindley, DV (1947). "Regression lines and the linear functional relationship". In: *Supplement to the Journal of the Royal Statistical ...* 9.2, pp. 218–244. URL: http://www.jstor.org/stable/2984115.

Linnet, K (1993). "Evaluation of regression procedures for methods comparison studies." In: *Clinical chemistry* 39.3, pp. 424–32. ISSN: 0009-9147. URL: http://www.ncbi.nlm.nih.gov/pubmed/8448852.

Liu, Chuanhai and Db Rubin (1998). "Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data". In: *Statistica Sinica* 8, pp. 729–747. URL: http://www3.stat.sinica.edu.tw/statistica/password.asp?vol=8\&num=3\&art=6.

Loeliger, Hans Andrea (2004). "An introduction to factor graphs". In: *IEEE Signal Processing Magazine* 21, pp. 28–41. ISSN: 10535888. DOI: 10.1109/MSP.2004.1267047.

Ludbrook, John (2010). "Linear regression analysis for comparing two measurers or methods of measurement: but which regression?" In: *Clinical and experimental pharmacology & physiology* 37.7, pp. 692–9. ISSN: 1440-1681.

DOI: 10.1111/j.1440-1681.2010.05376.x. URL: http://www.ncbi.nlm.nih.gov/pubmed/20337658.

Markovsky, I, A Kukush, and S Van Huffel (2006). "On errors-in-variables estimation with unknown noise variance ratio". In: pp. 172–177. URL: http://eprints.ecs.soton.ac.uk/18587/.

Markovsky, Ivan and Sabine Van Huffel (2007). "Overview of total least-squares methods". In: *Signal Processing* 87.10, pp. 2283–2302. ISSN: 01651684. DOI: 10.1016/j.sigpro.2007.04.004. URL: http://linkinghub.elsevier.com/retrieve/pii/S0165168407001405.

Monti, S and GF Cooper (1998). "A multivariate discretization method for learning Bayesian networks from mixed data". In: *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 404–413. URL: http://dl.acm.org/citation.cfm?id=2074142.

Murphy, Kevin P (1998). "Inference and Learning in Hybrid Bayesian Networks". In: *Russell The Journal Of The Bertrand Russell Archives* January. URL: http://www.cs.ubc.ca/~murphyk/Papers/cg.pdf.

Neil, Martin et al. (2008). "Modelling dependable systems using hybrid Bayesian networks". In: *Reliability Engineering & System Safety* 93, pp. 933–939. URL: http://linkinghub.elsevier.com/retrieve/pii/S0951832007001044.

Peng, Jie et al. (2008). "Partial Correlation Estimation by Joint Sparse Regression Models". In: *Journal of the American Statistical Association*, pp. 1–63. ISSN: 0162-1459. DOI: 10.1198/jasa.2009.0126. arXiv: 0811.4463. URL: http://amstat.tandfonline.com/doi/abs/10.1198/jasa.2009.0126http://arxiv.org/abs/0811.4463.

Skrondal, Anders and Sophia Rabe-hesketh (2005). *Structural Equation Modeling : Categorical Variables*. Tech. rep., pp. 1–8. DOI: 10.1002/0470013192.. URL: http://onlinelibrary.wiley.com/doi/10.1002/0470013192.bsa596/full.

Song, Le, Mladen Kolar, and Eric P Xing (2009). "Time-Varying Dynamic Bayesian Networks". In: *Advances in Neural Information Processing Systems*, pp. 1–9. URL: http://books.nips.cc/papers/files/nips22/NIPS2009_0858.pdf.

Talhouk, Aline, Arnaud Doucet, and Kevin Murphy (2012). "Efficient Bayesian Inference for Multivariate Probit Models With Sparse Inverse Correlation Matrices". In: *Journal of Computational and Graphical Statistics* 21, pp. 739–757. ISSN: 1061-8600. DOI: 10.1080/10618600.2012.679239. URL: http://amstat.tandfonline.com/doi/abs/10.1080/10618600.2012.679239$\backslash$nhttp://www.tandfonline.com/doi/abs/10.1080/10618600.2012.679239.

Tipping, Michael E and Christopher M Bishop (1999). "Probabilistic principal component analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 611–622.

Tucker, A and X Liu (2008). "A Bayesian network approach to explaining time series with changing structure". In: *Intelligent Data Analysis* 8, pp. 469–480. URL: http://hdl.handle.net/2438/2903.

Wentzell, PD and DT Andrews (1997). "Maximum likelihood principal component analysis". In: *Journal of ...* 11.October 1996, pp. 339–366. URL: http://www.researchgate.net/publication/251516790_Maximum_likelihood_principal_component_analysis/file/50463529c7a704025d.pdf.

Wilhelm, Stefan and Manjunath B G (2014). *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*. R package version 1.4-9. URL: http://CRAN.R-project.org/package=tmvtnorm.

Zhao, J. H., Philip L H Yu, and Qibao Jiang (2008). "ML estimation for factor analysis: EM or non-EM?" In: *Statistics and Computing* 18, pp. 109–123. ISSN: 09603174. DOI: 10.1007/s11222-007-9042-y.

Zhou, Xingcai and Xinsheng Liu (2007). "The Monte Carlo EM method for estimating multinomial probit latent variable models". In: *Computational Statistics* 23.2, pp. 277–289. ISSN: 0943-4062. DOI: 10.1007/s00180-007-0091-7. URL: http://link.springer.com/10.1007/s00180-007-0091-7.

Zou, Hui (2006). "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American Statistical Association* 101, pp. 1418–1429. ISSN: 0162-1459. DOI: 10.1198/01621450-6000000735.