Bayesian analysis of structural equation models using parameter expansion

Séverine Demeyer, Jean-Louis Foulley, Nicolas Fischer, Gilbert Saporta

1

Bayesian analysis of structural equation models using parameter expansion

Séverine Demeyer

Laboratoire National de Métrologie et d'Essais, LNE, 1 rue Gaston Boissier, Paris, France

Jean-Louis Foulley

INRA-GABI-PSGEN, Department of Animal Genetics, Domaine de Vilvert, 78350 Jouy-en-Josas, France

Nicolas Fischer

Laboratoire National de Métrologie et d'Essais, LNE, 1 rue Gaston Boissier, Paris, France

Gilbert Saporta

Chaire de Statistique Appliquée & CEDRIC, CNAM, 292 rue Saint Martin, Paris, France

CONTENTS

Structural Equation Models with latent variables (SEM) are hypothetical constructs used to represent causality relationships in data, where the observed correlation structure is transferred into the correlation structure of latent variables. In this paper a Bayesian analysis of SEM is proposed using parameter expansion to overcome identifiability issues. An original use of posterior draws from latent variables is proposed to model expert knowledge in uncertainty analysis.

1.1 Introduction

1.1.1 From latent variables to Structural Equation Models (SEM)

This paper relies on the ambivalent nature of latent variables. Their unobserved nature makes them either auxiliary variables used as computational tricks or latent concepts associated with observed variables.

This paper combines the power of these two aspects of latent variables.

Precisely latent auxiliary variables have proven to be efficient computational tools when applied to the EM algorithm, see Dempster, Laird, Rubin [2], and to data augmentation, see Tanner and Wong [10], and even more efficient when implemented in EM with parameter expansion (PX), see Liu, Rubin, Wu [6] and in MCMC with PX-Gibbs sampling, see Liu and Wu [7] and Van Dick [11].

On the other side, using meaningful latent variables is very popular in applied domains like marketing, psychology, sociology, education where interest lies in quantifying unobservable characteristics or aptitudes of individuals like satisfaction, self esteem, alienation, aptitude at school from studies.

Such latent constructs may also involve several meaningful latent variables associated with observed variables, thus focusing the interest to the relationships between these latent variables. The causal relationships between these latent variables are supposed to reflect the structure in the observed variables. Hence the terminology structural equation models (SEM).

In other words SEM are multivariate latent variable models used to represent causal latent structures in the data. The observed (manifest) variables are associated with latent variables in the outer (measurement) model and causality links are assumed between latent variables in the inner (structural) model, see figure 1.1.

1.1.2 Motivation of a Bayesian approach of SEM

The Bayesian approach of this paper has been motivated by our own practice of SEM to take advantage of the information conveyed in structural latent variables, processed outside the SEM. To that respect we are especially interested in the prediction of the structural latent variables.

Bayesian estimation of SEM meets this requirement providing draws from the joint posterior distribution of latent variables that are directly reusable outside SEM.

1.1.3 Motivation of a Parameter Expansion (PX) framework

In this paper PX is used to overcome identifiability issues due to the unobserved nature of latent variables. Identifiability issues of SEM are overcome by setting a scale for the latent variables. This issue has been addressed by Skrondal and Rabe-Hesketh [9] who propose to either scale latent variables in terms of a chosen manifest variable in each block (*anchoring*) or standardize latent variables (*scaling*).

A Bayesian approach of SEM has already been proposed by Lee [5] under *anchoring*, making however the imputation of structural latent variables somewhat tedious.

Using parameter expansion instead allows to easily sample the covariance matrix of latent variables as a correlation matrix (see sections 1.3.1 and 1.3.3) thus overcoming identifiability issues.

1.2 Specification of SEM for Mixed Observed Variables

1.2.1 Measurement (outer) model

Let $\boldsymbol{Y_i}$ the row vector of mixed continuous, binary and ordered categorical observed outcomes for individual i on the p manifest variables, divided into q disjoint blocks indexed by $k = 1 \dots q$ and n_k the number of observed variables within block k. Each block is assumed to reflect a unidimensional concept, summarized into a unique continuous latent variable. Let $\boldsymbol{Z_i}$ the row vector of q continuous latent variables for individual i.

SEM with mixed observed variables is defined within the framework of generalized linear models where binary and ordered categorical observed variables (indexed by $j = 1 \dots n_k$) are modelled as latent responses following Albert and Chib [1], using probit links functions and threshold values.

Let $\boldsymbol{Y_i^*} = \left\{ Y_{ikj}^*,\, k = 1 \dots q,\, j = 1 \dots n_k \right\}$ be the row vector of latent responses defined in expressions 1.3 and 1.4.

The measurement model relates each latent response vector to its associated structural latent variable in a reflexive model (because each observed variable reflects its latent variable) where conditional independence of observed variables is assumed given latent variables.

Using matricial notations the outer model is written for individual i as

$$\boldsymbol{Y_i^*} = \boldsymbol{\mu} + \boldsymbol{Z_i\theta} + \boldsymbol{E_i},\, 1 \leq i \leq n \tag{1.1}$$

where $\boldsymbol{E_i}$ is the measurement error term distributed $\boldsymbol{E_i} \sim \mathcal{N}(0, \boldsymbol{\Sigma_\varepsilon})$ with $\boldsymbol{\Sigma_\varepsilon}$ diagonal and $\boldsymbol{\theta}$ is the $q \times p$ matrix of regression coefficients.

To illustrate the formulas, with $q = 3$, $p = 6$, $n_1 = 2$, $n_2 = 2$ and $n_3 = 3$ (see the graphical model section 1.4.5) $\boldsymbol{\theta}$ is the matrix

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{22} & \theta_{23} & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{34} & \theta_{35} & \theta_{36} \end{pmatrix} \tag{1.2}$$

If $\boldsymbol{Y_{kj}}$ is continuous then it coincides with its quantified version $\boldsymbol{Y_{kj}^*}$.

If $\boldsymbol{Y_{kj}}$ is binary or ordered categorical then $\boldsymbol{Y_{kj}^*}$ is defined in the following univariate probit models.

A probit link for binary variables is used to model the probability of success $p(Y_{ikj} = 1) = \Phi(\mu_{kj} + \theta_{kj} Z_{ik})$.

The univariate probit model for binary outcomes is written

$$\begin{aligned} Y_{ikj} &= 1_{\left\{Y_{ikj}^* \geq 0\right\}} \\ Y_{ikj}^* &\sim \mathrm{N}(\mu_{kj} + \theta_{kj} Z_{ik}, 1) \end{aligned} \tag{1.3}$$

A probit link for ordered categorical variables is used to model the cumulated probabilities $p(Y_{kj} \leq c) = \Phi(\gamma_{kj,c} + \theta_{kj} Z_k)$.

The univariate probit models for ordered categorical outcomes is written

$$\begin{aligned} Y_{ikj} = c &\Longleftrightarrow \gamma_{kj,c-1} < Y_{ikj}^* \leq \gamma_{kj,c} \\ Y_{ikj}^* &\sim \mathrm{N}(\theta_{kj} Z_{ik}, 1) \end{aligned} \tag{1.4}$$

where, to ensure identifiability of thresholds, $\gamma_{kj,0} = -\infty$, $\gamma_{kj,1} = 0$ and $\gamma_{kj,n_{kj}} = \infty$ (n_{kj} the number of categories of question kj).

If $\boldsymbol{Y_i^*}$ and $\boldsymbol{Z_i}$ were observed, the measurement model (1.1) would reduce to a linear regression model.

1.2.2 Structural (inner) model: alternative modelling

Denoting $\boldsymbol{H_i}$ the endogenous latent variables and $\boldsymbol{\Xi_i}$ the exogeneous latent variables, the structural equations are simultaneous equations given by

$$\boldsymbol{H_i} = \boldsymbol{H_i\Pi} + \boldsymbol{\Xi_i\Gamma} + \boldsymbol{\Delta_i} \tag{1.5}$$

where $\boldsymbol{Z_i} = \boldsymbol{H_i\Xi_i}$, $\boldsymbol{\Pi}$ is the $q_1 \times q_1$ matrix of regression coefficients between endogeneous latent variables , Γ is the $q_2 \times q_1$ matrix of regression coefficients between endogeneous and exogeneous latent variables. $\boldsymbol{\Delta_i}$ is the error term distributed $\boldsymbol{\Delta_i} \sim \mathcal{N}(0, \boldsymbol{\Sigma_\delta})$ with $\boldsymbol{\Sigma_\delta}$ diagonal, independent with $\boldsymbol{\Xi_i}$ and $\boldsymbol{\Xi_i}$ is distributed $\mathcal{N}(0, \boldsymbol{\Phi})$.

Since a Bayesian approach allows to work with the joint distribution of latent variables, it is equivalent to work with the correlation matrix of latent variables, under the assumption of multinormality for the conditional distribution of $\boldsymbol{Z_i}$, so that the inner model considered in this paper is given by

$$\boldsymbol{Z_i}|\boldsymbol{R_Z} \sim N(\boldsymbol{0}, \boldsymbol{R_Z}) \tag{1.6}$$

with $\boldsymbol{R_Z}$ a correlation matrix.

In addition ${\boldsymbol{R_Z}}^{-1}$ contains the regression parameters of all possible regressions between latent variables.

1.3 Bayesian Estimation of SEM with Mixed Observed Variables

1.3.1 Implementation of parameter expansion

The implementation of parameter expansion in this paper mimics the implementation of parameter expansion in PX-EM algorithms as defined in Liu,Rubin,Wu [6] and so differs from the usual implementation for MCMC algorithms described in [7].

Parameter expansion [6] consists in working with unidentified parameters in the complete data model $f(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta})$ and indexing expanded latent variables $\boldsymbol{W}$ and expanded data models $p(\boldsymbol{Y}, \boldsymbol{W}|\boldsymbol{\theta}, \boldsymbol{\alpha})$ each corresponding to a value of the expansion parameter $\boldsymbol{\alpha}$, so that the observed likelihood $f(\boldsymbol{Y}|\boldsymbol{\theta})$ is preserved, that is, satisfies

$$f(\boldsymbol{Y}|\boldsymbol{\theta}) = \int f(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{Z} = \int p(\boldsymbol{Y}, \boldsymbol{W}|\boldsymbol{\theta}, \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{W} \tag{1.7}$$

Usually, the transformation indexed by the expansion parameter is a C^1 diffeomorphism (one-to-one mapping).

In this paper, the variances of structural latent variables $\boldsymbol{Z} = \boldsymbol{Z_1}, \ldots, \boldsymbol{Z_q}$ are expansion parameters. Under scaling constraints (that is in the complete data model), recall that $\boldsymbol{Z} \sim N(0, \boldsymbol{R_Z})$ where $\boldsymbol{R_Z}$ is a correlation matrix. Introducing variance parameters $\alpha_1, \ldots, \alpha_q$ and defining $\boldsymbol{\alpha} = diag(\alpha_1, \ldots, \alpha_q)$, creates expanded latent variables $\boldsymbol{W} = \boldsymbol{\alpha}^{\frac{1}{2}}\boldsymbol{Z}$ in the expanded data model indexed by $\boldsymbol{\alpha}$ where $\boldsymbol{W} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_W})$ with $\boldsymbol{\Sigma_W} = \boldsymbol{\alpha}^{\frac{1}{2}}\boldsymbol{R_Z}\boldsymbol{\alpha}^{\frac{1}{2}}$ a covariance matrix.

Identifiability issues are easily overcome in the parameter expansion setting: drawing a correlation matrix in the complete data model only involves sampling a covariance matrix in the expanded data model and applying the reverse transformation to the covariance matrix $\boldsymbol{R_Z} = \boldsymbol{\alpha}^{-\frac{1}{2}}\boldsymbol{\Sigma_W}\boldsymbol{\alpha}^{-\frac{1}{2}}$.

The same applies with residual variances of latent responses, where the expansion parameter α is the residual variance, see Meza, Jaffrézic, Foulley [8].

Parameter expansion involves computation of conditional posterior distributions of original and expansion parameters, where the expansion parameters are computed in the expanded data model and the original parameters are computed in the complete data model.

1.3.2 Imputation of latent variables

Latent responses are computed following Albert and Chib [1] based on models 1.3 and 1.4 for binary and ordered categorical variables respectively.

For a binary observed variables, Y^*_{ikj} is drawn from

$$Y^*_{ikj}|\mu_{kj},\theta_{kj},Z_{ik},Y_{ikj} \sim \mathrm{NT}\left(\mu_{kj}+\theta_{kj}Z_{ik},1;0,\infty\right) \text{ si } Y_{ikj}=1 \tag{1.8}$$

$$Y^*_{ikj}|\mu_{kj},\theta,Z_{ik},Y_{ikj} \sim \mathrm{NT}\left(\mu_{kj}+\theta_{kj}Z_{ik},1;-\infty,0)\right) \text{ si } Y_{ikj}=0 \tag{1.9}$$

where $NT\left(\mu,1;a,b\right)$ stands for the normal distribution $N\left(\mu,1\right)$ left truncated at a and right truncated at b.

For ordered categorical variables, the latent response Y^*_{ikj} is drawn from

$$Y^*_{ikj}|,\theta,Z_{ik},Y_{ikj} \sim \mathrm{NT}\left(\theta_{kj}Z_{ik},1;\gamma_{kj,Y_{ikj}-1},\gamma_{kj,Y_{ikj}}\right) \tag{1.10}$$

Given $\boldsymbol{\Theta}=\{\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\Sigma_\varepsilon},\boldsymbol{R_Z}\}$ the conditional posterior distribution of latent variables is expressed as

$$\left[\boldsymbol{W_i}|\boldsymbol{Y_i^*},\boldsymbol{\Theta}\right] \propto \left[\boldsymbol{Y_i^*}|\boldsymbol{Z_i},\boldsymbol{\Theta}\right]\left[\boldsymbol{Z_i}|\boldsymbol{\Theta}\right] \tag{1.11}$$

$$\propto \left[\boldsymbol{Y_i^*}|\boldsymbol{Z_i},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\Sigma_\varepsilon}\right]\left[\boldsymbol{Z_i}|\boldsymbol{R_Z}\right] \tag{1.12}$$

where $\boldsymbol{Y_i^*}|\boldsymbol{Z_i},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\Sigma_\varepsilon} \sim N\left(\boldsymbol{\mu}+\boldsymbol{\theta Z_i},\boldsymbol{\Sigma_\varepsilon}\right)$ is the likelihood of individual i computed from the measurement model (1.1) and $\boldsymbol{Z_i}|\boldsymbol{R_Z} \sim N\left(\boldsymbol{0},\boldsymbol{R_Z}\right)$ is the joint distribution of latent variables.

Then it can be easily shown that

$$\boldsymbol{W_i}|\boldsymbol{Y_i^*},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\Sigma_\varepsilon},\boldsymbol{R_Z} \sim \mathcal{N}\left(\boldsymbol{D\theta\Sigma_\varepsilon^{-1}}\left(\boldsymbol{Y_i^*}-\boldsymbol{\mu}\right),\boldsymbol{D}\right) \tag{1.13}$$

where $\boldsymbol{D^{-1}}=\boldsymbol{\theta\Sigma_\varepsilon^{-1}\theta^t}+\boldsymbol{R_Z^{-1}}$.

1.3.3 Simulation of the covariance matrix of structural latent variables

The covariance matrix Σ_W of structural latent variables is computed in the expanded data model under the following conjugate prior distribution

$$\Sigma_W \sim \text{Inverse-Wishart}_{\nu_0}\left(\left(\nu_0\boldsymbol{S_0}\right)^{-1}\right) \tag{1.14}$$

where ν_0 is a degree of freedom and $\boldsymbol{S_0}$ is our prior guess on covariance matrix Σ_W. A weakly informative prior is given by $\nu_0=q$ or $q+1$.

The posterior distribution of Σ_W is given by

$$\Sigma_W|\boldsymbol{W} \sim \text{Inverse-Wishart}_{\nu_0+n}\left(\left(\nu_0\boldsymbol{S_0}+\boldsymbol{W^tW}\right)^{-1}\right) \tag{1.15}$$

1.3.4 Conditional posterior distributions in the measurement model

Conditional posterior distributions of regression parameters

The model relating each latent response to its associated latent variable is a linear regression model, whose conjugate prior distribution is factorized as

$$\left[\mu_{kj}, \theta_{kj}, \sigma^2_{kj}\right] = \left[\mu_{kj}, \theta_{kj} | \sigma^2_{kj}\right] \left[\sigma^2_{kj}\right] \tag{1.16}$$

Under the conjugate prior distributions

$$\mu_{kj}, \theta_{kj} | \sigma^2_{kj} \sim N\left(\begin{pmatrix}\mu_{0kj}\\ \theta_{0kj}\end{pmatrix}, \sigma^2_{kj} \boldsymbol{H_0^{-1}}\right) \tag{1.17}$$

$$\sigma^2_{kj} \sim \text{Inverse-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2} s_0^2\right) \tag{1.18}$$

with $\boldsymbol{H_0^{-1}}$ the 2×2 prior covariance matrix of (μ_{kj}, θ_{kj}), and s_0^2 our prior guess for the residual variance σ^2_{kj}. A weakly informative prior is given by $\nu_0 = 1$ or 2.

Computation gives, with $\boldsymbol{X_k} = (\boldsymbol{1}, \boldsymbol{Z_k})$, $\boldsymbol{\beta_{kj}} = (\mu_{kj}, \theta_{kj})$, $\boldsymbol{\beta_{0kj}} = (\mu_{0kj}, \theta_{okj})$

$$\begin{aligned}\boldsymbol{\beta_{kj}} | \sigma^2_{kj}, \boldsymbol{Y^*_{kj}}, \boldsymbol{Z_k} \\ \sim N\left(\left(\boldsymbol{X_k^t X_k} + \boldsymbol{H_0}\right)^{-1}\left(\boldsymbol{X_k^t Y^*_{kj}} + \boldsymbol{H_0 \beta_0}\right), \sigma^2_{kj}\left(\boldsymbol{X_k^t X_k} + \boldsymbol{H_0}\right)^{-1}\right)\end{aligned} \tag{1.19}$$

$$\sigma^2_{kj} | \boldsymbol{Y^*_{kj}}, \boldsymbol{Z^*_k} \sim \text{Inverse-Gamma}\left(\frac{1}{2} + \frac{n}{2}, \frac{1}{2} + \frac{n}{2}\tilde{s}^2_{kj}\right) \tag{1.20}$$

$$n\tilde{s}^2_{kj} = \left(\boldsymbol{Y^*_{kj}} - \boldsymbol{X_k \beta_{kj}}\right)^t \left(\boldsymbol{Y^*_{kj}} - \boldsymbol{X_k \beta_{kj}}\right) + (\boldsymbol{\beta_{kj}} - \boldsymbol{\beta_{0kj}})^t H_0 (\boldsymbol{\beta_{kj}} - \boldsymbol{\beta_{0kj}}) \tag{1.21}$$

Conditional posterior distributions of thresholds (for categorical observed variables)

Thresholds are defined in model 1.4 where threshold $\gamma_{kj,c}$ separates modalities c and $c+1$. Assuming flat prior distributions $[\gamma_{kj,c}] \propto 1$, the posterior distribution of threshold $\gamma_{kj,c}$ for $2 \leq c \leq n_{kj} - 1$ is given by

$$\begin{aligned}\gamma_{kj,c} | Y_{kj}, Y^*_{kj}, \{\gamma_{kj,c'},\ c' \neq c\} \\ \sim \text{Unif}\left(\max\left\{Y^*_{kj} : Y_{kj} = c\right\}, \min\left\{Y^*_{kj} : Y_{kj} = c+1\right\}\right)\end{aligned} \tag{1.22}$$

which retrieves a stochastic EM estimate of threshold values. An alternative proposition would be to assume, as in Foulley and Jaffrézic [3], that the $\Delta_{kj,c} = \gamma_{kj,c} - \gamma_{kj,c-1}$ are uniformly distributed on the range $[0, \delta]$.

1.3.5 PX-Gibbs sampling

PX-Gibbs algorithm for estimating SEM with mixed outcomes involves two PX schemes yielding a three steps algorithm described as follows, whose steps are similar to the homologous steps implemented in PX-EM.

- **Step 1:** PX implementation in the probit models to generate latent responses matching the constraint of residual variance fixed at unity, given structural latent variables and current values of parameters in the complete data model.

– Draw latent responses in the expanded data model
$Y_{ikj}^{*(t+1)} = Y_{ikj}$ if Y_{ikj} is continuous
$Y_{ikj}^{*(t+1)} \sim f\left(Y_{ikj}^{*}|\mu_{kj}^{(t)}, \theta_{kj}^{(t)}, \sigma_{kj}^{2}=1, Z_{ik}^{(t)}, Y_{ikj}\right)$ if Y_{ikj} is binary

$$Y_{ikj}^{*(t+1)} \sim f\left(Y_{ikj}^{*}|\mu_{kj}^{(t)}, \theta_{kj}^{(t)}, \sigma_{kj}^{2}=1, \gamma_{kj,Y_{ikj}}^{(t)}, \gamma_{kj,Y_{ikj}+1}^{(t)}, Z_{ik}^{(t)}, Y_{ikj}\right)$$
if Y_{ikj} is ordered categorical

where f is a generic notation defined in expressions 1.8, 1.9 and 1.10.

– Draw the expansion parameters of the probit models

$$\sigma_{kj}^{2} \sim f\left(\sigma_{kj}^{2}|\mu_{kj}^{(t)}, \theta_{kj}^{(t)}, Y_{ikj}^{*(t+1)}, Z_{ik}^{(t)}\right)$$

where f is defined in expression 1.20.

– Compute latent responses in the complete data model

$$Y_{ikj}^{*(t+1)} \leftarrow Y_{ikj}^{*(t+1)} / \sqrt{\sigma_{kj}^{2}}$$

- **Step 2:** PX implementation in the structural model to generate structural latent variables matching the identifiability constraint of the covariance matrix being actually a correlation matrix, given latent responses with unit variance current values of parameters in the complete data model.

– Draw latent responses in the expanded data model

$$\boldsymbol{W_i^{(t+1)}} \sim f\left(\boldsymbol{W_i}|\boldsymbol{\mu^{(t)}}, \boldsymbol{\theta^{(t)}}, \boldsymbol{\Sigma_\varepsilon = I_q}, \boldsymbol{Y^{*(t)}}, \boldsymbol{R_Z^{(t)}}\right)$$

according to formula 1.13.

– Draw the expansion parameters (correlation matrix of structural LV) according to formula 1.15

$$\boldsymbol{\Sigma_Z} \sim f\left(\boldsymbol{\Sigma_Z}|\boldsymbol{W}\right)$$

– Compute structural LV in the complete data model

$$\boldsymbol{R_Z^{(t+1)}} = \left[\text{diag}\left(\boldsymbol{\Sigma_Z}\right)\right]^{-\frac{1}{2}} \boldsymbol{\Sigma_Z} \left[\text{diag}\left(\boldsymbol{\Sigma_Z}\right)\right]^{-\frac{1}{2}}$$
$$\boldsymbol{Z^{(t+1)}} = \boldsymbol{W^{(t+1)}} \left[\text{diag}\left(\boldsymbol{\Sigma_Z}\right)\right]^{-\frac{1}{2}}$$

- **Step 3:** Computation of the outer parameters from their posterior conditional distribution in the complete data model under both constraints.

$$\mu_{kj}^{(t+1)}, \theta_{kj}^{(t+1)} \sim f\left(\mu_{kj}, \theta_{kj}|\boldsymbol{Y_{kj}^{*(t)}}, \boldsymbol{Z_k^{(t)}}\right)$$
$$\gamma_{kj,c}^{(t+1)} \sim f\left(\gamma_{kj,c}|\gamma_{kj,c-1}^{(t)}, \gamma_{kj,c+1}^{(t)}, \boldsymbol{Y_{kj}^{*}}^{(t)}, \boldsymbol{Y_{kj}}^{(t)}\right)$$

1.4 Application: Modelling Expert Knowledge in Uncertainty Analysis

1.4.1 Context of interlaboratory comparisons

This section shows an original application of SEM for the first time in uncertainty analysis, in the field of interlaboratory comparisons. Interlaboratory comparisons are external quality controls designed to help laboratories improve their measurement process by measuring the same quantity and comparing their results.

Quality indicators are the consensus value of the comparison, which is the estimated value of the quantity computed from the results of the laboratories, its associated uncertainty and measurement bias, which is for a given laboratory the difference between its result and the consensus value.

If measurement bias were computed with respect to the true value of the quantity then monitoring measurement bias over time would be meaningful and trends could be detected. Instead, measurement bias intrinsically depends on results through the consensus value. Hence the need to robustify the consensus value to make it less dependent to observed data.

Current practice to try to overcome this dependence in computing consensus value and its associated uncertainty involve either

- robust algorithms to be less dependent to outliers, called *robust method*,
- computing the consensus value from a subset of expert laboratories, called *expert laboratories*.

The originality of our approach lies in combining advantages of both current approaches into an alternative approach, acting as a post processing of robustified results based on a management of expert knowledge using SEM. This new method is called *robustified consensus value*.

1.4.2 SEM to model prior distribution of measurement bias

Our use of SEM is to model expert knowledge to score laboratories according to the quality of their practice, based on a ranking of categories for each observed variable from the worst to the best practice. Latent variables can be interpreted as components of the overall quality of laboratories and used as prior information on measurement bias with SEM actually modelling the structure of bias.

This broader framework is still Bayesian in that latent variables represent pre existing information independent from measurement results used to update knowledge on all the indicators of the comparison: measurement bias, the consensus value and its associated uncertainty.

1.4.3 Combining SEM with measurement results

Latent variables are transformed into weights w_i reflecting the quality of practice to combine with measurement results. Among others a logistic transform can be applied to the sum s_i of latent variables

$$w_i = \frac{\exp s_i}{1 + \exp s_i} \tag{1.23}$$

A standardized robust algorithm from standard NF ISO 13528 [4] is first applied to

measurement results to treat outliers. Raw results are thus transformed into a winsorized sample. This step is not inconsistent with the approach in that a good laboratory will have a high weight even if it is an outlier with respect to the normal distribution. Besides, standards strongly recommend to treat outliers.

At each iteration of the PX-Gibbs algorithm used to estimate SEM, a weighted mean of the winsorized measurement results x_i^R and its variance are computed from the latent variables through the weights

$$x_p^{(t)} = \sum w_i^{(t)} x_i^R \tag{1.24}$$

$$u^2\left(x_p^{(t)}\right) = \sum w_i^{2(t)} V\left(x_i^R\right) \tag{1.25}$$

where $V\left(x_i^R\right)$ is the winsorized variance of the sample.

According to the ergodicity theorem, MCMC draws consequently yield full posterior distributions of the weighted mean and its variance.

1.4.4 Robustifying the consensus value

The consensus value is modelled in a hierarchical model whose first level is normal centred on the weighted mean with variance being the variance of the weighted mean. Two other levels represent the sampling variabilities of the weighted mean and its variance from the Markov Chains released when estimating SEM.

The marginal posterior distribution of the consensus value is computed from Monte Carlo draws according to the following hierarchical model

$$x_c \sim N\left(\mu_w, \sigma_w^2\right) \tag{1.26}$$

$$\mu_w \sim N\left(\mu_{w_0}, \sigma_{w_0}^2\right) \tag{1.27}$$

$$\sigma_w^2 \sim \text{Inverse-Gamma}\left(\frac{\alpha_0}{2}, \frac{\alpha_0}{2} S_0\right) \tag{1.28}$$

where μ_{w_0} and $\sigma_{w_0}^2$ are identified from the posterior distribution of the weighted mean, and α_0 and S_0 are identified from the posterior distribution of the variance of the weighted mean.

Since the first level is integrated out the variance parameter, the marginal posterior distribution of the consensus value is a Student distribution.

1.4.5 Results

The method was applied to the measurement of concentrations of water pollutants.

This analysis was performed on a small number of laboratories (18) regularly involved in this comparison and willing to take part to this study. The auxiliary information was investigated by a questionnaire designed by selected experts from universities and environmental laboratories.

The SEM resulting from expert processing is represented in figure 1.1 where posterior distributions have been computed under the following prior distributions $\mu_{kj}, \theta_{kj} | \sigma_{kj}^2 \sim N\left(\begin{pmatrix} 0 \\ 0.5 \end{pmatrix}, \sigma_{kj}^2 I_2\right)$,

$\sigma_{kj}^2 \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$, $\Sigma_W \sim \text{Inverse-Wishart}_3\left((I_3)\right)^{-1}$ with I_2 and I_3 the 2×2 and 3×3 identity matrices respectively.

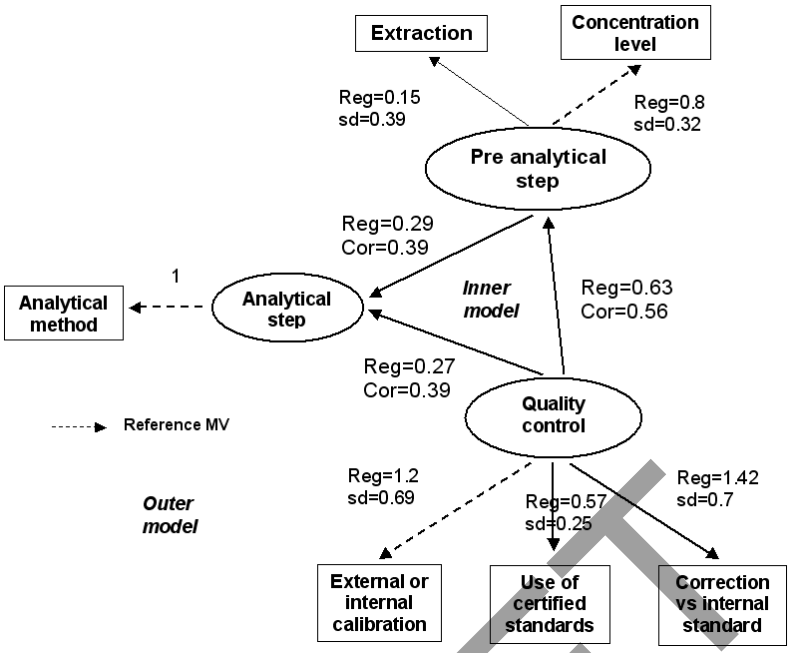

FIGURE 1.1
Estimates of SEM used to model expert knowledge in water pollutants field. *Reg* is the estimated regression coefficient, *sd* is the associated standard deviation and *cor* is the correlation coefficient. The dotted arrows represent reference variables defining the sign of latent variables.

TABLE 1.1 Results (x_i) and robustified results (x_i^R)

x_i	31.0	50.0	32.6	57.5	50.0	36.0	40.0	20.0
x_i^R	31.0	40.6	32.6	40.6	40.6	36.0	40.0	27.1

36.0	28.1	34.5	42.4	25.0	33.3	32.0	39.0	30.0	31.0
36.0	28.1	34.5	40.6	27.1	33.3	32.0	39.0	30.0	31.0

Results and robustified results are given in the table below

The results from the three competing methods are compared with the reference value, called *reference* and its uncertainty provided by LNE in terms of *consensus value*, the *associated uncertainty* and the 95% *confidence interval* and summarized in table 1.2.

Interpretation of results:

Results for the three methods used to compute the consensus value show consistency between them (intervals overlap) and with the reference value because the four central estimates belong to all the confidence intervals.

Due to the small number of laboratories, estimates of SEM are of poor quality, with relatively high standard deviations, so that this application cannot be used in the general purpose of testing relationships between *pre analytical step*, *analytical step* and *quality control*.

The implementation of the methodology seems all the same very promising and points out the benefits of additional information to improve existing methods, among them a

TABLE 1.2
Consensus value, associated uncertainty and confidence intervals.

Method	Cons. Val.	As. Uncert.	95 % Conf. Int.
Reference	33	1.72	$[29.55, 36.45]$
Robust method	34.46	1.64	$[31.18, 37.75]$
Expert labs	34	1.6	$[30.8, 37.2]$
Robustified CV	33.82	1.62	$[30.57, 37.06]$

larger number of laboratories, and results with uncertainties, necessary to quantify sources of measurement bias.

1.5 Conclusion and Perspectives

This work applied to uncertainty analysis provides practitioners with a powerful and flexible statistical tool based on Structural Equation Modelling of expert knowledge on measurement bias to improve the treatment of interlaboratory comparison data.

The new method relies on current robust standardized methods to propose a fully Bayesian modelling of interlaboratory comparisons data involving a Bayesian estimation of SEM.

The complete Bayesian framework allows to easily handle missing or censored data as well as a hierarchical structure in the results (e.g. measurements by country) and provides a rigorous framework for model comparison and validation.

The benefits of such a statistical approach are long term and the approach should accompany laboratories in the process of improving their measurements, along with the development of new reference methods by National Metrology Institutes.

Acknowledgements

The authors particularly thank Eric Parent from AgroParisTech for his involvement from the beginning of this project and his useful comments.

The research within this EURAMET joint research project receives funding from the European Communitys Seventh Framework Programme, ERA-NET Plus, under Grant Agreement No. 217257.

Bibliography

[1] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm "with discussion". *Journal of the Royal Statistical Society, Serie B*, 39(1):1–38, 1977.

[3] J. L. Foulley and F. Jaffrézic. Modelling and estimating heterogeneous variances in threshold models for ordinal discrete data via winbugs/openbugs. *Computer Methods and Programs in Biomedicine*, 97(1):19–27, 2010.

[4] ISO/TC69. Méthodes statistiques utilisées dans les essais d'aptitude par comparaisons interlaboratoires. ISO 13528:2005. International Organization for Standardization (ISO), Geneva, Switzerland, 2005.

[5] S. Y. Lee. *Structural Equation Modelling: A Bayesian Approach*. Wiley (Wiley Series in Probability and Statistics), 2007.

[6] C. Liu, D. B. Rubin, and Y. N. Wu. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85:755–770, 1998.

[7] J.S Liu and Y. N. Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, 1999.

[8] C. Meza, F. Jaffrézic, and J. L. Foulley. Estimation in the probit normal model for binary outcomes using the SAEM algorithm. *Computational Statistics and Data Analysis*, 53(4):1350–1360, 2009.

[9] A. Skrondal and S. Rabe-Hesketh. *Generalized latent variable modeling*. Chapman& Hall/CRC, 2004.

[10] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[11] D. A. van Dyk and X. L. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.