332

{CROSS-SECTIONAL / LONGITUDINAL} REDUCED RANK REGRESSION ANALYSIS BY ALTERNATING {LEAST SQUARES / MAXIMUM LIKELIHOOD}

JAN DE LEEUW
DEPARTMENTS OF MATHEMATICS AND PSYCHOLOGY
UCLA

In this talk we discuss a restricted class of path analysis models, and some particular methods of estimating structural parameters together with optimal transformations or quantifications of the variables.

We shall no go into the discussion about the 'causality' of these models. There is nothing inherently causal here, and path models are as 'causal' as principal component analysis or multiple regression (or computing a mean value, for that matter). Causal terminology ('explains', 'determines', 'effect') will be avoided as much as humanly possible. Or, alternatively, defined in purely technical terms.

There is more discussion on this point in De Leeuw (Psychometrika, 1985), De Leeuw (in Dijkstra(ed), 1986), De Leeuw (Statistica Neerlandica, 1988, Psychometrika, 1988), ...

A path diagram is a graph, with the variables in the study as corners and the relationships between the variables as edges.

We shall give a simple way of translating graphs into conditional independence statements, conditional independence statements into equations, equations into loss functions, and loss functions into algorithms.

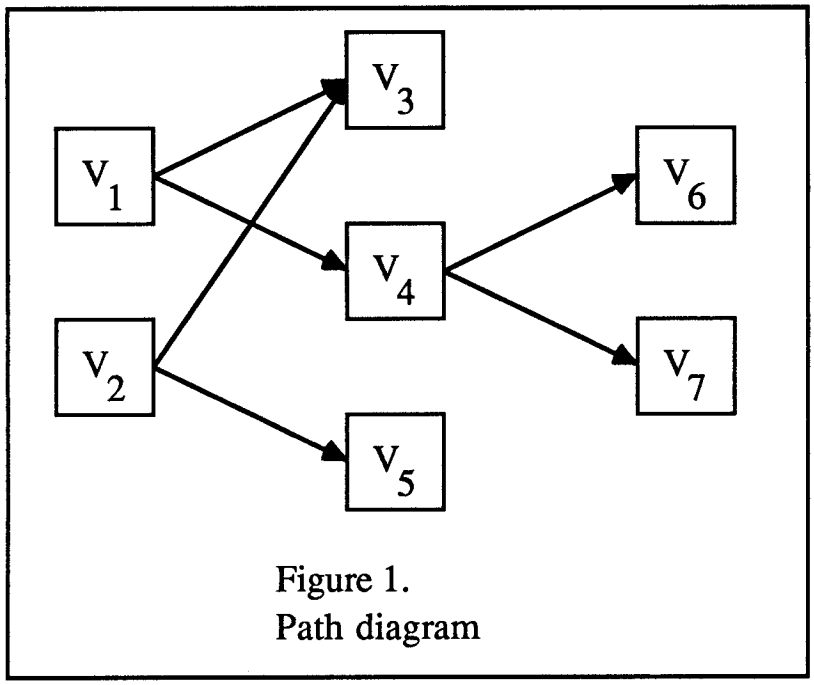


Figure 1.
Path diagram

If there is an arrow from V_1 to V_2 then we say that V_1 is a *direct cause* of V_2 (and V_2 is a *direct effect* of V_1).

If there is an path from V_1 to V_2 then we say that V_1 is a *cause* of V_2 (and V_2 is an effect of V_1).

A model is *transitive* if no variable is a cause of itself.

In transitive models we define a *level assignment* for the variables.

Exogeneous variables (which have no causes) get level 0. The *level* of an endogeneous variable is one larger than the maximum level of its direct causes.

V_1 is a *predecessor* of V_2 (and V_2 a *successor* of V_1) if the level of V_1 is less than that of V_2.

Table 1.
Causal relations in Figure 1

	level	causes	direct causes	predecessors
Var 1	0	* * * *	* * * *	* * * *
Var 2	0	****	****	* * * *
Var 3	1	{1,2}	{1,2}	{1,2}
Var 4	1	{1}	{1}	{1,2}
Var 5	1	{2}	{2}	{1,2}
Var 6	2	{1,4}	{4}	{1,2,3,4,5}
Var 7	2	{1,4}	{4}	{1,2,3,4,5}

Assumptions: the graph is tied to a notion of orthogonality (independence).

1: Given the direct causes, a variable is independent of its other predecessors.

2: Given the predecessors, two variables of the same degree are independent.

Observe: independence can be orthogonality and its can be real (probabilistic) independence. In the case of the normal distribution, the two are the same.

First in informal probabilistic notation (could refer to densities or to discrete distributions)

$$p_{7654321} = p_{76|54321}p_{54321} = p_{76|4}p_{543|21}p_{21} =$$

$$= p_{7|4}p_{6|4}p_{3|21}p_{4|1}p_{5|2}p_{12}.$$

In the purely qualitative case this is a log–linear model, with likelihood

$$\Delta = \Sigma\ n_{7654321} \ln p_{7654321} = \text{etc}$$

In the purely quantitative linear case we translate to

$$v_7 = \beta_{74}v_4 + \varepsilon_7,$$

$$v_6 = \beta_{64}v_4 + \varepsilon_6,$$

$$v_5 = \beta_{52}v_2 + \varepsilon_5,$$

$$v_4 = \beta_{41}v_1 + \varepsilon_1,$$

$$v_3 = \beta_{31}v_1 + \beta_{32}v_2 + \varepsilon_3,$$

with all ε_j orthogonal to v_1 and v_2, and to each other. Here multinormal ML becomes regression.

But this does not exhaust all possibilities.

We can have mixed qualitative and quantitative cases, we can have nonlinear path models with independence or nonnormal linear ones with orthogonality, etc.

We can have that v_7 and v_6 (sinks in the model) are discrete and ordinal, and that $p(v_6 = t|v_4)$ and $p(v_7 = t|v_4)$ follow some choice model or categorical variable model.

We can have v_1 and v_2 categorical and the rest numerical, which makes it possible to use dummies to model $p_{765453|12} = p_{7|4}p_{6|4}p_{3|21}p_{4|1}p_{5|2}$.

We can have binary variables, ordinal variables, etc. in the model, as long as we know how to model $p_{var|direct\ causes}$ in each case.

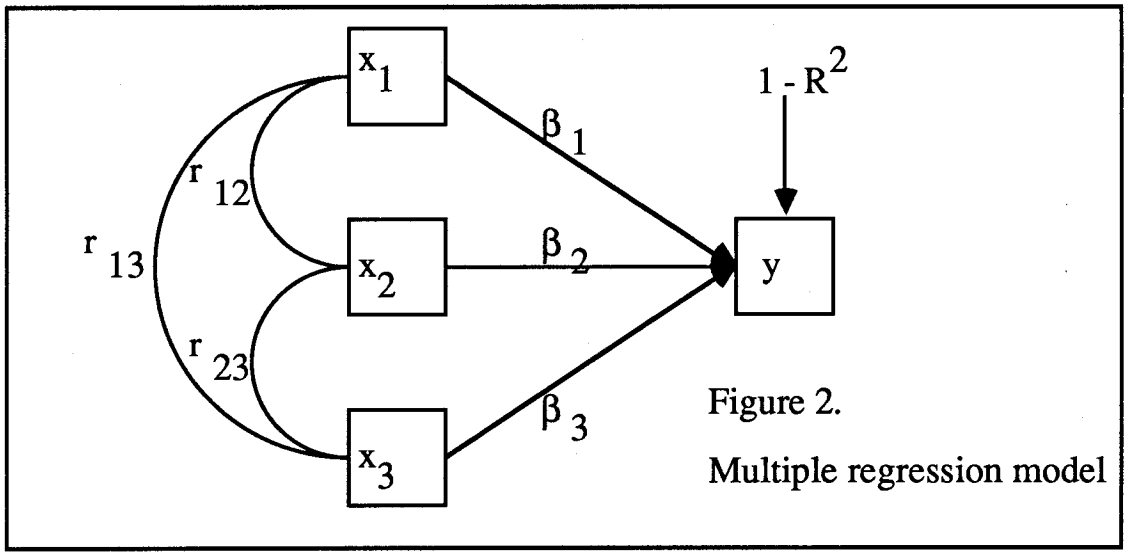


Figure 2.

Multiple regression model

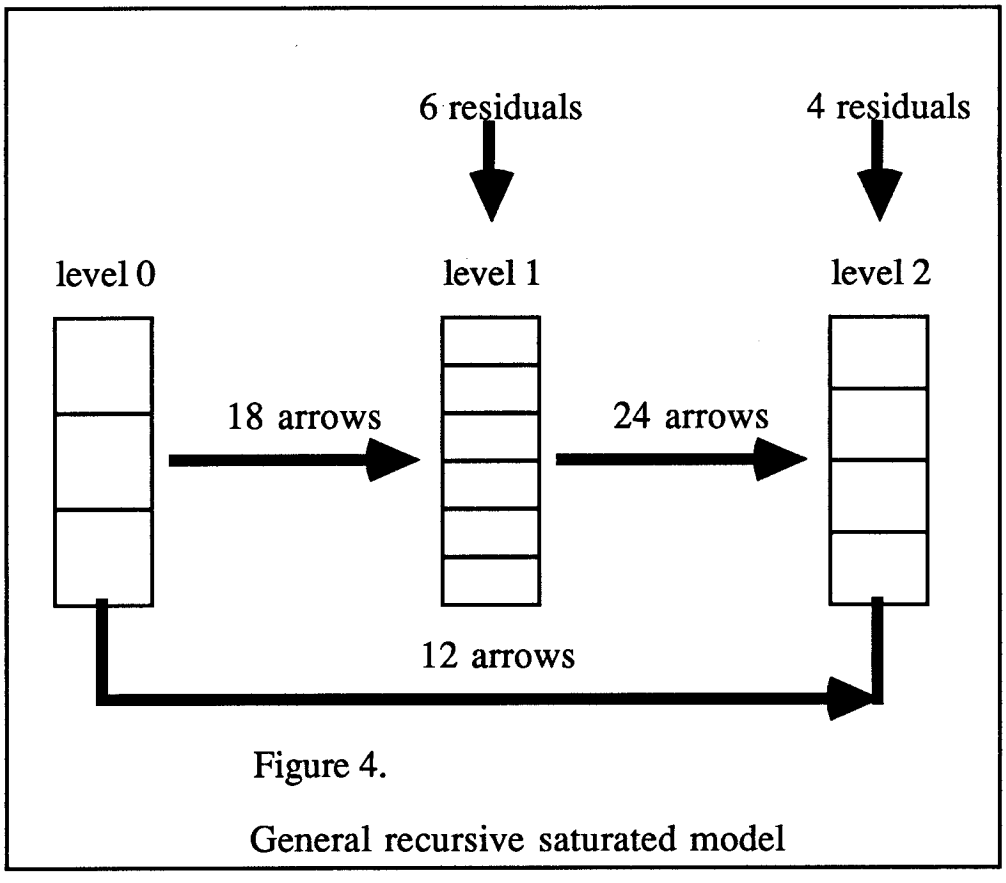


Figure 4.

General recursive saturated model

The whole field of path analysis becomes more interesting if we allow for *latent variables*. Our discussion of latent variables is somewhat nonstandard. We imbed them in the general framework of *measurement levels* or *incomplete information.*

In a particular situation we may have to deal with variables with are numerical, variables with are ordinal, variables which are nominal, and variables which are latent. For numerical variables we now their precise values, for ordinal variables we know the order of their values, nominal variables give information about the equality of certain observations on that variable, and for latent variables we do not even have this. We have nothing, the only thing we 'know' is that there is a variable somewhere, at the place indicated by the path model.

Let us now look at some simple path models with latent variables. The other variables can be anything: ordinal, numerical, or nominal, but not latent. We meet some old friends.

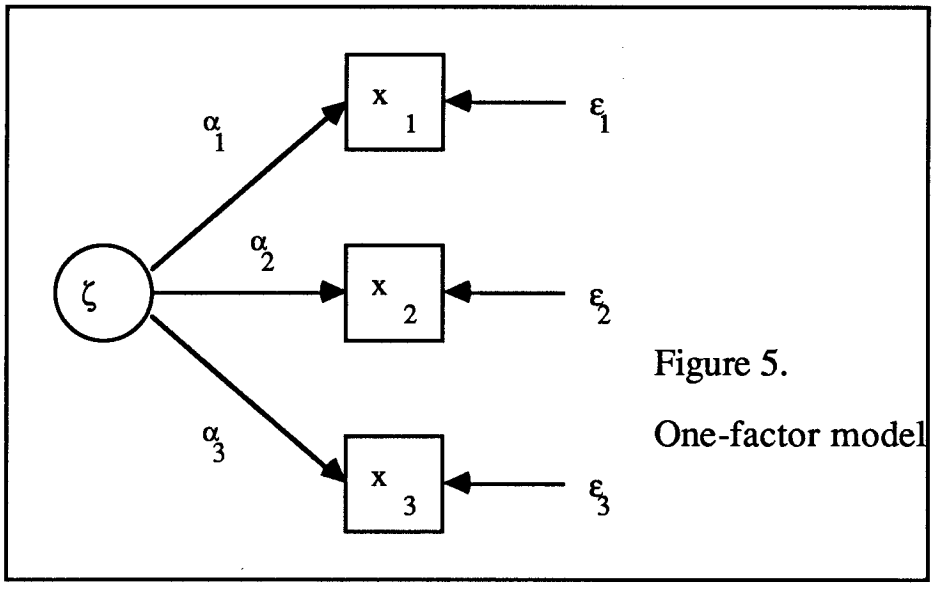


Figure 5.

One-factor model

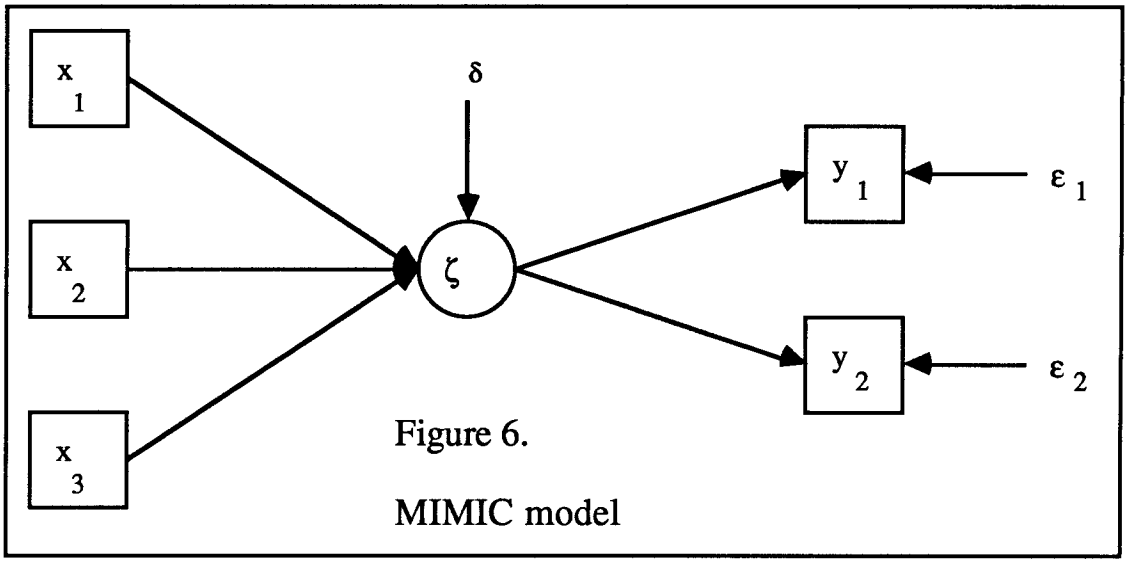


Figure 6.

MIMIC model

Probabilistic translation (factor analysis)

$p_{321\zeta} = p_{321|\zeta}p_{\zeta} = p_{3|\zeta}p_{2|\zeta}p_{2|\zeta}p_{\zeta}$

and thus

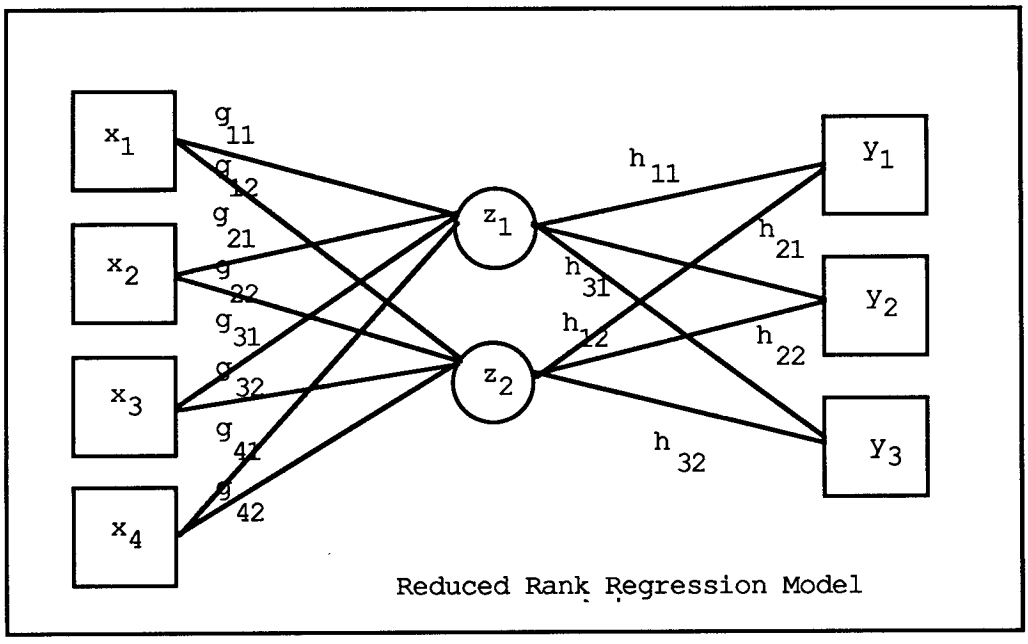
$p_{321} = \int p_{3|\zeta}p_{2|\zeta}p_{2|\zeta}p_{\zeta}\, d\zeta.$

Idem MIMIC

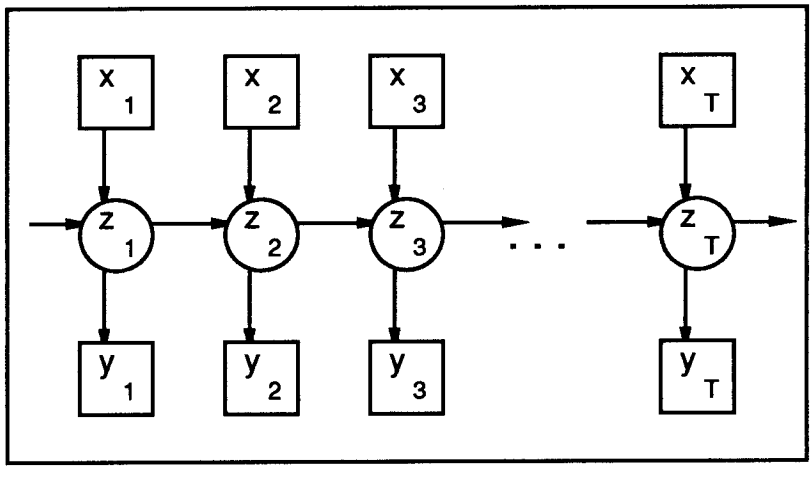
$p(y_1y_2|x_1x_2x_3) = \int p(y_1y_2|\zeta x_1x_2x_3)p(\zeta|x_1x_2x_3)\, d\zeta =$

$\int p(y_1|\zeta)p(y_2|\zeta)p(\zeta|x_1x_2x_3)\, d\zeta.$

Now again this can be specialized by adding linearity, homoscedasticity, normality, specific nonlinear (tobit, probit, logit) regression, and so on. Completely categorical defines latent class analysis, completely linear defines reduced rank regression.



A linear version



A dynamic version

The dynamic versions says, in probabilistic terms,

$$p(y|x) = \int \Pi_{t=1}^{T} p(y_t|z_t)p(z_t|z_{t-1}x_t)\, dz_t,$$

and in linear terms

$$z_t = Fz_{t-1} + Gx_t + \varepsilon,$$
$$y_t = Hz_t + \delta,$$

which is the familiar (Kalman) *linear dynamic system.*

If there is no input (no x) we call this *dynamic factor analysis* in the linear case, in the purely categorical case we call it the *latent Markov chain model.* And, again, many alternative specifications are possible for categorical, ordinal, truncated, censored, missing data.

Perhaps this is also the point to mention a trick, which is familiar in psychometrics (although it dates back to Pearson's tetrachoric and polychoric and polyserial correlation coefficients).

Suppose we do not know how to model $p_{y|x}$, for instance because y is nominal (religion) and x is ordinal (pro–choice, strongly, ...). The psychometric trick is to add two continuous latent variables η and ξ to the model, and to say

$$p(xy\eta\xi) = p(y|x\eta\xi)p(\eta|x\xi)p(x|\xi)p(\xi) =$$

$$p(y|\eta)p(\eta|\xi)p(x|\xi)p(\xi),$$

or

$$p(xy) = \int p(y|\eta)p(\eta|\xi)p(x|\xi)p(\xi)\, d\eta d\xi,$$

and

$$p(y|x) = \int p(y|\eta)p(\eta|\xi)p(\xi|x)\, d\eta\, d\xi$$

The second part of the talk is about fitting. It will not discuss many technical details, because they are boring (see the two boring UCLA statistics papers).

There are two basic general fitting algorithms for the linear case (which is the general case if the data are multinormal). There are algorithms for the more general cases as well, but they are a bit more specialized (fitting latent Markov chains by EM, fitting factor analysis by using empirical characteristic functions, etc).

The first principle is *alternating least squares*, the second *alternating maximum likelihood*. They are called alternating, because they alternate (optimal) transformation (quantification, imputation, scaling, reexpression) of the variables with fitting of the structural parameters (path coefficients, residual variances).

Let's do ALS first. We translate the path diagram to linear equations. Take MIMIC. This gives:

$$\delta_{LS} = \Sigma_{j=1}^{m} ||y_j - \alpha_j \zeta||^2 + ||\zeta - \Sigma_{s=1}^{t} \beta_s x_s||^2.$$

This is different from (but related to)

$$\delta_{LS} = \Sigma_{j=1}^{m} ||y_j - \alpha_j \Sigma_{s=1}^{t} \beta_s x_s||^2.$$

In ALS we do not eliminate the latent variables, because they have the same role as the observed variables, which may have other types of incomplete information anyway.

We use unweighted least squares (which could be modified), and we alternate over as many subsets as possible, sometimes using majorization. See ALS report.

Suppose

$$\delta = SSQ(Y - ZA') + \omega^2 SSQ(Z - XB)$$

for $\omega \to \infty$ this is an eigenvalue problem, equivalent to minimizing SSQ(Y – XBA').

For intermediate ω miniziming over A and B for fixed Z is easy enough. Minimizing over Z (satisfying Z'Z = I) for fixed A and B is not as easy. It amounts to maximizing
tr Z'(YA + XB)
which is a so–called Procrustus problem. Also maximizing over misisng information in X and B is easy.

For the dynamic model we use

$$\delta = SSQ(Y - ZH) + \omega^2 SSQ(Z - BZF - XG).$$

Again the limits $\omega \to 0$ and $\omega \to \infty$ make sense, but finding optimal Z for fixed (F,G,H) is now far from simple.

Proof: rewrite the previous result

$$\ln \int p(x,z,\theta)\, dz \geq \ln \int p(x,z,\xi)\, dz + \int p(z|x,\xi) \ln p(x,z,\theta)\, dz -$$

$$\int p(z|x,\xi) \ln p(x,z,\xi) dz.$$

or

$$L(x,\theta) \geq L(x,\xi) + \Delta(x,\theta,\xi) - \Delta(x,\xi,\xi).$$

Maximize $\Delta(x,\theta,\xi)$, this gives θ^+. Then

$$L(x,\theta^+) \geq L(x,\xi) + \Delta(x,\theta^+,\xi) - \Delta(x,\xi,\xi) \geq$$

$$L(x,\xi) + \Delta(x,\xi,\xi) - \Delta(x,\xi,\xi) = L(x,\xi).$$

Maximizing

$$\int p(z|x,\xi) \ln p(x,z,\theta)\, dz$$

is quite simple, because $\ln p(x,z,\theta)$ is of the form

$$\ln \det \Sigma(\theta) + (x \mid z)'\Sigma(\theta)^{-1}(x \mid z),$$

where $\Sigma(\theta)$ is the dispersion of the joint distribution of (x,z), which is usually simple (think of MIMIC).

Optimal scaling in AML uses ideas from many sources.

If we regres a discrete ordinal variable on a continuous latent variable we use Pearson's polychoric/polyserial methods, which are also used by Muthen and Joreskog in their programs.

If we regres a continuous ordinal variable on a discrete latent variable we use the ideas of Box and Cox, generalized suitably. This is in the paper.

Many additional types of regression remain to be worked out. We already know what to do with nominal on continuous (generalized logistic, Rasch, probit, ...), but other possibilities remain in the dark.

For AML we use the fact that the marginal likelihood of the observations contains an integral, which integrates out the unknown information (over a set of random variables, in stead of over just one). Suppose

$\Delta = \Sigma_{i=1}^{n} \ln \int p(x_i,z_i,\theta)dz_i.$

Now

$p(x,z,\theta) = p(x,z,\xi)\{p(x,z,\theta)/p(x,z,\xi)\}$

thus

$\ln \int p(x,z,\theta)\, dz \,/\, \int p(x,z,\xi)\, dz =$

$\ln \int p(x,z,\xi)\{p(x,z,\theta)/p(x,z,\xi)\}dz \,/\, \int p(x,z,\xi)\, dz \geq$

$\int p(x,z,\xi) \ln \{p(x,z,\theta)/p(x,z,\xi)\}dz \,/\, \int p(x,z,\xi)\, dz.$

Suppose we maximize

$\int p(z|x,\xi) \ln p(x,z,\theta)\, dz$

over θ, then it follows that we increase $\ln \int p(x,z,\theta)\, dz$.