LECTURE NOTES

DECOMPOSITION OF MULTIVARIABLES USING STATE-SPACE MODELS

Jan de Leeuw
UCLA Statistics

June 23, 1994

Contents

1 Multivariables and Profiles **1**
- 1.1 Variables 1
- 1.2 Multivariables 3
- 1.3 Coding 5
- 1.4 Profile Frequencies 6

2 Models **9**
- 2.1 Why Models ? 9
- 2.2 Fixed and Random Variables 10
- 2.3 Stability of Profile Frequencies 13
- 2.4 Fitting Models 14
- 2.5 Distance between Profile Frequencies . 15

3 Mixture Models for Profile Frequencies **19**
- 3.1 The Notion of State 20
- 3.2 The EM algorithm 22

4 Factor Analysis **24**
- 4.1 Latent Class Analysis 27
- 4.2 Latent Trait Analysis 28
- 4.3 Linear Factor Analysis 31
- 4.4 Ordinal and Nominal Variables 33

5 Input and Output **36**
5.1 Regression 36
5.2 Fixed and Random Predictors 38
5.3 The MIMIC Model 39

6 Time in the Picture **41**
6.1 The State Space Model 41
6.2 AR and MA 41
6.3 Using EM 41
6.4 The Kalman Filter 41
6.5 Event Histories and Point Processes . 41

7 We Go Into Space **43**
7.1 Spatial Grids 43
7.2 Kriging 43
7.3 Spatial EM 43

1 Multivariables and Profiles

1.1 Variables

A *variable* is a mapping ϕ defined on a *domain* Ω with values in a *target* ∇.

Elements of the domain are called *individuals* or *objects*, elements of the target are called *categories* or *values*.

Targets can be sets of real numbers, sets of natural numbers, ordered sets, or arbitrary sets.

If a variable corresponds with actual data (and not with a "model") then the domain $\Omega = \{\omega_1, \cdots, \omega_n\}$ is finite. This makes the image $\phi(\Omega) \subseteq \nabla$ finite as well.

In fact, for actual data we can suppose without loss of generality that ∇ is finite.

Some examples of variables:

$$\begin{aligned}
\phi : \text{Individuals} &\Rightarrow \text{IQ scores}\\
\phi : \text{Surveyees} &\Rightarrow \text{Agree with Thing}\\
\phi : \text{Students} &\Rightarrow \text{Religion}\\
\phi : \text{Timepoints} &\Rightarrow \text{Height Mrs. Entity}\\
\phi : \text{Time} \times \text{Person} &\Rightarrow \text{Height}\\
\phi : T \times X \times Y \times Z &\Rightarrow \text{Aftershock}
\end{aligned}$$

1.2 Multivariables

A *multivariable* is a finite sequence $\Phi = \{\phi_1, \cdots, \phi_m\}$ of variables with a common domain Ω and with targets $\nabla_1, \cdots, \nabla_m$.

Let $\nabla_\otimes = \nabla_1 \otimes \cdots \otimes \nabla_m$. Then a multivariable on Ω can also be interpreted as a variable on Ω with target $\nabla_\otimes$.

An element of $\nabla_\otimes = \nabla_1 \otimes \cdots \otimes \nabla_m$ is called a *profile*. Each profile is a sequence of m categories, with category j coming from variable j. Thus a multivariable maps the objects or individuals into the set of profiles. Or

$$\Phi : \Omega \Rightarrow \nabla_\otimes.$$

Often card$\{\Phi(\Omega)\}$ is small compared to card$\{\nabla_\otimes\}$, i.e. many of the profiles are unused.

Here is the GALO example, Groningen 1959.

1. Gender
   (a) Boys
   (b) Girls
2. IQ
   - values between 60 and 144
3. Teachers Advice
   (a) No further education
   (b) Extended primary education
   (c) Manual labour education
   (d) Agricultural education
   (e) General Education
   (f) Secondary school for girls
   (g) Pre-university
4. Fathers profession
   (a) Unskilled labour
   (b) Schooled labour
   (c) Lower white collar
   (d) Shopkeepers
   (e) Middle white collar
   (f) Professional
5. School
   - numbers 1-37

1.3 Coding

There are two essentially different ways to code profiles.

Suppose ∇_j has k_j elements. Define $g_{j\ell}$ to be a binary vector with k_j elements given by

$$\{g_{j\ell}\}_\nu = \begin{cases} 1 & \text{if } \ell = \nu \\ 0 & \text{otherwise} \end{cases}$$

A profile $< \ell_1, \cdots, \ell_m >$ can now be coded as a vector $g_\oplus = g_{1\ell_1} \oplus \cdots \oplus g_{m\ell_m}$ of length $\sum_{j=1}^m k_j$. Or, alternatively, as the $k_1 \times \cdots \times k_m$ array $g_\otimes = g_{1\ell_1} \otimes \cdots \otimes g_{m\ell_m}$.

The vector $g_\oplus$ has m elements equal to one, one for each variable. The array $g_\otimes$ has exactly one element equal to one, indicating where the profile is in the m-dimensional grid.

1.4 Profile Frequencies

In the Gifi system for multivariate analysis the basic coding is the additive coding $g_{\oplus}$ of the profiles. In this course we will use the multiplicative coding $g_{\otimes}$.

There are $\prod_{j=1}^{m} k_j$ possible profiles, indexed by $i \in \mathcal{I}$. Each profile occurs with frequency n_i in the data, $\sum_{i \in \mathcal{I}} n_i = n$.

The *profile frequencies* of a multivariable are a mapping of $\nabla_{\otimes}$ into the natural numbers $\{0, 1, 2, \cdots\}$. The interpretation is that the mapping assigns frequencies to each of the profiles.

Observe that we can recover the variables from the profile frequencies if, and only if, the order of the observations does not matter.

We use n_i for the frequency of profile i. Often we also use the profile relative frequencies given by

$$p_i = \frac{n_i}{n}.$$

The relative frequencies are in the simplex $\mathcal{P}$, which is the set of all vectors with $\prod_{j=1}^{m} k_j$ non-negative elements adding up to one.

In order to study stability and related statistical problems, we also have to define the related quantities π_i and $\underline{p}_i$. These quantities do not refer to the data, but to the model. We shall discuss them in a separate secion.

Here is a tiny example of a multivariable and its profile frequencies.

object	religion	height in feet	gender
john	buddhist	6	male
jan	mormon	5	female
joe	mormon	6	male
jim	buddhist	6	male
jack	buddhist	6	male
jill	buddhist	5	female
jebb	mormon	5	male
jane	mormon	5	female

religion	height in feet	gender	n	p
buddhist	5	male	0	.000
buddhist	5	female	1	.125
buddhist	6	male	3	.375
buddhist	6	female	0	.000
mormon	5	male	1	.125
mormon	5	female	2	.250
mormon	6	male	1	.125
mormon	6	female	0	.000

2 Models

2.1 Why Models ?

A *model* $\mathcal{M}$ is a subset of space $\mathcal{P}$. It can consist of s single point (simple model) or of the whole space (saturated model).

Models are useful, because they

- summarize prior knowledge;
- provide a language for discourse;
- reduce data volume;
- enhance stability.

Models are harmful, because they

- incorporate prejudices and idols;
- introduce bias.

Models are used to *filter* and *summarize* observed arrays.

2.2 Fixed and Random Variables

In many cases models do not only involve specifying a subset of array space, but they also specify something of the form

$$\text{Data} = \text{Structure} + \text{Deviation}.$$

What is this “deviation” ? It is something that will be different if the experiment is replicated. This does not imply that the experiment actually can or will be replicated, on the contrary, we rely more heavily on the model when actual replication is difficult or impossible.

And what is the “structure” ? This is also known als the *truth*. It is the part of the outcome that stays the same over replications. Again, this is a theoretical construct, which can never be verified of falsified in any strict sense. Without it, however, the whole notion of stability and cumulative science does not make much sense.

The standard model for variability around a true value uses the notion of a *random variable*. It is important to realize that random variables do not model data, they model a hypothetical sequence of experiments, sometimes known as a *replication framework*.

The truth corresponds with the expected value of the random variable, and with the statistical notion of unbiasedness. The deviations correpond with the variance of the random variable, and with the statistical notion of standard error (or confidence interval).

Now define $\underline{p}_n$ as the replication framework for p, i.e. it is the variation in the profile frequencies if we replicate the experiment a large number of times, and define $\pi = \mathbf{E}(\underline{p}_n)$ as the **Truth.**

We can also distinguish between fixed variables (also called *factors*) and random variables. The fixed factors remain the same in our hypothetical sequence of experiments. They do not have error, they are there as they are by *design*. The random variables have an error component, and vary over replications.

Random variation introduces a probability distribution over the profiles. Since we assume discreteness, the notion of replications actually generates a product multinomial structure. Each profile i consist of a part i_1 corresponding with the fixed variables and a part i_2 corresponding with the random variables. Thus $\pi(i) = \pi(i_1, i_2) = \pi(i_2|i_1)$. Each value of i_1 defines a multinomial distribution over the values of i_2.

2.3 Stability of Profile Frequencies

We have

$$\text{prob}(\underline{p}_n = p_n) = \prod_{i_1 \in \mathcal{I}_1} \mathcal{C}(i_1) \prod_{i_2 \in \mathcal{I}_2} \pi(i_2|i_1)^{n(i_2|i_1)}.$$

In the same way

$$\mathbf{V}\left(\underline{p}_n(i)\right) = n(i_1)\pi(i_2|i_1)(1 - \pi(i_2|i_1)),$$

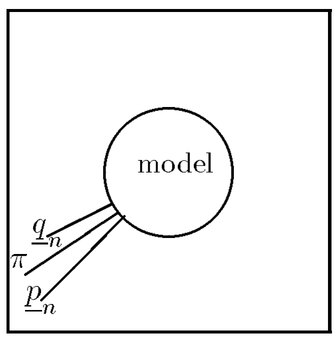
and, for $i \neq i'$,

$$\mathbf{C}\left(\underline{p}_n(i), \underline{p}_n(i')\right) =$$

$$\begin{cases} -n(i_1)\pi(i_2|i_1)\pi(i_2'|i_1') & \text{if } i_1 = i_1' \\ 0 & \text{otherwise} \end{cases}$$

2.4 Fitting Models

We *project* the observed array on the model. Or, to put it differently, we find the closest point on the model.



Observe that in this case the model is not "true", and observe that projection improves stability (although it introduces bias).

2.5 Distance between Profile Frequencies

There are many different ways to define distance between conforming vectors of frequencies. One that is particulary natural is based on *weighted least squares*.

If $p(i)$ is the observed vector of frequencies, and $\pi(i)$ is the expected vector (given the model), then

$$\Delta_{WLS}(p, \pi) = n \sum_{i \in \mathcal{I}} w(i)(p(i) - \pi(i))^2.$$

The weights $w(i)$ should be chosen in some "reasonable" way, to reflect the stability of the observed frequencies.

For the time being we suppose there are only random variables.

We now that the dispersion of the $\underline{p}_n$ in matrix form is equal to $V = \Pi - \pi\pi'$, with Π the diagonal matrix with the $\pi(i)$ on the diagonal. This matrix is singular, because its rows and columns add up to zero. For the generalized inverse we find

$$[V + \frac{1}{n}ee']^{-1} = V^{+} + \frac{1}{n}ee',$$

and thus

$$V^{+}(p - \pi) = \Pi^{-1}(p - \pi).$$

This suggests

$$\Delta_{XS}(p,\pi) = n(p-\pi)'V^{+}(p-\pi) = n(p-\pi)'\Pi^{-1}(p-\pi),$$

or, in scalar notation,

$$\Delta_{XS}(p,\pi) = n\sum_{i\in\mathcal{I}}\frac{(\pi(i)-\pi(i))^2}{\pi(i)}.$$

Here is another one:

$$\Delta_{ML}(p,\pi) = 2n\sum_{i\in\mathcal{I}} p(i)\log\frac{p(i)}{\pi(i)}.$$

By writing

$$p(i) = \pi(i) + (p(i) - \pi(i))$$

we see that

$$\begin{aligned}\log\frac{p(i)}{\pi(i)} &= \log\left\{1 + \frac{p(i)-\pi(i)}{\pi(i)}\right\} = \\ &\approx \frac{p(i)-\pi(i)}{\pi(i)} - \frac{1}{2}\frac{(p(i)-\pi(i))^2}{\pi(i)^2},\end{aligned}$$

which shows

$$\Delta_{ML}(p,\pi) \approx \Delta_{XS}(p,\pi).$$

Another popular and elegant one:

$$\begin{aligned}\Delta_{HD}(p,\pi) &= 4n\sum_{i\in\mathcal{I}}(\sqrt{\pi(i)}-\sqrt{p(i)})^2 = \\ &= 8n(1-\sum_{i\in\mathcal{I}}\sqrt{p(i)\pi(i)}).\end{aligned}$$

Write

$$\sqrt{p(i)\pi(i)} = \pi(i)\sqrt{1+\frac{p(i)-\pi(i)}{\pi(i)}}$$

$$\approx \quad \pi(i)+\frac{1}{2}(p(i)-\pi(i))-\frac{1}{8}\frac{(p(i)-\pi(i)^2}{\pi(i)},$$

and again

$$\Delta_{HD}(p,\pi)\approx\Delta_{XS}(p,\pi).$$

3 Mixture Models for Profile Frequencies

Mixture models assume that each vectur of frequencies is a mixture of "simple" vectors. Thus we write

$$\pi(i)=\sum_{s=1}^{r}\pi(s)\pi(i|s),$$

or

$$\pi(i)=\int_{-\infty}^{+\infty}\pi(s)\pi(i|s)ds,$$

and we assume the $\pi(i|s)$ are "simple", in some sense. The $\pi(s)$ are the *mixing proportions*. It is as if we really have a two-dimensional table $\pi(i,s)$, for which we only observe the marginal $\pi(i)$.

3.1 The Notion of State

States, or state variables, are also known as *latent variables*. The general idea is that the state characterizes the system, in the sense that we knew the most important aspects of the system if we know the state. Unfortunately, the state cannot be measured directly, but it has to be inferred.

Some of the major examples show more clearly what this means. If we know the intelligence, the IQ tests do not really provide additional information. If we know the true value of a quantity, then the repeated measurements merely show what the error of measurement is. If we know the state of a dynamic system, then there is no need to enquire into its past, because the state has all the information necessary for prediction.

Many people question the notion of latent variables. We silence them by quoting a Nobel Price laureate.

Quotation

There is now a school of mathematical Physicists which objects to the introduction of ideas which do not relate to things which can actually be observed and measured. ··· I hold that if the introduction of a quantity promotes clearness of thought, then even if at the moment we have no means of determining it with precision, its introduction is not only legitimate but desirable. The immeasurable of to-day may be the measurable of to-morrow.

J.J. Thomson, January 29, 1930

3.2 The EM algorithm

Let us study projecting on the model, using $\Delta_{ML}(p,\pi)$. This is known as *maximum likelihood estimation*, because it amounts to maximizing the log-likelihood

$$\mathcal{L}(\pi) = \sum_{i\in\mathcal{I}} n(i)\log\pi(i) = \sum_{i\in\mathcal{I}} n(i)\log\sum_{s=1}^{r}\pi(s)\pi(i|s).$$

Suppose $\tilde{\pi}(s)$ and $\tilde{\pi}(i|s)$ are our current best estimates. Then

$$\log\frac{\pi(i)}{\tilde{\pi}(i)} = \log\frac{\sum_{s=1}^{r}\tilde{\pi}(s)\tilde{\pi}(i|s)\frac{\pi(s)\pi(i|s)}{\tilde{\pi}(s)\tilde{\pi}(i|s)}}{\sum_{s=1}^{r}\tilde{\pi}(s)\tilde{\pi}(i|s)},$$

which shows

$$\begin{aligned}\log\pi(i) &\geq \log\tilde{\pi}(i) + \\ \sum_{s=1}^{r}\tilde{\pi}(s|i)\log\pi(i,s) &- \sum_{s=1}^{r}\tilde{\pi}(s|i)\log\tilde{\pi}(i,s).\end{aligned}$$

The only term on the right hand side that depends on $\pi(i,s)$ is $\sum_{s=1}^{r}\tilde{\pi}(s|i)\log\pi(i,s)$. By maximizing this term, we maximize the right-hand-side, and this maximum will be larger than $\tilde{\pi}(s)$.

Because of the inequality on the previous page, we see that this implies that we increase $\pi(i)$ as well. Thus, in each step, we maximize

$$\begin{aligned}&\sum_{i\in\mathcal{I}} n_i\sum_{s=1}^{r}\tilde{\pi}(s|i)\log\pi(s)\pi(i|s)\\ =\;& n\sum_{s=1}^{r}\tilde{p}(s)\log\pi(s)\\ +\;& n\sum_{s=1}^{r}\tilde{p}(s)\sum_{i\in\mathcal{I}}\tilde{p}(i|s)\log\pi(i|s).\end{aligned}$$

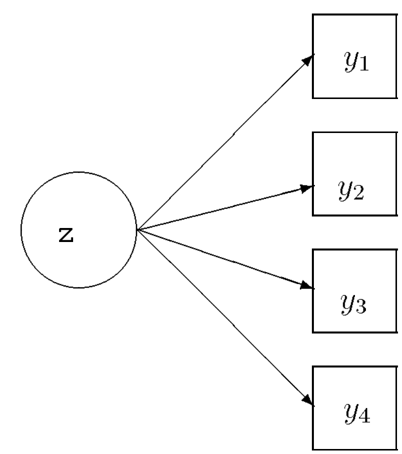
with $\tilde{p}(i,s) = p_i\tilde{\pi}(s|i)$, etc. This is usually much simpler than the original problem, but we have to solve it a large number of times.

4 Factor Analysis

We now give a brief and unconventional introduction to factor analysis. In our presentation of the model, it covers latent class analysis, Rasch models, and various other special cases as well. In fact, the important thing to understand is that basically these are all the same model.

Factor analysis is a cross-sectional state space model without input variables.

This point of view is far from new. It was discussed with varying degree of generality in the forties by Lazarsfeld and Guttman, in the fifties by Anderson and Koopmans, and in the sixties and seventies by McDonald, Lord, and others.



The basic assumption for factor analysis is *local or conditional independence*. Given the value of the state variable, the observed variables are independent. To use some canonical examples:

- Given a person's intelligence, the scores on various intelligence tests are independent;
- Given a person's ability, his results on the items of a test are independent.

This is just another way of saying that intelligence tests only have intelligence in common, and test items only have the fact that they measure ability in common.

Thus the "simplicity" in the factor analysis model is independence. This is not strictly necessary. We can also have a factor analysis in which "simple" means no second-order interactions.

4.1 Latent Class Analysis

Suppose there is only a finite number of states. For notational simplicity, suppose there are three observed variables, i.e. our array is three-dimensional with elements p_{ijk}. The model is

$$\pi_{ijk} = \sum_{s=1}^{r} \pi_s \pi^A_{i|s} \pi^B_{j|s} \pi^C_{k|s}.$$

The EM algorithm is simply

$$\begin{aligned} \pi_s &\leftarrow \tilde{p}_s \\ \pi_{i|s} &\leftarrow \tilde{p}_{i|s} \\ \pi_{j|s} &\leftarrow \tilde{p}_{j|s} \\ \pi_{k|s} &\leftarrow \tilde{p}_{k|s}. \end{aligned}$$

This is a simple as it gets. We do *iterative proportional fitting* on the augmented array.

4.2 Latent Trait Analysis

Suppose the state is a single continuous variable such as intelligence, or ability), and the (three) variables are all binary (such as correct-false items). Then, for profile $< 1, 1, 1 >$ for example,

$$\pi(111) = \int_{-\infty}^{+\infty} \pi(s)\pi_1(s)\pi_2(s)\pi_3(s)ds,$$

where $\pi_1(s)$ is short for the probability of a correct response on variable 1 given ability s. The function $\pi_j(s)$ is called the trace-line of item j. More generally, we can write

$$\pi(y_{i1}, \cdots, y_{im}) = \int_{-\infty}^{+\infty} \pi(s) \prod_{j=1}^{m} \pi_j(s)^{y_{ij}} (1 - \pi_j(s))^{1-y_{ij}} ds.$$

What can we say about the EM algorithm in this case ? Well, in the first place we need to distinguish some cases. If $\pi(s)$ is not further specified, and estimated, then we do *semiparametric maximum marginal likelihood* estimation. We call it semiparametric, because often the trace lines will have a prescribed parametric form, such as

$$\begin{aligned}
\pi_j(s) &= \frac{1}{1 + \exp\{-(s - \mu_j)\}}, \\
\pi_j(s) &= \frac{1}{1 + \exp\{-\frac{s-\mu_j}{\sigma_j}\}}, \\
\pi_j(s) &= \frac{1}{\sigma_j\sqrt{2\pi}} \int_{-\infty}^{s} \exp\{-\frac{1}{2}\frac{(t - \alpha_j)^2}{\sigma_j^2}\}dt.
\end{aligned}$$

We can also consider the *nonparametric situation* in which neither the functional form of the $\pi(s)$, nor that of the $\pi_j(s)$ is specified (except for monotonicity of the $\pi_j(s)$). This actually brings us back to latent class analysis.

For the general latent trait model

$$\log \pi(i|s) = \sum_{j=1}^{m} y_{ij} \log \left\{ \frac{\pi_j(s)}{1-\pi_j(s)} \right\} + \log\{1-\pi_j(s)\},$$

and thus

$$\sum_{i\in\mathcal{I}} \tilde{p}(i|s) \log \pi(i|s) =$$
$$\sum_{j=1}^{m} u_j(s) \log \pi_j(s) + (1-u_j(s)) \log\{1-\pi_j(s)\},$$

with

$$u_j(s) = \sum_{i\in\mathcal{I}} \tilde{p}(i|s) y_{ij}.$$

The EM algorithms sets $\pi(s)$ equal to $\tilde{p}(s)$, and sets $\pi_j(s)$ equal to a monotonic version of $u_j(s)$. Because of the monotone regression, this creates step functions as solutions. The problem becomes slightly more complicated if we also require that the trace lines do not cross.

4.3 Linear Factor Analysis

Suppose we have

$$\pi(y_1, \cdots, y_m) = \int_{-\infty}^{+\infty} \pi(s) \prod_{j=1}^{m} \pi(y_j|s) ds,$$

and the regressions of the observed variables on the state variables are linear and homoscedastic, i.e.

$$\mathbf{E}(\underline{y}_j|s) = \sum_{h=1}^{f} \alpha_{jh} s_h,$$
$$\mathbf{V}(\underline{y}_j|s) = \sigma_j^2$$

Then

$$\mathbf{C}(\underline{y}_j, \underline{y}_\ell) = \sum_{h=1}^{f}\sum_{g=1}^{f} \alpha_{jh}\alpha_{\ell g}\omega_{hg} + \delta^{j,\ell}\sigma_j^2.$$

This is the famous Thurstone Multiple Factor Analysis Model, with $f = 1$ the special Spearman case.

The factor analysis model with a continuous state space, and a “continuous” observed space is not too different from the laten class model. In principle the same EM algorithm could be used. Unfortunately if the variables are “continuous” the profiles have low frequencies (usually either zero or one), and there are too many parameters for stable estimation. Thus we must use a parametric model, which assume for instance that $\pi(s)$ and/or $\pi(y_j|s)$ are normal. But then

$$-2\log\pi(y|s)=\sum_{j=1}^{m}\log\sigma_j^2+\frac{(y_j-\sum_{h=1}^{f}\alpha_{jh}s_h)^2}{\sigma_j^2},$$

and thus we see that the EM algorithm is based on the simple result

$$-2\int_{-\infty}^{+\infty}\pi(s)\log\pi(y|s)=\sum_{j=1}^{m}\log\sigma_j^2+$$
$$+\frac{\mathbf{E}\left((y_j-\sum_{h=1}^{f}\alpha_{jh}s_h)^2\right)}{\sigma_j^2}.$$

4.4 Ordinal and Nominal Variables

We see that factor analysis models are easy to build for both binary outcomes and numerical outcomes. They are slightly more involved for ordinal and nominal outcomes. There are two basic machanisms to make the extensions.

The main trick is to study, instead of the m variables $\underline{y}_1,\cdots,\underline{y}_m$ a set of m pairs of variables $(\underline{y}_1,\underline{\eta}_1),\cdots,(\underline{y}_m,\underline{\eta}_m)$. The $\underline{y}_j$ are observed, the $\underline{\eta}_j$ are again latent or in state space.

We now need to connect the members of the pair. The first possibility is to assume a deterministic relationship $\underline{y}_j=F_j(\underline{\eta}_j)$, where F_j is completely or partially known. The second possibility is to assume that $\underline{y}_j$ only depends on $\underline{\eta}_j$ but that the connection is probabilistic.

Three examples of the deterministic approach. The first is the Box-Cox method. We set

$$F(\eta,\lambda) = \begin{cases} \frac{\eta^\lambda - 1}{\lambda} & \text{if } \lambda > 0 \\ \log(\eta) & \text{if } \lambda = 0 \end{cases}$$

The second example sets

$$F(\eta,\alpha) = \begin{cases} 1 & \text{if } \alpha_0 < \eta \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < \eta \leq \alpha_2 \\ \cdots & \cdots \\ k & \text{if } \alpha_{k-1} < \eta \leq \alpha_k \end{cases}$$

And finally we can set $F(\eta,\alpha,\beta)$ equal to a (monotone) spline with knots α and B-spline coefficients β. This is all meant for ordinal or even "continuous" variables.

We give a single example of the probabilistic approach. Suppose $\underline{y}$ is discrete, and has values $1,\cdots,k$. Then we can set

$$\text{prob}(\underline{y} = \ell | \underline{\eta} = \eta) = \frac{\exp\{\alpha_\ell + \beta_\ell \eta\}}{\sum_{v=1}^k \exp\{\alpha_v + \beta_v \eta\}},$$

and, of course,

$$\text{prob}(y) =$$
$$\int_{-\infty}^{+\infty} \text{prob}(y|\eta)\text{prob}(\eta) d\eta =$$
$$\int_{-\infty}^{+\infty} \text{prob}(\eta) \prod_{j=1}^m \text{prob}(y_j|\eta_j) d\eta.$$

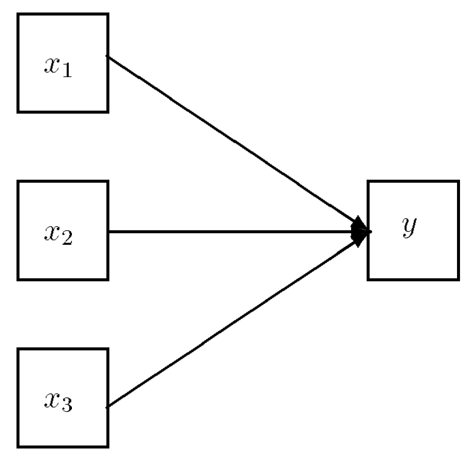
5 Input and Output

5.1 Regression

In factor analysis (and the related techniques we discussed above) all observed variables enter symmetrically into the model. They all play the same role. The state variable is different (it is "causally prior") but it is not observed.

In *input-output models*, also known as *regression models,* the situation is more or less reversed. There is observable input and observable output, or *independent* and *dependent* variables.

In its simplest form the model is given in the following figure, but since this is not a state space or latent variable model, we shall not discuss it in detail.

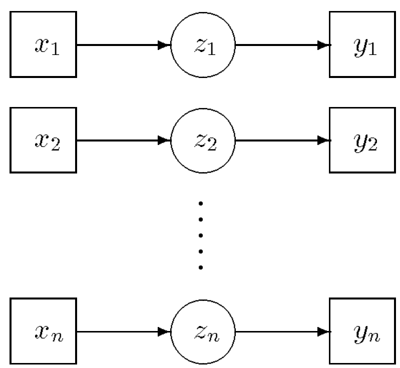


5.2 Fixed and Random Predictors

Observe that a distinction we made earlier becomes relevant here. Do we consider the input as fixed or as random ? This depends entirely on what we consider to be a replication of the experiment. If a replication means identical input but different output (as in experimental design) the the input is fixed, if a replication will give different output as well as input (observational study) then input is random.

From the modeling point of view we often model $\pi(x, y)$ as $\pi(y|x)\pi(x)$ and of course in $\pi(y|x)$ we have fixed x, at least formally.

5.3 The MIMIC Model



6 Time in the Picture

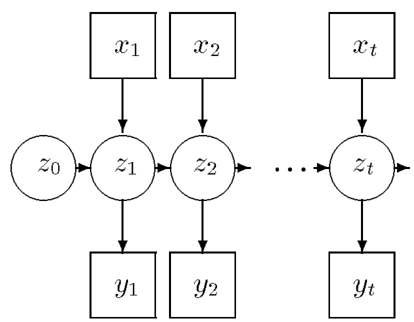
6.1 The State Space Model

6.2 AR and MA

6.3 Using EM

6.4 The Kalman Filter

6.5 Event Histories and Point Processes



7 We Go Into Space

7.1 Spatial Grids

7.2 Kriging

7.3 Spatial EM