Graphical Models For Complex Health Data (P8124)

Daniel Malinsky

Columbia University
dsm2128@cumc.columbia.edu

Missing Data, Latent Variables, EM, etc.

Missing data patterns

Consider a data set of observations $X^1, ..., X^n$ where rows have various missing values.

X_1	X_2	X_3	X_4...
0	1	?	1...
0	?	0	?...
1	0	?	?...
?	1	0	0...

Define for each X_i a random variable (missingness indicator) M_i such that $M_i^j = 1$ if X_i^j is observed and $M_i^j = 0$ if X_i^j is missing ("?").

Missing data patterns

X_1	X_2	X_3	X_4...	M_1	M_2	M_3	M_4....
0	1	?	1...	1	1	0	1...
0	?	0	?...	1	0	1	0...
1	0	?	?...	1	1	0	0...
?	1	0	0...	0	1	1	1...

Let X_{obs} denote the observed components/entries of X and X_{mis} denote the missing components.

Why is missing data an important problem for analysis?

It depends on *why* the data is missing, i.e., what is the *missingness mechanism*. If the missing values are absent for systematic reasons – related to the object of analysis – ignoring the missing data may lead to very misleading conclusions: your estimates may be biased, CIs may undercover, etc.

Let's consider some simple examples.

Average height

Consider a single variable X corresponding to adult height in some population. You're interested in $\mu = \mathbb{E}[X]$ which is the average height. The MLE estimate $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{j=1}^{n} X^j$.

However, if your sample is not a complete random sample of the whole population – e.g., taller people are more likely to be sampled – then your estimate will biased.

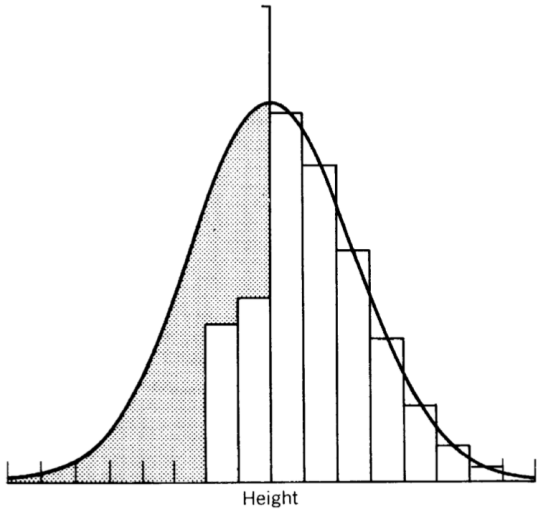


Figure 1.3. Observed and population distributions of historical heights. Population distribution is normal, observed distribution is represented by the histogram, and the shaded area represents missing data.

These are historical observations of height using military records.

Opinion polls

Opinion pollsters try to obtain representative samples of likely voters to estimate the proportion who prefer candidate A to candidate B. However, sampling methods are imperfect. Consider sampling voters by calling randomly selected landline phone numbers in a district. Even if landline owners are representative of the voting population at large (they are not!), you will still have the problem that many prospective voters will not answer, or refuse to give their opinion. If supporters of candidate A are more likely to give no response, you may severely underestimate that candidate's likelihood of winning the election.

Dropout in clinical trials

Even in a very controlled data-gathering procedure, missing data can cause severe problems. Consider a randomized clinical trial for the effectiveness of drug A on health outcome Y. Participants are recruited into a trial and randomly assigned $A = 1$ (treatment) or $A = 0$ (control). However, many participants will fail to show up for the followup evaluation, and so for those participants Y is missing. If people fail to show up for underlying health reasons – precisely because they were too sick – then the observed outcomes Y_{obs} will seem better/healthier on average, and you may false attribute that to the proposed treatment. Or it may be that the treatment had extreme side-effects for some participants, which caused them to not show up or to stop taking it.

"Drop out" of clinical trials and observational studies is a *major* problem in evaluating treatment efficacy.

What do we have to know/assume about the missingness mechanism to get unbiased results?

Treating each M_i as a random variable (collected in the vector M), we may ask how the distribution of M depends on X:

$$p(m|x)$$

This is the distribution induced by the missingness mechanism. Note that $X = (X_{obs}, X_{mis})$ so this is a distribution which (possibly) depends on latent variables – thus, whatever we assume about $p(m|x)$, we cannot know on the basis of the data alone.

$$p(x) = \frac{p(x, M = 1)}{p(M = 1|x)}$$

$p(M = 1|x)$ is the key to learning about features of the distribution we care about, $p(x)$. The numerator $p(x, M = 1)$ is not a problem because it only looks at observed rows.

Assumptions on the missingness mechanism

1) $p(m|x) = p(m)$: $M \perp\!\!\!\perp X$ "Missing Completely at Random" (MCAR)

- Probability of missingness is entirely independent of the variables X.
- Think of data missing due to random errors not related to the object of study
- Strongest assumption.

Assumptions on the missingness mechanism

1) $p(m|x) = p(m)$: $M \perp\!\!\!\perp X$ "Missing Completely at Random" (MCAR)

- Probability of missingness is entirely independent of the variables X.
- Think of data missing due to random errors not related to the object of study
- Strongest assumption.

2) $p(m|x) = p(m|x_{obs})$: $M \perp\!\!\!\perp X_{mis}|X_{obs}$ "Missing at Random" (MAR)

- Probability of missingness is independent of missing variables given what's observed.
- X_{obs} carries "all the relevant information" about missingness.
- Weaker, but still sometimes too strong assumption.

Assumptions on the missingness mechanism

1) $p(m|x) = p(m)$: $M \perp\!\!\!\perp X$ "Missing Completely at Random" (MCAR)

- Probability of missingness is entirely independent of the variables X.
- Think of data missing due to random errors not related to the object of study
- Strongest assumption.

2) $p(m|x) = p(m|x_{obs})$: $M \perp\!\!\!\perp X_{mis}|X_{obs}$ "Missing at Random" (MAR)

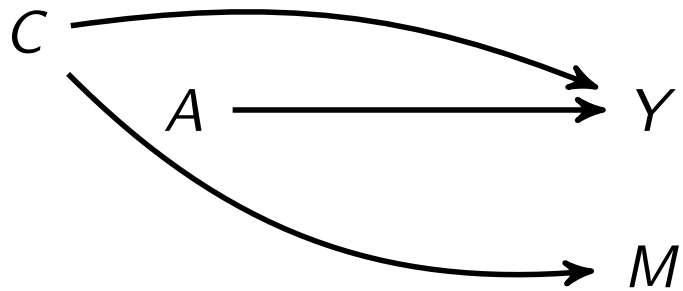
- Probability of missingness is independent of missing variables given what's observed.
- X_{obs} carries "all the relevant information" about missingness.
- Weaker, but still sometimes too strong assumption.

3) Anything else: neither MCAR nor MAR. "Missing Not at Random" (MNAR)

- Distribution depends on missing values.
- Weakest assumption. But may lead to non-identifiability of the target distribution.

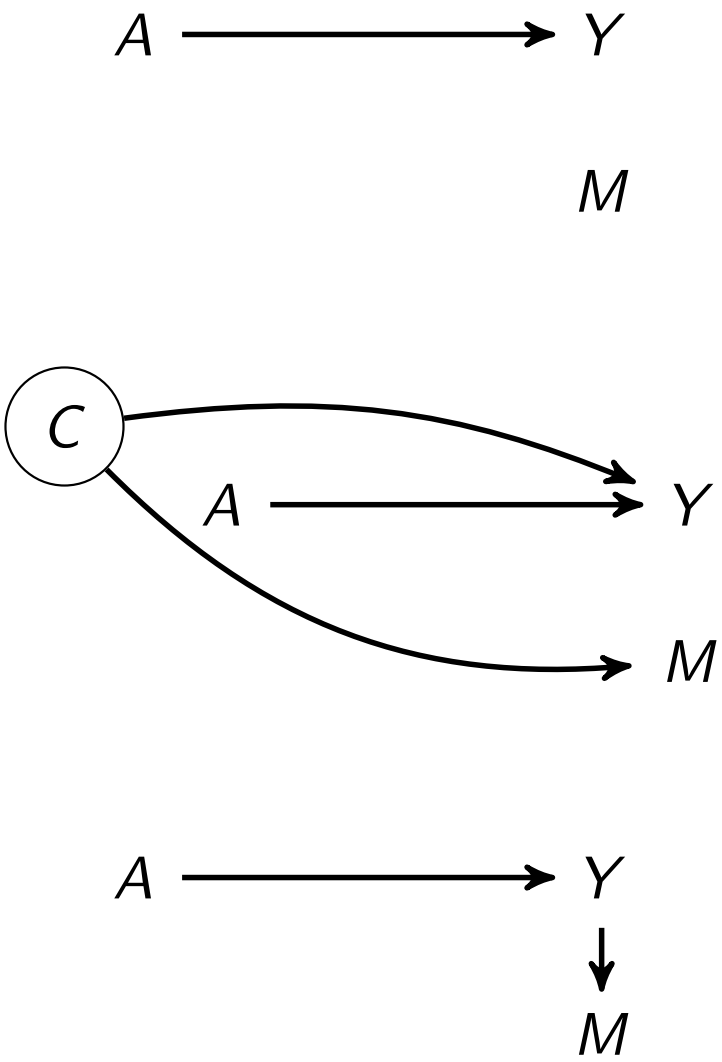
Simple randomized trial

$X = (A, Y)$ and M corresponds to "drop-out."



Simple randomized trial

$X = (A, Y)$ and M corresponds to "drop-out."



Methods for dealing with missing data

MCAR

- Complete-case analysis

MAR

- Maximum likelihood with EM algorithm
- Multiple imputation
- Reweighting estimators, semi-parametric methods

MNAR

- Not always identified: requires additional assumptions
- Reweighting estimators, semi-parametric methods in certain cases

EM algorithm

The Expectation-Maximization (EM) algorithm is a general approach to MLE in settings with a combo of observed (X) and latent (Z) variables.

The observed data log-likelihood can be written $\ell(\theta) = \sum_{j=1}^{n} \log \sum_z p(x^j, z^j; \theta)$. We have an iterative procedure. Begin with an initial guess of θ^0.

E-step: $Q(\theta, \theta^{t-1}) \equiv \mathbb{E}[\ell(\theta)|X = x, \theta^{t-1}]$
(expectation wrt $Z \mid X = x, \theta^{t-1}$)

M-step: $\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$

t indexes iterations. We repeat this procedure until convergence.

EM algorithm

1. Initialize θ
2. Compute probability of each value of Z given θ
3. Use computed values of Z to get an improved estimate of θ by MLE
4. Repeat until convergence

Can prove that this procedure monotonically increases the observed data likelihood: $\ell(\theta^t) \geq \ell(\theta^{t-1})$.

Missing data EM

The basic idea of using EM for missing data is to treat X_{mis} as latent, and X_{obs} as observed. Then EM gives you a way of maximizing the observed data likelihood. We make the distinction between $\ell(\theta|X_{obs}, X_{mis}) \equiv \ell(\theta|X)$ and $\ell(\theta|X_{obs})$.

E-step:
$Q(\theta, \theta^{t-1}) \equiv \mathbb{E}[\ell(\theta|X)|X_{obs}, \theta^{t-1}] = \int \ell(\theta|X) p(x_{mis}|x_{obs}; \theta^{t-1}) dx_{mis}$

M-step: $\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$

We repeat this procedure until convergence.

Important: this strategy is popular, but not always valid. This version of EM "ignores" the missingness mechanism, only valid under MAR.

Multiple imputation

Alternative to an EM approach, which involves evaluating likelihoods, one may consider imputing the missing values themselves, and then running whatever "full data" analysis on the imputed data set as usual. This has a lot of appeal because after the imputation is done, no additional special adjustment needs to made for incomplete data.

However, imputing the data properly must be done with care, because badly imputed data can drastically skew/bias your results.

How is this done in practice? Let's consider a particularly popular approach called MICE (Multiple Imputation with Chained Equations).

Multiple imputation

X_1	X_2	X_3	X_4...	
0	1	?	1...	
0	?	0	?...	(*original data*)
1	0	?	?...	
?	1	0	0...	

X_1	X_2	X_3	X_4...	
0	1	1	1...	
0	0	0	1...	(*imputation*1)
1	0	0	1...	
1	1	0	0...	

X_1	X_2	X_3	X_4...	
0	1	0	1...	
0	1	0	1...	(*imputation*2)
1	0	0	1...	
0	1	0	0...	

Multiple imputation

- Start with initial imputed values for all missing observations (something naive, like the column average).
- For each column i, sample missing values from $p_i(x_{i,mis}|x_{i,obs}, x_{-i})$ using the rest of the columns, with some flexible modeling method.
- Iterate this procedure until predicted values do not change much.
- Return a "complete" imputed data set.

One should repeat this procedure multiple times, creating multiple imputed data sets, since each one will reflect the random chance of sampling from the imputation models. Then, proceed with full data analysis on each imputed data set to estimate $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$ for K imputations. Then, average the estimated parameters together: $\frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$. (Usually users choose $K = 5$ or 10.)

Multiple imputation

Note there are actually two broad approaches to sampling missing values.

One approach (frequentist) uses an initial parameter estimate $\hat{\theta}_i$ and samples from $p_i(x_{i,mis}|x_{i,obs}, x_{-i}; \hat{\theta}_i)$ at every step.

The other approach (Bayesian) samples θ_i^* from $p_i(\theta_i|x_{i,obs}, x_{-i})$ and then X_i from $p_i(x_{i,mis}|x_{i,obs}, x_{-i}, \theta_i^*)$ at every step.

The differences are not crucial here.

Multiple imputation: words of caution

The predictive distributions p_i must be chosen by the user, ideally in a way that is both *flexible* (doesn't make unrealistic parametric assumptions, like the missing data is a linear function of "observed" variables) and *compatible with a joint distribution*. This is a problem, because many ways of specifying a series of seemingly reasonable conditional models $p_i(x_{i,mis}|x_{i,obs}, x_{-i}; \theta_i)\ \forall i \in \{1, ..., p\}$ are not compatible with *any* joint distribution over $p(x)$. The properties of imputation with incompatible conditional distributions are not entirely understood: sometimes you can get consistent imputation estimates with (sort of) incompatible distributions, and sometimes not.

In practice, people typically use one of the default methods implemented in multiple imputation software packages. A popular method is called *predictive mean matching*. The following slide has a rough description.

Predictive Mean Matching

1. For cases with no missing data, estimate a linear regression of X_i on X_{-i}, producing a set of coefficients β.
2. Make a random draw from the "posterior predictive distribution" of β, producing a new set of coefficients β^*. Typically this would be a random draw from a multivariate normal distribution with mean β and the estimated covariance matrix of β (with an additional random draw for the residual variance).
3. Using β^*, generate predicted values for X_i for all cases, both those with data missing on X_i and those with data present.
4. For each case with missing X_i, identify a set of cases with observed X_i whose predicted values are "close" to the predicted value for the case with missing data.
5. From among those close cases, randomly choose one and assign its observed value to substitute for the missing value.
6. Repeat steps 2 through 5 for each completed data set.[1]

[1] This description due to Paul Allison.

Pros and cons of multiple imputation

Pros:

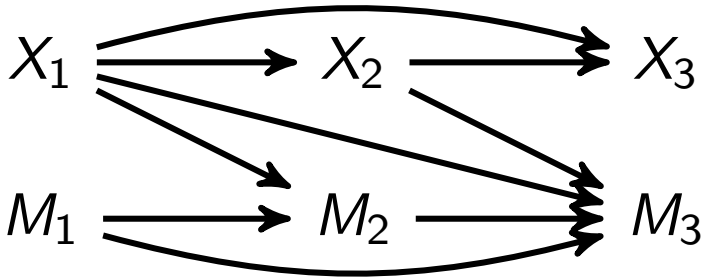
- Easy to do with available software.
- Can analyze completed data sets with regular complete-data methods, no need to know very much about missing data mechanism (except how to choose imputation model).
- Works pretty well in practice, in a range of settings.

Cons:

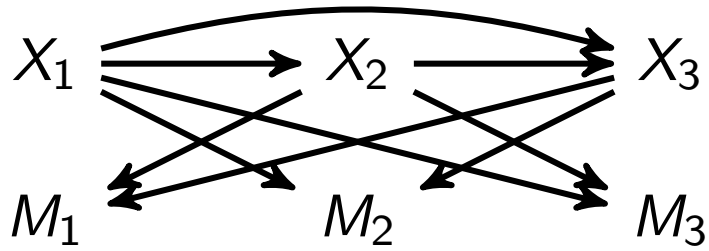
- Hard to specify imputation models in a way that is both flexible and compatible.
- Theoretical properties are not very well understood, and difficult to prove.
- Standard estimates of variance for your final parameter estimates $\hat{\theta}$ may not be correct, depending on how you calculate them. Also depends on using frequentist versus Bayesian sampling method.
- Can fail miserably if the true missingness mechanism is MNAR.

Independence assumptions for identifiability w/ MNAR

There are some MNAR missingness models under which $p(x)$ is nonparametrically identified. Some of these are conditional independence models, and so can be represented by graphs.



$$M_i \perp\!\!\!\perp X_i | X_{<i} \quad \forall i$$



$$M_i \perp\!\!\!\perp X_i | X_{-i} \quad \forall i$$

The No Self-Censoring Model

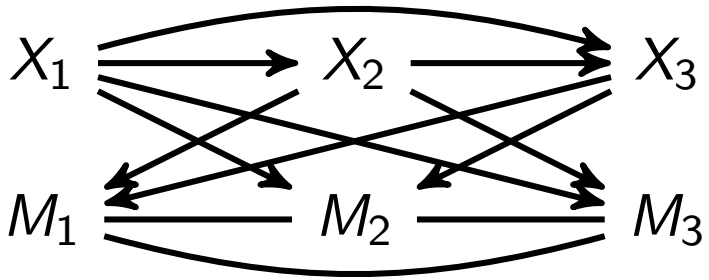


Figure: A chain graph representation of the "no self-censoring" independence model for 3 variables.

$$M_i \perp\!\!\!\perp X_i \mid M_{-i}, X_{-i} \quad \forall i$$

(a.k.a. "itemwise conditionally independent nonresponse")

Identifiability

Can prove that the missingness mechanism $p(m|x)$, and therefore $p(x)$, is (surprisingly!) identified from observed data under "no self-censoring" assumption.

Estimating some functional of $p(x)$ then can be accomplished by re-weighting the observed data $p(x, M = 1)$ by an estimate of $p(M = 1|x)$ based on this identifying formula. This is an example of *inverse probability weighting*.

With MNAR models, need to establish identifiability on a case-by-case basis, and then each may have different identifying formula for $p(M = 1|x)$.

Example

Under the NSC model for p missing variables, can show the following:

$$p(M = 1|x) = \frac{1}{Z(x)} \prod_{i=1}^{p} p(M_i = 1|M_{-i} = 1, x_{-i})$$

Where $Z(x)$ is a normalizing constant that is somewhat complicated to calculate but yet identified from the observed data (involves a mixture over all possible missingness patterns).

This implies that, for any target parameter $\beta = b(p(x))$ that is some smooth functional of the full data distribution, we can estimate β by solving a re-weighed estimating eq:

$$\sum_{j=1}^{n} \frac{\mathbb{I}(M^j = 1)}{p(M = 1|x^j; \widehat{\eta})} \phi_{\text{full}}(x^j; \beta) = 0$$

where $\phi_{\text{full}}(x; \beta)$ is the full-data estimating eq for β if there were no missing data.

Summary

When you have missing data, first you must determine an appropriate assumption/model for the missingness mechanism.

- If missingness status is completely independent of X, may do complete-case analysis.
- If missingness status is independent of missing information conditional on what is observed (MAR), there are various options: likelihood-based inference with EM, multiple imputation (usually easiest option), or semiparametric weighting estimators
- If missingness is not independent of missing information (MNAR), the full data dist. is typically not identifiable. However, may make some additional, context-specific assumptions and then use semiparametric weighting estimators in some circumstances
- Graphical models can be useful for studying/evaluating missingness assumptions

Things To Know About Latent Variables:

- You have no data on them (latent!), so any assumptions you make about LVs will be untestable
- LVs are a threat to valid causal inference ("ommited variable bias")
- LVs may be introduced to more compactly model complex joint distributions (reduce parameters)
- LVs may be used for lower-dimensional summaries of data
- LVs may be used for grouping observations into clusters
- LVs may be used for better prediction models
- LVs may correspond to factors/entities posited by scientific theories (e.g., psychological constructs) $\Rightarrow$ objects of interest in-themselves
- etc

Mixture Models

One use of LV modeling is mixture models: models that assume the observed density $p(x)$ is a convex combination of K cluster distributions

$$p(x;\theta) = \sum_{k=1}^{K} \lambda_k p_k(x;\theta_k)$$

where mixing weights λ_k satisfy $\lambda_k > 0$ and $\sum_{k=1}^{K} \lambda_k = 1$. You can imagine a sampling process

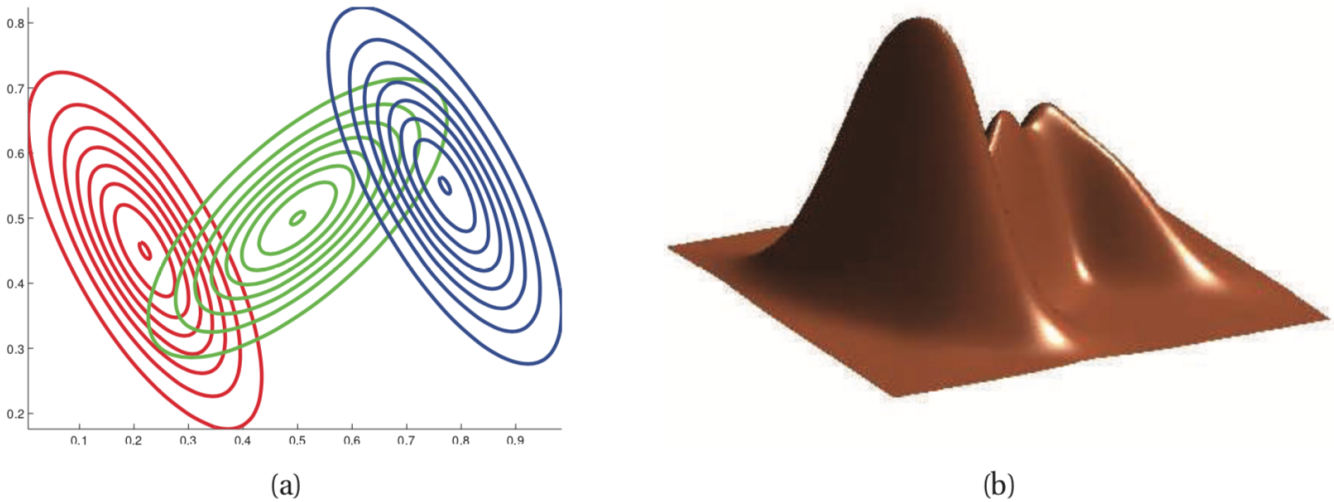
$$\begin{aligned} Z &\sim \text{Mult}(\lambda_1, ..., \lambda_K) \\ X|Z &\sim p_Z \end{aligned}$$

with Z discrete LV that indicates which "cluster" X is drawn from.

Mixture of Gaussians

$$p(x;\theta) = \sum_{k=1}^{K} \lambda_k p(x;\mu_k,\Sigma_k)$$

where each $p(x;\mu_k,\Sigma_k)$ is a Gaussian density



May replace Gaussians with other parametric families

Clustering

One common application of MMs is to assign data points to clusters. For data point x^j, calculate:

$$p(z^j = k|x^j;\theta) = \frac{p(z^j = k;\theta)p(x^j|z^j = k;\theta)}{\sum_{k'=1}^{K} p(z^j = k';\theta)p(x^j|z^j = k';\theta)}$$

Clustering

One common application of MMs is to assign data points to clusters. For data point x^j, calculate:

$$p(z^j = k|x^j;\theta) = \frac{p(z^j = k;\theta)p(x^j|z^j = k;\theta)}{\sum_{k'=1}^{K} p(z^j = k';\theta)p(x^j|z^j = k';\theta)}$$

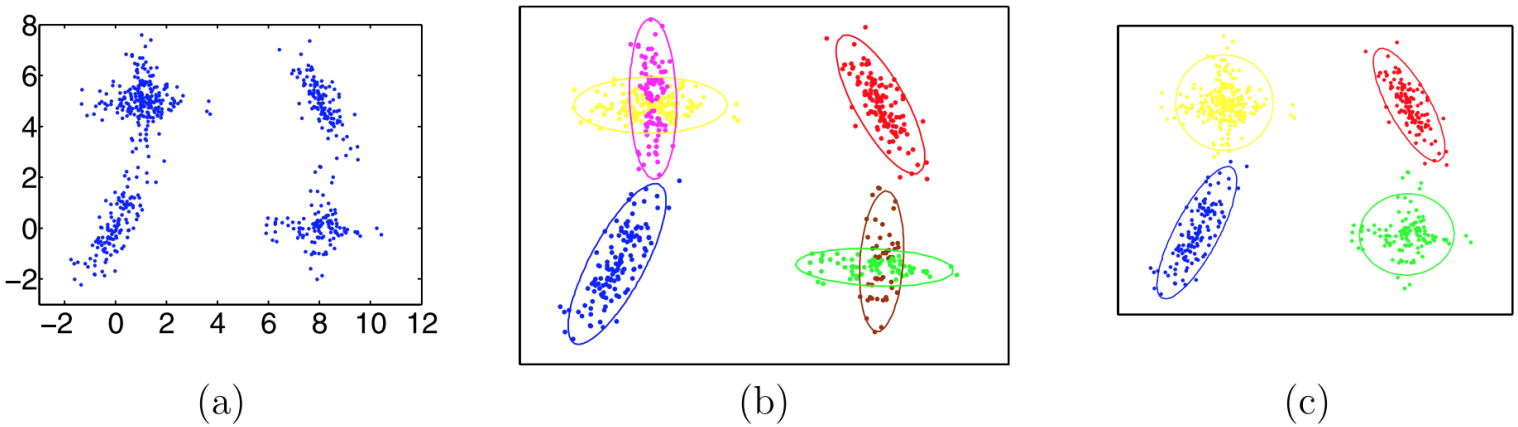
May choose cluster by

$$\arg\max_k p(z^j = k|x^j;\theta)$$

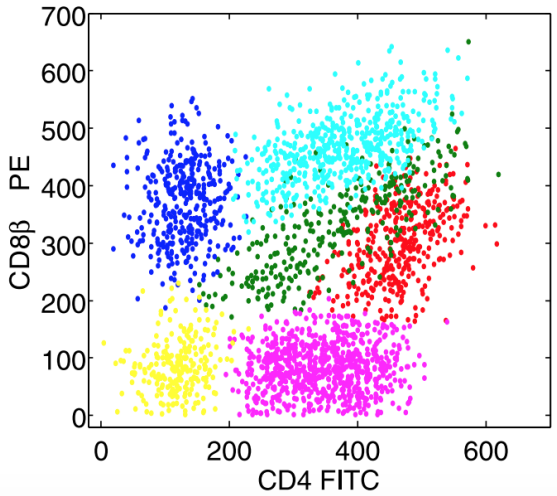
or

$$\arg\max_k \log p(x^j|z^j = k;\theta) + \log p(z^j = k;\theta)$$

Clustering with GMMs



Simulated data with $K = 6$ clusters ↑ Real flow cytometry data ↓



both from Baudry et al. (2010) "Combining mixture components for clustering." *Journal of Computational and Graphical Statistics*

Mixture of Experts

Mixture models can also be useful for prediction

$$Z|X = x \sim \text{Cat}(v(x))$$

$$Y|X = x, Z = k; \theta) \sim N(\beta_k^T x, \sigma_k^2)$$

This is a different linear regression model for each region of input space.

(Linear regression may be replaced by a more complicated "expert" model, e.g., a neural network.)

Estimating mixture models with EM

Generally, the parameters of a mixture model may be estimated with EM.

E-step: $Q(\theta, \theta^{t-1}) \equiv \mathbb{E}[\ell(\theta)|X = x, \theta^{t-1}]$
(expectation wrt $Z \mid X = x, \theta^{t-1}$)

M-step: $\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$

However, there can be identifiability problems if multiple parameter choices lead to the same observed distribution. One specific version of this is called **label degeneracy**, where we can swap labels among clusters without changing anything observable.

Latent Dirichlet Allocation for topic modeling

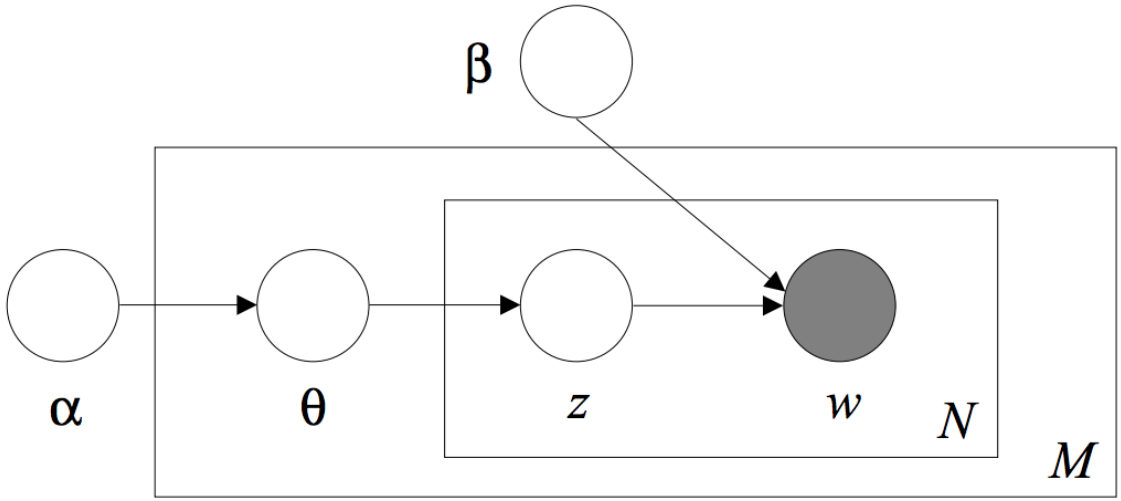
LDA is a smoothed mixture model for predicting "topics" of text documents

- A *word* w is an indicator from a finite vocabulary indexed by $\{1, ..., V\}$ with $w^v = 1$ for vth word and $w^u = 0$ for $u \neq v$
- A *document* is a sequence of N words $\mathbf{w} = (w_1, ..., w_N)$
- A *corpus* is a collection of M documents $D = \{\mathbf{w}_1, ..., \mathbf{w}_M\}$

Model:

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of the N words w_n:
   a. Choose a topic $Z_n \sim \text{Multinomial}(\theta)$
   b. Choose a word $w_n | Z_n = z_n \sim p(w_n | z_n; \beta)$

LDA



Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Blei et al (2003) "Latent Dirichlet Allocation." *Journal of Machine Learning Reserach*

LDA

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Blei et al (2003) "Latent Dirichlet Allocation." *Journal of Machine Learning Research*

Factor Analysis

$\mathbf{X}$ is an $n \times p$ data matrix in $\mathbb{R}^{n \times p}$. A Factor Analysis (FA) model decomposes the data matrix into a lower-dimensional part + noise:

$$\mathbf{X} = \mathbf{F}_q \mathbf{w}_q + \boldsymbol{\epsilon}$$
$$[n \times p] = [n \times q][q \times p] + [n \times p]$$

where $q < p$

Factor Analysis

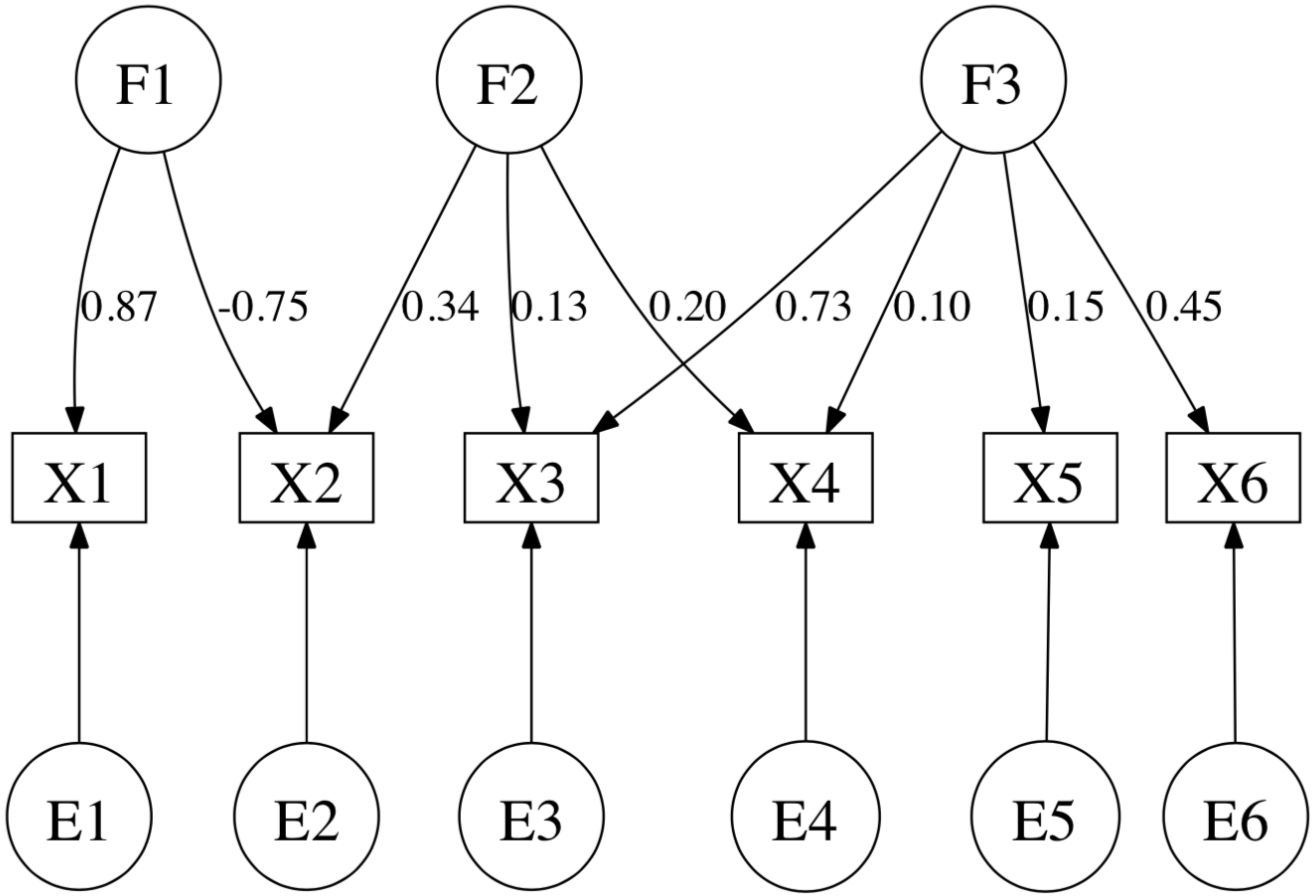
$$\mathbf{X} = \mathbf{F}_q \mathbf{w}_q + \boldsymbol{\epsilon}$$

This resembles a linear regression model. The q variables in $\mathbf{F}$ are considered latent, called "factor scores," with $\mathbf{w}$ a matrix of "factor loadings."

- We assume the variables X_i and F_i have mean zero and unit variance; the ϵ_i have mean zero. (Choice of scale, standardize)
- It is typical to assume that the F_i are mutually independent (and iid).
- It is typical to assume the ϵ_i are mutually independent and independent of the F_i (also iid).
- It is also typical to assume the LVs are jointly Gaussian, though this is not necessary.

Besides the choice of scale, the other assumptions are substantive: may be wrong for any given data set.

Factor Analysis



Assumes the dependence between Xs is *entirely* explained by the Fs.

Example from Shalizi (2019) *Advanced Data Analysis from an Elementary Point of View*

Covariance explained by the latents

$$
\begin{aligned}
Cov[X_1, X_2] &= \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2] = \mathbb{E}[X_1 X_2] \\
&= \mathbb{E}[(F_1 w_{11} + F_2 w_{21} + \epsilon_1)(F_1 w_{12} + F_2 w_{22} + \epsilon_2)] \\
&= \mathbb{E}[F_1^2 w_{11} w_{12} + F_1 F_2 (w_{11} w_{22} + w_{21} w_{12}) + F_2^2 w_{21} w_{22}] \\
&+ \mathbb{E}[\epsilon_1 \epsilon_2] + \mathbb{E}[\epsilon_1 (F_1 w_{12} + F_2 w_{22})] + \mathbb{E}[\epsilon_2 (F_1 w_{11} + F_2 w_{21})] \\
&= w_{11} w_{12} + w_{21} w_{22}
\end{aligned}
$$

by model assumptions.

Generally: $Cov[X_i, X_j] = \sum_{k=1}^{q} w_{ki} w_{kj}$ for $i \neq j$. We say observable i *loads onto* factor k when $w_{ki} \neq 0$.

Covariance matrix

Let ψ be the (diagonal) covariance matrix among the ϵs. Then

$$\Sigma \equiv \mathbb{E}[\frac{1}{n}\mathbf{X}^T\mathbf{X}] = \psi + \mathbf{w}^T\mathbf{w}$$

.
This is true even without assuming the variables are jointly Gaussian.

The FA model decomposes a (full rank) observed-data covariance matrix into something that is "low rank plus noise."

The Σ matrix is something we could estimate from the data... does that mean we can solve this equation to estimate $\mathbf{w}$? In general, no.

Unidentifiability

The FA model is unidentifiable. One way to see this is to consider an arbitrary rotation matrix $\mathbf{r}$, such that $\mathbf{r}^T\mathbf{r} = \mathbf{I}$ (identity). Define $\tilde{\mathbf{w}} \equiv \mathbf{wr}$. Then the likelihood function of the modified matrix is the same as for the unmodified matrix:

$$\begin{aligned}\Sigma &= \psi + \tilde{\mathbf{w}}^T\tilde{\mathbf{w}} \\ &= \psi + \mathbf{w}^T\mathbf{r}^T\mathbf{rw} \\ &= \psi + \mathbf{w}^T\mathbf{w}\end{aligned}$$

i.e., nothing has changed. Diff parameter settings give us the same observational consequences. Another way of seeing this is by counting degrees of freedom on both sides of the eq above, while imposing the assumptions (constraints) of the FA model. We get that the lhs has $p(p-1)$ dof in Σ (by symmetry, and unit diagonal) and $pq - q(q-1)/2$ degrees of freedom in $\mathbf{w}$ (by orthogonality). For some *specific* FA model specifications that place a sufficient number of constraints on the $\mathbf{w}$ parameters, the system of equations may be solvable.

Fitting identified FA models

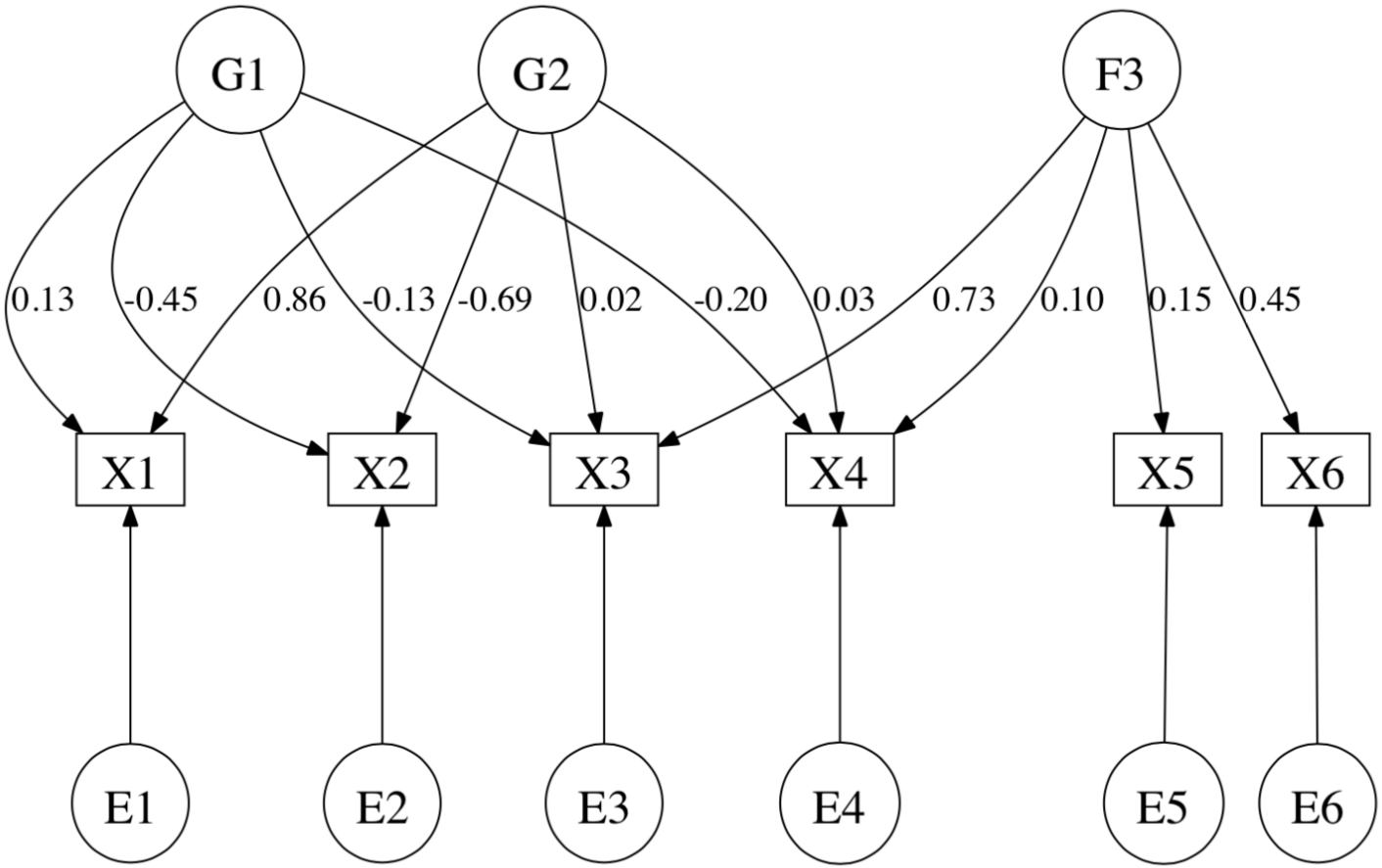
There are multiple ways to estimate the parameters: solve a system of eqs using empirical covariance matrix, use EM...

In any case, this will only make sense (will only have a unique solution) in cases where you've specified enough constraints to make the model identifiable.[2]

Sometimes a particular scientific theory (and interpretation of the latent factors) will yield sufficient constraints, and sometimes people use heuristics like forcing $\mathbf{w}$ to be lower-triangular (which endows no real interpretation of the factors, pretty arbitrary).

[2] If there is no unique maximizer of the likelihood, running EM will typically pick out one of a continuum of maximizers. You can still typically use these to get decent predictions on future observed data, but you should *not* treat these numbers as estimates of the true underlying parameters.

The rotation problem



$$\mathbf{r} = \begin{bmatrix} \cos 30 & -\sin 30 & 0 \\ \sin 30 & \cos 30 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Example from Shalizi (2019)

Goodness-of-fit

The rotated model will have *the exact same* goodness-of-fit (measured by some fitness test, like a likelihood ratio test or whatever[3]) as the "original" model $\Rightarrow$ different FA models *cannot* be distinguished by picking the model with the best "fit."

[3] Sometimes people calculate $R^2 = \frac{\sum_{j=1}^{q} \sum_{k=1}^{p} w_{jk}^2}{p}$ statistics to roughly measure "how much variance does my FA model explain." R^2 should *not* be used for model selection! See Shalizi (2019) for some explanation of why this is bad.

Summary of FA

- FA model can be viewed as a kind of data summary or reduction, summarizing the covariance relationships entirely by means of a small(er) number of latent variables. (Gaussian) FA is often described as a "low rank parameterization of a multivariate Gaussian distribution."
- FA model involves probabilistic modeling assumptions which may or may not be appropriate for a given setting.
- Can estimate parameters of a FA model with sufficient constraints to make model identifiable.
- FA can be used as a generative model which can be used to make predictions. If the Gaussian FA model is an accurate description of the DGP, then new observations will be drawn from a Gaussian with the specified (low rank + noise) covariance matrix.

Causal (latent entity) discovery

FA has its historical origins as an approach to causal discovery (or latent entity discovery), particularly in psychology. However, "Exploratory Factor Analysis" techniques are now known to be a particularly *bad* and *unreliable* way to do causal / latent discovery.[4]

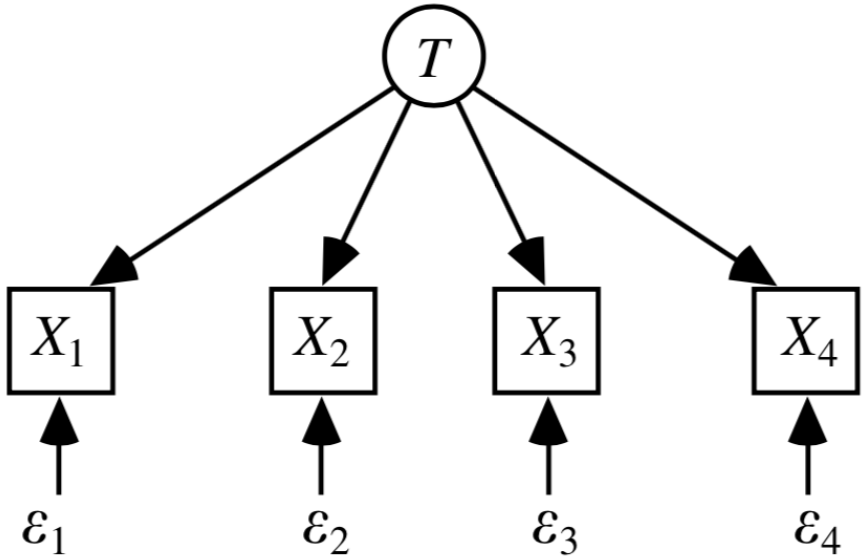
Typically these procede somewhat hueristically: first, determine number of LV by starting small and adding factors if they improve the "fit" of the model. When the fit stops improving, stop adding. Then, consider rotations of the parameters which satisfy some simplicity criterion, like lower-triangularity or sparsity.

You shouldn't do it! Such techniques have bad performance, not much better than chance. So, how would you do discovery *correctly*?

[4] You may use such techniques to select the *number* of latent factors more-or-less ok; just don't take anything but the number of factors too seriously, because of the rotation problem. "Exploratory FA" = model selection/search/discovery whereas "Confirmatory FA" = fitting a given model.

Historical origins

Charles Spearman (ca. 1904) was interested in the hidden structure of human intelligence. He noticed patterns of correlation among schoolchildren's test scores in different subjects. Specifically, test scores were correlated across subjects (math, English, history), which he posited could be explained by a single latent factor: *general intelligence*.



Historical origins

Spearman's model was basically

$$\mathbf{X} = \mathbf{G}\mathbf{w} + \boldsymbol{\epsilon}$$

where $\mathbf{G}$ is $[n \times 1]$ (one factor).

$$\sigma_{ij} \equiv Cov(X_i, X_j) = w_i w_j$$

Consider any 4 observed measures (test scores), i, j, k, l. We have:

$$\frac{\sigma_{ij}/\sigma_{kj}}{\sigma_{il}/\sigma_{kl}} = \frac{w_i w_j / w_k w_j}{w_i w_l / w_k w_l} = 1$$

so

$$\sigma_{ij}\sigma_{kl} = \sigma_{il}\sigma_{kj}$$

This is called a "tetrad equation" or "tetrad constraint" and it is empirically testable.

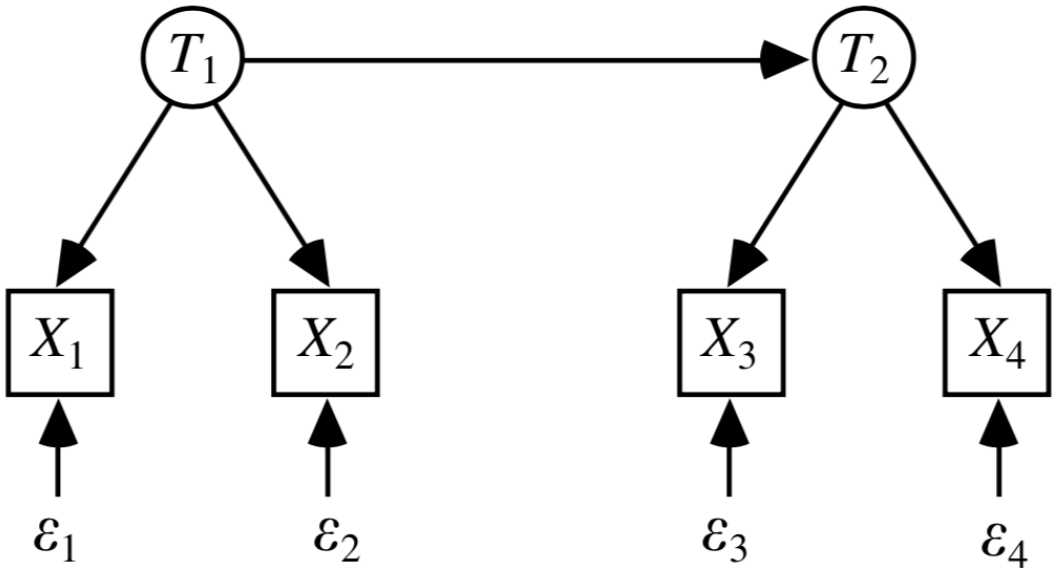
Historical origins

Spearman's observed that such constraints seemed to approximately hold in his data on schoolchildren test scores and concluded his "general intelligence" model was correct. (This was not a warranted conclusion!)

Later, such constraints seemed to hold in data on some combinations of scores but not others, which led future psychologists (e.g., L.L. Thurstone) to posit multiple factors, different "types" of intelligence.

Empirically distinguishing latent structures

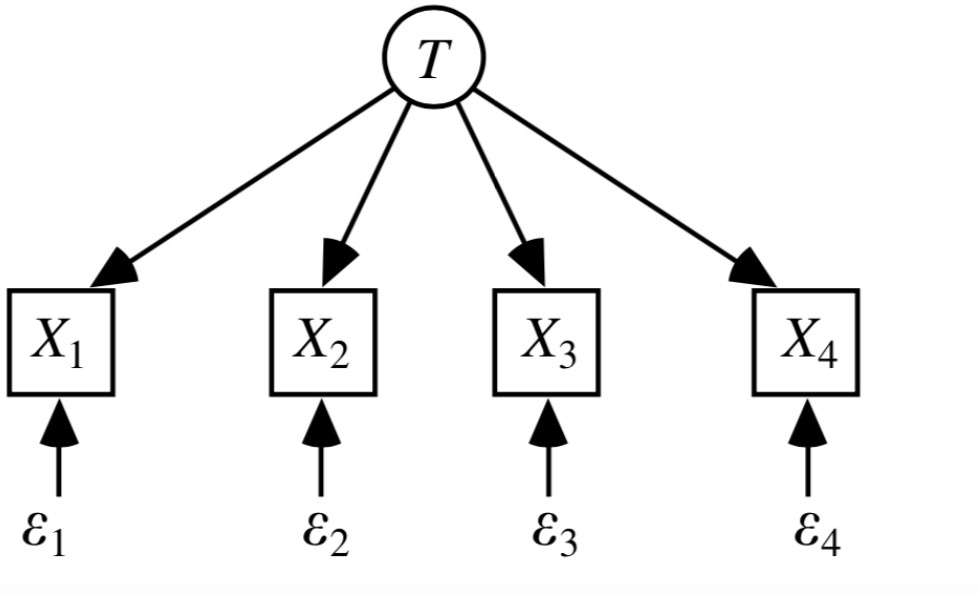
What's really *amazing* is that distinct structures can be distinguished on the basis of their implied tetrad constraints (assuming linearity).



implies only $\sigma_{13}\sigma_{24} = \sigma_{14}\sigma_{23}$

Empirically distinguishing latent structures

Whereas



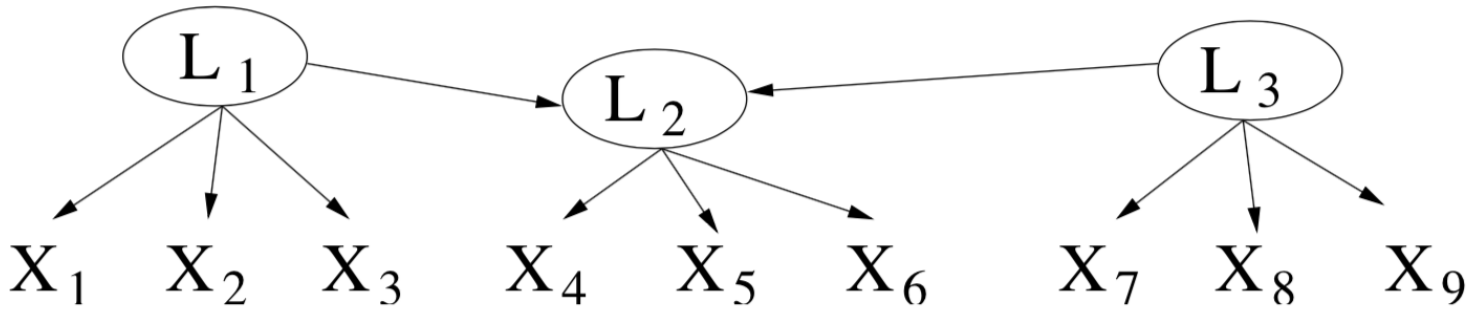
implies:

$\sigma_{13}\sigma_{24} = \sigma_{14}\sigma_{23}$

$\sigma_{12}\sigma_{34} = \sigma_{14}\sigma_{23}$

$\sigma_{13}\sigma_{24} = \sigma_{12}\sigma_{34}$

Empirically distinguishing latent structures



implies both that X_1, X_2, X_3 are independent of X_7, X_8, X_9 and $\sigma_{13}\sigma_{24} = \sigma_{14}\sigma_{23}$ (and others)

Observationally equivalent structures

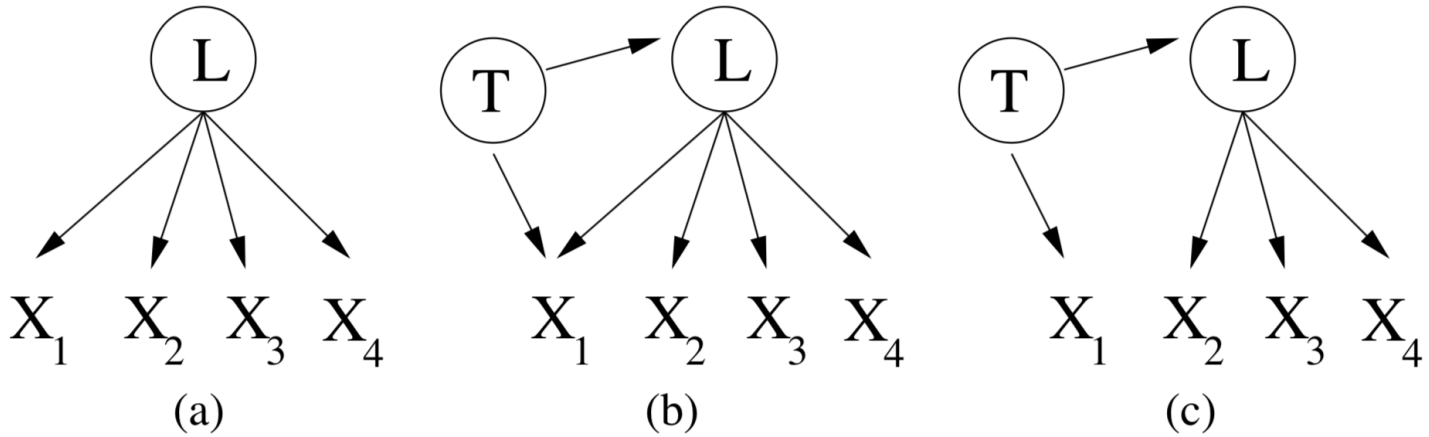


Figure 4: A linear latent variable model with any of the graphical structures above entails all possible tetrad constraints in the marginal covariance matrix of $X_1 - X_4$.

Just as with learning (fully observed) DAGs, we have to be careful about equivalent structures. In this case, we have to keep in mind that multiple structures may imply the same exact tetrad constraints.

Can we generalize these observations into some principles for latent model discovery?

Thoerem.[5] A directed acyclic graph $\mathcal{G}$ corresponding to a linear latent variable model entails tetrad constraint $\sigma_{ij}\sigma_{kl} = \sigma_{ik}\sigma_{jl}$ if and only if there is a vertex L (the choke point) in $\mathcal{G}$ that trek separates $\{X_i, X_l\}$ from $\{X_j, X_k\}$.

(Definition of trek separation on the next slide.)

[5] Called the "tetrad representation theorem." Due to Peter Spirtes in the late 1980s/early 1990s.

Trek separation

A trek in $\mathcal{G}$ from X_i to X_j $(i \neq j)$ is an ordered pair of directed paths (P_1, P_2) where P_1 has sink X_i, P_2 has sink X_j, and both P_1 and P_2 have the same source X_k. Note that one or both of P_1 and P_2 may consist of a single vertex, i.e., a path with no edges. A trek (P_1, P_2) is called *simple* if the only common vertex among P_1 and P_2 is the common source.

Let A, B, be two disjoint subsets of vertices V in $\mathcal{G}$, each with two distinct vertices as members. We let $T(A, B)$ and $S(A, B)$ denote the sets of all treks and all simple treks from a member of A to a member of B, respectively.

We say that the vertex X_i trek separates (or t-separates) A from B if for every trek in $S(A, B)$ P_1 contains X_i or for every trek in $S(A, B)$ P_2 contains X_i.

Rank reduction

A tetrad constraint is a rank reduction of a part of the covariance matrix, specifically a sub-matrix of the covariance matrix Σ over 4 variables X_i, X_j, X_k, X_l.

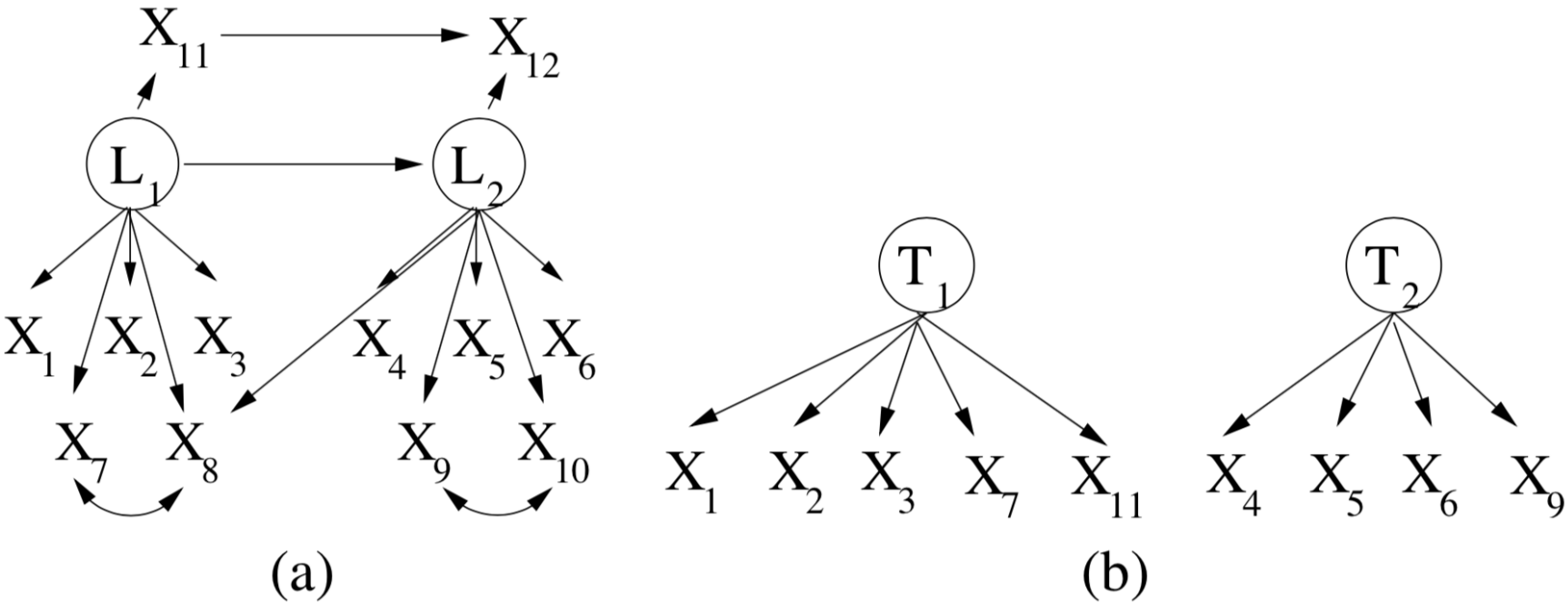
More recently, results in algebraic statistics have generalized the tetrad representation theorem connecting treks to more general rank reductions of submatrices of various sizes. (More on this later.)

BuildPureClusters

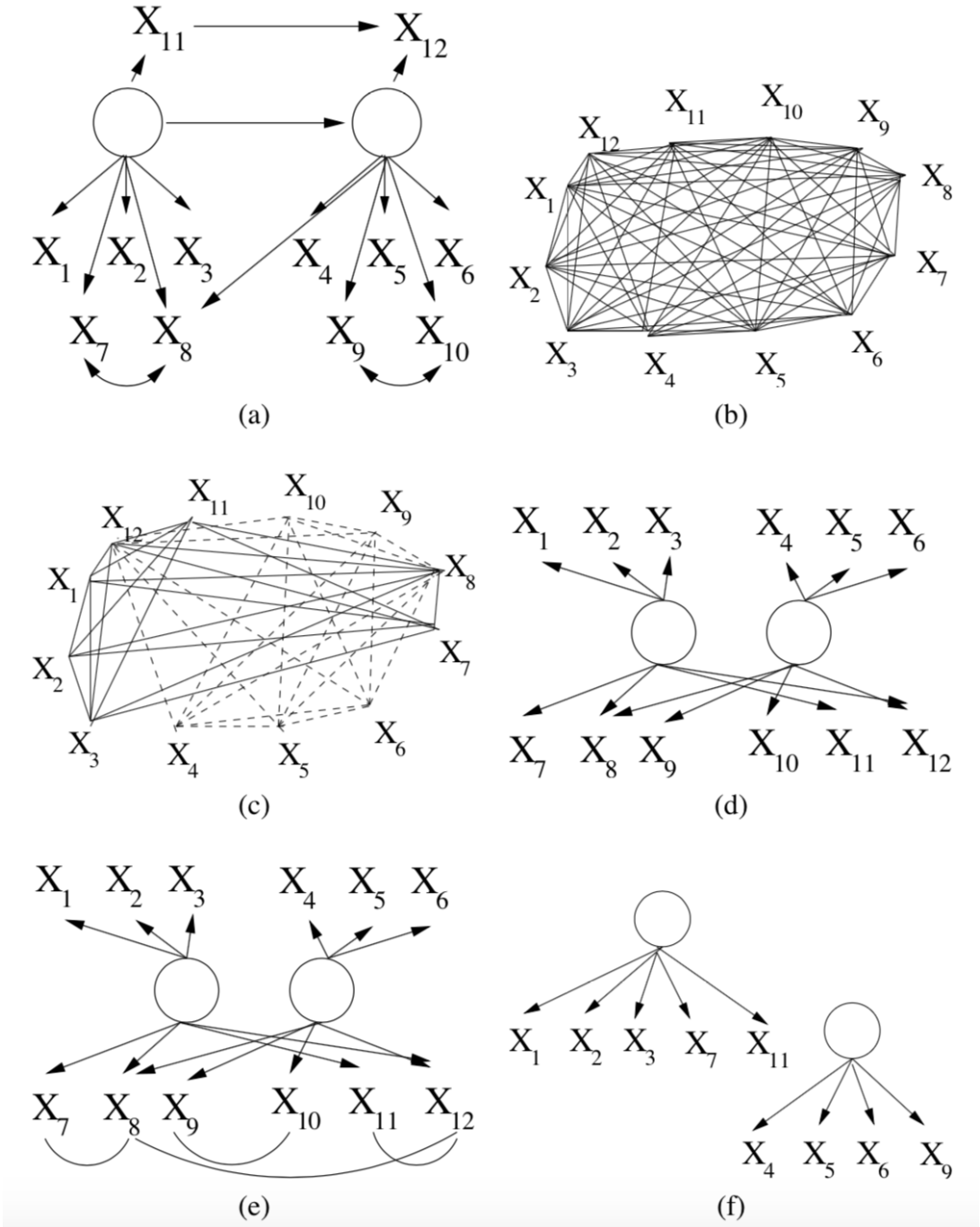
These graphical principles, combined with others based on d-separation, Markov equivalence, etc, have been exploited to design various learning algorithms with some provable gaurantees that they asymptotically discover true features of the underlying latent variable model (though not *everything* about the model). One example is a procedure called BuildPureClusters.[6]

BuildPureClusters focuses on extracting "pure measurement models" which are substructures in which each collection of observed variables (measures) has a single latent variable common parent, with no direct connections among the measures.

[6] Silva et al. (2006) "Learning the structure of linear latent variable models." *Journal of Machine Learning Research*



(a) Is the true structure and (b) is the pure measurement model discovered by BuildPureClusters


X11
X12
X1
X2
X3
X4
X5
X6
X7
X8
X9
X10
(a)
X11
X10
X9
X12
X8
X1
X7
X2
X6
X3
X4
X5
(b)
X11
X10
X9
X12
X1
X8
X7
X2
X6
X3
X4
X5
(c)
X1
X2
X3
X4
X5
X6
X7
X8
X9
X10
X11
X12
(d)
X1
X2
X3
X4
X5
X6
X7
X8
X9
X10
X11
X12
(e)
X1
X2
X3
X7
X11
X4
X5
X6
X9
(f)

Can we learn structure *among* the latents?

Latent variables may "cause" each other and/or exhibit patterns of conditional independence among themselves. Can we try to learn relationships among the unobserved?

Theorem. Let $\mathcal{G}$ be a pure linear latent variable model. Let L_1, L_2 be two latents in $\mathcal{G}$, and Q a set of latents in $\mathcal{G}$. Let X_1 be a measure of L_1, X_2 be a measure of L_2, and X_Q be a set of measures of Q containing at least two measures per latent. Then L_1 is d-separated from L_2 given Q in $\mathcal{G}$ if and only if the rank of the correlation matrix of $\{X_1, X_2\} \cup X_Q$ is less than or equal to $|Q|$ with probability 1 wrt the Lebesgue measure over the linear coefficients and error variances of $\mathcal{G}$.

PC-MIMBUILD, GES-MIMBUILD, FCI-MIMBUILD...

Given this result and the output of BuildPureClusters, one may use tests of rank reductions in the correlation/covariance matrix to decide d-separation relations among the latent variables. The combinations of BuildPureClusters with algorithms like PC, GES, etc. for structure learning among latents have been called PC-MIMBUILD, GES-MIMBUILD, etc.

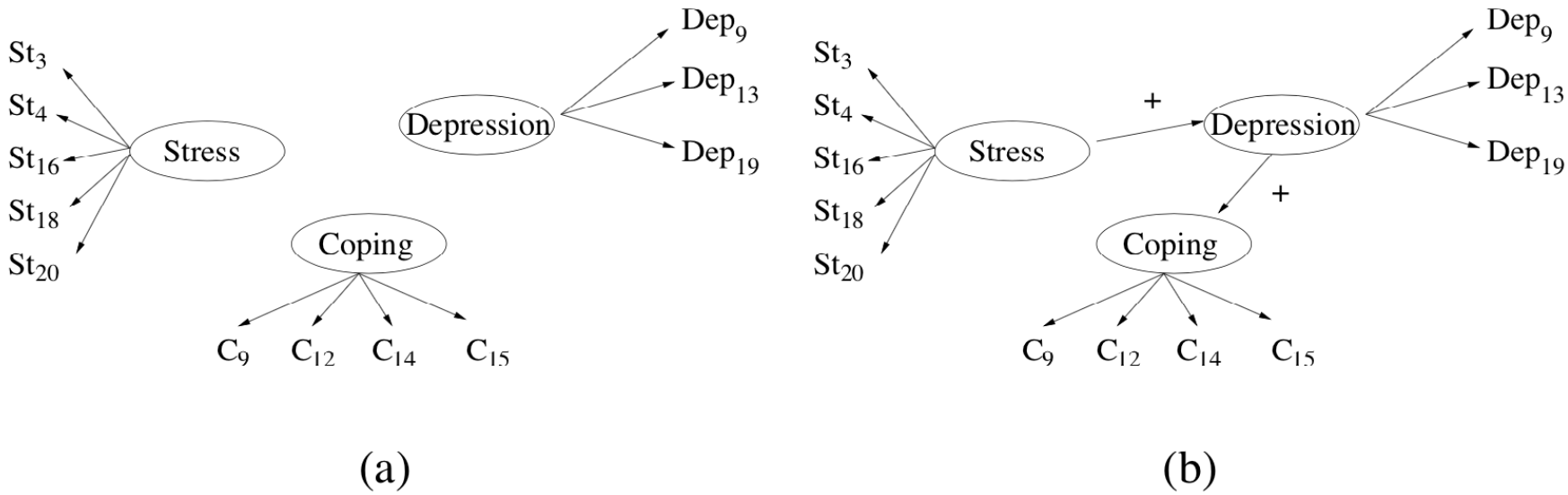


Figure 15: The output of BPC and GES-MIMBUILD for the coping study.

Results on data from 127 university students. "Coping" refers to religious/spiritual coping. A questionaire was designed to measure each latent variable with 20 or 21 questions. The final model incorporates the background knowledge/assumption that stress is not an effect of the other latent variables.

More recent generalizations...

A generalization of the previous definition of trek-separation:

Let A, B, C_A, C_B, be four subsets of vertices V in $\mathcal{G}$ which need not be disjoint. We say that the pair (C_A, C_B) trek separates (or t-separates) A from B if for every trek (P_1, P_2) from a vertex in A to a vertex in B, either P_1 contains a vertex in C_A or P_2 contains a vertex in C_B.

Thereom.[7] The submatrix $\Sigma_{A,B}$ has rank less than or equal to r for all covariance matrices consistent with the graph $\mathcal{G}$ if and only if there exist subsets $C_A, C_B, \subset V$ with $\#C_A + \#C_B \leq r$ such that (C_A, C_B) t-separates A from B.

$\Sigma_{A,B}$ is the submatrix of the covariance matrix with row indices A and column indices B.

[7] Sullivant et al. (2010) "Trek separation for Gaussian graphical models." *Annals of Statistics*

More recent generalizations...

These generalized rank constraints ("n-tad constraints") have been used to develop algorithms which are more flexible, allowing for multiple latents to cause subsets of the measured variables, allowing some "impurities" (direct connections among measures) and so forth. The algorithms get more complicated but the basic principles are the same.