Stat 521, Graphical Models and Parsimonious Dependence
Prof Harry Joe, Fall 2015, 2015.10.26–2015.12.05
Introductory lecture.

Outline

- Examples of graphical models (trees, vines, path diagrams, Bayesian networks, independence graphs): for many types of graphical objects, conditional independence relations can be deduced. For some graphs, two specific variables Y and Y' are conditionally independent given the variables that are in the middle of the path between Y and Y'. Consider graphs to summarize association, ignore possibility of causation; directed edges make sense for longitudinal or "expert judgment".
- Examples of parsimonious dependence for multivariate Gaussian dependence model: form of correlation matrix and precision matrix (inverse of covariance matrix). Simple definition $d \times d$ correlation matrix has $O(d)$ parameters as d increases.
- What extends to multivariate non-Gaussian.

Graphs, conditional independence, parsimony

Graphs, defined in different ways, **are set of nodes and edges; edges are line segments joining two nodes**. *Locally*, 2 nodes that connected have more dependence that 2 nodes that are not connected.

A graphical object provides information on conditional independence relations; it does not imply a multivariate distribution.

To get a multivariate model, quantities or (conditional) distributions are assigned to the edges of the graphical object.

Parsimonious dependence means that there are (i) few edges, say $O(d)$ edges in d variables, and (ii) lots of conditional independence relations.

Examples of parsimonious dependence structures for multivariate normal/Gaussian

S1. exchangeable (compound symmetric)

S2. AR(1), antedependence of order 1

S3. AR(p), antedependence of order p (linear order of variables, e.g. longitudinal)

S4. Markov tree (Markov order 1)

S5. p-truncated vine (partial order of variables)

S6. 1-factor (dependence explained by latent or unobserved variable)

S7. p-factor (dependence explained by several latent variables)

S8. combination: factor model with Markov structure on latent variables, examples are simplex structure and hidden Markov

S9. Bayesian network (parsimonious form for a (recursive) sequence of regressions)

S10. parsimonious form for regression of each variable on remaining variables

Why parsimonious?

Important when dimension d is large. If model has $O(d)$ parameters, maybe algorithms have computational complexity $O(d)$.

Why multivariate normality?

For some of these dependence models, good estimation may depend on sample size n large enough and not on whether $n < d$ or $n > d$ (d =dimension). If n is small, multivariate normality assumption is often used for mathematical convenience. With n or the order of several hundreds or in the thousands, univariate plots usually show deviation from normality (skewness or heavier tails than exponential, or both).

If variables are indivually non-normal, they can be converted to normal with probability integral transforms after fitting appropriate models. Then bivariate plots can be looked add. These bivariate plots can also show deviation from bivariate normality (tail asymmetry or tail dependence or both).

The normality assumption can rarely be justified based on variable description. From the Central Limit theorem (general form with independent random variables), the normality assumption is plausible only for variables that can be considered as sums or averages of many components (with no dominant component).

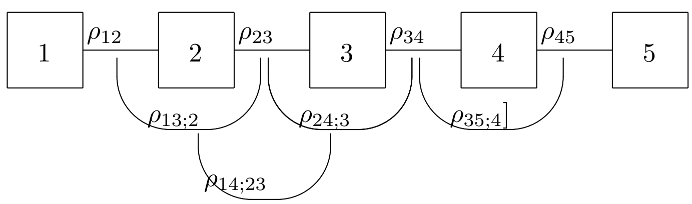
Nevertheless, dependence models based on multivariate normal can be viewed as first order models. In the copula literature, useful parsimonious dependence structures have been extended to non-Gaussian. The Gaussian dependence models must be understood in order to be extended.

Why graphical representation?

- Show links of variables with the strongest dependence
- Show links that may be "causal"
- Show how to algorithmically compute a joint probability distribution from components

---

Graph 1: Markov model for time series/longitudinal data
one form of drawing a **vine** (compare, the grape vine).



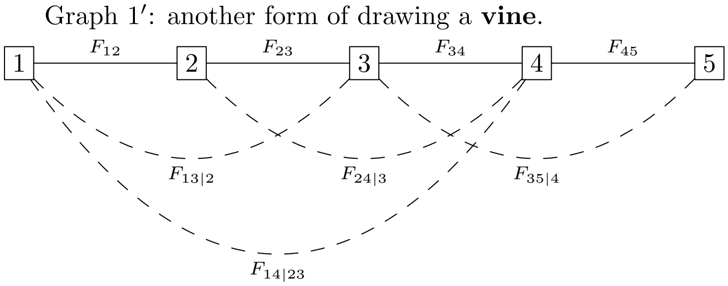
If stationary, $\rho = \rho_{12} = \rho_{23} = \rho_{34} = \cdots$ is lag 1 serial correlation, $\alpha_2 = \rho_{13;2} = \rho_{24;3} = \cdots$ is partial autocorrelation of lag 2.

Markov process, Gaussian AR(1): only edges are 12, 23, 34, 45 and variables two or more apart are conditionally independent, for example, partial autocorrelations $\rho_{13;2}, \rho_{24;3}, \rho_{14;23}$ are 0.

Markov order 2 process, Gaussian AR(2): only edges are 12, 23, 34, 45, $13;2$, $24;3$, $35;4$, and variables apart by ≥ 3 edges are conditionally independent, for example, $\rho_{14;23}, \rho_{25;34}, \rho_{15;234}$ are 0.

For non-Gaussian, correlations are replaced by bivariate marginal distributions and partial correlations are replaced by bivariate conditional distributions.

Graph 1′: another form of drawing a **vine**.

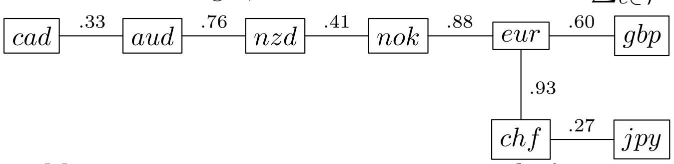


Same idea for non-linear ordering of variables:
Markov tree and truncated vines.

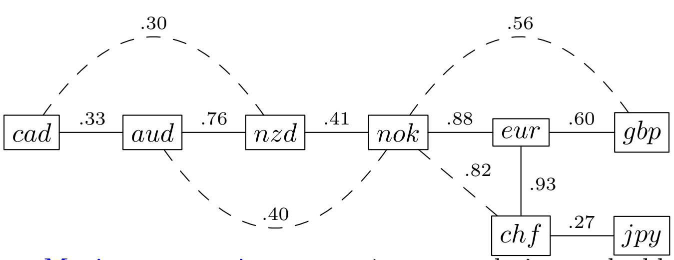
Foreign exchange example to motive non-linear ordering,
years 2000–2002, $n = 780$
Correlation matrix of normal scores of log returns of exchange rates (relative to USD).

	EUR	AUD	CAD	CHF	GBP	JPY	NOK	NZD
EUR	1.000	0.403	0.134	0.930	0.603	0.239	0.876	0.408
AUD	0.403	1.000	0.332	0.361	0.273	0.148	0.397	0.756
CAD	0.134	0.332	1.000	0.120	0.079	0.127	0.147	0.296
CHF	0.930	0.361	0.120	1.000	0.587	0.274	0.824	0.379
GBP	0.603	0.273	0.079	0.587	1.000	0.242	0.558	0.279
JPY	0.239	0.148	0.127	0.274	0.242	1.000	0.234	0.125
NOK	0.876	0.397	0.147	0.824	0.558	0.234	1.000	0.410
NZD	0.408	0.756	0.296	0.379	0.279	0.125	0.410	1.000

Graph 2: **Maximum spanning tree with absolute correlations on edges**; draw complete graph, put correlations on edges; find subtree that minimizes $\sum_{e\in\mathcal{T}} \log(1-\rho_e^2)$.



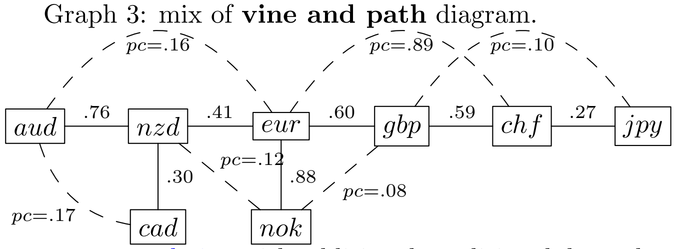
Maximum spanning tree or **1-truncated vine**, not a good fit (too parsimonious); chf $\perp\!\!\!\perp$ gbp given eur etc.. Note that Markov assumption implies other correlations are obtained from products along edges — this is not a good approx. for 0.76×0.41 not near 0.40, etc.



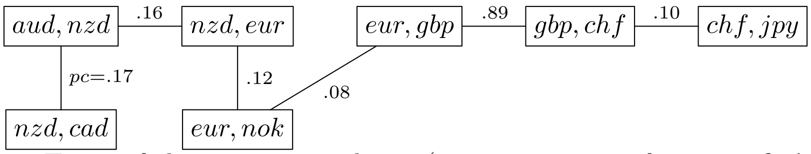
Maximum spanning tree or 1-truncated vine, embedded in graph,

---

Graph 3: mix of **vine and path** diagram.

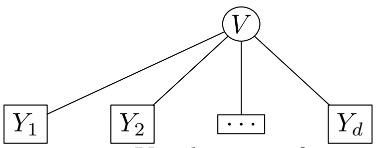


2-truncated vine with additional conditional dependence for variables that are 2-apart (compare AR(2)). Tree 1 is a modification of maximum spanning tree. Partial correlation nzd,nok|eur is 0.12 etc. If 2-truncated vine does not explain all dependence, add more edges of form aud,gbp|nzd,gbp. Solid edges: tree1; dashed edges: tree2, etc. **Vine is a sequence of trees satisfying some proximity conditions.**

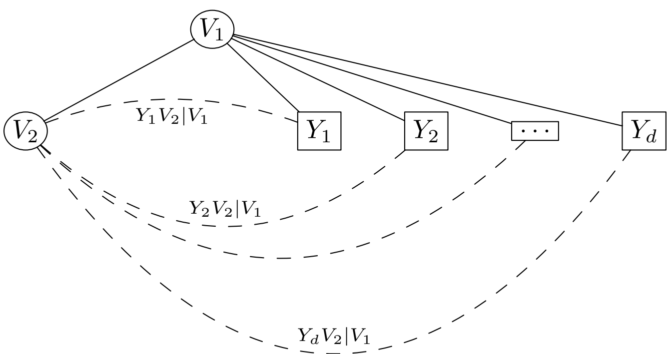


Tree 2 of above 2-truncated vine (common picture of **tree 2 of vine**); edges of tree1 are nodes of tree2. Draw graph of all edges that satisfy proximity condition — also have [nzd,eur]–[eur,gbp], partial correlation: 0.05; [nzd,cad]–[nzd,eur], partial correlation: 0.02

Graphs 4 and 5, Latent variable models or factor analysis:
mentioned in classical multivariate statistics books for parsimonious dependence. For graphs, **latent variables in circles, observed variables in rectangles** (psychometrics).

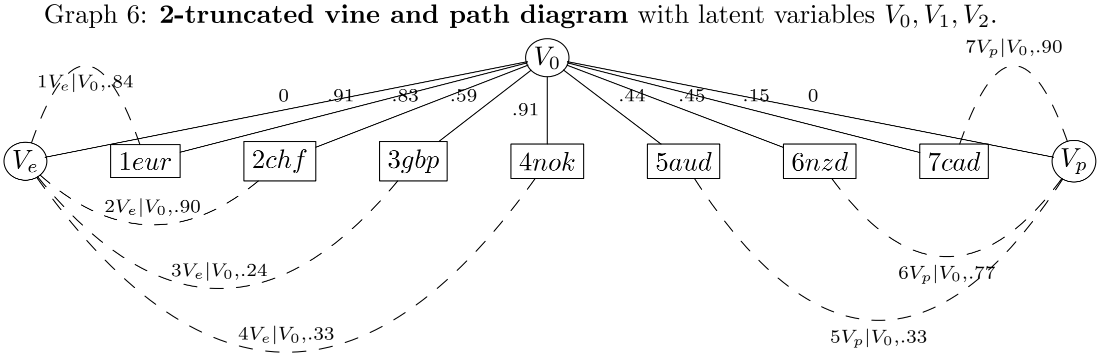


V is latent; 1-factor model (**Markov tree**) with edges $[VY_1], \ldots, [VY_d]$, $Y_j \perp\!\!\!\perp Y_k | V$, $j \neq k$

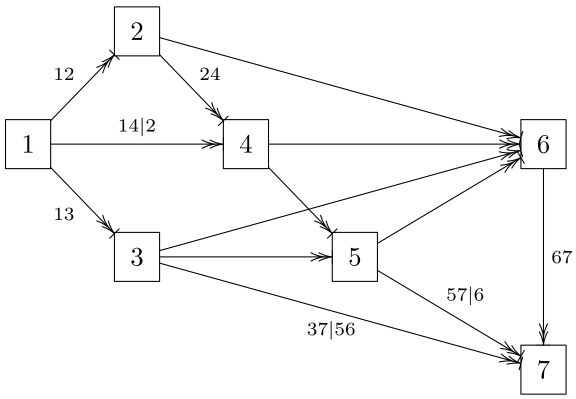


V_1, V_2 latent; 2-factor model (**2-truncated vine**) with edges $[V_1V_2], [V_1Y_1], \ldots, [V_1Y_d]$. for tree $\mathcal{T}_1$ and $[V_2Y_1|V_1], \ldots, [V_2Y_d|V_1]$ for tree $\mathcal{T}_2$. $Y_j \perp\!\!\!\perp Y_k | V_1, V_2$, $j \neq k$.

Graph 6: **2-truncated vine and path diagram** with latent variables V_0, V_1, V_2.



bi-factor with 2 groups (latent variables in circles, 4 European currencies in group 1, 3 Pacific currencies in group 2; good fit, edges have correlations and partial correlations. V_0 affects all currencies (relative to USD); V_1 affects European currencies, V_2 affects Pacific currencies.



Graph 7: DAG=directed acyclic graph.

Joint density $f_1 f_{2|1} f_{3|1} f_{4|12} f_{5|34} f_{6|2345} f_{7|356}$. One non-unique edge labeling (non-uniqueness matters for non-Gaussian only) is:

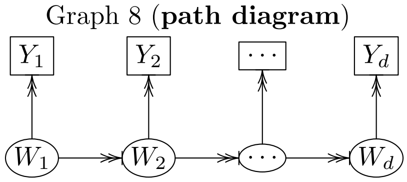
$$12; 13; 24, 14|2; 35, 45|3; 56, 36|5, 46|35, 26|345; 67, 57|6, 37|56$$

$\rho_{jk;S}$ for Gaussian, $C_{jk;S}$ for pair-copulas where $F_{jk|S} = C_{jk;S}(F_{j|S}, F_{k|S})$. **Conditional independence:** $Y_3 \perp\!\!\!\perp Y_2|Y_1$, $Y_4 \perp\!\!\!\perp Y_3|(Y_1, Y_2)$, $Y_5 \perp\!\!\!\perp (Y_1, Y_2)|(Y_3, Y_4)$, $Y_6 \perp\!\!\!\perp Y_1|(Y_2, Y_3, Y_4, Y_5)$, $Y_7 \perp\!\!\!\perp (Y_1, Y_2, Y_4)|(Y_3, Y_5, Y_6)$. Expert opinion or learning algorithm (numbers assigned to variable after graph).

Simplex structure for longitudinal or for variables that can be linearly ordered (latent AR). $Y_1, \ldots, Y_d$ are ordered by time or difficulty, etc. W_j is latent variable associated to Y_j. $\{W_1, \ldots, W_d\}$ is AD(1) or Markov order 1.

Concrete example from Guttman (1954): 9 verbal ability tests for 437 Chicago eight-grade school children; the $d = 9$ test scores Y_j are ordered by degree of difficulty.

Graph 8 (**path diagram**)



From graphs to structural covariance matrices.

Parsimonious dependence is mostly based on some form of conditional independence relations (Gaussian or non-Gaussian); for Gaussian, $d \times d$ correlation matrix is a function of fewer than $d(d-1)/2$ parameters.

For Gaussian, **conditional dependence** is represented by **partial correlations** (conditional independence relation corresponds to a partial correlation that is zero).

For some forms of graphical models, one can see the conditional independence relations from the graph or another summary.

For truncated vines, there is conditional independence of Y and Y' given the variables in the path between them. Edges represent correlations or partial correlations in the Gaussian case.

There are parsimonious dependence models that don't have graphical representations.

Examples of parsimonious dependence for multivariate Gaussian dependence model: form of $d \times d$ correlation matrix and precision matrix (inverse of covariance matrix). Simple definition $d \times d$ correlation matrix has $O(d)$ parameters as d increases.

1-parameter dependence models are: (i) [**S1**] exchangeable correlation matrix $\rho_{jk} = \rho$, $j \neq k$, $-1/(d-1) \leq \rho \leq 1$; (ii) [**S2**] AR(1) stationary $\rho_{jk} = \rho^{|j-k|}$, $j \neq k$, $-1 \leq \rho \leq 1$.

Dependence models with d or $d-1$ parameters: (iii) [**S6**] 1-factor dependence $\rho_{jk} = \alpha_j\alpha_k$ where α_j is the correlation of variable j with latent variable; (iv) [**S4**] tree for d variables has $d-1$ edges, $\rho_{jk} = \rho_e$ is correlation if $e = [jk]$ is an edge of the tree, otherwise $\rho_{jk} = \prod_{e\in P_{jk}} \rho_e$ where P_{jk} is the set of edges in the path from variable j to k.

Dependence models with $2d$ or $2d-3$ parameters: (v) [**S7**] 2-factor dependence (2 latent variables); (vi) [**S5**] 2-truncated vine: $d-1$ edges in tree 1, and tree 2 has $d-2$ edges for variables that are 2-apart in tree 1;

Dependence models with $3d$ or $3d-6$ parameters: (v) [**S7**] 3-factor dependence (3 latent variables); (vi) [**S5**] 3-truncated vine.

Also, Bayesian network [**S9**] with parent sets of at most 2 have $2d-3$ parameters are fewer; Bayesian network with parent sets of at most 3 have $3d-5$ parameters are fewer.

Combinations [**S8**] of 1-factor and Markov tree can have $2d-1$ parameters

---

Precision matrix, parsimony and independence graphs

Other parsimonious dependence models [**S10**] are based on assuming 0s in the **precision (inverse covariance/correlation) matrix** $\mathbf{\Sigma}^{-1} = (\sigma^{jk})_{1\le j,k\le d}$.

A (conditional) **independence graph**, according to Whittaker (1990), **has edges between two variables** j,k **only if** $\sigma^{jk} \neq 0$. In the multivariate Gaussian case, $\sigma^{jk} = 0$ implies that variables j and k are conditionally independent given variables in the set $T(j,k) = 1,\dots,d\backslash\{j,k\}$.

A more meaningful interpretation comes from the regressions of each variable on the remaining variables.

In-class exercises

1. Check where precision matrix has zeros for some special cases:
AR(1), AR(2), Markov tree, 1-factor, Bayesian network. Hence draw the independence graphs for the preceding examples.

For multivariate Gaussian, positions are zeros in precision matrix are easily determined from a sequence of recursive regression equations (or products of conditional densities of the form in a Bayesian network).

2. Exchangeable and AR(1) dependence; compare the various graphical objects: independence graph, vine (possibly with latent), path diagram, Bayesian network.
Precision matrix of exchangeable correlation: $[1+(d-2)\rho)](1-\rho)^{-1}[1+(d-1)\rho]^{-1}$ on diagonal and $-\rho(1-\rho)^{-1}[1+(d-1)\rho]^{-1}$ on off-diagonal, with $-(d-1)^{-1} \le \rho \le 1$.

Summary for correlation matrices of 1-factor and Markov tree (the two simplest structures)

1-factor with loading vector $(\alpha_1,\dots,\alpha_d)$:

$$\begin{pmatrix} 1 & \alpha_1\alpha_2 & \alpha_1\alpha_3 & \cdots & \alpha_1\alpha_d \\ \alpha_2\alpha_1 & 1 & \alpha_2\alpha_3 & \cdots & \alpha_2\alpha_d \\ \alpha_3\alpha_1 & \alpha_3\alpha_2 & 1 & \cdots & \alpha_3\alpha_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_d\alpha_1 & \alpha_d\alpha_2 & \alpha_d\alpha_3 & \cdots & 1 \end{pmatrix} \quad \text{compare rows } j,k, \text{ in other columns: similarly ordered.}$$

Markov tree:

$$\begin{pmatrix} \cdots & j & \cdots & k & \cdots & m & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 1 & \cdots & \rho_{jk} & \cdots & \rho_{jm} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \rho_{kj} & \cdots & 1 & \cdots & \rho_{km} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \rho_{mj} & \cdots & \rho_{mk} & \cdots & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

for any triple of variables Y_j, Y_m, Y_k with a path covering all three, weakest correlation is product of other 2. Easiest to check for linear ordering (D-vine) or star ordering (C-vine).

Exercise: derive from sequence of linear (regression) equations

Note: the techniques used to solve these exercises or techniques that will be later developed in this course, will be useful to research areas besides graphical models.

Book references I. Graphical models.

- Whittaker (1990). Graphical Models in Applied Multivariate Statistics. Wiley. [independence graphs]
- Cox and Wermuth (1996). Multivariate Dependencies: Models, Analysis and Interpretation. Chapman&Hall. [precision matrix, chain graphs]
- Kurowicka and Cooke (2006). Uncertainty Analysis with High Dimensional Dependence Modelling. Wiley. [vines, Bayesian networks, briefly more general indep. graph]
- Mulaik (2009). Linear Causal Modeling with Structural Equations. Chapman&Hall/CRC. [path diagrams, latent variable models]
- Hojsgaard, Edwards and Lauritzen, (2012). Graphical Models with R. Springer. [Bayesian networks]
- Pourahmadi (2013). High-dimensional Covariance Estimation. Wiley. [precision matrix, sparse PCA]
- Joe (2014). Dependence Modeling with Copulas. Chapman&Hall/CRC. [vines, path diagrams, factor models]
- Scutari and Denis (2015). Bayesian Networks with Examples in R. Chapman&Hall/CRC.

II. Gaussian time series and classical multivariate

- Brockwell and Davis (1987). Time Series: Theory and Methods. Springer.
- Johnson and Wichern (2002). Applied Multivariate Statistical Analysis, fifth edition. Prentice Hall,