Random graph methods

October 16, 2018

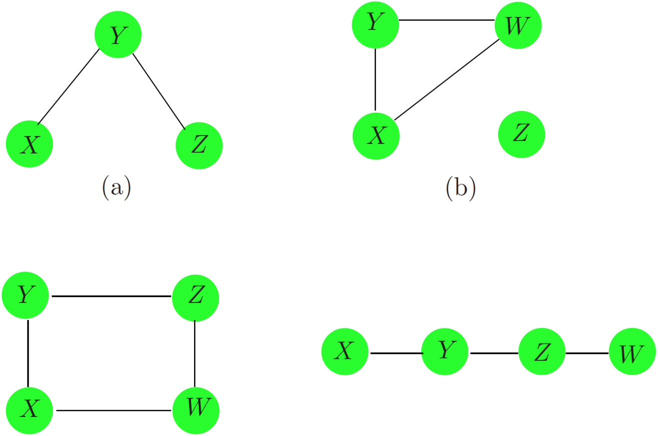
Graphs and Trees – a poetic point of view

> A dead tree, cut into planks and read from one end to the other, is a kind of line graph, with dates down one side and height along the other, as if trees, like mathematicians, had found a way of turning time into form.

Alice Oswald, British Poet

Unidirectional graph

- A graph consists of a set of **vertices (nodes)**, along with a set of **edges** joining some pairs of the vertices.



(a)

(b)

(c)

(d)

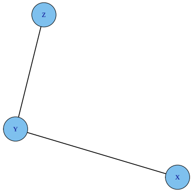
Graph – a map of random dependencies

- Let each vertex correspond to (represent) a random variable.
- The graph gives a visual way of understanding the joint distribution of the entire set of random variables.
- In this approach, the absence of an edge between two vertices has a special meaning: the corresponding random variables are conditionally independent, given the other variables (represented by other vertices).
- Such a graph does not tell a full story about the model but helps understand dependencies and search for them.
- If one specifies the model than the graphs plus some parameters for the distributions completely defines the model.

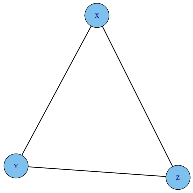
Simple examples

- **Example:** Let X_1, X_2, X_3 be independent random variables.

What is the graph for $X = X_1 + X_2$, $Y = X_2$, and $Z = X_2 + X_3$?



What is the graph for $X = X_1 + X_2$, $Y = X_1 + X_3$, and $Z = X_2 + X_3$?



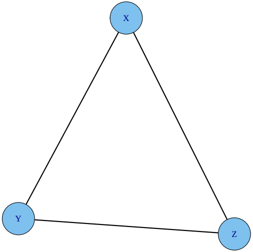
What is the graph for $X = X_1$, $Y = X_2$, and $Z = X_1 + X_2 + X_3$?

?

How to plot graphs in **R**

```
install.packages("igraph") #only if not installed before!!!
library(igraph)

edges = matrix(c("Y","Z","X","Z","X","Y"), nrow=3, ncol=2, byrow=T)
g = graph.edgelist(edges, directed=FALSE)
plot(g, edge.width=2, vertex.size=30, edge.color='black')
```



Specific models for distributions

- Without further specification of the model is difficult to say what kind of dependence one have.
- Interpretation of graphs is difficult unless some distributional structure is imposed.
- One needs specify models for distributions to make complete answers.
- Two models are popular:
    - For continuous variables: **Gaussian** models
    - For discrete variables: **Ising** model (**Boltzman machines**)

Fundamentals

- We assume that the observations have a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- There are several important properties of Gaussian distributions:
  - The distribution is specified by pairwise covariances plus means.
  - Conditional distributions are always Gaussian.
  - The covariances of conditional distributions do not depend on the values of variables on which conditioning is taken but only on $\boldsymbol{\Sigma}$.
  - The independence of two variables (the lack of edge between corresponding nodes) means that the conditional covariance (given all other variables) is zero – these conditional covariances are called **partial covariances** .
  - The inverse of covariance $\boldsymbol{\Sigma}$, often called the **precision matrix**

$$\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$$

tells when partial covariances are zero (lack of an edge):

zero in the precision matrix is equivalent to zero of the corresponding partial correlation

Partial covariances vs. the precision matrix

Partial correlation:

- Partial correlation can be formulated in terms of the projections of the observations to the subspaces
  - Let X_i and X_j be two coordinates in $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{X}_{\widehat{ij}}$ be the vector of all remaining coordinates.
  - Y_i - the residual from the orthogonal (the least square) projection of X_i to $\mathbf{X}_{\widehat{ij}}$.
  - Y_j - the residual from the orthogonal projection of X_j to $\mathbf{X}_{\widehat{ij}}$.
  - $n \times n$ matrices of partial covariances and partial correlations

$$\mathbf{PC} = [\langle Y_i, Y_j \rangle]\,, \quad \mathbf{R} = [\rho_{ij}] = \left[\frac{\langle Y_i, Y_j \rangle}{\|Y_i\| \|Y_j\|} \right]$$

Precision matrix:

- The inverse $\mathbf{\Theta}$ of covariance $\mathbf{\Sigma}$ of X_i's – the **precision matrix**

$$\rho_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

Formulation of the problem

- We can view our model as a graph with edges marked with values of partial covariances and the vertices marked by the mean values.
- By splitting conceptually the model into
  1. Graph that represents dependencies,
  2. Means
  3. Partial covariances associated with each edge

  we can divide the main problem of fitting the data to Gaussian density into three parts
  1. Estimate the means at each vertex
  2. Estimate the structure of the graph
  3. Given an estimate structure of the graph estimate partial covariances
- The means can be simply estimated by the mean values of variable corresponding to this vertex
- Estimating the rest is difficult

Given the structure estimate the covariances

- Given a number N of values of X's, we would like to estimate the correlations (partial correlations) corresponding to an undirected graph that is representing the non-zero partial correlations.
- Suppose first that the graph is complete (fully connected).
    - It is well known that the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is the sample covariance matrix

      $$\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

    - So in this case, the estimate is straightforward
- Suppose now that there are some edges missing in the actual graph of the partial covariances.
- The problem of finding an estimate given these constraints is non-trivial.

Multivariate normal (Gaussian) distribution

Everyone believes in Gauss distribution: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact.

Quote attributed to Henri Poincaré by de Finetti. However, Cramer attributes the remark to Lippman and quoted by Poincaré) *Gabriel Lippman* – a Nobel prize winner in physics, *Henri Poincaré* – a mathematician, theoretical physicist, engineer, and a philosopher of science

- The multivariate normal or Gaussian random vector $\mathbf{X} = (X_1, \ldots, X_p)$ is given by density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

  that is characterized by: a vector parameter $\boldsymbol{\mu}$ and a matrix parameter $\mathbf{\Sigma}$.

- The notation $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ should be read as *"the random vector* $\mathbf{X}$ *has multivariate normal (Gaussian) distribution with the vector parameter* $\boldsymbol{\mu}$ *and the matrix parameter* $\mathbf{\Sigma}$*."*

Multivariate normal (Gaussian) distribution – properties

We often drop the dimension p from the notation writing $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- The vector parameter $\boldsymbol{\mu}$ is equal to the mean of $\mathbf{X}$ and the matrix parameter $\boldsymbol{\Sigma}$ is equal to the covariance matrix of $\mathbf{X}$.
- Any coordinate X_i of $\mathbf{X}$ is also normally distributed, i.e. X_i has $\mathcal{N}(\mu_i, \sigma_i^2)$.
- If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A}$ is a $q \times p$ (non-random) matrix, $q \leq p$, (and the matrix $\mathbf{A}$ is of the rank q), then

$$\mathbf{AX} \sim \mathcal{N}_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

Subsetting from coordinates of MND

Any vector made of a subset of different coordinates of X is also multivariate normal with the corresponding vector mean and covariance matrix.

More precisely, if $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

are partitioned into sub-vectors $\mathbf{X}_1 : q \times 1$ and $\mathbf{X}_2 : (p-q) \times 1$ then with

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\mathbf{X}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \text{ and } \mathbf{X}_2 \sim \mathcal{N}_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

Conditional distributions

If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

$$\mathbf{X} = \left[\begin{array}{c} \mathbf{X}_1 \\ \mathbf{X}_2 \end{array} \right]$$

are partitioned into sub-vectors $\mathbf{X}_1 : q \times 1$ and $\mathbf{X}_2 : (p-q) \times 1$ then with

$$\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{array} \right] \quad \text{and} \quad \boldsymbol{\Sigma} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right]$$

the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$, is

$$\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2 \sim \mathcal{N}_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

Regression reinterpretation of conditional distributions

Vector $\mathbf{X}_1$ given $\mathbf{X}_2$ forms a regression model

$$\mathbf{X}_1 = \mathbf{a} + \mathbf{D}\mathbf{X}_2 + \epsilon,$$

where

- The constant term $\mathbf{a} = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2$
- The design matrix $\mathbf{D} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$
- The error term $\epsilon \sim \mathcal{N}_q(0, \boldsymbol{\Sigma}_{11} - \boldsymbol{D}\boldsymbol{\Sigma}_{21})$

Special case $\mathbf{X}_1 = (X_i, X_j)$ – calculating partial covariances

Partial covariance matrix

Recall that the partial covariance 2×2 matrix $\mathbf{\Sigma}_{ij}$ of (X_i, X_j) is given at the covariance of their distribution conditionally all other variables:

$$(X_i, X_j) = (a_i, a_j) + \mathbf{D}\mathbf{X}_2 + \epsilon,$$

- The constant term $(a_i, a_j) = (\mu_i, \mu_j) - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2$, where $\mathbf{\Sigma}_{12}$ is made of the ith and jth rows of of $\mathbf{\Sigma}$ without the ith and jth coordinates in these rows, thus it is $2 \times (p-2)$ matrix, $\mathbf{\Sigma}_{22}$ the covariance matrix with out the ith and jth columns and rows, thus it is a $(p-2) \times (p-2)$ matrix, $\boldsymbol{\mu}_2$ the mean values with the μ_i and μ_j values dropped.
- The $2 \times (p-2)$ design matrix $\mathbf{D} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$,
- The error term $\epsilon \sim \mathcal{N}_q(0, \mathbf{\Sigma}_{ij})$, $\mathbf{\Sigma}_{11} - \boldsymbol{D}\mathbf{\Sigma}_{21}$, $\mathbf{\Sigma}_{21}$ is the transpose of $\mathbf{\Sigma}_{12}$
- The(i, j)th partial correlation θ_{ij} is the correlation in the covariance matrix $\mathbf{\Sigma}_{ij}$, i.e. of the diagonal term divided by square roots of the diagonal terms.

Estimation of the partial correlations

We divided the problem of estimation of the given model into parts

1. Estimate the means at each vertex
2. Estimate the structure of the graph
3. Given an estimate structure of the graph estimate partial covariances

We briefly discuss the third part.

Organisation of observations

The observations in a sample are arranged in a $n \times p$ matrix $\mathbf{X}$ where n is the number of experimental units (the size of the sample) and p is the number of variables.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \dots & x_{jk} & \dots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np} \end{bmatrix}$$

Vector notation

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_j^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

Row *j* in this matrix

$$\mathbf{x}_j^T = [x_{j1}\ x_{j2}\ \ldots\ x_{jk}\ \ \ldots, x_{jp}]$$

is a *p*-dimensional observation.

Sample mean vector

Given a sample mean for a variable *i*

$$\bar{x}_i = \frac{1}{n} \sum_{k=1}^{n} x_{ki}$$

we define the sample mean vector as

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Sample covariance matrix

Given sample covariances

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

between variables *i* and *j* we define the (sample) covariance matrix

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}.$$

Sample correlation matrix

Given sample correlations

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

between variables *i* and *j*, we define the (sample) correlation matrix

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}.$$

Estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be n independent observations of $\mathbf{X}$ and let

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{X}_k$$

Then

$$\mathbb{E}(\bar{\mathbf{X}}) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}(\mathbf{X}_k) = \mathbb{E}(\mathbf{X}_k)$$

Estimation of $\boldsymbol{\mu}$

The mean vector $\bar{\mathbf{X}}$ is an unbiased estimator of $\boldsymbol{\mu}$.

Estimation of $\boldsymbol{\Sigma}$

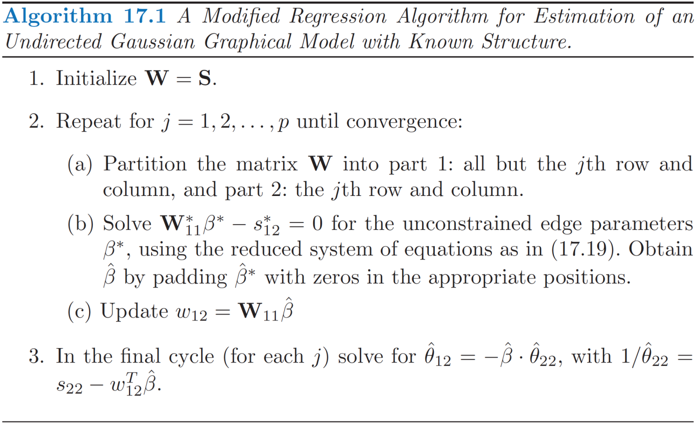
It holds that

$$\mathbb{E}(\mathbf{S}) = \boldsymbol{\Sigma}$$

Estimation of $\mathbf{\Sigma}^{-1}$ given zeros of partial covariances

- Estimating $\mathbf{\Sigma}^{-1}$ is equivalent to estimating the matrix of partial correlations, let us consider θ_{ij} as the (i,j)th-term of $\mathbf{\Sigma}^{-1}$.
- The maximizing likelihood under constraint that some of θ_{ij}s are zero can not, in general, solved analytically.
- Numerical algorithms have to be implemented.

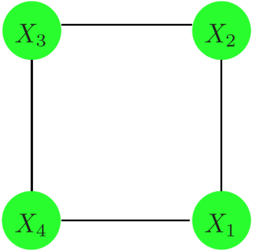
Algorithm – estimating partial correlations given a graph structure



Algorithm 17.1 *A Modified Regression Algorithm for Estimation of an Undirected Gaussian Graphical Model with Known Structure.*

```
1. Initialize $\mathbf{W} = \mathbf{S}$.
2. Repeat for $j = 1, 2, \ldots, p$ until convergence:
    (a) Partition the matrix $\mathbf{W}$ into part 1: all but the $j$th row and column, and part 2: the $j$th row and column.
    (b) Solve $\mathbf{W}_{11}^* \beta^* - s_{12}^* = 0$ for the unconstrained edge parameters $\beta^*$, using the reduced system of equations as in (17.19). Obtain $\hat{\beta}$ by padding $\hat{\beta}^*$ with zeros in the appropriate positions.
    (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$
3. In the final cycle (for each $j$) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = s_{22} - w_{12}^T \hat{\beta}$.
```

Example



$$\mathbf{S} = \begin{pmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{pmatrix}$$

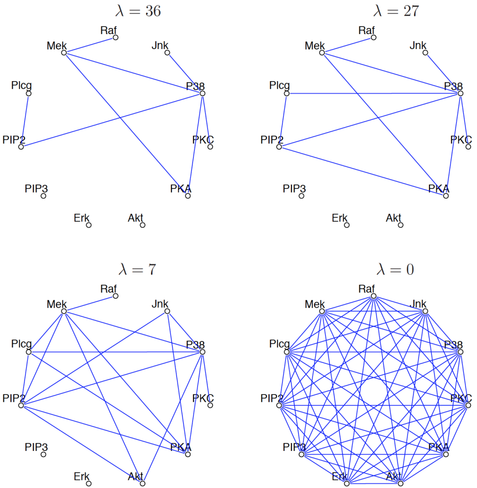
$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 10.00 & 1.00 & 1.31 & 4.00 \\ 1.00 & 10.00 & 2.00 & 0.87 \\ 1.31 & 2.00 & 10.00 & 3.00 \\ 4.00 & 0.87 & 3.00 & 10.00 \end{pmatrix}, \quad \hat{\mathbf{\Sigma}}^{-1} = \begin{pmatrix} 0.12 & -0.01 & 0.00 & -0.05 \\ -0.01 & 0.11 & -0.02 & 0.00 \\ 0.00 & -0.02 & 0.11 & -0.03 \\ -0.05 & 0.00 & -0.03 & 0.13 \end{pmatrix}$$

Estimating the graph structure – Lasso method

- In most cases we do not know which edges to omit from the graph
- One would like to try to discover this from the data itself.
- The lasso penalty by maximizing the penalized log-likelihood

$$\log \det \boldsymbol{\Theta} - \mathrm{trace}(\boldsymbol{S}\boldsymbol{\Theta}) - \lambda \|\boldsymbol{\Theta}\|_1$$

Example – the flow-cytometry



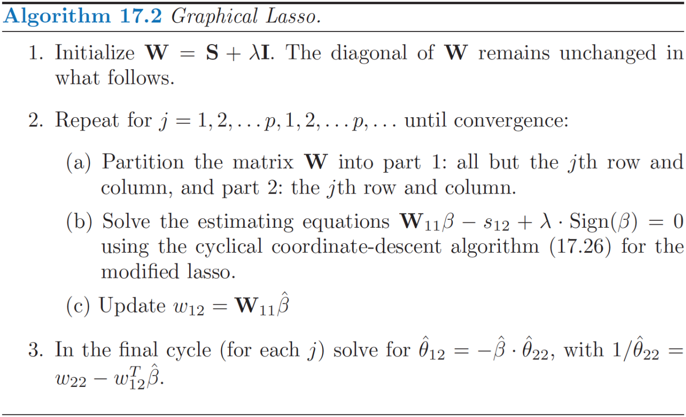
$\lambda = 36$

$\lambda = 27$

$\lambda = 7$

$\lambda = 0$

Algorithm – estimating the graph structure



Algorithm 17.2 *Graphical Lasso.*

1. Initialize $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}$ remains unchanged in what follows.

2. Repeat for $j = 1, 2, \dots p, 1, 2, \dots p, \dots$ until convergence:

   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.

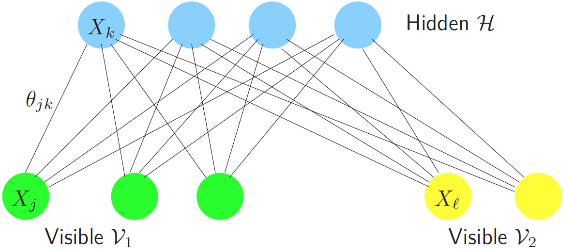
   (b) Solve the estimating equations $\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0$ using the cyclical coordinate-descent algorithm (17.26) for the modified lasso.

   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$

3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}$.

General comments

- Undirected Markov networks with all discrete variables are popular
- Pairwise Markov networks with binary variables are the most common – Ising models
- The values at each node can be observed (‘visible’) or unobserved (‘hidden’) – the so called Boltzman machines assume no interactions between hidden nodes
- The nodes are often organized in layers, similar to a neural network.
- These models are useful both for unsupervised and supervised learning, especially for structured input data such as images, but have been hampered by computational difficulties.



Some details

- Denoting the binary valued variable at node j by X_j, the Ising model for their joint probabilities is given by

  $$p(X, \boldsymbol{\Theta}) = \exp\left[\sum_{j,k} \theta_{jk} X_j X_k - \Phi(\boldsymbol{\Theta})\right]$$

  with $X \in 0, 1^p$.
- Only pairwise interactions are modeled.
- The Ising model was developed in statistical mechanics, and is now used more generally to model the joint effects of pairwise interactions.
- $\Phi(\boldsymbol{\Theta})$ is the log of the partition function, and is defined by

  $$\Phi(\boldsymbol{\Theta}) = \log \sum_{x} \exp\left(\sum_{i,j} \theta_{jk} x_j x_k\right)$$

  The partition function ensures that the probabilities add to one over the sample space.

Example – handwritten digits

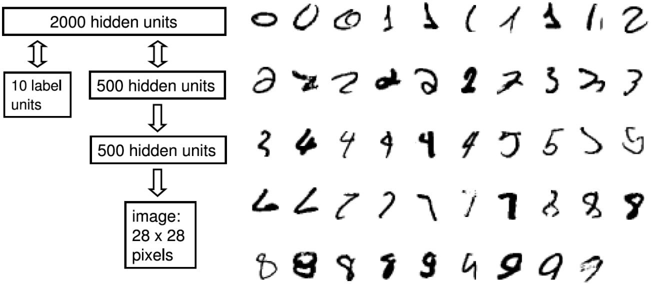


FIGURE 17.7. *Example of a restricted Boltzmann machine for handwritten digit classification. The network is depicted in the schematic on the left. Displayed on the right are some difficult test images that the model classifies correctly.*

Software

- Packages in R: *Igraph*, *lars*, *glasso*
- Other: *Julia*, *NetLogo*