Statistics for Neural Netters

Warren S. Sarle
SAS Institute
saswss@unx.sas.com

January 26, 1994

Statistics:

- Theory of decision making under uncertainty
- Making inferences about a population based on information contained in a sample

Population: the set of all potential measurements in which you are interested

Sample: the set of measurements that are actually available

Parameter: a function of the population values

Statistic: a function of the sample values

A statistic is to a sample as
a parameter is to a population.

Some Statistical Ideas

- Estimation
- Hypothesis testing
- Model selection
- Bayesian inference
- Decision theory
- Diagnostics
- Exploratory data analysis
- Graphical displays
- Experimental design
- Nonparametric inference
- Robust inference
- Time series

Distributions, Probabilities, Densities

- A population of values can be described in terms of its **cumulative distribution function**, which gives the proportion of the population less than each possible value.
- A discrete population can also be described by a **probability function** giving the proportion of the population equal to each possible value.
- A continuous population can often be described by a **density function**, which is the derivative of the cumulative distribution function. A density function can be approximated by a histogram giving the proportion of the population lying within each of a series of intervals of values. A probability density function is like a histogram with an infinite number of infinitely small intervals.
- A value randomly selected from a population is a **random variable** (RV).

Sampling Assumptions

Samples can be selected in a variety of ways.

- Many statistical methods assume that the data constitute a **simple random sample**, which means that the sample was selected in such a way that all possible samples were equally likely to be selected.
- An alternative assumption is that the values are independent and have identical distributions (iid).

Statistical Inference

Statistics from a sample can be used to make **inferences**, or reasonable guesses, about the parameters of a population.

For example, if you take a random sample of thirty children from a school, the mean height for those thirty children is a reasonable guess, or **estimate**, of the mean height of all the children in the school.

If the sample is random, other statistics, such as the standard error, can provide information about how good the estimate is likely to be.

Basic Formulas

Value from sample or population: x_i

Sample size: n

Population size: N

Sample mean: $\bar{x} = \frac{\Sigma x_i}{n}$

Population mean: $\mu = \frac{\Sigma x_i}{N}$

Sample variance: $s2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$

Population variance: $\sigma 2 = \frac{\Sigma (x_i - \bar{x})^2}{N}$

Standard deviation $= \sqrt{\text{variance}}$

Sampling Distributions

If you repeatedly draw samples of size n from a population and compute the mean of each sample, then the sample means themselves have a distribution.

Consider a new population consisting of the means of all the samples that could possibly be drawn from the original population.

The distribution of this new population is called a **sampling distribution**.

Sampling Distributions (2)

If the original population has mean μ and standard deviation σ, then the sampling distribution of the mean has:

- mean μ
- standard deviation $\sigma/\sqrt{n}$

The standard deviation of the sampling distribution of the mean is called the **standard error of the mean**.

The standard error of the mean provides an indication of the accuracy of a sample mean as an estimator of the population mean.

The larger the sample, the smaller the standard error.

Central Limit Theorem

If the original population has a normal distribution, then the sampling distribution of the mean is also normal.

If the original distribution is not normal but does not have excessively long tails, then the sampling distribution of the mean can be approximated by a normal distribution for large sample sizes.

Confidence Intervals

The probability that a normal RV is within 2 standard deviations of the mean is about 95%.

In repeated sampling, the probability that a sample mean is within 2 standard errors of the population mean is approximately 95%.

The probability that the random interval from $\bar{x} - 2\sigma/n$ to $\bar{x} + 2\sigma/n$ contains the population mean is approximately 95%. This interval is called a **confidence interval** (CI) for the mean.

If σ is unknown, a 95% CI can be computed as $\bar{x} - t_{.025}s/n$ to $\bar{x} + t_{.025}s/n$ where $t_{.025}$ is computed from Student's t distribution.

Prediction Intervals

A CI measures the uncertainty in estimating the population mean.

A **prediction interval** (PI) measures the uncertainty in predicting a new observation sampled from the same population.

The PI is $\bar{x} - 2\sigma(1 + 1/n)$ to $\bar{x} + 2\sigma(1 + 1/n)$

Jargon

Neural	Statistical
Pattern	Observation, Case
Input	Independent variables, explanatory variables, predictors, regressors
Output	Predicted values
Training values Target values	Dependent variables, Observed values, Responses
Errors	Residuals
Noise	Error term

Jargon (2)

Neural	Statistical
Weight	Parameter estimate, Regression coefficient
Bias	Intercept
Training, Learning, Adaptation	Estimation, Model fitting

Simple Linear Regression

n Sample size

x Independent variable, input

y Dependent variable, target

β Parameters

b Parameter estimates, weights

ϵ Error, noise: mean 0, st. dev. σ

r Residual, error

$\hat{y}$ Predicted value, output

Population (true) model for $i = 1$ to n:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Estimated model for $i = 1$ to n:

$$\begin{aligned} y_i &= b_0 + b_1 x_i + r_i \\ \hat{y}_i &= b_0 + b_1 x_i \end{aligned}$$

Multivariate Multiple Linear Regression

n	Sample size
X	$n \times p$ matrix of independent variables
Y	$n \times q$ matrix of dependent variables
B	$p \times q$ matrix of parameters
$\hat{B}$	$p \times q$ matrix of parameter estimates
E	Error, noise: mean 0, covariance matrix Σ
R	Residual
$\hat{Y}$	Predicted values

Population (true) model:

$$Y = XB + E$$

Estimated model:

$$\begin{aligned} Y &= X\hat{B} + R \\ \hat{Y} &= X\hat{B} \end{aligned}$$

Linear Regression Formulas

Least-squares parameter estimates:
$\hat{B} = (X'X)^{-}(X'Y)$

Predicted values: $\hat{Y} = X\hat{B}$

Residuals: $R = Y - \hat{Y}$

Estimate of error covariance matrix:
$S = (R'R)/(n-p)$

Prediction error for the kth dep. var.: $\sqrt{S_{kk}}$

Standard error of $\hat{B}_{jk}$: $\sqrt{(X'X)^{-}_{jj} S_{kk}}$

95% CI for B_{jk}: $\hat{B}_{jk} \pm t_{.025}\sqrt{(X'X)^{-}_{jj} S_{kk}}$

Leverage for $X_{i_}$ (ith row of X):
$h_i = X_{i_}(X'X)^{-}X'_{i_}$

Standard error of $\hat{Y}_{ik}$: $\sqrt{h_i S_{kk}}$

95% CI for mean of Y_{ik}: $\hat{Y}_{ik} \pm t_{.025}\sqrt{h_i S_{kk}}$

95% PI for Y_{ik}: $\hat{Y}_{ik} \pm t_{.025}\sqrt{(1+h_i) S_{kk}}$

Badness of Fit

Total sum of squares: $TSS = \Sigma(y_i - \bar{y})2$

Error sum of squares: $SSE = \Sigma(y_i - \hat{y}_i)2$

Criteria for adequacy of the estimated model in the sample:

Biased mean squared error: SSE/n

Proportion of variance accounted for:
$R2 = 1 - \frac{SSE}{TSS}$

Proportional improvement in prediction:
$PIP = 1 - \sqrt{\frac{SSE}{TSS}} = 1 - \sqrt{1 - R^2}$

Generalization

Criteria for adequacy of the true model in the population:

Unbiased mean squared error: $MSE = \frac{SSE}{(n-p)}$

Adjusted $R2$: $1 - \frac{MSE}{TSS/(n-1)}$

Adjusted PIP: $1 - \sqrt{\frac{MSE}{TSS/(n-1)}}$

Criteria for adequacy of the estimated model in the population:

Predictive MSE: $PMSE = SSE\frac{(n+p)}{(n-p)n}$

Predictive $R2$: $1 - \frac{PMSE}{TSS/(n-1)}$

Predictive PIP: $1 - \sqrt{\frac{PMSE}{TSS/(n-1)}}$

Nonlinear Regression

Model:

$$y_i = f_i(B) + \epsilon_i$$

For Gauss-Newton algorithm, let:

$$\begin{aligned} r_i &= y_i - f_i(\hat{B}) \quad \text{residuals} \\ X_{ij} &= \left.\frac{\partial f_i(B)}{\partial B_j}\right|_{\hat{B}} \quad \text{Jacobian matrix} \end{aligned}$$

Choose initial values for $\hat{B}$. Repeat:

1. Fit linear regression $r_i = \Sigma X_{ij}\Delta_j + \text{error}$.
2. If $R2$ from this regression is small, stop.
3. Update $\hat{B}$ by adding Δ

The linear regression formulas for standard errors and confidence intervals apply approximately for nonlinear regression using the linear approximation given above.

Maximum Likelihood

Let $f(x_i, \beta)$ be the density or probability function for an observation x_i from a distribution with parameter vector β.

The **likelihood** of a sample is $L(\beta) = \Pi f(x_i, \beta)$.

The **maximum likelihood** (ML) estimate of β is the value $\hat{\beta}$ such that $L(\hat{\beta})$ is maximized.

For large samples under suitable regularity conditions, the sampling distribution of the ML estimator $\hat{\beta}$ is approximately multivariate normal with mean β and covariance matrix
$(X'X)-/(n-p)$ where X is the Jacobian matrix of $-\log L(\beta)$ at $\hat{\beta}$.

Generalized Linear Models

A GLIM is a perceptron with no hidden layer and certain types of activation and loss functions based on ML estimation. The inverse of the activation function is called the **link** function. Some common GLIMs:

Distribution	Inverse Link	Loss function
Normal	Identity	$(y-\hat{y})^2$
Binomial	Logistic	$y\log\frac{y}{\hat{y}}+(1-y)\log\frac{1-y}{1-\hat{y}}$ ll
Poisson	Exponential	$y\log\frac{y}{\hat{y}}-(y-\hat{y})$
Gamma	Reciprocal	$-\log\frac{y}{\hat{y}}+\frac{y-\hat{y}}{\hat{y}}$

The binomial loss function above assumes that all cases are based on the same number of trials. Otherwise, the cases should be weighted according to the number of trials.

Unsupervised Learning

The basic idea of "unsupervised learning" is that there are no target values. This idea is wrong. It is the **inputs** that are absent from the model. Typical methods for "unsupervised learning" attempt to optimize a loss function based on reproducing the target values from a reduced represenatation.

The most common statistical varieties of "unsupervised learning" are:

- Principal components: reduce the number of dimensions
- Cluster analysis: reduce the number of cases

Principal Components

Model:

$$Y = XB + E$$

where both X and B are to be estimated by least squares (i.e. minimize trace$(E'E)$) and $p \leq q$, where p and q are the number of columns in X and Y respectively. This model is an example of **dimensionality reduction**, since we are finding a p-dimensional approximation X to the q-dimensional data set Y. Usually Y is **centered**, meaning that the mean is subtracted from each column. The principal component solution is obtained by a singular value decomposition. Other solutions can be obtained by rotating the singular vectors.

K-means Clustering

Model:

$$Y = XB + E$$
$$X_{ij} \in \{0, 1\}$$
$$\Sigma_j X_{ij} = 1$$

where X and B are to be estimated by least squares. Each column of X is an indicator vector for a cluster. B contains the cluster means. An algorithm with finite convergence is:

Initialize X

Repeat until no further change:

1. Set B equal to the cluster means
2. Set X by assigning each case to the cluster with the nearest mean as determined by Euclidean distance

Adaptive vector quantization is a nonconvergent algorithm for fitting the k-means model.

Hidden layers

Another principal components model:

$$Y = YCB + E$$

where C is a $q \times p$ matrix of parameters.

This model is analogous to encoding and autoassociative networks. It is the most common statistical model with a hidden layer. Other statistical models with hidden layers include:

- Maximum redundancy analysis
- Generalized additive models
- Projection pursuit regression

Neural-Statistical Connections

Neural	Statistical
Higher-order nets	Interactions
Functional-link nets	Transformations
Probabilistic neural nets	Kernel discriminant analysis
General regression neural nets	Kernel regression analysis

Problems with Neural Nets as Statistical Models

Singularity (use generalized inverses)

Underindentification (break permutational symmetry?)

Misspecification (use Halbert White's formulas for the covariance matrix of the estimates)

How nonlinear is a neural net? (simulation)

How large a sample is required for asymptotics to hold? (simulation)

References

Borowiak, D.S. (1989) **Model Discrimination for Nonlinear Regression Models**, Marcel-Dekker: NY.

Gallant, A.R. (1987) **Nonlinear Statistical Models**, Wiley: NY.

Gifi, A. (1990) **Nonlinear Multivariate Analysis**, Wiley: Chichester.

Haerdle, W. (1990) **Applied Nonparametric Regression**, Cambridge Univ. Press: Cambridge.

Huber, P.J. (1981), **Robust Statistics**, Wiley: NY.

McCullagh, P. and Nelder, J.A. (1989) **Generalized Linear Models**, 2nd ed., Chapman & Hall: London.

McLachlan, G.J. (1992) **Discriminant Analysis and Statistical Pattern Recognition**, Wiley: NY.

Myers, R.H. (1986), **Classical and Modern Regression with Applications**, Duxbury Press: Boston.

Ripley,B.D. (1993) "Statistical Aspects of Neural Networks", in O.E. Barndorff-Nielsen, J.L. Jensen and W.S.

Kendall, eds., **Networks and Chaos: Statistical and Probabilistic Aspects**, Chapman & Hall: London.

Ross, G.J.S. (1990) **Nonlinear Estimation**, Springer-Verlag: NY.

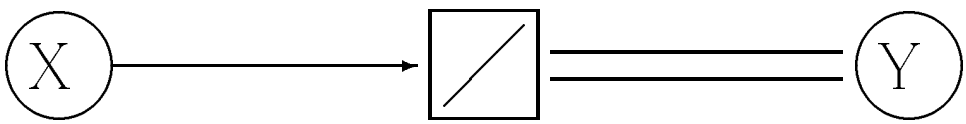
Seber, G.A.F and Wild, C.J. (1989) **Nonlinear Regression**, Wiley: NY.

Weisberg, S. (1985), **Applied Linear Regression**, NY: Wiley

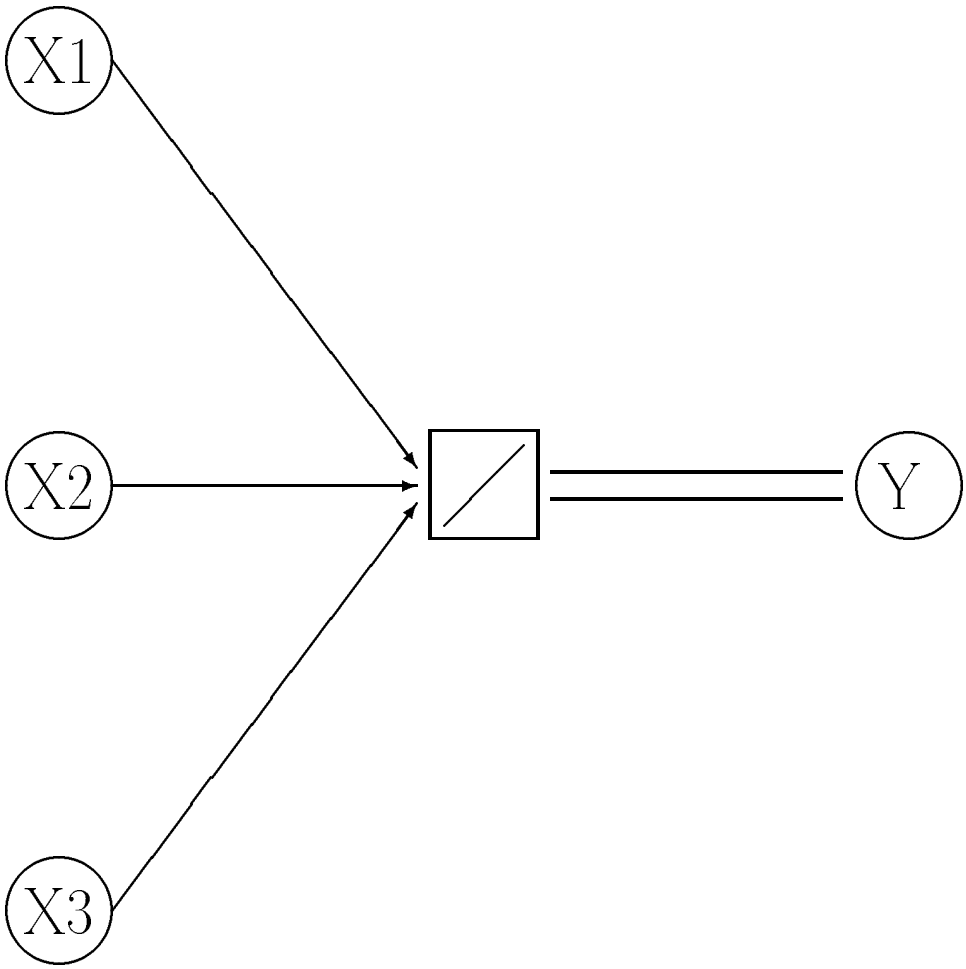
Weiss, S.M. and Kulikowski, C.A. (1991) **Computer Systems That Learn**, Morgan Kaufmann.

White, H. (1992), **Artificial Neural Networks: Approximation and Learning Theory**, Blackwell.

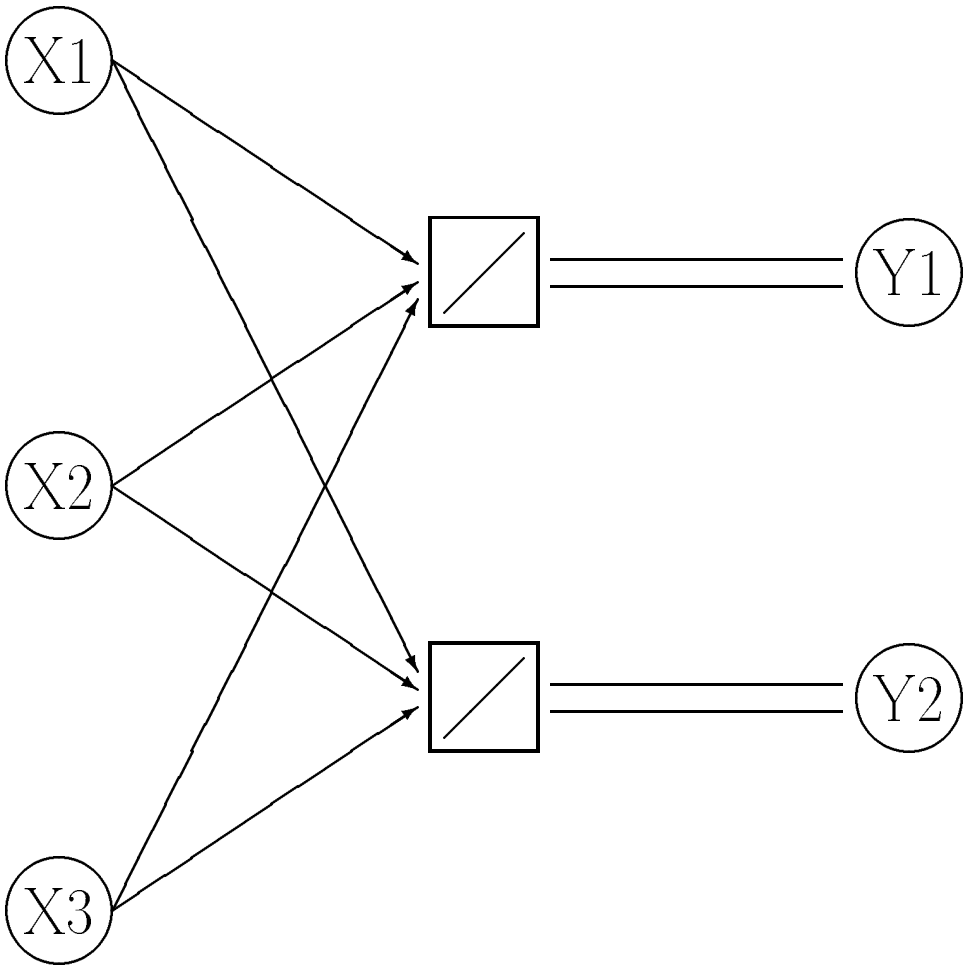
Simple Linear Regression



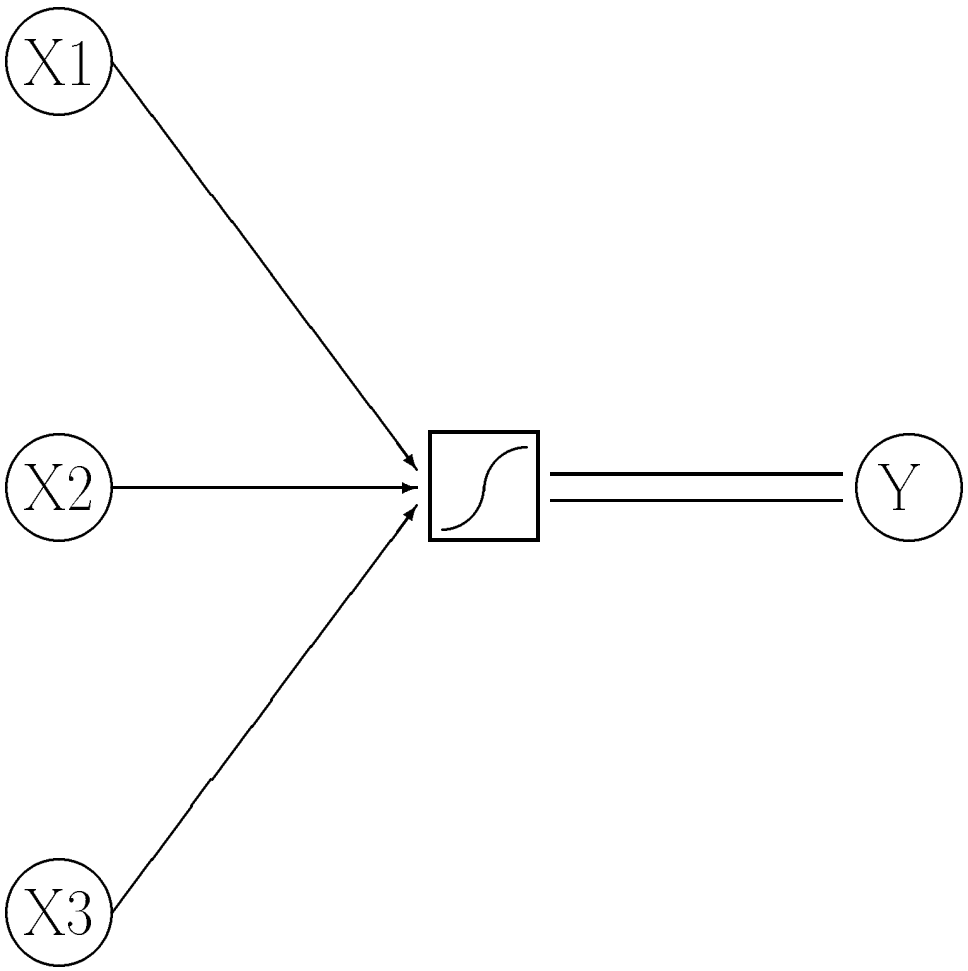
Multiple Linear Regression



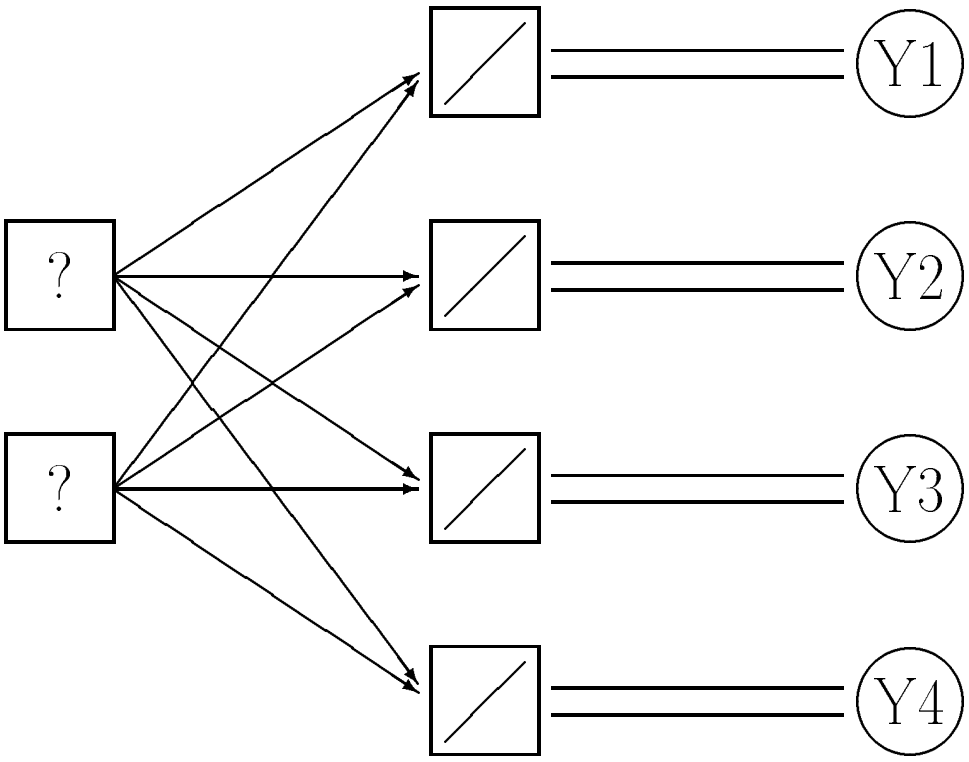
Multivariate Linear Regression



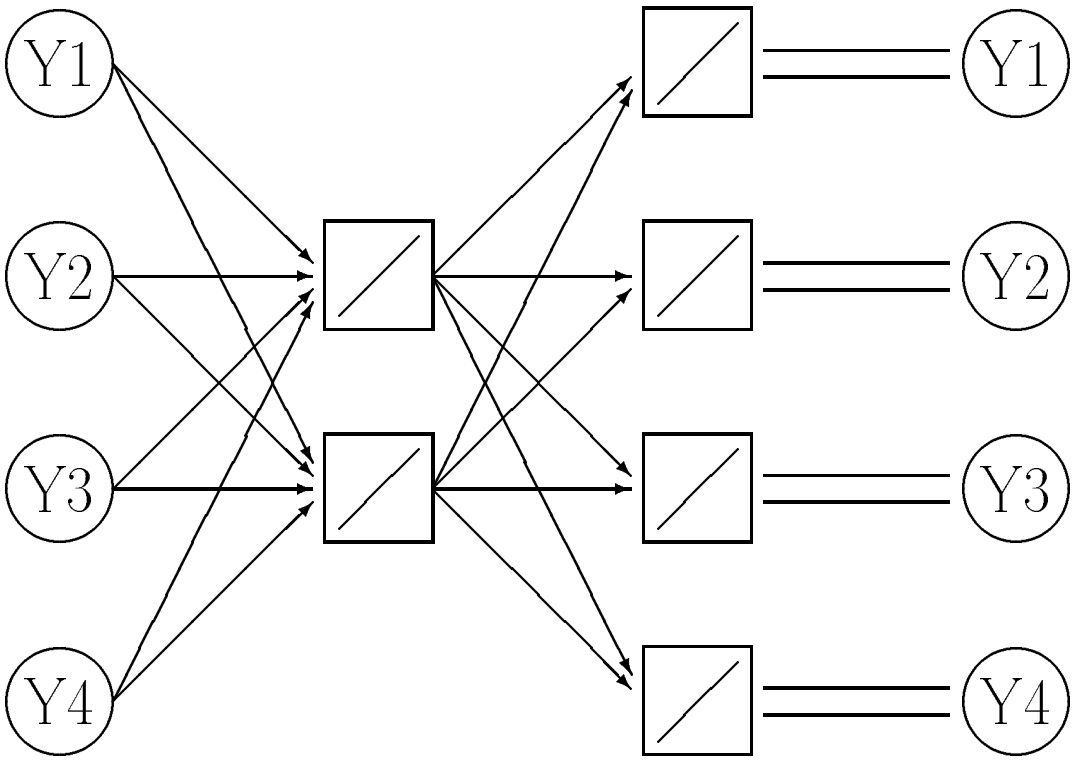
Logistic Regression



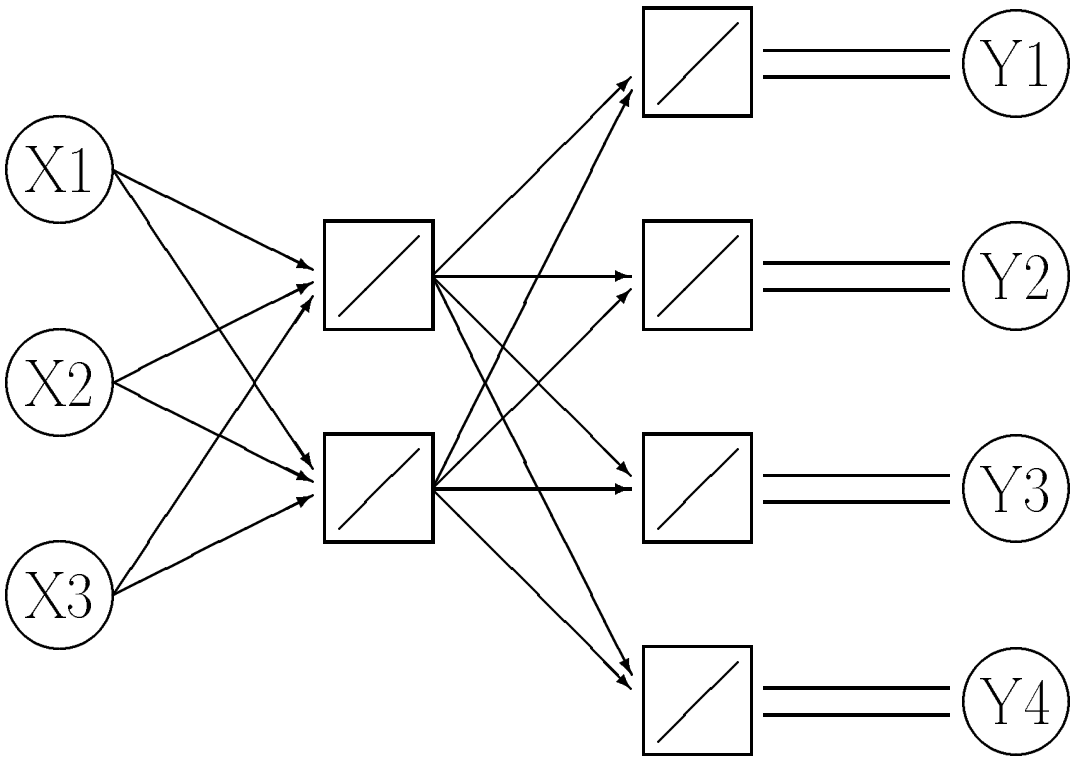
Principal Component Analysis



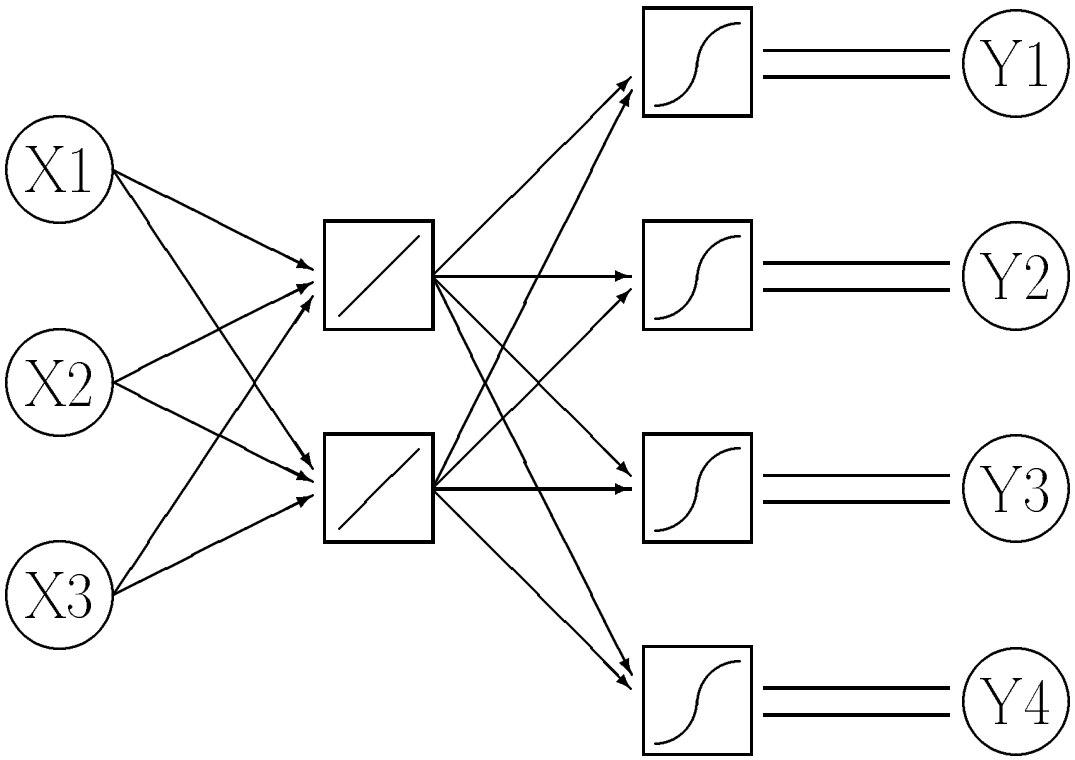
Principal Component Analysis Again



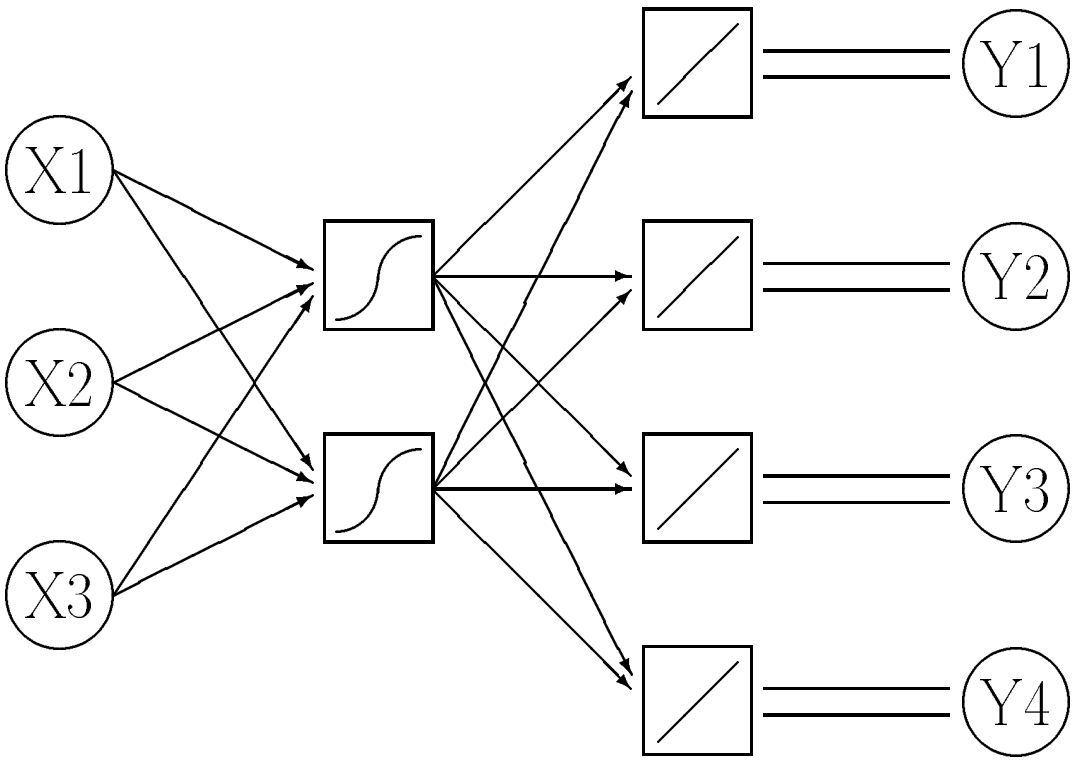
Maximum Redundancy Analysis



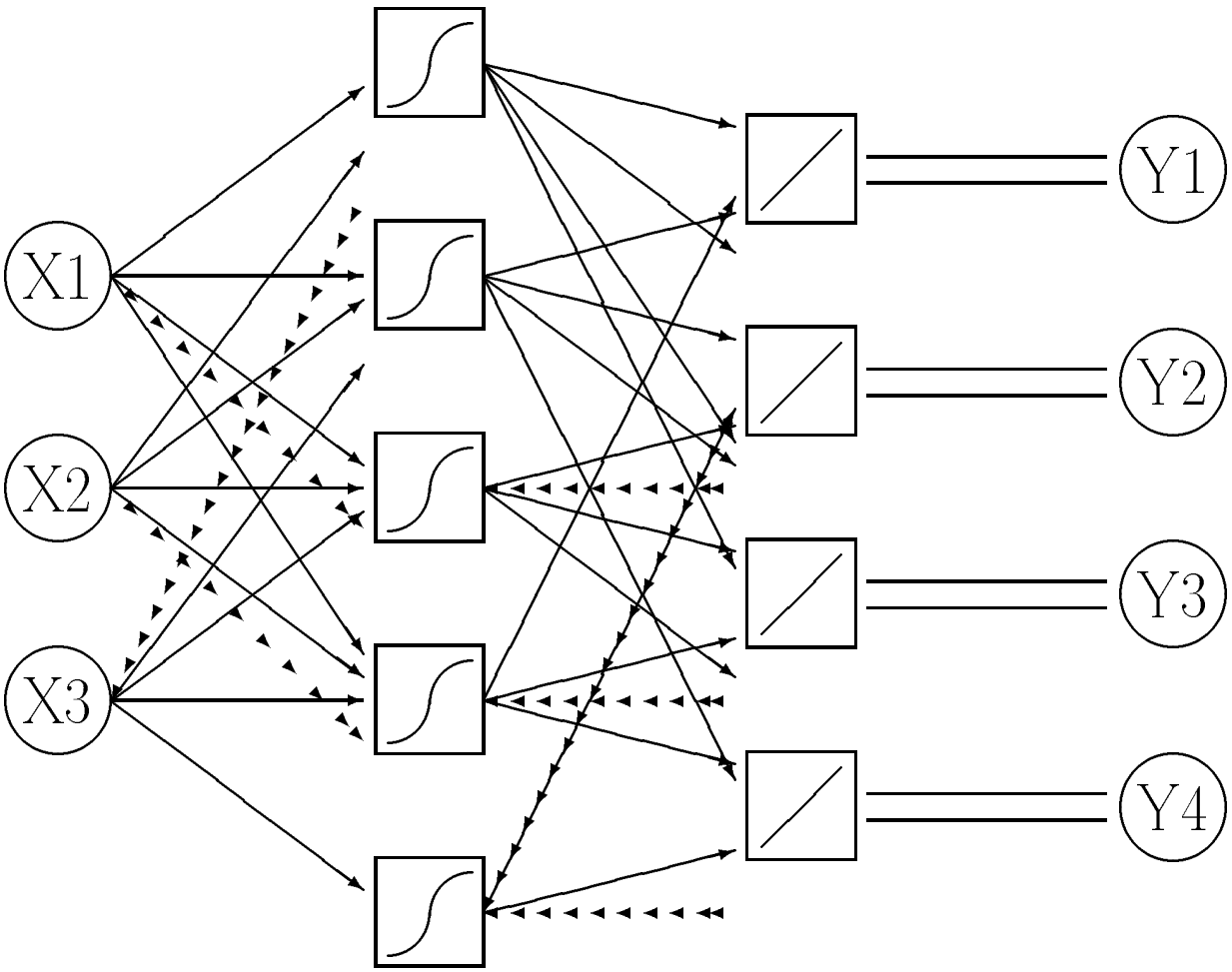
Logistic Maximum Redundancy Analysis



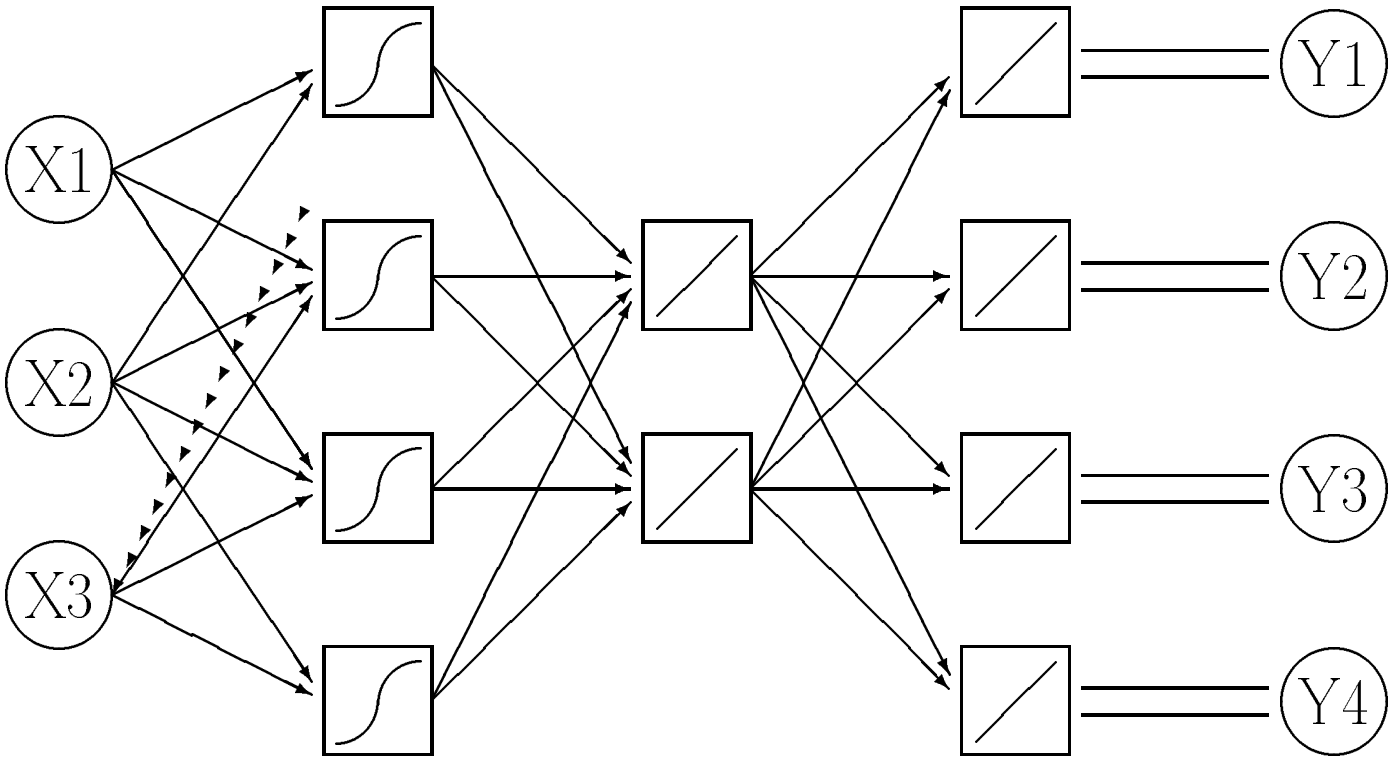
Nonlinear Regression (1)



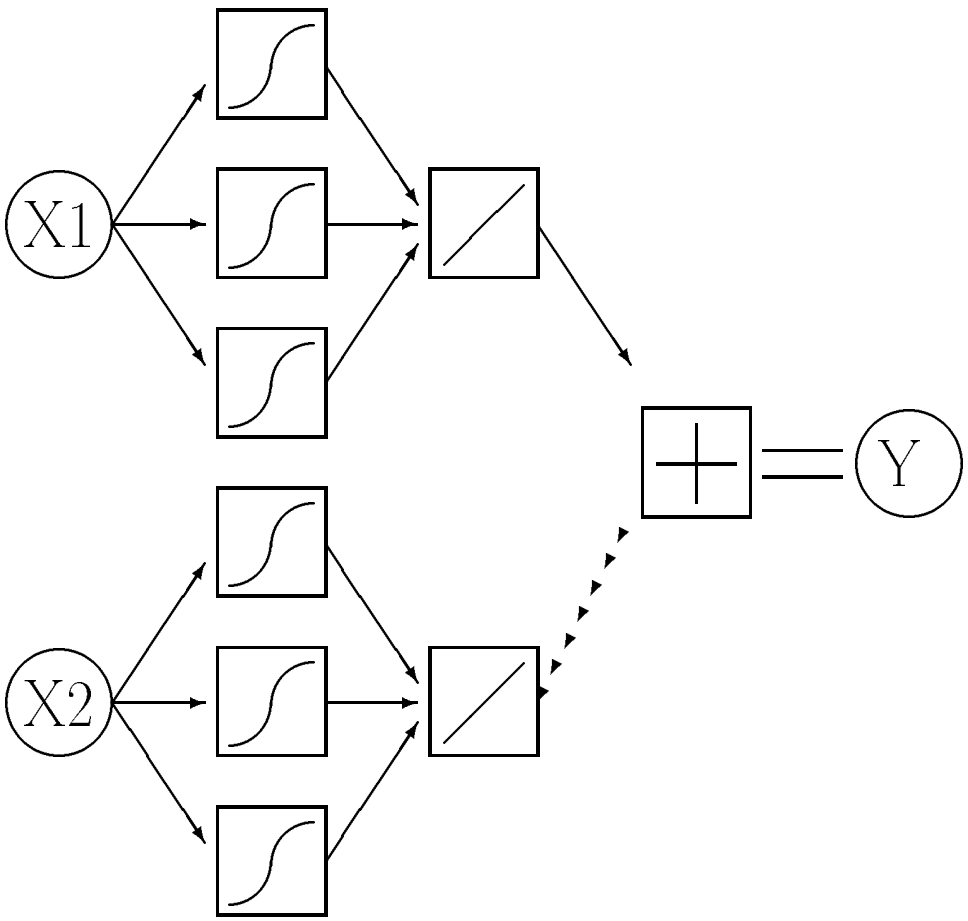
Nonlinear Regression (2)



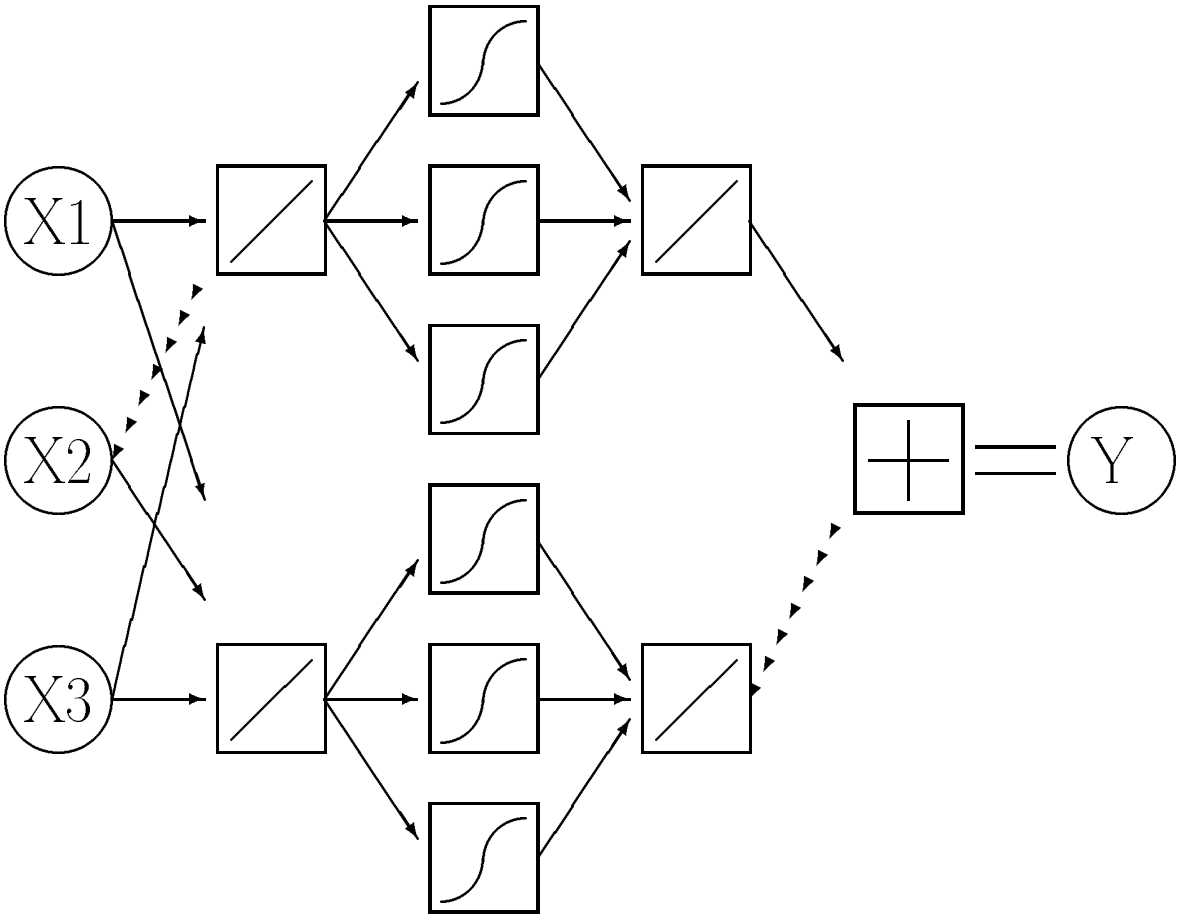
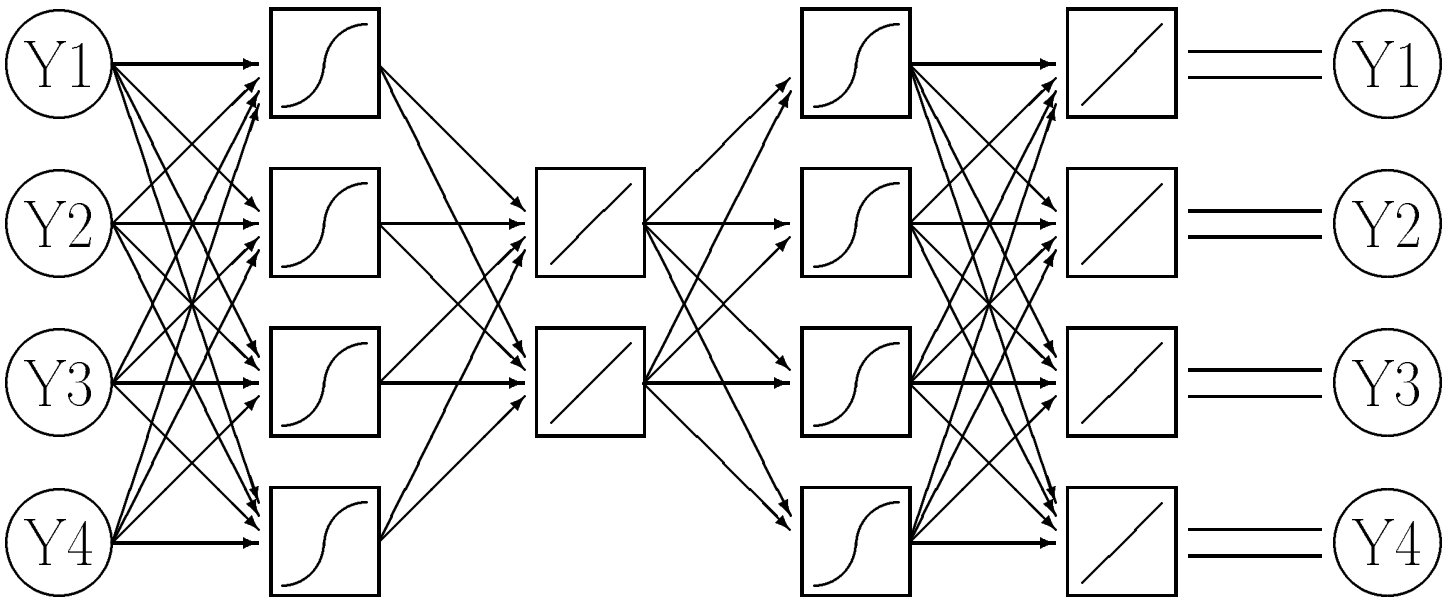
Nonlinear Maximum Redundancy Analysis



Generalized Additive Model



Projection Pursuit Regression



Nonlinear Analog of Principal Components