Statistical Inference and Methods

David A. Stephens

Department of Mathematics
Imperial College London

d.stephens@imperial.ac.uk
http://stats.ma.ic.ac.uk/~das01/

6th December 2005

Objectives

- Data Analyses
- Methods of Statistical Inference
- Classes of Models
- Statistical Computation Techniques

Data Analyses

- Summary/exploratory
- Inferential
- Predictive

Methods of Statistical Inference

- Frequentist
- Likelihood
- Quasi-likelihood
- Estimating Equations
- Generalized Method of Moments
- Bayesian

Classes of Models

- Univariate, independent
- Multivariate, independent
- Regression
- Generalized Regression
- Univariate, dependent (Time Series)
- Multivariate, dependent

Statistical Computation

- Numerical Methods
- Kalman Filter
- Monte Carlo
- Markov chain Monte Carlo

Outline of Syllabus

Session 1

1 **Probabilistic and Statistical Modelling**
  - Forms of Data
  - Probability and probability distributions
  - Multivariate modelling
  - Least-squares and Regression
  - Stochastic Processes

Session 2

2 Inference

- Likelihood theory
- Quasi-likelihood/Estimating Equations
- Generalized Method of Moments
- Bayesian theory

Session 3

3 Time Series Analysis

- ARIMA/Box-Jenkins Modelling
- Forecasting
- Spectral Methods
- Long memory
- Nonstationarity
- Unit roots

Session 4

4 Multivariate Time Series

- Vector ARIMA
- Cointegration

Session 5

5 Statistical Computation

- Monte Carlo
- Importance Sampling
- Quasi Monte Carlo
- Markov chain Monte Carlo
- Sequential Monte Carlo

Session 6

6 Filtering

- Kalman Filter
- Particle Filter

Session 7

7 Volatility Modelling

- ARCH/GARCH
- Stochastic volatility
- Multivariate Methods

Session 8

8 Panel Data

- Models for Longitudinal Data

Part I

Session 1: Probabilistic Modelling

Random quantity denoted X

Probability model denoted $f_X(x;\theta)$ (pdf) or $F_X(x;\theta)$ (cdf)

$$F_X(x) = \int_{-\infty}^{x} f_X(t;\theta)\, dt$$

Finite dimensional parameter θ

Data $x_1, x_2, \ldots, x_n$ available

Repeated observations of random variables $X_1, X_2, \ldots, X_n$.

Different assumptions about the data collection mechanisms lead to different probability models.

Crucial assumptions relate to dependencies between the variables.

(a) Scalar random variables, mutually independent
- repeated observation of the same quantity
- observations do not influence/affect each other.
- the *random sample* assumption
- UNIVARIATE ANALYSIS

(b) Vector random variables, mutually independent
- repeated observation of the same set of quantities or *features*
- observations do not influence/affect each other.
- possible dependence between features
- MULTIVARIATE ANALYSIS

(c) Predictor/Response
- repeated observation of the paired variables
- systematic (causal) relationship between variables.
- REGRESSION

(d) Repeated Measures
- small number of repeated observations of the same set of quantities on the same experimental units
- possible dependence between repeated observations
- MULTIVARIATE ANALYSIS

(e) Scalar, repeated observation, time-ordered

- long sequences of repeated measurement of single quantity.
- time ordering structures dependence between variables
- TIME SERIES ANALYSIS

(f) Vector-valued, repeated observation, time-ordered

- long sequence of vector observation
- time ordering structures dependence between variables
- MULTIVARIATE TIME SERIES

- Dependence
- Latent Structure
- Periodicity
- System changes
- Nonstationarity

Objectives of data analysis:

- Summary
- Comparison
- Inference
- Testing
- Model Assessment
- Prediction/Forecasting

Why do we bother with probabilistic modelling ?

- because we are forced to deal with *uncertainty* due the *lack of perfect information*
- because we wish to represent the uncertainty in our analyses correctly
- because we wish to act in a *coherent* fashion in combining or updating our knowledge or opinion
- because we want to carry out *prediction*

Probability is the only framework that offers coherent treatment of uncertainty.

Probability Models: Common Univariate Distributions

- Discrete distributions
  - Binomial
  - Geometric
  - Poisson
- Continuous distributions
  - Exponential
  - Gamma (Chisquared)
  - Beta
  - Normal
  - Student-t
  - Fisher-F

- Binomial distribution

$$f_X(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

for parameter $\theta > 0$, and positive integer $n > 0$.

Number of successes in n independent and identical 0/1 trials.

- Poisson distribution

$$f_X(x;\lambda) = \frac{\exp\{-\lambda\}\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots$$

for parameter $\lambda > 0$.

Most common model for count data.

- Gamma distribution

$$f_X(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp\{-\beta x\} \qquad x > 0$$

for parameters $\alpha, \beta > 0$, where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\exp\{-x\}\,dx = (\alpha-1)\Gamma(\alpha-1).$$

Special Case: if $\alpha = \nu/2$ for positive integer ν, and $\beta = 1/2$,

$$Gamma(\nu/2, 1/2) \equiv Chisquared(\nu)$$

- Normal (Gaussian) distribution

$$f_X(x;\mu,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

for parameters μ, σ where $\sigma > 0$.

Most commonly used model for data analysis.

Models linked to the Normal:

- Chisquared
- Student-t
- Fisher-F
- Laplace

Distributions linked via *transformation*.

Multivariate distributions: versions of

- Binomial (*Multinomial*)
- Gamma (*Multivariate Gamma, Wishart*)
- Beta (*Dirichlet*)
- Normal (*Multivariate Normal*)
- Student-t

exist.

Multivariate Normal Distribution

Suppose that vector random variable $\mathbf{X} = (X_1, X_2, \ldots, X_k)^\mathsf{T}$ has a multivariate normal distribution with pdf given by

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\mu},\Sigma) = \left(\frac{1}{2\pi}\right)^{k/2} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where Σ is the $k \times k$ (positive definite, non-singular) variance-covariance matrix

Consider the case where the expected value $\boldsymbol{\mu}$ is the $k \times 1$ zero vector; results for the general case are easily available by transformation.

Consider partitioning $\mathbf{X}$ into two components $\mathbf{X}_1$ and $\mathbf{X}_2$ of dimensions d and $k-d$ respectively, that is,

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

We attempt to deduce

(a) the marginal distribution of $\mathbf{X}_1$, and
(b) the conditional distribution of $\mathbf{X}_2$ given that $\mathbf{X}_1 = \mathbf{x}_1$.

First, write

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where Σ_{11} is $d \times d$, Σ_{22} is $(k-d) \times (k-d)$, $\Sigma_{21} = \Sigma_{12}^{\mathsf{T}}$, and

$$\Sigma^{-1} = V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

so that $\Sigma V = I_k$ (I_r is the $r \times r$ identity matrix) gives

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} = \begin{bmatrix} I_d & 0 \\ 0 & I_{k-d} \end{bmatrix}$$

$$\begin{aligned} \Sigma_{11}V_{11} + \Sigma_{12}V_{21} &= I_d && (1) \\ \Sigma_{11}V_{12} + \Sigma_{12}V_{22} &= 0 && (2) \\ \Sigma_{21}V_{11} + \Sigma_{22}V_{21} &= 0 && (3) \\ \Sigma_{21}V_{12} + \Sigma_{22}V_{22} &= I_{k-d}. && (4) \end{aligned}$$

From the multivariate normal pdf, we can re-express the term in the exponent as

$$\mathbf{x}^{\mathsf{T}}\Sigma^{-1}\mathbf{x} = \mathbf{x}_1^{\mathsf{T}}V_{11}\mathbf{x}_1 + \mathbf{x}_1^{\mathsf{T}}V_{12}\mathbf{x}_2 + \mathbf{x}_2^{\mathsf{T}}V_{21}\mathbf{x}_1 + \mathbf{x}_2^{\mathsf{T}}V_{22}\mathbf{x}_2. \qquad (5)$$

We can write

$$\mathbf{x}^{\mathsf{T}}\Sigma^{-1}\mathbf{x} = (\mathbf{x}_2 - \mathbf{m})^{\mathsf{T}}M(\mathbf{x}_2 - \mathbf{m}) + \mathbf{c} \qquad (6)$$

and by comparing with equation (5) we can deduce that, for quadratic terms in $\mathbf{x}_2$,

$$\mathbf{x}_2^{\mathsf{T}}V_{22}\mathbf{x}_2 = \mathbf{x}_2^{\mathsf{T}}M\mathbf{x}_2 \qquad \therefore \qquad M = V_{22} \qquad (7)$$

for linear terms

$$\mathbf{x}_2^{\mathsf{T}}V_{21}\mathbf{x}_1 = \mathbf{x}_2^{\mathsf{T}}M\mathbf{m} \qquad \therefore \qquad \mathbf{m} = V_{22}^{-1}V_{21}\mathbf{x}_1 \qquad (8)$$

and for constant terms

$$\mathbf{x}_1^{\mathsf{T}}V_{11}\mathbf{x}_1 = \mathbf{c} + \mathbf{m}^{\mathsf{T}}M\mathbf{m} \qquad \therefore \qquad \mathbf{c} = \mathbf{x}_1^{\mathsf{T}}(V_{11} - V_{21}^{\mathsf{T}}V_{22}^{-1}V_{21})\mathbf{x}_1 \qquad (9)$$

That is

$$\mathbf{x}^\mathsf{T}\Sigma^{-1}\mathbf{x} = (\mathbf{x}_2 - V_{22}^{-1}V_{21}\mathbf{x}_1)^\mathsf{T} V_{22}(\mathbf{x}_2 - V_{22}^{-1}V_{21}\mathbf{x}_1) + \mathbf{x}_1^\mathsf{T}(V_{11} - V_{21}^\mathsf{T} V_{22}^{-1} V_{21})\mathbf{x}_1, \quad (10)$$

a sum of two terms, where the first can be interpreted as a function of $\mathbf{x}_2$, given $\mathbf{x}_1$, and the second is a function of $\mathbf{x}_1$ only.

Hence

$$f_\mathbf{X}(\mathbf{x}) = f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2|\mathbf{x}_1) f_{\mathbf{X}_1}(\mathbf{x}_1) \quad (11)$$

where

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2|\mathbf{x}_1) \propto \exp\left\{-\frac{1}{2}(\mathbf{x}_2 - V_{22}^{-1}V_{21}\mathbf{x}_1)^\mathsf{T} V_{22}(\mathbf{x}_2 - V_{22}^{-1}V_{21}\mathbf{x}_1)\right\} \quad (12)$$

giving that

$$\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1 \sim N\left(V_{22}^{-1}V_{21}\mathbf{x}_1, V_{22}^{-1}\right) \quad (13)$$

and

$$f_{\mathbf{X}_1}(\mathbf{x}_1) \propto \exp\left\{-\frac{1}{2}\mathbf{x}_1^\mathsf{T}(V_{11} - V_{21}^\mathsf{T} V_{22}^{-1} V_{21})\mathbf{x}_1\right\} \quad (14)$$

giving that

$$\mathbf{X}_1 \sim N\left(0, (V_{11} - V_{21}^\mathsf{T} V_{22}^{-1} V_{21})^{-1}\right). \quad (15)$$

But, from equation (2), $\Sigma_{12} = -\Sigma_{11}V_{12}V_{22}^{-1}$, and then from equation (1), substituting in Σ_{12},

$$\Sigma_{11}V_{11} - \Sigma_{11}V_{12}V_{22}^{-1}V_{21} = I_d$$

so that

$$\Sigma_{11} = (V_{11} - V_{12}V_{22}^{-1}V_{21})^{-1} = (V_{11} - V_{21}^\mathsf{T} V_{22}^{-1}V_{21})^{-1}.$$

Hence

$$\boxed{\mathbf{X}_1 \sim N(0, \Sigma_{11}),} \quad (16)$$

that is, we can extract the Σ_{11} block of Σ to define the marginal variance-covariance matrix of $\mathbf{X}_1$.

From equation (2), $V_{12} = -\Sigma_{11}^{-1}\Sigma_{12}V_{22}$, and then from equation (4), substituting in V_{12}

$$-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}V_{22} + \Sigma_{22}V_{22} = I_{k-d}$$

so that

$$V_{22}^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \Sigma_{22} - \Sigma_{12}^{\mathsf{T}}\Sigma_{11}^{-1}\Sigma_{12}.$$

Finally, from equation (2), taking transposes on both sides, we have that $V_{21}\Sigma_{11} + V_{22}\Sigma_{21} = 0$. Then pre-multiplying by V_{22}^{-1}, and post-multiplying by Σ_{11}^{-1}, we have

$$V_{22}^{-1}V_{21} + \Sigma_{21}\Sigma_{11}^{-1} = 0 \qquad \therefore \qquad V_{22}^{-1}V_{21} = -\Sigma_{21}\Sigma_{11}^{-1},$$

so we have, substituting into equation (13), that

$$\boxed{\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1 \sim N\left(-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{x}_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right).} \qquad (17)$$

Summary
Any marginal, and any conditional distribution of a multivariate normal joint distribution is also multivariate normal.

These results are very important in *regression modelling* to allow study of properties of estimators and predictors.

The Central Limit Theorem

The Normal distribution is commonly used in statistical calculations to approximate the distribution of sum random variables. For example, common estimators include the *sample mean* $\overline{X}$ and *sample variance* s^2

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

The Central Limit Theorem Characterizes the distribution of such variables (under certain regularity conditions)

THEOREM (Lindeberg-Lévy)
Suppose $X_1, ..., X_n$ are i.i.d. random variables with mgf M_X, with $E_{f_X}[X_i] = \mu$ and $Var_{f_X}[X_i] = \sigma^2 < \infty$.

Then

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{\mathfrak{L}} Z \sim N(0,1)$$

as $n \longrightarrow \infty$,irrespective of the distribution of the X_is.

That is, the distribution of Z_n tends to a *standard normal distribution* as n tends to infinity.

This result allows us to construct the following approximations:

$$Z_n \;\dot{\sim}\; N(0,1)$$

$$T_n = \sum_{i=1}^{n} X_i \;\dot{\sim}\; N(n\mu, n\sigma^2)$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \;\dot{\sim}\; N(\mu, \sigma^2/n)$$

Regression Modelling

Suppose we have

- **response** Y
- **predictors** $X_1, X_2, \ldots, X_D$

we want to explain the variation in Y via a function of $X_1, X_2, \ldots, X_D$.

The observed value of Y can be modelled as

$$Y = g(X, \beta) \circ \epsilon$$

where

- X is a **design matrix** of predictors
- β is $K \times 1$ parameter vector
- g is some **link** function
- ϵ is a random (residual) error vector
- $\circ$ is a operator defining the measurement error scale (typically additive or multiplicative)

Most typically, $\circ$ is addition, and the random error term is presumed Normally distributed.

The model can be simplified further if it can be written

$$Y = g(X)\beta + \epsilon$$

that is, **linear** in the parameters.

Inference for this model is straightforward. Another common assumption has the elements of error vector ϵ as identically distributed and independent random variables (**homoscedastic**).

All of these simplifying assumptions can be relaxed:

- homoscedasticity (yields GENERALIZED REGRESSION)
- independence (yields MULTIVARIATE REGRESSION)
- linearity (yields NON-LINEAR REGRESSION)
- normality (yields GENERALIZED LINEAR MODELLING)

Stochastic Processes

Can think of repeated observation of the system $X_1, X_2, \ldots,$

- representing a sequence of observations of a process evolving in DISCRETE time usually at fixed, equal intervals.
- representing a sequence of discrete-time observations of a process evolving in CONTINUOUS time

X could be **univariate** or **multivariate**. We wish to use time series analysis to characterize time series and understand structure.

Possibilities

State (possible values of X)	Time	Notation
Continuous	Continuous	$X(t)$
Continuous	Discrete	X_t
Discrete	Continuous	
Discrete	Discrete	

Denote the process by $\{X_t\}$. For fixed t, X_t is a random variable (r.v.), and hence there is an associated cumulative distribution function (cdf):

$$F_t(a) = P(X_t \leq a),$$

and

$$E[X_t] = \int_{-\infty}^{\infty} x \, dF_t(x) \equiv \mu_t \qquad Var[X_t] = \int_{-\infty}^{\infty} (x - \mu_t)^2 \, dF_t(x).$$

We are interested in the relationships between the various r.v.s that form the process. For example, for any t_1 and $t_2 \in T$,

$$F_{t_1,t_2}(a_1, a_2) = P(X_{t_1} \leq a_1, X_{t_2} \leq a_2)$$

gives the bivariate cdf. More generally for any $t_1, t_2, \ldots, t_n \in T$,

$$F_{t_1,t_2,\ldots,t_n}(a_1, a_2, \ldots, a_n) = P(X_{t_1} \leq a_1, \ldots, X_{t_n} \leq a_n)$$

We consider the subclass of **stationary processes**.

COMPLETE/STRONG/STRICT stationarity

$\{X_t\}$ is said to be completely stationary if, for all $n \geq 1$, for any

$$t_1, t_2, \ldots, t_n \in T$$

and for any τ such that

$$t_1 + \tau, t_2 + \tau, \ldots, t_n + \tau \in T$$

are also contained in the index set, the joint cdf of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ is the same as that of $\{X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_n+\tau}\}$ i.e.,

$$F_{t_1,t_2,\ldots,t_n}(a_1, a_2, \ldots, a_n) = F_{t_1+\tau,t_2+\tau,\ldots,t_n+\tau}(a_1, a_2, \ldots, a_n),$$

so that the probabilistic structure of a completely stationary process is invariant under a shift in time.

SECOND-ORDER/WEAK/COVARIANCE stationarity

$\{X_t\}$ is said to be second-order stationary if, for all $n \geq 1$, for any

$$t_1, t_2, \ldots, t_n \in T$$

and for any τ such that $t_1 + \tau, t_2 + \tau, \ldots, t_n + \tau \in T$ are also contained in the index set, all the joint moments of orders 1 and 2 of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ exist and are finite. Most importantly, these moments are identical to the corresponding joint moments of $\{X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_n+\tau}\}$. Hence,

$$E[X_t] \equiv \mu \qquad Var[X_t] \equiv \sigma^2 \qquad (= E[X_t^2] - \mu^2),$$

are constants independent of t.

If we let $\tau = -t_1$,

$$E[X_{t_1}X_{t_2}] = E[X_{t_1+\tau}X_{t_2+\tau}] = E[X_0X_{t_2-t_1}],$$

and with $\tau = -t_2$,

$$E[X_{t_1}X_{t_2}] = E[X_{t_1+\tau}X_{t_2+\tau}] = E[X_{t_1-t_2}X_0].$$

Hence, $E[X_{t_1}X_{t_2}]$ is a function of the absolute difference $|t_2 - t_1|$ only, similarly, for the **covariance** between X_{t_1} & X_{t_2}:

$$\begin{aligned} Cov[X_{t_1}, X_{t_2}] &= E[(X_{t_1} - \mu)(X_{t_2} - \mu)] \\ &= E[X_{t_1}X_{t_2}] - \mu^2. \end{aligned}$$

For a discrete time second-order stationary process $\{X_t\}$ we define the **autocovariance sequence** (**acvs**) by

$$\begin{aligned} s_\tau &\equiv Cov[X_t, X_{t+\tau}] \\ &= Cov[X_0, X_\tau]. \end{aligned}$$

NOTES:

- τ is called the lag.
- $s_0 = \sigma^2$ and $s_{-\tau} = s_\tau$.
- The autocorrelation sequence (acs) is given by

$$\rho_\tau = \frac{s_\tau}{s_0} = \frac{Cov[X_t, X_{t+\tau}]}{\sigma^2}.$$

- Since ρ_τ is a correlation coefficient, $|s_\tau| \le s_0$.

- The variance-covariance matrix of equispaced X's, $(X_1, X_2, \ldots, X_N)^T$ has the form

$$\begin{bmatrix} s_0 & s_1 & \ldots & s_{N-2} & s_{N-1} \\ s_1 & s_0 & \ldots & s_{N-3} & s_{N-2} \\ \vdots & & \ddots & & \\ s_{N-2} & s_{N-3} & \ldots & s_0 & s_1 \\ s_{N-1} & s_{N-2} & \ldots & s_1 & s_0 \end{bmatrix}$$

which is known as a symmetric Toeplitz matrix – all elements on a diagonal are the same. Note the above matrix has only N unique elements, $s_0, s_1, \ldots, s_{N-1}$.

- A stochastic process $\{X_t\}$ is called Gaussian if, for all $n \geq 1$ and for any $t_1, t_2, \ldots, t_n$ contained in the index set, the joint cdf of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ is multivariate Gaussian.
- 2nd-order stationary Gaussian $\Rightarrow$ complete stationarity
  - follows as the multivariate Normal distribution is completely characterized by 1st and 2nd moments
  - not true in general.
- Complete stationarity $\Rightarrow$ 2nd-order stationary in general.