AST 2003/04: Assignments

To quote from the Unit description:

Assessment Methods. Assessment is based upon two pieces of coursework, one for each of the two assignments selected, and a short presentation on one of your two assignments. The presentation will count 20% towards your final mark; it will be assessed jointly by the lecturers for the unit. There will be no formal written examination.

Award of Credit Points. Credit points will be awarded if you attain a final mark of 40 or more. If you obtain a mark between 30 and 39, you may still be awarded credit points if you have attended at least 90% of the lectures (unless excused for good reason) and made reasonable attempts at the two chosen assignments and at the presentation.

I propose that the deadline for submitting assignments should be 4 May 2004, the Tuesday of the third week of the summer term, thus minimising the impact on your revision for exam courses. The oral presentations will be held shortly after the May/June examinations. This document describes the assignments set by each of the 4 lecturers. (The research skills module is not directly examined.) Each assignment contains more than one numbered question – please note that you are assessed on all of the questions in the two modules you choose, we are not offering you choices within modules. If anything is unclear about these assignments, please ask the individual lecturer concerned.

P. J. Green, 11 February 2004

A. Weak convergence (Stas Volkov)

1. Write an essay on the history of the central limit theorem. (You would have to search for this in the books, on the web, etc.)

2. (a) For a sequence of random variables define (1) convergence almost surely, (2) in probability, (3) in $\mathbb{L}^p$, (4) in distribution.

   Prove that (1) implies (2), (3) implies (2) and that (2) implies (4). Give your own counterexamples that: (2) does not imply (1), (2) does not imply (3), and (4) does not imply (2).

   (b) Formulate the Lindeberg–Feller central limit theorem (in the form it was given in class). Sketch its proof, carefully justifying what you assume and specifying what lemmas you are using.

   Choose some assumption of the Lindeberg–Feller theorem, and now suppose that this assumption does not hold. For this "truncated" set of assumptions, construct a "counterexample" to the theorem, thus showing that the omitted condition was essential.

   (c) Define a *tail event*. Formulate Kolmogorov's 0–1 law. Give *three* different examples of application of this theorem, different from the following one. (Example: if X_i, $i = 1, 2, \ldots$ are independent random variables, and $S_n = \sum_{i=1}^{n} X_i$, then since the event "S_n converges" is a tail event, then either $\lim_{n\to\infty} S_n$ exists a.s. or does not exist a.s.)

   (d) What properties of characteristic functions do you know? List a few (2–3) characteristic functions and check all these properties. If you are given a characteristic function $\varphi(t)$ of some random variable X, what is a sufficient condition for that random variable to be continuously distributed? [question continues over page]

Suppose $X_1, X_2, \ldots$ are i.i.d. with a common characteristic function $\varphi(t) = e^{-|t|^c}$. What values of c are acceptable for this to be a characteristic function?
Find the distribution of $S_n/n^{1/c}$, where $S_n = \sum_{i=1}^n X_i$. Comment on your finding.

B. Principles of stochastic approximation (Christophe Andrieu)

1. Linear regression

We observe two sequences $\{X_n\} \subset \mathbb{R}^{n_x}$ (the *inputs*), and $\{Y_n\} \subset \mathbb{R}^{n_y}$ (the *outputs*) and assume that

$$Y_n = \overline{\theta}^{\mathrm{T}} X_n + e_n$$

for some unknown $\overline{\theta} \in \mathbb{R}^{n_x}$ where $\{X_n\} \subset \mathbb{R}^{n_x}$ and $\{e_n\} \subset \mathbb{R}^{n_y}$ are assumed to be two independent sequences of i.i.d. random variables with respective probability densities π and p. The aim of this part of the assignment is to propose and analyse a general algorithm to estimate $\overline{\theta}$ from the training set $\{X_n, Y_n\}$.

(a) Assume that we define the optimal θ as being the minimiser of

$$J(\theta) := \mathbb{E}_p(\Psi(Y - \theta^{\mathrm{T}} X)) = \int_{\mathbb{R}^{n_y}} \Psi(e) p(e) de,$$

where $\Psi :\rightarrow \mathbb{R}^+$ is a twice differentiable function. We will denote ψ the derivative of Ψ and will further assume that $\mathbb{E}_p(|\psi|) < \infty$. Ψ can be interpreted as being a loss function used here to measure the quality of fit of the linear regression.

(i) Examples include $\Psi(x) = x^2$ and $\Psi(x) = -\log(p(x))$. What well known criterion does this later choice correspond to?

(ii) Describe a stochastic approximation algorithm (without reprojections) to minimise $J(\theta)$ of the form

$$\theta_{i+1} = \theta_i + \gamma_{i+1} H(\theta_i, Z_{i+1})$$

for appropriate $\{\theta_i\}$, field H and $\{Z_i\}$ that you will define. What is the mean field $h(\theta)$ of this algorithm?

(b) We assume from now on that ψ satisfies the condition

$$x\mathbb{E}_p(\psi(x + e_1)) > 0 \text{ for } x \neq 0. \tag{1}$$

Consider the function $w : \mathbb{R}^{n_y} \rightarrow \mathbb{R}^+$

$$w(\theta) = \frac{|\theta - \bar{\theta}|^2}{2}.$$

(i) Show that it is a valid Lyapunov function for the algorithm above.

(ii) From now on we assume that $\mathbb{E}_\pi(|X_1|^4) < \infty$ and $\mathbb{E}_p(|e_1|^2) < \infty$. Show that $< \nabla w, h(\theta) >= 0$ iff $\theta = \bar{\theta}$.

(iii) Check that $(A1)$ from the lecture notes is satisfied under conditions that you will outline.

(c) Now if p is log-concave and $\Psi = -\log(p)$ and such that $\psi(0) = 0$ then show that Eq. (1) is satisfied.

(d) Consider for some decreasing sequence of stepsizes $\{\rho_i\}$ satisfying

$$\sum_{i=1}^{+\infty} \rho_i^2 < +\infty,$$

H and h as defined above the sequence $\{u_i\}$

$$\begin{aligned} u_{i+1} &= u_i + \rho_{i+1} H(u_i, Z_i) \\ &= u_i + \rho_{i+1} h(u_i) + \rho_{i+1}\xi_{i+1} \end{aligned}$$

started from $u_0 \in \mathcal{K}$, where $\mathcal{K}$ is some compact set $\mathcal{K} \subset \mathbb{R}^{n_y}$. Define the first instant $\{u_i\}$ exits $\mathcal{K}$

$$\sigma(\mathcal{K}) := \inf\{k : u_k \notin \mathcal{K}\}.$$

Assuming that

$$|\psi(x)| \leq C(1 + |x|)$$

for some constant $C < +\infty$, prove that for any compact set $\mathcal{K}$ defined as above the condition

$$\lim_{n\to\infty} \sum_{k=1}^{n} \rho_i \xi_i I(n \geq \sigma(\mathcal{K})) < +\infty$$

is almost surely satisfied. You may find the following theorem on the convergence of martingales useful.

Theorem 1 *Let $\mathcal{F}_n$ be an increasing sequence of σ-fields and M_n an $\mathcal{F}_n$-martingale. Let $\Delta_n := M_n - M_{n-1}$. If for some $p \in [1, 2]$*

$$\sum_{n=1}^{\infty} \mathbb{E}[|\Delta_n|^p | \mathcal{F}_{n-1}] < \infty \quad \textit{almost surely},$$

then $\{M_n\}$ converges almost surely.

(e) Conclude that $\{\theta_n\}$ is almost surely bounded and converges to $\bar{\theta}$ under some additional conditions on $\{\gamma_i\}$. Reading of parts of [2] might help to answer this part of the assignment.

Note that in fact the boundedness can here be proved directly, and that there is no need for reprojections. One can study for example the sequence $\{u_n = \mathbb{E}(|\theta_n - \bar{\theta}|^2) + 1\}$ and conclude.

2. Adaptive MCMC methods.

Markov chain Monte Carlo methods are techniques to sample from a probability distribution π defined, say, on $\mathsf{X} = \mathbb{R}^{n_x}$. An example of such a Markov chain with transition probability, say $P_\theta(x, y)$, is as follows. Assume that the chain is currently at $x \in \mathbb{R}^{n_x}$ then a candidate $x + z$ is proposed, where $z \sim \mathcal{N}(0, \theta^2)$. This candidate is accepted with probability

$$\alpha(x, x + z) = 1 \wedge \frac{\pi(x + z)}{\pi(x)},$$

i.e. $y = x + z$, or rejected, in which case $y = x$. The value of θ is known to be crucial for the algorithm to perform well. When π is continuous, too small a θ will lead to a high acceptance probability and too large a θ will lead to a chain that does not accept many transitions. In either cases the exploration of the target distribution π is not satisfactory. Some theoretical argument suggest that optimal (in a sense not described here) θ's correspond to specific values of the expected acceptance probability

$$\alpha(\theta) = \mathbb{E}_{\pi q_\theta}(\alpha(X, X+Z)) = \int_{\mathsf{X}^2} \pi(x) q_\theta(x,y)\alpha(x,y)dxdy.$$

The aim of this part of the assignment is to propose and study stochastic approximation algorithms which automatically adapt θ in order to achieve a predefined value α^* of the expected acceptance probability. Several strategies based on stochastic approximation have been proposed in the literature [1], [3]. The aims of the project are: (a) to understand the algorithms (b) discuss their validity and limitations (c) suggest alternatives (d) implement some of the algorithms on toy examples (e) analyse the algorithms using the results found in [2].

References

[1] C. Andrieu and C.P. Robert, “Controlled MCMC for optimal sampling”, Tech. Rep. University of Bristol, 2001.

[2] C. Andrieu, É. Moulines and P. Priouret, “Stability of stochastic approximation under verifiable conditions”, Tech. Rep. University of Bristol, 2003-2004.

[3] Y. Atchadé and J. Rosenthal, “On Adaptive Markov chain Monte Carlo Algorithms”, available from the MCMC preprint service.

C. Multiscale methods (Guy Nason)

1. (a) A wavelet is defined in terms of a mother wavelet, $\psi(x)$, $x \in \mathbb{R}$ by

$$\psi_{jk}(x) = 2^{j/2}\psi\left(2^j x - k\right),$$

for $j, k \in \mathbb{Z}$. The L_2 norm of a function f on $\mathbb{R}$ is defined to be

$$||f|| = ||f||_2 = \left\{ \int_{-\infty}^{\infty} f(x)^2 \, dx \right\}^{1/2}.$$

Show that for any mother wavelet ψ we have

$$||\psi_{jk}|| = ||\psi||.$$

In other words, that all wavelets have the same L_2 norm as the mother.

(b) The Haar mother wavelet is defined as

$$\psi(x) = \begin{cases} 1 & x \in [0, 1/2), \\ -1 & x \in [1/2, 1). \end{cases}$$

For Haar show that $||\psi|| = 1$.

(c) Show that for Haar wavelets $\{\psi_{jk}(x)\}_{j\in\mathbb{Z}, k\in\mathbb{Z}}$ is an orthonormal system of functions.

(d) Suppose that $f(x)$ is a probability density function with associated probability density function $F(x)$ with the following orthogonal wavelet representation:

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{jk}\psi_{jk}(x).$$

Show that

$$d_{lm} = \int_{-\infty}^{\infty} f(x)\psi_{lm}(x)\, dx,$$

for all $l, m \in \mathbb{Z}$.

If ψ is the Haar mother wavelet show further that

$$d_{jk} = 2^{j/2}\left[2F\left\{2^{-j}(k+1/2)\right\} - F(2^{-j}k) - F\left\{2^{-j}(k+1)\right\}\right],$$

for all $j, k \in \mathbb{Z}$.

(e) Suppose now that $f(x)$ is the probability density function of the *exponential distribution*. Show that the Haar wavelet coefficients on $[0, \infty)$ are given by

$$d_{jk} = 2^{j/2}S(2^{-j}k)F^2(2^{j-1})$$

where $S(x) = 1 - F(x)$ is the survivor function of the exponential distribution.

2. Write an essay about how to estimate curves from noisy data using wavelet shrinkage. Your essay should be in two parts:

(a) a general introduction to the principles of wavelet shrinkage including a brief description of the discrete wavelet transform, methods of thresholding and primary resolution.

(b) a more detailed investigation of one particular thresholding policy out of the following: ideal, SURE, cross-validation, FDR or Bayesian techniques.

You will need to download and read papers from the scientific literature to find out about these concepts. One way to do this is a forward citation search on the Nason and Silverman paper below.

The general introduction should be approximately one-third of the length of the whole. The maximum length of the essay should be 12 sides of A4 script.

Your essay should include some practical examples of wavelet shrinkage using the `WaveThresh3` software and any *real* data set. Further details regarding the `WaveThresh3` software can be found on

```
http://www.stats.bris.ac.uk/~wavethresh
```

and particularly on the help page for the `threshold.wd()` function.

You can download WaveThresh3 for R for Windows from the WaveThresh homepage.

See also the paper Nason, G.P. and Silverman, B.W. (1994) The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*, **3**, 163–191.

D. Statistical methods in epidemiology (Jonathan Sterne)

1. Case-control studies are studies in which risk factors (covariates) are assessed retrospectively in diseased cases and disease-free control subjects. Define $D = 0$ for controls, $D = 1$ for cases, and let $x' = (x_{i1}, \ldots, x_{ip})$ be the vector of covariates for subject i. Show that logistic regression may be used to make the same inferences about associations between the covariates and disease that could be made if the data had been obtained in a prospective study.

2. Discuss the reasons for matching in case-control studies, and the implications of matching for the analysis of case-control studies.

3. Identify a case-control study published recently in the International Journal of Epidemiology or the American Journal of Epidemiology. (Both these journals are available online at `www.bris.ac.uk/is`). Write a short essay describing the results of this study. Include the following information:

   - Full reference for the study
   - How were the cases defined and ascertained?
   - How were the controls defined and selected?
   - Did the study use a matched design? If so, describe it.
   - What were the exposure variables whose association with the outcome was investigated?
   - What confounding variables were measured? How did controlling for these variables affect the estimate of the exposure-outcome association?
   - What were the conclusions of the study?

References

[1] N.E. Breslow and N.E. Day. (1980). Statistical Methods in Cancer Research. Volume I the analysis of case-control studies.

[2] D. Clayton and M. Hills. (1993) Statistical Models in Epidemiology. Oxford University Press, Oxford.

[3] R.L. Prentice and R. Pyke. (1979) Logistic disease incidence models and case-control studies. Biometrika 1979; 66: 403-11