Biometrika (2015), *pp.* 1–5
Printed in Great Britain
doi: 10.1093/biomet/asv033

Big data and precision

BY D. R. COX

Nuffield College, Oxford OX1 1NF, U.K.
david.cox@nuffield.ac.ox.uk

SUMMARY

So-called big data are likely to have complex structure, in particular implying that estimates of precision obtained by applying standard statistical procedures are likely to be misleading, even if the point estimates of parameters themselves may be reasonably satisfactory. While this possibility is best explored in the context of each special case, here we outline a fairly general representation of the accretion of error in large systems and explore the possible implications for the estimation of regression coefficients. The discussion raises issues broadly parallel to the distinction between short-range and long-range dependence in time series theory.

Some key words: Components of variance; Large data; Long-range dependence; Multilevel model; Time series.

1. INTRODUCTION

Big data raise several essentially statistical issues. There may be concern over data quality and the standardization of definitions and with the rationale for inclusion in the data base. Importantly also, there is a distinction between investigations in which the research questions are at least broadly defined from the start and those in which there needs to be a wide search to isolate any information of interest.

Here, however, we concentrate on a narrower objective. With very large amounts of data, direct use of standard statistical methods, including simulation-based approaches, will tend to produce estimates of apparently very high precision, essentially because of strong explicit or implicit assumptions of at most weak dependence underlying such methods.

In particular applications it may be feasible to represent the main sources of variation in an explicit model and thereby produce both improved estimates and more relevant assessments of precision. In the present paper we outline in general terms a formulation that indicates at least some of the ways in which error may accrue in relation to data size.

The discussion has quite a strong link with the contrast in time series analysis between long-range and short-range dependence. We therefore start with a brief review of that distinction.

2. DEPENDENCE IN TIME SERIES ANALYSIS

We consider a stationary time series $\{Y(t)\}$ in continuous time with mean zero and autocorrelation and spectral density functions

$$\rho(h) = \mathrm{corr}\{Y(t),\ Y(t+h)\}, \quad f(\omega) = (2\pi)^{-1} \int_{-\infty}^{\infty} \exp(-\mathrm{i}\omega h)\rho(h)\,\mathrm{d}h,$$

respectively. There is long-range dependence if

$$\int_0^{\infty} \rho(h)\,\mathrm{d}h$$

diverges or, equivalently, if $f(\omega)$ is singular at the origin. In more specific terms, we may have the autocorrelation function behaving proportionally to h^{a-1} with $0 < a < 1$ for large h, and have $f(\omega)$ proportional to ω^{-a} for small ω, the latter leading to the alternative name for long-range dependence, namely $1/f$ noise, where f, replacing ω, represents frequency.

Another important interpretation is for means or totals of sections. Let

$$C_{Y,m}(t) = \int_t^{t+m} Y(u)\,\mathrm{d}u$$

be the total of a section of length m, with $\bar{C}_{Y,m}(t)$ denoting the corresponding mean. Then, when there is long-range dependence, $\mathrm{var}(\bar{C}_{Y,m})$ behaves for large m like m^{a-1} with $0 < a < 1$.

Equivalently, there is the property of self-similarity, namely that the correlation between the means of adjacent sections of length m satisfies

$$\mathrm{corr}\{\bar{C}_{Y,m}(t),\ \bar{C}_{Y,m}(t+m)\} = 2^a - 1.$$

For broad reviews of these issues, see Cox (1984) and Beran (1994).

3. A model for big data

We now represent the accretion of sources of variability in abstract and idealized form. The account is intended, in particular, to illustrate in broad outline the impact of a complex pattern of sources of variability on the precision of such statistics as means and regression coefficients. Realistic discussion of individual applications requires explicit identification of the sources of variability.

For example, the compilation of data on patients suffering from a specific chronic condition might start with a record of visits at a particular clinic and be extended to cover other hospitals, other regions and other countries. As the data evolve in time and geographically, additional sources of variability enter that may be connected with patient features, imperfect standardization of measurement procedures, and so on.

This leads us to consider big data as evolving in a possibly notional time-frame. At various time-points new sources of variability enter, possibly operating on longer and longer time scales. Contributions on different time scales are regarded as statistically independent. We consider two different possibilities. In the first, the different sources are considered for simplicity to have the same essential structure, whereas in the second they operate on a longer and longer time scale. We assume that the different contributions enter in a time-dependent Poisson process, in general having an atom at time zero to represent a system initially of relatively simple structure. Once a source of variability enters the system, it operates permanently, although other possibilities could be considered as well. We start with a single Gaussian observation at each time t, time being taken as scalar for simplicity. We consider the data as being defined in continuous time, but this is for convenience only. To treat independent and identically distributed observations, we suppose the autocorrelations to be zero and the total time of observation to represent sample size.

For nonnegative λ, let $B(\cdot\,;\lambda)$ be independent stationary Gaussian processes of zero mean, unit variance and lag-h autocorrelation $\rho_B(\kappa h)$, called base processes. For the first structure mentioned above, we have $\kappa = 1$, i.e., all the base processes have the same structure. The main alternative is that $\kappa = \lambda$, so that the autocorrelations for a given λ have the form $\rho_B(\lambda h)$. We regard $B(\cdot\,;\lambda)$ as having zero contribution to $Y(t)$ before time λ. The standard deviation of its contribution to $Y(t)$ is denoted by $\sigma_Y(\lambda)$; that is,

$$Y(t) = \int_0^t B(t;\lambda)\sigma_Y(\lambda)\,\mathrm{d}N(\lambda),$$

where $N(\cdot)$ is a Poisson process of rate $\nu(\lambda)$, in general having an atomic component at $\lambda = 0$.

It follows that the variance $V_Y(t)$ of $Y(t)$ is

$$V_Y(t) = \int_0^t \sigma_Y^2(\lambda)\nu(\lambda)\,\mathrm{d}\lambda = \int_0^t \tau_Y(\lambda)\,\mathrm{d}\lambda, \tag{1}$$

say. Thus $\tau_Y(t)$ is the rate of increase in local variance at time t.

Similarly, central to the correlation structure of the process, we have for $h \geqslant 0$ that

$$\operatorname{cov}\{Y(t), Y(t+h)\} = \int_0^t \tau_Y(\lambda)\rho_B(\kappa h)\,\mathrm{d}\lambda. \tag{2}$$

If $\kappa = 1$, corresponding to all base processes being of the same form, this becomes $\rho_B(h)V_Y(t)$. If, on the other hand, $\kappa = \lambda$, then if $\rho_B(h)$ is approximately kh^{a-1}, for large h the right-hand side of (2) becomes

$$kh^{a-1}\int_0^t \tau_Y(\lambda)\lambda^{a-1}\,\mathrm{d}\lambda.$$

In many cases, the behaviour of $\tau_Y(\cdot)$ will ensure that the integral is convergent.

4. Some properties as a base for estimation

For the remainder of the discussion we suppose that $\kappa = 1$, so that the base processes are independent versions of the same stochastic process with a scaling factor depending on λ. We treat the mean as known to be zero and estimate average variance after time t from

$$\tilde{V}(t) = t^{-1}\int_0^t Y^2(z)\,\mathrm{d}z.$$

Then, by (1),

$$E\{\tilde{V}(t)\} = \frac{1}{t}\int_0^t \mathrm{d}u \int_0^u \tau_Y(\lambda)\,\mathrm{d}\lambda = \bar{V}_Y(t),$$

the average variance of $Y(\cdot)$ over the interval $(0, t)$.

Now consider the estimation of the mean of the process by

$$\bar{Y}(t_0) = \frac{1}{t_0}\int_0^{t_0} Y(u)\,\mathrm{d}u = C_Y(t_0)/t_0,$$

say. Then, on writing $\operatorname{var}\{C_Y(t_0)\} = V_{C_Y}(t_0)$, we have that

$$\begin{aligned} V_{C_Y}(t_0) &= \int_0^{t_0} V_Y(t)\,\mathrm{d}t + 2\int_0^{t_0}\mathrm{d}t_1\int_{t_1}^{t_0} V_Y(t_1)\rho_B(t_2 - t_1)\,\mathrm{d}t_2 \\ &= \int_0^{t_0} V_Y(t)\,\mathrm{d}t + 2\int_0^{t_0} V_Y(t_1)\mathrm{P}_B(t_0 - t_1)\,\mathrm{d}t_1. \end{aligned} \tag{3}$$

Here, the integrated autocorrelation of the base processes is

$$\mathrm{P}_B(t) = \int_0^t \rho_B(z)\,\mathrm{d}z.$$

There are now two rather different aspects for consideration. One is the general nature of the dependence of the variance on t_0, and the other is the specific comparison of the variance for a given large t_0 with that for independent and identically distributed observations.

To study the first aspect, it is convenient to introduce Laplace transforms; for example, define

$$V_C^*(s) = \int_0^\infty \exp(-st)V_C(t)\,\mathrm{d}t.$$

Then, after using the relation between the Laplace transform of a function and that of its indefinite integral, it follows that

$$V_{C_Y}^*(s) = \frac{\tau^*(s)}{s^2}\{1 + 2\rho_B^*(s)\}.$$

There is extensive theory relating the behaviour of functions of t for large arguments and the behaviour of their Laplace transforms for small arguments s. We merely note that if for large t a function has the form kt^{-a} for $0 < a < 1$, then for small s the Laplace transform is approximately $ks^{a-1}\Gamma(1-a)$. If, on the other hand, the function of t tends to zero more rapidly, for example following the previous form with $a > 1$ or exhibiting exponential decay, then the Laplace transform is bounded near $s = 0$. We need the inverse statements to study $V_{C_Y}(t)$, and a rigorous statement involves the regularity conditions required for Tauberian theory, i.e., for passing from the properties of averages, Laplace transforms, to those of functions. For a general summary see, for example, Feller (1966, ch. 13).

The essence is that if either or both of $\tau_Y(t)$ and $\rho_B(h)$ decay more slowly than $1/t$ and $1/h$, respectively, then $\mathrm{var}\{C_Y(t_0)\}$ increases as t_0^{2-a} for $0 < a < 1$, so that the variance of the mean decays at rate t_0^{-a}, i.e., more slowly than $1/t_0$.

The second issue, the comparison with observations of independent random variables, depends on the ratio of the second term in (3) to the first. If the ratio converges to a positive constant, we have what may be called simple overdispersion, with the possibility also of underdispersion if the limit is negative. If, however, $\rho_B(h)$ decays more slowly than $1/h$, then the ratio between the variance of the mean and that for independent observations increases indefinitely.

5. Regression

We now extend the discussion to least squares regression. We make the approximation that the explanatory variable $X(t)$ of interest can be totally orthogonalized with respect to other explanatory variables, so that all the underlying base processes are simultaneously orthogonalized. We thus consider simple linear regression of $Y(t)$ on $X(t)$ based on the statistic

$$C_{XY}(t_0) = \int_0^{t_0} X(t)Y(t)\,\mathrm{d}t. \tag{4}$$

To study this, we suppose that the generating process for $X(t)$ has similar form to that for $Y(t)$, with the same underlying Poisson process and with

$$X(t) = \int_0^t A(t;\lambda)\sigma_X(\lambda)\,\mathrm{d}N(\lambda).$$

The autocorrelation structure of the processes $A(\cdot\,;\cdot)$ is specified by $\rho_A(\cdot)$ and the cross-dependencies by

$$\rho_{BA}(\lambda; h) = \mathrm{cov}\{B(t;\lambda),\ A(t-h;\lambda)\},$$

which we assume for simplicity to be the same for all λ, with processes for different λ being independent. The regression coefficient corresponding to this correlation is

$$\beta_{BA}(\lambda) = \rho_{BA}(\lambda; 0)\sigma_Y(\lambda)/\sigma_X(\lambda).$$

Then

$$E\{C_{XY}(t_0)\} = \int_0^{t_0} (t_0 - \lambda)\tau_X(\lambda)\beta_{BA}(\lambda)\,\mathrm{d}\lambda,$$

whereas the denominator of the regression coefficient has expectation

$$\int_0^{t_0} (t_0 - \lambda)\tau_X(\lambda)\,\mathrm{d}\lambda.$$

That is, in general a weighted average of the base-process regression coefficients is estimated, typically with higher weight attached to the components with smaller λ. A very special case is where the regression coefficient is, say, β_0 for the initial base process at $\lambda = 0$ and zero otherwise. Then the standard parameter

estimate of the least squares regression coefficient is attenuated, and its probability limit is

$$\frac{\tau_X(0)\beta_0}{\tau_X(0) + \int_{0+}^{t_0}(1 - \lambda/t_0)\tau_X(\lambda)\,\mathrm{d}\lambda}.$$

To calculate the variance of (4), we condition not only on the observed orthogonalized process $X(\cdot)$ but also on the defining base processes $A(\cdot\,;\lambda)$. The deviation of $Y(t)$ from its regression on the whole $X(\cdot)$ process is denoted by $Y_{\cdot X}(t)$, with corresponding notation for its properties. Then

$$\mathrm{var}\{C_{XY}(t)\} = \int_0^t W_{Y\cdot X}(u)V_{XX}(u)\,\mathrm{d}u + 2\int_0^t W_{Y\cdot X}(u)V_{XX}(u)\mathrm{P}_{Y\cdot X}(t-u)\,\mathrm{d}u, \tag{5}$$

where

$$W_{Y\cdot X}(u) = \mathrm{var}\{Y_{\cdot X}(u)\}, \quad \mathrm{P}_{Y\cdot X}(u) = \int_0^u \rho_{B\cdot A}(v)\rho_{XX}(v)\,\mathrm{d}v.$$

The discussion now largely parallels that of (3). An important distinction, however, is that the term induced in (5) by autocorrelations depends on the properties of both variables. In particular, if $X(\cdot)$ represents a randomized treatment, then $\mathrm{P}_{Y\cdot X}(u) = 0$ and the standard results for regression coefficients apply regardless of autocorrelation in $\{Y(\cdot)\}$. More generally, whereas a positive second term in (5) will indicate overdispersion, the divergence of the integral requires slow convergence of both correlation factors contributing to $\mathrm{P}_{Y\cdot X}(u)$, and so is relatively less likely than in the estimation of the mean.

To apply these ideas in a specific application requires, if at all feasible, a model specific to that application. Failing that, the discussion suggests that regression coefficients may well have major overdispersion relative to the errors assessed assuming independence. The possibility of anomalous dependence on sample size seems less likely than in the corresponding problem of estimating a mean. The most serious possibility of misinterpretation arises when the regression coefficient takes very different values in the different base processes.

Acknowledgement

It is a pleasure to thank two referees for their constructive comments.

References

Beran, J. (1994). *Statistics for Long-Memory Processes*. London: Chapman and Hall.

Cox, D. R. (1984). Long-range dependence: A review. In *Statistics: An Appraisal*, Ed. H. A. David & H. T. David. Ames, Iowa: Iowa State University Press, pp. 55–74.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications*, vol. II. New York: Wiley.

[*Received August* 2014. *Revised May* 2015]