Biometrika (2020), 103, 1, pp. 1–9
Printed in Great Britain
Advance Access publication on 31 July 2018

BBBB

By Linsui Deng
Institute of Statistics and Big Data, Renmin University of China,
Beijing, China
denglinsui@ruc.edu.cn

Summary

BBBBBBB

Some key words: AAA;BBB;CCCCC; DDDDD.

1. Introduction

With the avalibility of massive data, reliable variable selection is a fundamental challenge arising in many fields like genetics, technology, and astronomy. For example, if we are interested in the gene relating to a specific disease, large amounts of gene are examined and we want to select those truly associating with the disease. Formalizing it in statistics, determining whether a particular gene relates to the disease is a hypothesis testing problem while discovering which genes relate to the disease is a multiple testing problem. An ideal testing is discovering as much hypothesis as possible while controlling the false discoveries. There are several criteria measuring false discovery: in frequentist view, the familywise error rate(FWER), the probability of making any false discovery (Simes (1986); Janson & Su (2016)) is in the strong sense and the false discovery rate(fdr), the ratio of false discovery and total discovery (Benjamini & Hochberg (1995)) is in the weak sense; while in Bayesian approach, local false discovery rate(lfdr) is considered (Efron et al. (2001)). When the problem scale is large, controlling FWER is too conservative to limit the potent discovery. At the same time, the lfdr is suitable for large scale problem but the model assumption is needed. Therefore, this paper propose a novel approach to control false discovery rate.

The elementary approach to control false discovery rate is Benjamini and Hochberg(BH) procedure (Benjamini & Hochberg (1995)). Breifly speaking, with p-value associating to each hypothesis, BH procedure develops a data-adaptive threshold and rejects those hypothesis whose p-value is below the threhsold. Benjamini & Yekutieli (2001) justfied that BH procedure strictly control false discovery rate when the test statistics satisfy positive regression dependent strcture corresponding to the true null hypothesis. It also divides the threhsold of BH procedure with a logorithm term to adapt arbitrary dependency structure. Storey (2002) and Genovese & Wasserman (????) determined the threshold ensuring an consistent fdr estimator under desired level. These methods increases the power because the proportion of null hypothesis among all hypothesis is included. Recently, Lei & Fithian (2016), Li & Barber (2018) and Lei et al. (2020) improved BH procedure by incorporating prior information in hypothesis testing.

Barber & Candès (2015) created artificial copies of explanation variables and compared their performance with original copies .

2. Methodology

For each author please give one postal address, including a department, postcode and country, and one e-mail address; these should be the best permanent addresses current at time of publication. Acknowledgements to other institutions should be put with other acknowledgements at the end of the paper. Names of states should be given in full, thus: California rather than CA, São Paulo rather than SP. Use U.S.A. and U.K. Note that England, Scotland and Wales should not be used.

3. Length

The average length for papers published in recent years is around 14 sides. The probability of acceptance drops sharply beyond this length, particularly if it is felt that a paper is long in relation to its original content. Authors should endeavour to write as concisely as possible, consistent with clarity. Long or standard derivations should be omitted, referenced elsewhere, or made available in a supplementary document on the Biometrika web site; see page 8. Essential technical details may be placed in an appendix.

The maximum length for a paper in the Miscellanea section is 8 journal sides.

4. Style

4.1. Sections, subsections and paragraphs

If subsections are used to divide a section, no text should appear before the first subsection; all text should appear within numbered subsections. Subsubsections are not used.

The end of a paragraph is marked in the .tex file by a blank line. Extra characters such as \\ at the end of lines or paragraphs should not be used. Bad line breaks are corrected during the production process.

4.2. Spelling, abbreviations and special symbols

English spelling is used, with Oxford "-ize" endings.

Verbal phrases inside dashes, or in italic or bold type, should not be used and phrases inside brackets should be used sparingly. Quotation marks should be used only for direct quotations, which should be attributed. Footnotes should be avoided except for tables.

Abbreviations should be avoided wherever possible. Exceptions to this are common non-mathematical abbreviations such as DNA and HIV, which appear as ordinary upper-case letters, and, in exceptional cases, where the use of an abbreviation clearly improves the readability of the paper.

Do not create abbreviations to describe methods. Thus 'our method is more efficient than Wellner and Zhang's method' should replace 'our new method is better than method WZ'.

Special symbols like $\xrightarrow{\mathrm{d}}$, $\forall$, $\exists$, := and =: should not be used. The symbol | should not be used in text as shorthand, and in mathematics the TEX symbol \mid should be used to denote conditioning, rather than | or \vert.

Symbols comprising several letters such as aic or ar(p) may be used as mathematical objects if previously defined. They may not be used as abbreviations for English words; thus 'the ar model' should be 'the autoregressive model'. In such cases small capital letters, for example the TEX syntax \textsc{aic} for aic, are used; consistency is best assured by defining a macro at the start of the .tex file.

One of the most common reasons that publication of scientifically acceptable papers is delayed is authorial failure to adhere to journal policy on abbreviations, so it may be worthwhile to explain why Biometrika eschews them. The purpose of scientific writing is to convey ideas as clearly and directly as possible. Abbreviations militate against this: a reader who does not know them will spend time looking back through the paper to find what they mean, and they lead to sloppy mechanistic writing. A sentence such as 'MLE for a GLMM may be performed using the BFGS, NR, CG or EM algorithms, but MCMC is an alternative' forces the reader to waste energy on parsing acronyms rather than focusing on the underlying ideas.

4.3. English

English sentences containing mathematical expressions or displayed formulae should be punctuated in the usual way: in particular please check carefully that all displayed expressions are correctly punctuated. Displayed expressions should be preceded by a colon only if grammatically warranted. Do not place a colon in the middle of a clause.

Words in common terms such as central limit theorem or Brownian motion are capitalized only if they are derived from proper names: thus bootstrap, lasso and mean square error rather than Bootstrap, Lasso and Mean Square Error.

Hyphens - (- in TEX), n-dashes – (--), m-dashes — (---), and minus signs − ($-$) have different uses. Hyphens are used to join two words, or in the double-barrelled name of a single person (e.g. non-user, Barndorff-Nielsen); n-dashes are used in ranges of numbers or to join the names of two different people (1–7, Neyman–Pearson); and minus signs are used in mathematics (e.g. -2). m-dashes are not used in Biometrika. Parenthetical remarks, like this subordinate clause, are placed between two commas.

Two bugbears: the phrase 'note that' can almost always be deleted, and the phrase 'is given by' should be cut to 'is' in a sentence such as 'The average is given by $\bar{X} = n^{-1}(X_1 + \cdots + X_n)$.

4.4. Mathematics

Equation numbers should be included only when equations are referred to; the numbers must be placed on the right. Long or important mathematical, not verbal, expressions should be displayed, i.e., shown on a separate line. Short formulae should be left in the text to save space where possible, but must not be more than one line high and not contain reduced-size type. For example $\frac{\mathrm{d}y}{\mathrm{d}x}$ must not be left in the text, but should be written $\mathrm{d}y/\mathrm{d}x$ or it should be displayed. Likewise write $n^{1/2}$ not $n^{\frac{1}{2}}$. Also $\binom{a}{b}$ and suchlike expressions must not be left in the text. Equations involving lengthy expressions should, where possible, be avoided by introducing suitable notation.

Symbols should not start sentences. Distinctive type, e.g., boldface, for matrices and vectors is not used in Biometrika. Vectors are assumed to be column vectors, unless explicitly transposed. The use of an apostrophe to denote matrix or vector transposition should be avoided; it is preferable to write A^{T}, a^{T}. Capital script letters may be used sparingly, typically to denote sets, but care should be taken as some are hard to distinguish.

Please arrange brackets in the order $[\{( )\}]$, iterating as necessary, and follow the usual conventions for e, exp, use of solidus, square root signs and so forth as in a recent issue. The sign $\sqrt{\ }$ is not used, and the sign $\surd$ is used only sparingly; powers of complicated quantities should be represented as $(mnpq)^a$.

Multiple overbars such as $\bar{\bar{x}}$ must be avoided, as must $\widehat{ab}, \widehat{(a+b)}, \widehat{\text{var}}, \overline{ab}, \overline{(a+b)}$ and symbols with underbars. Subscripts and superscripts, and second-order sub- and superscripts, should be aligned horizontally. Avoid sub- and superscripts of third, and greater, order.

Please use: $\text{var}(x)$ not var x or $\text{Var}(x)$; cov not Cov; pr for probability not Pr or P; tr not trace; $E(X)$ for expectation not EX or $\mathcal{E}(X)$; $\log x$ not $\log_e x$ or $\ln x$; rth not r-th or r^{th}. Please avoid: '$\cdot$' or '$\cdot$' for product; a/bc, which should be written $a/(bc)$ or $a(bc)^{-1}$. Use the form $x_1, \ldots, x_n$ not $x_1, x_2, \ldots x_n$ and $\sum_{i=1}^{n}$ not $\sum_1^n$. Zeros precede decimal points: 0.2 not .2.

The use of '$\cdots$' and '...' is ... in lists, such as $y_1, \ldots, y_n$, and $\cdots$ between binary operators, giving $y_1 + \cdots + y_n$. Ranges of integers are denoted $i = 1, \ldots, n$, whereas $0 \leq x \leq 1$ is used for ranges of real numbers.

Biometrika deprecates the appearance of words in displayed equations, which should be formatted as

$$\bar{Y} = n^{-1} \sum_{j=1}^{n} Y_j, \quad S^2 = \sum_{j=1}^{n} (Y_j - \bar{Y})^2; \tag{1}$$

note the punctuation and space between the expressions. Displays such as (1) should take no more space than necessary, being placed on a single line where possible. Displayed mathematical expressions should be punctuated thus: indexed equations and similar quantities in text are formatted as $y_j = x_j^{\mathrm{T}} \beta + \varepsilon_j$ $(j = 1, \ldots, n)$, and are displayed as

$$y_{ij} = x_j^{\mathrm{T}} \beta_i + \varepsilon_{ij} \quad (i = 1, \ldots, m;\ j = 1, \ldots, n).$$

References to sequences of equations are (1)–(3), not (1–3).

4.5. Figures

Figures are a common source of delay during production, usually because elementary guidelines have not been respected. General comments may be found in the document 'RSSGraphs.pdf' enclosed with the Biometrika formatting files, and more detail is given in standard references such as Cleveland (1993, 1994) or Tufte (1983).

All the elements of a graph, including axis labels, should be large enough to be read easily, so the graph should be given a shape that will use the page space well. The use of large symbols, such as $\times$, for points should be avoided. If both axes of a panel show the same quantities, the panel should usually be square. Many graphs are made using the statistical environment R (R Development Core Team, 2012). If so, they should be made at roughly the size at which they will appear in the journal. Usually graphs reduced from A4 or US page sizes must be remade to ensure their legibility.

Check that all the axes are labelled correctly and include units of measurement. Axis labels should have the format 'Difference of loglikelihoods': only the initial letter of the first word is upper-case. The numbers on the vertical axis should be parallel to the horizontal axis, and should be in the same font as the text; normally the change of font is left to the production process, but it is helpful if the numbers are placed horizontally.

A panel should not contain an inset defining the line-types and symbols; this description should appear in the caption.

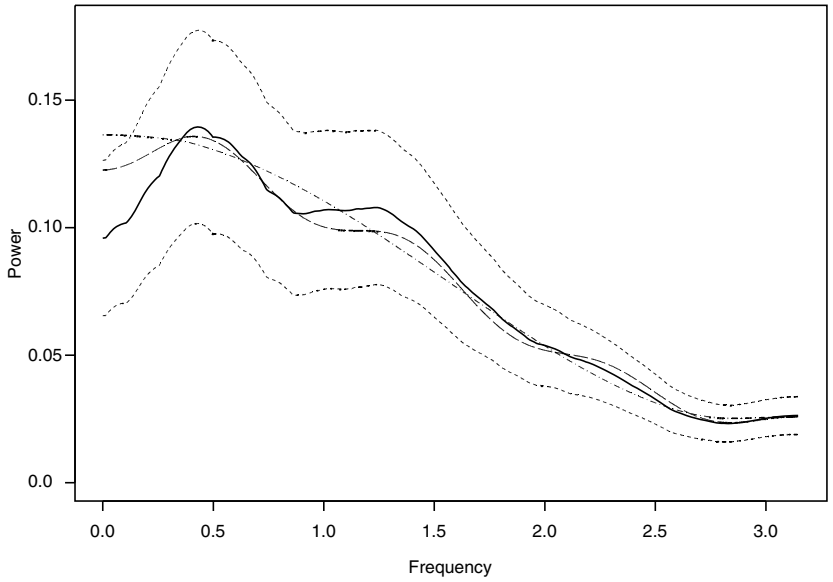

Fig. 1. A graph showing the truth (dot-dash), an estimate (dashes), another estimate (solid), and 95% pointwise confidence limits (small dashes).

Please submit figures in greyscale wherever possible. Biometrika publishes in colour where this is essential, but care should be taken to ensure that any colours chosen will be distinguishable on the cream paper used for the journal, i.e., yellow should be avoided.

Figures should be referred to consecutively by number. Use of the LaTeX \label and \ref commands to refer to figures and tables helps to reduce errors and so is preferred. Figure 1 is a reference to a figure at the start of a sentence, whereas subsequent references are abbreviated, for example to Fig. 1.

4.6. Tables

Tables should be referred to consecutively by number. Table is not abbreviated to Tab.

Check that the arrangement makes effective use of the Biometrika page. Layouts that have to be printed sideways should be avoided if possible. For this reason tables should not be more than 92 characters wide, including decimal points and brackets (1 character), and minus and other signs and spaces (at least 2 characters). Rules are not used in Biometrika tables, which should be arranged to be clear without them.

Often tables can be improved by multiplying all the entries by a power of ten, so that 0.002 and 0.02 become 2 and 20 respectively, for example; this will often both save space and convey the message of the table more effectively. Table 1 uses the definition \def~{\hphantom{0}} to insert invisible spaces into columns of the table; see the source code for this document.

Very often tables containing results of Monte Carlo simulations use more digits than can be justified by the size of the simulation, and space can be gained and clarity improved by appropriate rounding. Standard errors or some other measure of precision should be given for Monte Carlo results. Often it suffices to give a phrase such as 'The largest standard error for the results in column 2 is 0.01.' in the caption to the table.

Table 1. Perceptions about racial groups in the U.S. population

	2000 Census percent of U.S. population	Mean percent estimated for U.S. population		
		White Rs	Black Rs	Hispanic Rs
White	75	59	56	60
Black	12	30	38	40
Asian	4	16	21	30
American Indian	1	13	17	23
More than two races	2	41	48	50
Hispanic	13	23	27	42

U.S., United States of America; R, respondent.

4.7. Captions to figures and tables

The caption to a figure should contain descriptions of lines and symbols used, but these should not be duplicated in the text, which should give the interpretation of the figure. The figure should not contain an inset. Verify that the caption and the graph agree, that every line and symbol is described correctly and that all lines or symbols in the graph are described in the caption.

Any abbreviations used in the body of a table should be explained in a footnote to it.

Figure captions always end with a full stop. The last sentence of a table title does not have a full stop.

4.8. References

References in the text should follow the current style used in Biometrika. It is preferable to use BibTeX if possible, as in this guide. In citing references use 'First author et al.' if there are three or more authors. The list of references at the end should correspond to those in the text, and be in exactly current Biometrika form.

References to books should be to the latest edition; a page, section or chapter number is nearly always necessary. References to books of papers should include title of book, editor(s), first and final page numbers of paper, where published and publisher.

Complete lists of authors and editors should be given; in exceptional cases they may be abbreviated at the discretion of the editor.

PhD theses, unpublished reports and articles can be referred to in the text, using a phrase such as 'as shown in a 2009 Euphoric State University Department of Statistics PhD thesis by M. Zapp' or 'the proofs may be found in an unpublished 2003 technical report available from the first author', but should not be included in the References except where they have been accepted for publication, and unless they appear in a permanent repository such as arXiv; in this case the most recent version of the work is cited like a paper, e.g., Berrendero (2015).

URLs should be presented using the \texttt{} font in LaTex. Avoid giving URLs of pages that are likely to become obsolete quickly. Technical details for published papers should be prepared as Supplementary Material, so that they remain permanently available. Likewise software should be submitted as Supplementary Material; it should be adequately documented, e.g., by including a README file to accompany R code.

Cox (1972) is an example of an active citation, and an example of a passive citation is (Heard et al., 2006). The abbreviations for their journals should be noted.

4.9. Theorem-like environments

Biometrika does not use LaTeX list environments such as itemise, description, or enumerate. In this subsection we illustrate the use of theorem-like environments.

Definition 1 (Optional argument). This is a definition.

Assumption 1 (Another optional argument). This is an assumption.

Proposition 1. This is a proposition.

Lemma 1. This lemma precedes a theorem.

Proof. This is a proof of Lemma 1. Perhaps it should be placed in the Appendix. □

Theorem 1. This is a theorem.

Some text before we give the proof.

Proof of Theorem 1. The proof should be here. □

Example 1. This is an example.

Some text before the next theorem.

Theorem 2 (Optional argument). Another important result.

Corollary 1. This is a corollary.

Remark 1. This is a remark.

Step 1. This is a step.

Condition 1. This is a condition.

Property 1. This is a property.

Restriction 1. This is a restriction.

Algorithm 1. A simple algorithm.

Set $s = 0$
For $i = 1$ to $i = n$
    Set $t = 0$
    For $j = 1$ to $j = i$
        $t \leftarrow t + x_{ij}$
    $s \leftarrow s + t$
Output s

5. Discussion

This is the concluding part of the paper. It is only needed if it contains new material. It should not repeat the summary or reiterate the contents of the paper.

Acknowledgement

Acknowledgements should appear after the body of the paper but before any appendices and be as brief as possible subject to politeness. Information, such as contract numbers, of no interest to readers, must be excluded.

Supplementary material

Further material such as technical details, extended proofs, code, or additional simulations, figures and examples may appear online, and should be briefly mentioned as Supplementary Material where appropriate. Please submit any such content as a PDF file along with your paper, entitled 'Supplementary material for Title-of-paper'. After the acknowledgements, include a section 'Supplementary material' in your paper, with the sentence 'Supplementary material available at Biometrika online includes ...', giving a brief indication of what is available. However it should be possible to read and understand the paper without reading the supplementary material.

Further instructions will be given when a paper is accepted.

Appendix 1

General

Any appendices appear after the acknowledgement but before the references, and have titles. If there is more than one appendix, then they are numbered, as here Theorem A1.

Theorem A1. This is a rather dull theorem:

$$a + b = b + a; \tag{A1}$$

a little equation like this should only be displayed and labelled if it is referred to elsewhere.

Lemma A1. If $\alpha_j > 2$, $\eta_j/\alpha_j = O(j^{-m})$ $(j = 1, \ldots, \infty)$ and $m > 1/2$, then $P_l(\mathcal{C}) = 1$.

Appendix 2

Technical details

Often the appendices contain technical details of the main results.

Theorem B1. This is another theorem full of gory details.

Lemma B1. If $\delta > 2$, $\rho > 0$, $\alpha_j(\delta) = \delta^j$ and $\eta_j(\rho) = \rho$ for $j = 1, \ldots, \infty$, then $P_l(\mathcal{C}) = 1$, where P_l has density p_{mgdP} in (4) with hyperparameters $\alpha_j(\delta)$ and $\eta_j(\rho)$ $(j = 1, \ldots, \infty)$. Furthermore, given $\epsilon > 0$, there exists a positive integer $k(p, \delta, \epsilon) = O\{\log^{-1}\delta \log(p/\epsilon^2)\}$ for every Ω such that for all $r \geq k$, $\alpha_j(\delta) = \delta^j$, $\eta_j(\rho) = \rho$ $(j = 1, \ldots, r)$ and $\Omega^r = \Lambda^r \Lambda^{r\mathrm{T}} + \Sigma$, we have that $\mathrm{pr}\{\Omega^r \mid d_\infty(\Omega, \Omega^r) < \epsilon\} > 1 - \epsilon$ where $d_\infty(A, B) = \max_{1 \leq i,j \leq p} |a_{ij} - b_{ij}|$.

Appendix 3

Often the appendices contain technical details of the main results:

$$a + b = c. \tag{C1}$$

Remark C1. This is a remark concerning equations (A1) and (C1).

Lemma C1. The conditional density model $\mathcal{M}$ of § 3 is sequentially strongly convex with $H_k(p)(z) \equiv p\left(a_k \mid \bar{l}_k, \bar{a}_{k-1}\right)$.

References

Barber, R. F. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. The Annals of Statistics 43, 2055–2085.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29, 1165–1188.

Berrendero, J. R., C. A. . T. J. L. (2015). On the use of reproducing kernel hilbert spaces in functional classification. arXiv , 1507.04398v3.

Cleveland, W. S. (1993). Vizualizing Data. Summit: Hobart Press.

Cleveland, W. S. (1994). The Elements of Graphing Data. Summit: Hobart Press, revised ed.

Cox, D. R. (1972). Regression models and life tables (with Discussion). J. R. Statist. Soc. B 34, 187–220.

Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151–1160.

Genovese, C. & Wasserman, L. (????). A stochastic process approach to false discovery control. The Annals of Statistics 32.

Heard, N. A., Holmes, C. C. & Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. J. Am. Statist. Assoc. 101, 18–29.

Janson, L. & Su, W. (2016). Familywise error rate control via knockoffs. Electron. J. Statist. 10, 960–975.

Lei, L. & Fithian, W. (2016). Adapt: An interactive procedure for multiple testing with side information. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80.

Lei, L., Ramdas, A. & Fithian, W. (2020). A general interactive framework for false discovery rate control under structural constraints. Biometrika .

Li, A. & Barber, R. F. (2018). Multiple testing with the structure-adaptive benjamini-hochberg algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology) .

R Development Core Team (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.

Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B (Methodological) 64, 479–498.

Tufte, E. R. (1983). The Visual Display of Quantitative Information. Cheshire: Graphics Press.

[Received on 2 *January* 2017. Editorial decision on 1 *April* 2017]