華醫學報 第 54 期
2021 年 6 月
頁數 1-10

The Extension of Quantity Group

Chiahua Ling

Department of Accounting, Shih Chien University

Abstract

We use a new approach which simulates the idea of Bayesian method with a difference. In traditional Bayesian method one considers a prior distribution, may be a conjugate one, which is centered at super parameters. Our method of data augmentation is supposed to help for small samples or cases where a very few observations are available. We use prior distributions, centered at the given observations (calling them to be super parameters), to generate a larger artificial dataset which may be termed as second generation dataset. This larger second generation dataset is then used to draw inferences. The powerful computers, coupled with suitable numerical algorithms, caused an increased interest in nonlinear models (such as neural networks) as well as the creation of new types, such as generalized linear models and multilevel models. Increased computing power has also led to the growing popularity of computationally intensive methods based on resampling, such as permutation tests and the bootstrap, while techniques such as Gibbs sampling have made use of Bayesian models more feasible.

Keywords: Conjugate distribution, Gibbs sampling, Order statistics, Prior distribution

Chiahua Ling (凌嘉華) E-mail: ling@g2.usc.edu.tw
Associate Professor, Department of Accounting, Shih Chien University

華醫學報 第 54 期
2021 年 6 月
頁數 1-10

數量群的延伸

凌嘉華

實踐大學會計學系

摘要

這篇研究我們使用和貝氏技術些微不同的方法，傳統的貝氏方法是假設有先驗分配(可能是共軛分配)和以超參數為中心。我們用小樣本中非常少的觀察值作資料增加擴大，用以觀察值(假設它們是超參數)為中心的先驗分配來產生一個較大的資料集(稱作第二產生資料集)，這個第二產生資料集可以用來做統計推論，電腦及其演算法導致非線性模型（如神經網路）和新式演算法（如廣義線性模式、等級線性模型）的大量應用，電腦效能的增強使得需要大量計算的再取樣演算法成為時尚，如置換檢定、自助法。Gibbs 取樣法也使得貝氏模型更加可行。

關鍵詞：共軛分配、次序統計量、先驗分配、吉布斯取樣法

Introduction

Increased computing power has also led to the growing popularity of computationally intensive methods based on resampling, such as permutation tests, the EM algorithm (see Dempster, Laird and Rubin (1997)), and the bootstrap (see Efron (1982)), while techniques such as Gibbs sampling (see Casella and George (1992)) have made use of Bayesian models more feasible. To use a sample as a guide to an entire population, it is important that it truly represents the overall population. Computational resources and availability of affordable computational facilities have changed the face of research in mathematical sciences including statistical science. Many theoretical results can now be verified or be motivated with the help of simulation and/or numerical computations before they are proved analytically. The purpose of this note is also to propose a simple technique with the help of computational resources which can enable to understand a small dataset with more clarity.

Both methods, the bootstrap and the jackknife, estimate the variability of a statistic from the variability of that statistic between subsamples, rather than from parametric assumptions. For the more general jackknife, the delete-m observations jackknife, the bootstrap can be seen as a random approximation of it. Both yield similar numerical results, which is why each can be seen as approximation to the other. Although there are huge theoretical differences in their mathematical insights, the main practical difference for statistics users is that the bootstrap gives different results when repeated on the same data, whereas the jackknife gives exactly the same result each time. The bootstrap is preferred (e.g., studies in physics, economics, biological sciences).

Usually the jackknife is easier to apply to complex sampling schemes than the bootstrap. Complex sampling schemes may involve stratification, multiple stages (clustering), varying sampling weights (non-response adjustments, calibration, post-stratification) and under unequal-probability sampling designs. The bootstrap estimate of model prediction bias is more precise than jackknife estimates with linear models such as linear discriminant function or multiple regression.

The idea of our proposed technique is very simple, and borrows the idea of standard Bayesian method. It has a flavor of parametric bootstrap since the technique is built up on a parametric model for the idea, and the dataset has been used twice. In a nutshell the proposed technique is described as follows. Assume we

have iid observations $X_1, X_2, \ldots, X_n$ from a distribution $f(x \mid \theta)$. When n is large, efficient inferences on θ is possible with the help of classical methods. But when n is small (very small) then one may seek the help of other methods notably the Bayesian one. In such a case one assumes a suitable prior distribution $\theta \sim \pi(\theta)$, often a conjugate one, to draw inferences on θ. What we are proposing here is that given the original dataset $(X_1, X_2, \ldots, X_n)$, call it first stage observations, generate the augmented dataset or the second stage dataset X_{ij} (j = 1, 2, …, m_i ; i = 1, 2, … , n) as X_{ij} iid $f(\cdot|x_i)$, j = 1, 2, …, m_i. Once this is done, then draw the inferences on θ based on $X_{ij} (1 \le \mathrm{j} \le \mathrm{m_i}; 1 \le i \le n)$ with a new sample size $m = m_1 + m_2 + \ldots + m_n$ which can be made much larger than the original sample size n. Computer-intensive resampling/bootstrap methods are feasible when calculating reference intervals from non-Gaussian or small reference samples. Parametric methods are preferable to resampling methods when the distributions of observations in the reference samples is Gaussian or can transformed to that distribution even when the number of reference samples is less than 120. Resampling methods are appropriate when the distribution of data from the reference samples is non-Gaussian and in case the number of reference individuals and corresponding samples are in the order of 40. At least 500-1000 random samples with replacement should be taken from the results of measurement of the reference samples.

In the following section we describe the above mentioned method for the logistic distribution. This is still an ongoing work and we hope to expand it for other distributions as well.

Logistic Distribution

Now let us consider the logistic model where $X_1, X_2, \ldots, X_n$ are iid logistic(μ , σ) with

$$f(x;\mu,\sigma) = \frac{e^{-\frac{(x-\mu)}{\sigma}}}{\sigma\left(1+e^{-\frac{(x-\mu)}{\sigma}}\right)^2} \qquad -\infty < x < \infty.$$

$$, -\infty < \mu < \infty, \sigma > 0.$$

After observing x_i's , the second stage observations are generated as

$$X_{11}, X_{12}, \cdots, X_{1m_1} iid.Logistic(X_1);$$
$$\vdots$$
$$X_{n1}, X_{n2}, \cdots, X_{nm_n} iid.Logistic(X_n);$$

As a demonstration, we generate n = 15 iid observations from logistic(0,1) distribution (taking μ =0, σ =1). The following figure gives the pdf curve superimposed on the relative frequency histogram of the original sample.

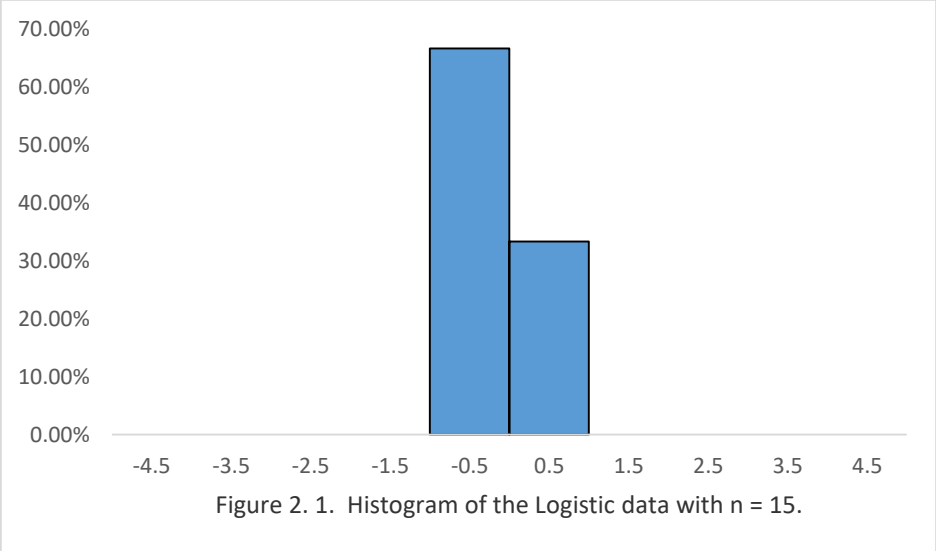

Figure 2. 1. Histogram of the Logistic data with n = 15.

As a demonstration, we generate n = 15 iid observations from logistic(0,1) distribution (taking μ =0, σ =1). The above figure gives the pdf curve superimposed on the relative frequency histogram of the original sample.

Using $m_1 = m_2 = \ldots = m_{15} = 10 = m_0$ (say) = 10, we now generate $m = nm_0 = 150$ second stage observations from the above first stage observations. The following diagram shows the relative histogram

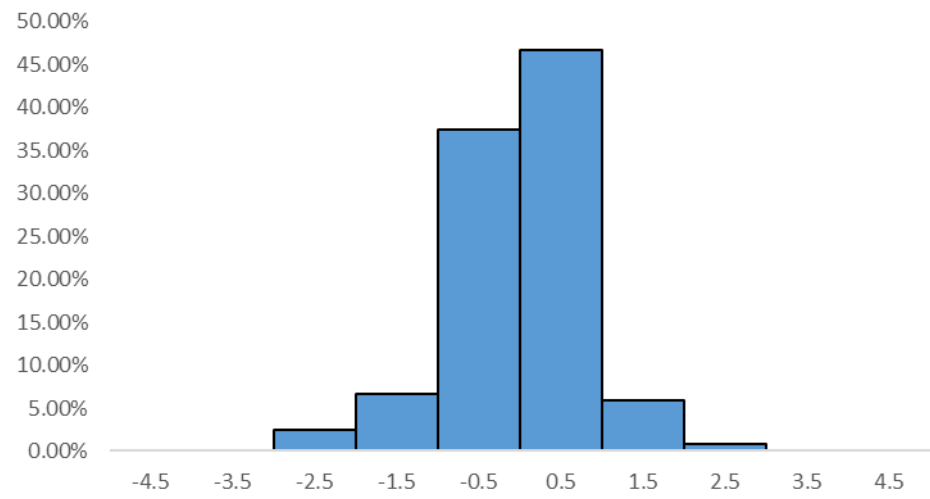

Figure 2. 2. Histogram of the Logistic data with n = 120.

A modified sampling scheme can be used where $m_i's$ can be chosen suitably. Since the model is positively skewed, we'll give more emphasis on the smaller observations which have higher probabilities to appear. So, first order the observations $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. Then take m_i proportional to (n + 1 - i). For instance, take (n + 1 - i) ; i.e. , m_1 = 15, m_2 = 14, … , m_{15} = 1, and m = 120. So, generate $X_{i1}, X_{i2}, \cdots, X_{im_i}$ iid Logistic$(X_{(i)})$. Thus the second stage sample size is m = 120 and the following figure shows the relative histograms.

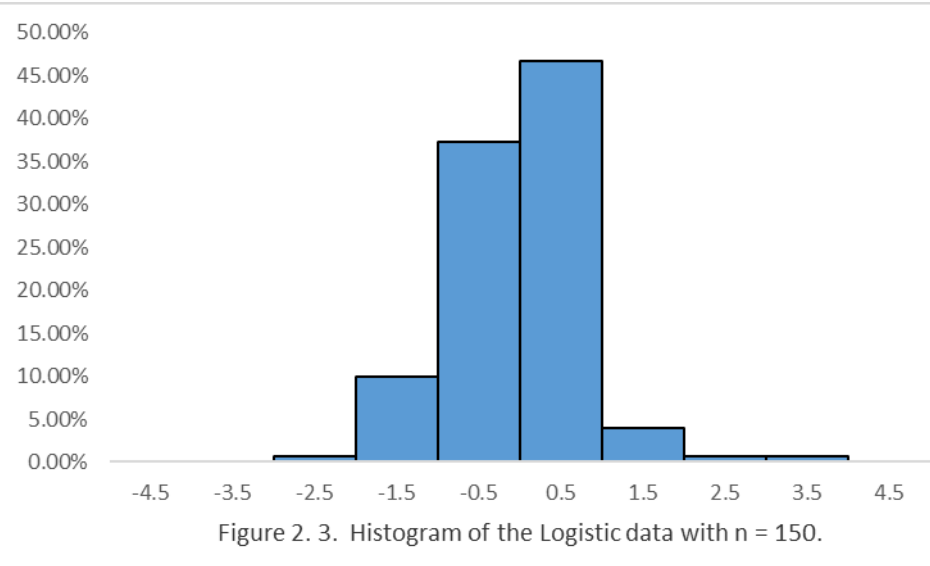

Figure 2. 3. Histogram of the Logistic data with n = 150.

Remark 2.1. If n = 2k, then the second stage sampling scheme can be modified accordingly. The sample median is then $(X_{(k)} + X_{(k+1)})/2$. One can take $m_i \propto 1/(|(k+1/2)-i|+1)$, $1 \le i \le n$.

Remark 2. 2. How do we estimate θ with the second stage data ?

Note that with the above modified sampling scheme. $X_{i1}, \cdots, X_{im_i}$ are iid $\text{Exp}(X_{(i)})$. Therefore, $X_{i\cdot} = \sum_{j=1}^{m_i} x_{ij} \mid x_{(i)}$ follows Gamma $(x_{(i)}, m_i)$ with pdf $(x_{(i)}{}^{m_i}/\Gamma(m_i))\exp(-x_{(i)}x_{i\cdot})x_{i\cdot}{}^{m_i-1}$. So, from the first stage data, we have the second stage sufficient statistic $(X_{1\cdot}, X_{2\cdot}, \cdots, X_{n\cdot})$. $X_i \sim \text{Exp}(\theta)$, $x_{i\cdot} \mid x_i \sim$ Gamma (x_i, m_i). So, jointly $(x_i, x_{i\cdot}) \sim$ pdf $(\theta x_i{}^{m_i}/\Gamma(m_i))\exp\{-x_i(\theta + x_{i\cdot})\}x_{i\cdot}{}^{m_i-1}$. Therefore, marginally, $x_{i\cdot}$ follows $g_\theta(x_{i\cdot}) = m_i x_{i\cdot}{}^{m_{i-1}}\theta(\theta + x_{i\cdot})^{-(m_i+}$. Using $g(x_{i\cdot})$ as likelihood for each $x_{i\cdot}$, one can get the joint likelihood as $L_A = \prod_{i=1}^{n} g_\theta(x_{i\cdot})$, called the likelihood of the augmented data. The value of θ which maximizes L_A (or $\ln L_A$) is denoted by $\hat{\theta}_{(2)}$ which satisfies the equation

$$\text{N} = \hat{\theta}\sum_{i=1}^{n}(m_i + 1)/(\hat{\theta} + x_{i\cdot}). \qquad \ldots (2.1)$$

In the following we observe the bias and SE of the above $\hat{\theta}_{(2)}$ along with those of $\hat{\theta}_{(1)}$ here $\hat{\theta}_{(1)} = (1/\bar{x})$ = the MLE of θ based on the original sample. Using N = 10^4, we compute

$$\text{B}(\hat{\theta}_{(1)}) \approx \sum_{j=1}^{N}\hat{\theta}_{(1)}^{j}/N - 1 \;; \qquad \text{B}(\hat{\theta}_{(2)}) \approx \sum_{j=1}^{N}\hat{\theta}_{(2)}^{j}/N - 1;$$

$$\mathrm{SE}(\hat{\theta}_{(1)}) \approx \sqrt{\sum_{j=1}^{N}(\hat{\theta}_{(1)}^{(j)} - \hat{\theta}_{(1)}^{(\cdot)})^2 / N} \quad ; \quad \mathrm{SE}(\hat{\theta}_{(2)}) \approx \sqrt{\sum_{j=1}^{N}(\hat{\theta}_{(2)}^{(j)} - \hat{\theta}_{(2)}^{(\cdot)})^2 / N} \quad ;$$

The computed values are

$\mathrm{B}(\hat{\theta}_{(1)}) = 0.00153071$; $\mathrm{B}(\hat{\theta}_{(2)}) = -0.628425$;

$\mathrm{SE}(\hat{\theta}_{(1)}) = 0.2389213$; $\mathrm{SE}(\hat{\theta}_{(2)}) = 0.1789137$.

Conclusions and Further Applications

In this study we proposed the method of standard Bayesian technique with a difference. It has a flavor of parametric bootstrap since the technique is built up on a parametric model for the idea, and the dataset has been used twice which can be made much larger than the original sample size n. Our approach of data augmentation is supposed to help for small samples or cases where a very few observations are available. We use some computational resources, and may be useful in applied problems. The method is dependent on computational resources, and may be useful in applied statistical problems, and powerful computers, coupled with suitable numerical algorithms, caused an increased interest in nonlinear models (such as neural networks) as well as the creation of new types, such as generalized linear models and multilevel models. Increased computing power has also led to the growing popularity of computationally intensive methods based on resampling, such as permutation tests and the bootstrap, while techniques such as Gibbs sampling have made use of Bayesian models more feasible. The computer revolution has implications for the future of statistics with a new emphasis on "experimental" and "empirical" statistics. Bootstrapping is where you sample a value from a population of data and then replace that value before drawing another value. If your data has rare extreme values, bootstrapping will undervalue these observations. If these extreme values are partly a result of mistakes, then this might be good. If these are a true part of the underlying distribution then this will be bad because it will underestimate the variability in the underlying population. Randomization is where you withdraw values from a population of data without replacement. If one uses all of the data, then the rare original observations will be as common in each randomization as they were in the original data. In the future, we can use the proposed method for Laplace and Pareto distributions.

References

Casella, G. and Berger, R. L. (2001). Statistical Inference, 2nd ed. Brooks/Cole Cengage Learning.

Casella, G. and George, E. (1992). Weibulllaining the Gibbs sampler. *The American Statistican*, 46, (pp. 167-174).

Chung, E. and Romano, P. J. (2011). "Asymptotically valid and exact permutation testsbased on two-sample UU-statistics". Technical Report 2011-09, Dept. Statistics,Stanford Univ.

Chung, E. and Romano, P. J. (2013). "Supplement to "Exact and asymptotically robust permutation tests". DOI:10.1214/13-AOS1090SUPP.

Chung, EY and Romano, JP (2013). "Exact and asymptotically robust permutation tests". The Annals of Statistics. 41 (2): 487–507.

Clayton, D., (2003). "Conditional likelihood inference under complex ascertainment using data augmentation." Biometrika, 90: 976–981.

Collingridge, Dave S. (2012). "A Primer on Quantitized Data Analysis and Permutation Testing", Journal of Mixed Methods Research. 7 (1): 81–97.

David, Herbert A. and Nagaraja, Haikady N. (2003). Order Statistics, 3rd ed. *Wiley Series in Probability and Statistics*. John Wiley & Sons.

Delaigle, A., Hall, P. and Jin, J. (2011). "Robustness and accuracy of methods for high dimensional data analysis based on Student's tt-statistic". J. R. Stat. Soc. Ser. BStat. Methodol. 73 283–301.

Del Moral, Pierre (2013). "Mean field simulation for Monte Carlo integration". Monographs on Statistics & Applied Probability, Chapman & Hall/CRC Press. p.626.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society for Industrial and Applied Mathematics*, Philadelphia, Pa., USA.

Everitt, B.S. (2002) The Cambridge Dictionary of Statistics, UP. ISBN 0-521-81099-X

Faraway, J. J. (1992). On the Cost of Data Analysis. *Journal of Computational and Graphical Statistics* 1, (pp. 213-29).

Johnson, N.L., Kotz S. and Balakrishnan, N. (1994) Continuous Univariate Distributions, Wiley. ISBN 0-471-58495-9. p. 60

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer-Verlag.

Kamakura, W.A., Wedel, M., Rosa, F.D. and Mazzon, J.A. (2003). "Cross Selling through Database Marketing: A Mixed Data Factor Analyzer for Data

Augmentation and Prediction," International Journal of Research in Marketing, Vol. 20, No. 1, pp.45-65.

Kosuke, I. and David, A.V.D. (2005). "A Bayesian Analysis of the Multinomial Probit Model Using Marginal Data Augmentation," Journal of Econometrics, Vol.124,No.2, pp.311-344.

Liu, Jun S. and Wu, Ying Nian. (1999). "Parameter Expansion for Data Augmentation," Journal of the American Statistical Association, Vol. 94, No. 48. pp, 1264-1274.

Minguillon, J. and Pujol, J. (2001). "JPEG standard uniform quantization error modeling with applications to sequential and progressive operation modes" (PDF). Journal of Electronic Imaging. 10 (2): 475–485.

Norton, Robert M. (May 1984). "The Double Exponential Distribution: Using Calculus to Find a Maximum Likelihood Estimator". The American Statistician. American Statistical Association. 38 (2): 135–136.

Omelka, M. and Pauly, M. (2012). "Testing equality of correlation coefficients in a potentially unbalanced two-sample problem via permutation methods". J. Statist. Plann. Inference 142 1396–1406.

Pauly, M. (2011). "Discussion about the quality of F-ratio resampling tests for comparing variances". TEST 20 163–179.

Pelosi, M. K. and Sandifer, T. M. (2002). Airspace Data Set, *Doing Statistics For Business: Data, Inference, and Decision Making*. 2nd ed. John Wiley & Sons, Inc., New York.

Tanner, M. A. (1991), "Tools for Statistical Inference. Observed Data and Data Augmentation Methods," New York: Springer- Verlag.

Tanner, T. A. and Wong, W. H. (1987), "The Calculation of Posterior Distribution by Data Augmentation," Journal of the American Statistical Association, 82, 528-550.

Van Dyk, D. A., and X. -L. Meng, (2001). "The art of data augmentation" (with discussion). J. Comput. Graph. Stat. 10: 1–111.

Wei, G. C. G., and Tanner, M. A. (1990). "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm." Journal of the American Statistical Association, 85, 699-704.