SciMedCentral
Annals of Biometrics & Biostatistics

Short Communication

Statistical Methods of Handling Noise in Data Processing

Hoa Le[1,2], Uyen Pham[1], Nguyen Thanh Nguyen[2], and Pham The Bao[2]*

[1]Department of Mathematics and Economic Statistics, University of Economics and Law, Vietnam
[2]Department of Computer Science, University of Science Ho Chi Minh City, Vietnam

***Corresponding author**

Pham The Bao, Department of Computer Science, Mathematics and Computer Science Faculty, University of Science Ho Chi Minh City, Vietnam, Tel: 84-28-38350098; Email: ptbao@hcmus.edu.vn

Submitted: 20 October 2017

Accepted: 13 December 2017

Published: 16 December 2017

ISSN: 2333-7109

Keywords

- Noise in data
- Frequentist statistics
- Bayesian statistics
- Time series

Abstract

Statistical noise is usually a main concern in collecting data. Technical malfunction of devices or asynchronous data collection could easily lead to noise appearance. In this paper, we provide some methods for handling noise through the development stages in statistics. While the traditional frequentist approaches could lead to errors in forecasting, methods using Bayesian Statistics “framework” are proposed to deal with noise in data, and issues that need to be improved in these methods are also mentioned.

INTRODUCTION

Data processing in frequentist statistics generally follows the parametric and nonparametric methods [1]. In parametric methods, we begin with an assumption of normal distribution for population. However, real data hardly follow normal distribution which makes it more difficult to process with noise.

Nonparametric methods do not require an assumption about the specific form of the population’s probability distribution. Nevertheless, the problems in these methods would be based on the measures of central tendency (equivalent to mean) used in estimation and parameter test and the measures of variation (equivalent to standard deviation).

Problems in frequentist statistics all have one thing in common: the *p* - value. Controversy surrounding use of *p* - value hypothesis tests is drawing a lot of attention statistics circles, as in [2-4].

Indeed, using frequentist Statistics to process data could lead to mistaken inferences since the calculated results rely on the samples observation which causes the tests or the estimation efficiency to be unstable under small changes of the underlying distribution, even if there are some alternative methods in [5]. Therefore, if data contain outliers, as in [6], classical estimates such as sample mean, sample variance, sample covariances and correlations, or the least squares fit of a regression model will be misleading. In order to restrain the mistakes, Maronna et al. [6], suggested robust parameter estimates when there are outliers.

However, when estimating parameter by robust methods, e.g. the median as central tendency of data, we might deal with integration mess.

As it might be seen, from frequentist point of view, estimated parameters (parametric or nonparametric) are considered as constants while in reality, these estimates always change. Therefore, we need to change our mind that these parameters should be represented as probability distributions, and this exactly what Bayesian Statistics assumption is, where parameters are considered as random variables [7].

Bayesian Statistics, nonetheless, has difficulties in choosing appropriate prior distribution such that the posterior distribution best matches reality. In addition, noise in data can make likelihood function and prior distribution deviated from reality. Even though Bolstad and Curran in [8] have proposed Robust Bayesian Methods, using median as central tendency of data can also result in imprecise inference.

In this paper, we suggest an efficiency evaluation of a statistical method through the performance of forecasting observations, which is from Shmueli’s point of view [9].

In the next sections, we will give a brief summary of some methods of forecasting for stationary time series, including Frequentist Statistics in section 2, Robust Statistics in section 3, and Bayesian Statistics and Inferential models in section 4, accompanied by misleading that one’s may encounter when processing data with noise. The last section will be the conclusion.

STATIONARY TIME SERIES AND FORECASTING METHODS BASED ON FREQUENTIST STATISTICS

Some of the first forecasting methods that are appropriate for a time series include moving averages, weighted moving averages and exponential smoothing. Besides, a brief summary of methods that are appropriate for time series exhibiting a horizontal or a trend pattern is also provided in this section [1].

Averages method: The forecasts of all future value are equal to the mean of the historical data (2-1).

$$\hat{y}_T + h / T = \bar{y} = \frac{y_1 + y_2 + \ldots + y_T}{T}, \qquad (2\text{-}1)$$

where $y_1, y_2, \ldots y_T$ are historical data. This method will be illustrated through a real data set which is available as a Supplementary Appendix at IJE Online [10] (Figure 1) represents how the data is distributed in time series.

Drift method: This method uses the average of the most recent k data values in the time series as the forecast for the next period (2-2).

$$\hat{y}_T + h / T = y_T + \frac{h}{T-1}\sum_{T=2}^{T}(y_t - y_{t-1}) = y_T + h\left(\frac{y_T - y_1}{T-1}\right) \qquad (2,2)$$

Seasonal Naive method: Forecast for time $T + h$ is written as (2-3).

$$U \; U \quad \hat{y}_T + h / T = y_{T+h-km}$$

Where

m is the seasonal period,

k is the least that satisfies $k \geq \frac{h-1}{m}$ $\theta = (\mu, \sigma^2)$

Linear Trend Regression: It is presented by (2-4).

$T_T = b_0 + b_1 t,$ (2-4)

Wł ere

T_t is the linear trend forecast in period t,

b_0 is the intercept of the linear trend line,

b is the slope of linear trend line,

t is the time period.

Nonlinear quadratic trend equation: The form of nonlinear quadratic trend equation is (2-5).

$$T_t = b_0 + b_1 t + b_2 t^2 \qquad (2\text{-}5)$$

Exponential trend equation: The form of exponential trend equation is (2-6).

$$T_t = b_0\left(b_1\right)^t \qquad (2\text{-}6)$$

We will measure forecast accuracy by the following formula, according to [1] we have (2-7),

Forecast Error = Actual Value – Forecast (2-7)

Obviously, if we use all values of data with noise to forecast, the above methods will evidently result in errors.

Our tested data consists of ozone density $\mu g / m^3$, temperature (ºC) relative humidity (%) and number of deaths in London from 2002 to 2006. In most cases, frequency methods in forecast would be reliable if the data are normal distributed or approximately normal distributed. However, in this dataset, only the number of deaths data follows normal distribution. The ozone and temperature are positive skewed, whereas the humidity data is negatively skewed (Figure 2).

We apply Moving Average methods to the time series with the original data or adjusted data which focus on upcoming trends before forecasting. Number of historical data used for forecasting is another issue which should be considered for comparison. In our test, we use $T = 30$ and $T = 365$, or the first month and year, respectively. Lower and upper bounds for prediction are 95% or 84% confidence. Based on particular data ranges of all attributes, we suggest using constant bound on each prediction for a better perspective between each method and guaranteeing reasonable intervals, which are $\pm 10, \pm 20$ for ozone, humidity, number of deaths attributes and $\pm 2, \pm 3$ for temperature. Table 1 shows some results of the possibility of a real data falling in the forecasting intervals by using Moving Average method. Table 2 on the next section also shows some results from similar method to Moving Average, but using median instead of mean.

FORECASTING METHODS BASED ON ROBUST STATISTICS

In case of noise in data, robust statistics can be used to remove noise through parameters in the location model.

The outcome x_i of each observation depends on μ of the unknown parameter and on some random error process in formula (3-1).

$$x_i = \mu + u_i, i = \overline{1, n}, \qquad (3\text{-}1)$$

Where the errors $u_1, u_2, \ldots u_n$ are independent and identically distributed randμm variables [6]

Therefore, estimates of have the form as the median. There is also another model for outliers which is called *f at - taled* or *f at - tailed* distributions [6].

An extended case when each x_i is represented by a model with two unknown parameters by (3-2).

$$x_i = \mu + \sigma u_i \qquad (3\text{-}2)$$

Where u_i has density f_0, and Maronna et al. [6], also has shown that using bootstrap method can have better results. However, simulating more data will lead to difficulty of large dataset, which results in high complexity of data organizing algorithm.

It is worthy of note that the data can be in continuous or discrete form. Robust Statistics will have better results than Frequentist Statistics if the data is continuous. On the contrary, in discrete form, our work of examining a song whether it is a plagiarism has shown that using median will make features to be divided into two sections which might result in misleading. The same situation holds when researching DNA's information.

BAYESIAN STATISTICS AND INFERENTIAL MODELS

Bayesian statistics

In Bayesian Statistics, assumes that the sample distribution is $f(x/\theta)$, in which $\pi(\theta)$, prior distribution on θ is available, then the posterior distribution is obtained by Bayesian formula (4-1) [7].

$$\pi(\theta / x) = \frac{f(x/\theta)\pi(\theta)}{\int f(x/\theta)\pi(\theta)d\theta} \qquad (4\text{-}1)$$

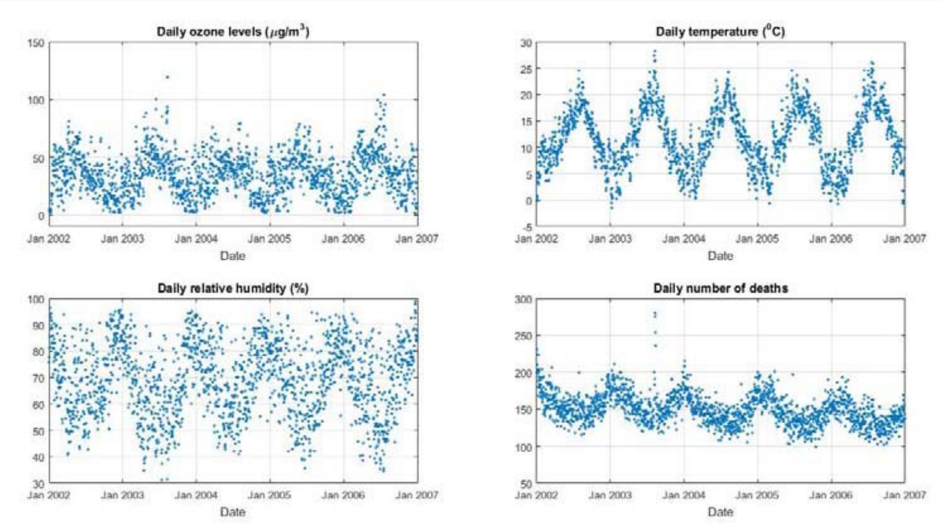

Figure 1 Electrospun nanofibers membrane of poly-ε-caprolactone visualization after 21 days of human Osteoblasts culture (Cells visualization in blue (nucleus /DAPI) and PLL^{FITC} labelled nanofibers in green): colonization and proliferation of osteoblasts into the nanofibers membrane.

The posterior distribution is the updating of the information available on θ, then, based on the information contained in $l(\theta / x)$ in which $\pi(\theta)$ represents the information in the priori before observing x.

The predictive distribution of y, when $y \sim g(y/\theta, x)$, is obtained by (4-2).

$$g(y/x) = \int g(y/\theta, x)\pi(\theta / x)d\theta \qquad (4\text{-}2)$$

The main key to Bayesian analysis is choosing the prior distribution [7], since knowing the prior distribution can lead in inference by minimizing posterior losses, computing higher posterior density region or integrating out parameters to find the predictive distribution. This is almost the most difficult since the prior information in practice is hardly precise enough which adversely affects the exact determination of the prior distribution.

Table 3 and Figure 3 show the results and plots of Bayesian forecasting method

When data contain noise, it is then necessary to choose reasonable prior distribution to have posterior distribution which best suits the data in order to get the most accurate result. Although Bolstad et al. [8], has used robust methods to reduce noise, as can be seen, we will have some issues of using median in computing as it might divide features into two sections which leads to incorrect results.

Inferential models: Ryan Martin and Chuanhai Liu introduced a new framework of statistical inference [11]. It is somewhat between classical and Bayesian approaches because it not only bases on observed data but also introduces a "semi-data driven" which might be considered as a "prior" idea.

An inferential model will give a "relation" between the variable of interest and its distribution.

Let X be the real-valed random variable of interest defined on some pro F_x ability space (Ω, A, P), and let $X = (X_{1,} X_2, ..., X_n)$ be the observed data drawn from X and wish, among other things, to discover the law governing its random evolution, and to predict its future values.

The distribution function $F(\cdot): \mathbb{R} \to [0,1]$ of X is defined as (4-3).

$$F(x) = P(X \le x) \qquad (4\text{-}3)$$

The relation actually link X with its distribution F (or F_x when needed)

A more "formal" relation between X and F is obtained via its quantile function $F^{-1}(\cdot): (0,1) \to \mathbb{R}$, which is defined as (4-4).

$$F^{-1}(\alpha) = \inf \{x \in \mathbb{R} : F(x) \ge \alpha\} \qquad (4\text{-}4)$$

It was named by (4-5).

$$X =^D F^{-1}(U) \qquad (4\text{-}5)$$

Where U is uniformly distributed on the interval (0,1).

Therefore, explicit "equations" relating X, its distribution F, and some "auxiliary" unobservable random variable U, could lead to a new framework for statistical inference.

For example, in parametric sampling models where the distribution function of the observable X is $F_\theta(\cdot), \theta \in \Theta$, the above "equation" takes the form $X = F_\theta^{-1}(U) = a(\theta, U)$, where the "association" function a $(\cdot,\cdot)$ is known, as well as the distribution of the unobservable U.

In normal model $\theta = (\mu, \sigma^2)$ for X, we write to (4-6).

$$X = \mu + \sigma Z \qquad (4\text{-}6)$$

Then $X = a(\theta, Z)$, where $\theta = (\mu, \sigma)$ and Z distributed as standard normal.

As another example, if X follows Bernoulli distribution then $X = a(\theta, U) = 1_{[0,\theta]}(U)$, with U being uniformly distributed on interval $[0,1]$. The above association is due to

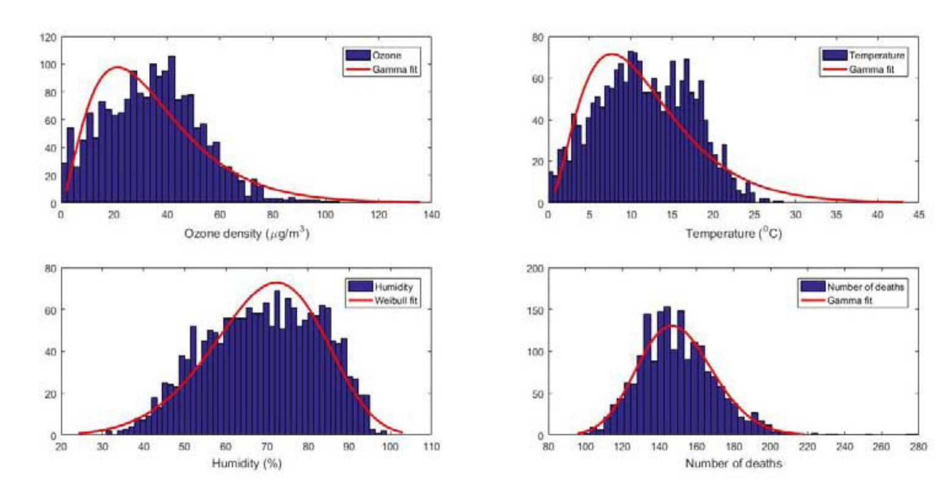

Figure 2 Histogram of 4 attributes in the London data.

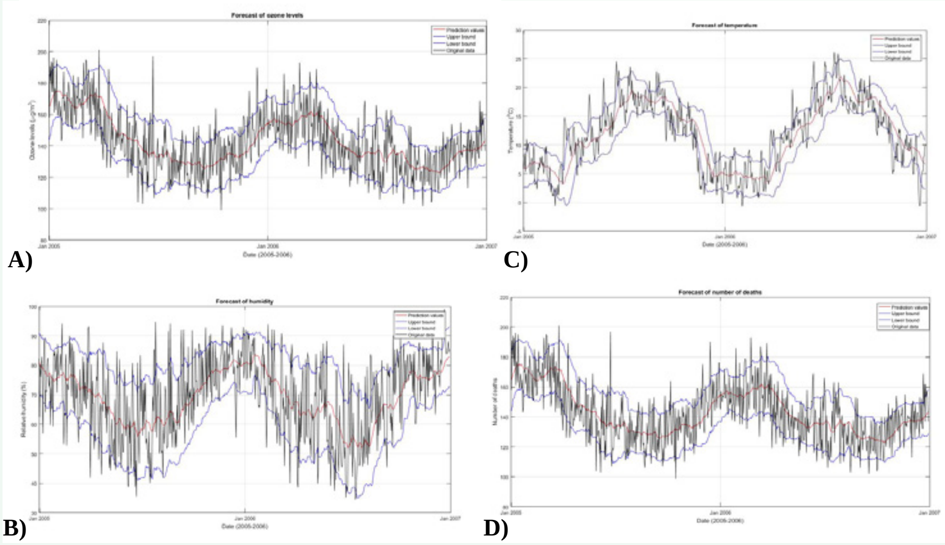

Figure 3 Forecasting results based on Bayesian Methods processed with raw data and confidence interval of 84%.

the fact that $1_{[0,\theta]}(U)$ is equal to X in distribution.

As for now, the identification problem from the "association" served in statistical inference is discovering *F* with observed data from X and the uniform random variable U. This can be seen as the reverse problem of simulations. Although X is observable, information on U is limited to only a known distribution. Thus, in order to take advantages of the inferential model, a "guess" of unobserved values of U is necessary. As a result, a predictive random set U is proposed. Under a "near Bayesian" point of view, U can be treated as a counterpart of the Bayesian prior but it is not subjective "prior" which is a drawback in Bayesian Statistics.

In short, the inferential model framework is based on a compromise between frequentist and Bayesian approaches to statistical analysis in which starting by an objective prior P_U and the subsequent analysis (using observed data) is like a posterior analysis. In other words, this frame work is termed "posterior analysis without prior".

DISCUSSION AND CONCLUSION

Noises appear when collecting data is an inevitable matter. To deal with it, we need efficient statistical methods to achieve the best result as possible. Robust method, with pros like eliminating outliers, could handle the constant bounds well. Bayesian method is generally better than Frequentist or Robust method, especially with its confidence interval of forecasting values. In addition, its capability of updating data over time could result in reflecting real trends better.

In order to have reasonable forecasting methods, properties of data in real test, such as qualitative, quantitative or bimodal distributed properties, and their distribution should be studied with caution. This paper has suggested some methods to process

Table 1: Results of time series forecasting by using Moving Average method.

Prediction Interval Attributes	$\mu\pm1.96\times\sigma$	$\mu\pm\sigma$	Large constant bound (±20 and ±3)	Small constant bound (±10 and ±2)
Ozone density	96.58%	89.04%	77.88%	48.36%
Temperature	97.95%	93.84%	63.36%	48.56%
Relative humidity	94.86%	72.88%	88.44%	59.73%
Number of deaths	94.86%	69.18%	74.32%	50.84%

Table 2: Results of forecasting using robust statistics with median M.

Prediction Interval Attributes	M±1.96×sd(M)	*M±sd(M)*	Large constant (±20 and ± 3)	Small constant (±10 and ±2)
Ozone density	90.62%	71.51%	84.18%	55.75%
Temperature	91.3%	80.14%	87.74%	70.62%
Relative humidity	87.74%	60.68%	89.73%	62.26%
Number of deaths	87.47%	58.36%	74.38%	52.84%

Table 3: Results of Bayesian Statistics.

Prediction Interval Attributes	$\mu\pm1.96\times\sigma$	$\mu\pm\sigma$	Large constant bound (±20 and ±3)	Small constant bound (±10 and ±2)
Ozone density	97.05%	91.78%	78.01%	48.29%
Temperature	98.22%	94.45%	82.74%	63.15%
Relative humidity	97.05%	80.75%	88.84%	59.73%
Number of deaths	99.04%	84.32%	74.32%	43.56%

data with noise, such as making use of all data in Frequentist Statistics, eliminating outliers in Robust Statistics, assuming parameters as random variables in Bayesian Statistics and Robust Bayesian Analysis or linking a random variable with its distribution through a description of "association". However, we still need improvements on our statistical techniques to handle noise better and get more accurate results in the future.

REFERENCES

1. Anderson DR, Sweeney DJ, Williams TA, Camm JD, Cochran JJ. Statistics for Business and Economics, Revised. Cengage. 2014.

2. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016; 70: 129-133.

3. Baker M. Statisticians issue warning over misuse of P values. Nature. 2016; 531: 151.

4. Altman N, Krzywinski M. Points of significance: P values and the search for significance. Nat Methods. 2017; 14: 3-4.

5. Maronna RA, Martin DR, Yohai VJ. Robust Statistics: Theory and Methods. John Wiley & Sons. 2006.

6. Robert CP. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Springer. 2007.

7. Bolstad WM, Curran JM. Introduction to Bayesian Statistics. John Wiley & Sons. 2017.

8. Shmueli G. To Explain or to Predict? Stat Sci. 2010; 25: 289-310.

9. Huber PJ. The 1972 Wald Lecture Robust Statistics: A Review. Ann Appl Stat. 1972; 43: 1041-1067.

10. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. Int J Epidemiol. 2013; 42: 1187-1195.

11. Martin R, Liu C. Inferential Models: Reasoning with Uncertainty. CRC Press. 2015.

Cite this article

Le H, Pham U, Nguyen NT, Bao PT (2017) Statistical Methods of Handling Noise in Data Processing. Ann Biom Biostat 4(1): 1027.