國立臺灣大學理學院應用數學科學研究所
碩士論文
Institute of Applied Mathematical Sciences
College of Science
National Taiwan University
Master Thesis

建立在伽瑪散度下穩健模型的調整參數之選取
Robust Model Fitting - Selection of Tuning Parameters in the Aspect of Gamma Clustering

蕭奕
Yi Hsiao

指導教授：杜憶萍博士
Advisor: I-Ping Tu, Ph.D.

中華民國 107 年 6 月
June, 2018

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

誌謝

感謝指導教授杜憶萍老師，讓我在大學的時候接觸統計這個領域，並因此對統計產生興趣，走上研究的道路。雖然路途上遭遇許多困難，但每次和老師討論時，能學習到用各種方式去看待不同問題，並找出問題的關鍵點，再加上不時的鼓勵，讓我一步一步幸運且順利地完成論文。

感謝大學及研究所的師長們，讓我在求學生涯中打下扎實的數學基礎。

感謝陪我唸書、一起討論課業準備考試的同學們，以及協助我完成論文和準備口試的學長姐，有你們陪伴學習更加有趣。

最後感謝我的父母，對我的決定總是給予支持，讓我更專注於課業。儘管不常回家，但總會讓我知道家永遠是最好的避風港。

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

摘要

Windham 在 1995 年的論文 Robustifying Model Fitting 提出權重分布方法，解決在存有異常數據情況下可有效的估計均值，這個權重分布使用了一個參數，這個參數會影響到均值的估計表現。Windham 同時提出此參數的選取方法，但我們發現在一些數值模擬例子中，這個參數選取方法表現並不佳。我們提出另一個選取方法，在數值模擬上有較佳的表現。除了一般的均值估計，我們也成功地把此方法應用在群聚分析。

關鍵字： 穩健估計、影響函數、伽瑪散度、群聚分析

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

Abstract

In 1995, Windham came out with an idea of weighted distribution in his thesis, Robustifying Model Fitting, and he used the idea to find a mean estimator when there are outliers in the original data. There is a tuning parameter in this estimator, and selecting the parameter will affect the mean estimate in the same data. In the same thesis, he also suggested a criterion of selecting the tuning parameter, but we found out that this criterion wasn't doing well in some simulations. Considering the problem, we propose another criterion which can derive a better mean estimator. Besides, we can also apply this method to clustering problem.

Keywords: robust estimate, influence function, gamma-divergence, clustering

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

Contents

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

List of Figures

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

List of Tables

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

Chapter 1

Introduction

1.1 What is Robust?

Maximum likelihood estimate (MLE) or method of moments are most common way to estimate parameter. Under some regularity conditions, maximum likelihood estimator is consistent. That is, the probability of the estimator converging to the true parameter is 1, as the number of samples goes to infinity. However, finite sample or outliers may hinder the consistent property of the estimate. We can think this through the fundamental spirit of maximizing the likelihood. The likelihood function is defined as the product of the density of the observed data (assume i.i.d.) including outliers. Considering this, it is unreasonable to maximize the likelihood for those outlier terms. And if there are more outliers or the outliers are further far away from the true data, the influence on the estimator become larger. For example, considering a normal distribution model, the maximum likelihood estimator for mean is the sample mean. This estimator can become arbitrary large if we add distant data points in the sample (Figure 1.1). Thus, we say that sample mean is not robust. Another common mean estimator, median, needs over half of the data points to have effect on the estimator, hence is robust. We may want to know how many portions of data can influence the estimator, or how much a data point influence the estimator, thus following are some measurements often used for quantify robustness.

doi:10.6342/NTU201900783

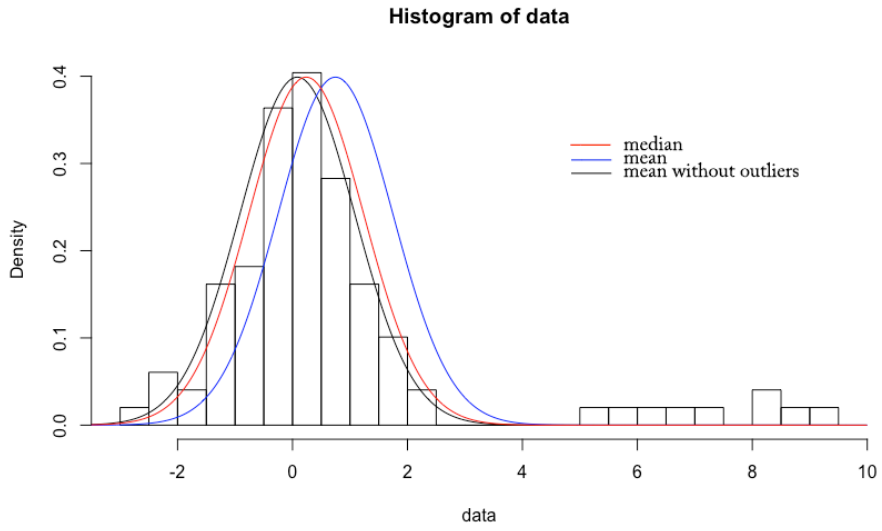


Figure 1.1: median, mean, robustified mean

Measurement of Robustness

- Influence function (Hampel, 1968) is defined as

$$IF(x;T,F) = \lim_{\varepsilon \to 0} \frac{T\{(1-\varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon},$$

where F is the distribution function, $T(F)$ is the parameter estimator, and $\delta_x(u)$ is a heaviside step function, which is 0 when $u < x$, and is 1 otherwise.

This definition means that we perturbate the distribution at a point x, and have interest in the variation rate of the estimator.

- Gross-error sensitivity is defined as

$$\gamma^*(T,F) = \sup_x |IF(x;T,F)|,$$

$\gamma^*(T,F)$ is the supreme of the absolute value of influence function, which measures the worst influence a data point can make on the estomator.

doi:10.6342/NTU201900783

- Breakdown point is defined as

$$\varepsilon^* = \inf\{\varepsilon : \sup_x |T(F) - T(F_\varepsilon)| = \infty\},$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_x$.

Let's give an example. Suppose F is the distribution, then sample mean can be expressed as $T(F) = \int udF$. Hence the influence function of $T(F)$ is

$$IF(x;T,F) = \lim_{\varepsilon\to 0} \frac{T\{(1-\varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon}$$

$$= \lim_{\varepsilon\to 0} \frac{\int ud((1-\varepsilon)F + \varepsilon\delta_x) - \int udF}{\varepsilon}$$

$$= \lim_{\varepsilon\to 0} \frac{\varepsilon\int ud\delta_x - \varepsilon\int udF}{\varepsilon} = x - T(F)$$

The gross-error sensitivity is

$$\gamma^*(T,F) = \sup_x \|IF(x;T,F)\| = \sup_x \|x - T(F)\| = \infty,$$

which is the worst case in the view of robustness.

And the breakdown point of sample mean is

$$\varepsilon^* = \inf\{\varepsilon : \sup_x |T(F) - T(F_\varepsilon)| = \infty\} = \inf\{\varepsilon : \sup_x |\varepsilon(x - T(F))| = \infty\} = 0,$$

which is also the worst case.

Now, we figure out that sample mean is not robust, so next section is some ideas to get robust.

1.2 How to get Robust?

One way to get robust is to improve maximum likelihood estimator, since it is maximum likelihood type, we called it M-estimator.

doi:10.6342/NTU201900783

1.2.1 M-estimator

- MLE: min $\sum -\log f(x_i;\theta)$, where f is the density function.

  Generalized MLE: min $\sum p(x_i;\theta)$, where p can be any function.

- If $p(x;\theta)$ is differentiable with respect to θ, then the M-estimator is to solve

$$E_F[\psi(x;\theta)] = 0,$$

  where $\psi(x;\theta) = \frac{\partial p(x;\theta)}{\partial \theta}$ is the score function.

- The influence function of the M-estimator

  Let $T(F)$ be the M-estimator with $E_F[\psi(x;T(F))] = 0$, for all distributions F.

  Hence,

$$E_{F_\varepsilon}[\psi(x;T(F_\varepsilon))] = 0$$

  To derive influence function, take derivative with respect to ε,

  $\frac{\partial}{\partial \varepsilon} E_{F_\varepsilon}[\psi(x;T(F_\varepsilon))] = 0$

  $\Rightarrow \frac{\partial}{\partial \varepsilon} \int \psi(u;T(F_\varepsilon)) d((1-\varepsilon)F(u) + \varepsilon\delta_x) = 0$

  If the integration and differentiation can be interchanged, then

$\Rightarrow \int \frac{\partial}{\partial \varepsilon}\{\psi(u;T(F_\varepsilon)) d((1-\varepsilon)F(u) + \varepsilon\delta_x)\} = 0$

$\Rightarrow \int \big(\frac{\partial}{\partial \varepsilon}\{\psi(u;T(F_\varepsilon))\} d((1-\varepsilon)F(u) + \varepsilon\delta_x) + \psi(u;T(F_\varepsilon)) d(\frac{\partial}{\partial \varepsilon}\{(1-\varepsilon)F(u) + \varepsilon\delta_x\})\big) = 0$

$\Rightarrow \int \big(\frac{\partial}{\partial \theta}[\psi(u;\theta)]_{\theta=T(F_\varepsilon)} dF_\varepsilon(u) \cdot \frac{\partial}{\partial \varepsilon}[T(F_\varepsilon)] + \psi(u;T(F_\varepsilon)) d(\delta_x - F(u))\big) = 0$

$\Rightarrow \int \big(\frac{\partial}{\partial \theta}[\psi(u;\theta)]_{\theta=T(F)} dF(u) \cdot \frac{\partial}{\partial \varepsilon}[T(F_\varepsilon)]_{\varepsilon=0} + \psi(u;T(F)) d(\delta_x - F(u))\big) = 0$

Note that $\frac{\partial}{\partial \varepsilon}[T(F_\varepsilon)]_{\varepsilon=0} = IF(x;T,F)$ and

$\int \psi(u;T(F)) d(\delta_x - F(u)) = \psi(x;T(F))$, we have

$IF(x;T,F) = \frac{\psi(x;T(F))}{-\int \frac{\partial}{\partial \theta}[\psi(u,\theta)]_{\theta=T(F)} dF(u)}$

Figure 1.2 shows various score functions of location estimators θ, and $\theta = 0$. Huber estimator (1964) is an M-estimator with score function $\psi_k(x;\theta) = \max(-k, \min(k, x-\theta))$, the black line is for $k = 1$. And Windham estimator (1995) is also an M-estimator with $\psi_c(x;\theta) = \sqrt{c+1}\exp(-\frac{c}{2}(x-\theta)^2)(x-\theta)$, the red curve is for $c = 0.5$ and blue

doi:10.6342/NTU201900783

curve is for $c = 1$.

For different function $p(x;\theta)$ or $\psi(x;\theta)$, the estimator derived has different robustness. This property can be implicated by Figure 1.2, traditional score function of MLE $\psi(x;\theta) = x - \theta$, the dashed line, counts every point x and going to arbitrary large when x goes to infinity, while other robustified estimators don't.

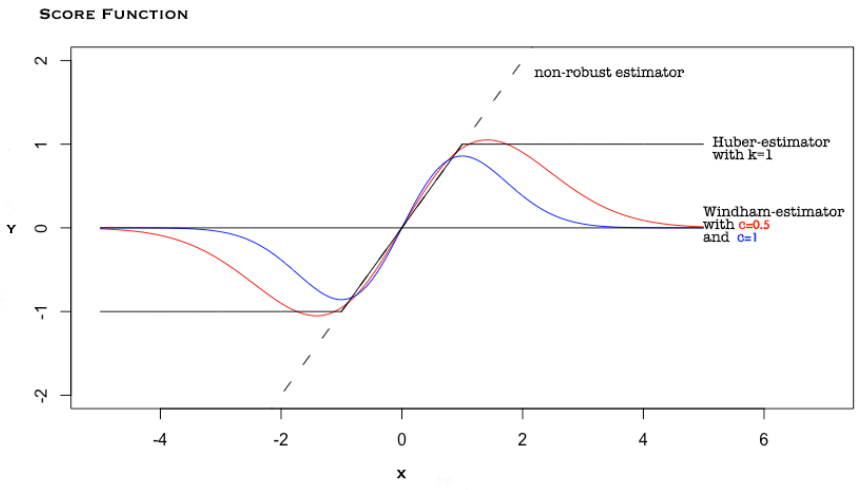


Figure 1.2: Score function of estimator, assuming the location parameter is 0.

1.2.2 Weighted data

Another way to get robust is to weight the data such that the point more likely to happen has higher weight, and less likely to happen has lower weight. So, thinking a weight with such property, one can immediately came up with density function, and Windham gave an extra exponent c to the power of the density function. For example, data $\{x_1, x_2, ..., x_n\}$ are from normal distribution F with mean 0 and unknown variance θ, then its weight function is

$$w(x_i;\theta) = K\phi^c(x_i;\theta),$$

where ϕ is the normal density function with mean 0 in the model, c is a fixed exponent term, K is a constant so that the sum of weight is 1.

Note that weighting density is to change the frequency of data points, not the value of data.

doi:10.6342/NTU201900783

Hence, the weighted sample variance is

$$v = \sum_{i}^{n} w(x_i;\theta)x_i^2$$

And the weighted population variance is

$$v^* = \int x^2 w^*(x;\theta)\phi(x;\theta)dx,$$

where

$$w^*(x;\theta) = K^*\phi^c(x;\theta), \tag{1.1}$$

$K^* = [\int \phi^c(x;\theta)dF]^{-1} = E_F[\phi^c(x;\theta)]^{-1}$ is a constant. In this model $dF = \phi(x;\theta)dx$, hence $\int w^*(x;\theta)dF = \int w^*(x;\theta)\phi(x;\theta)dx = 1$. Notice that $w^*(x;\theta)\phi(x;\theta)$, called the weighted density function, and is same as the normal density with mean 0 and variance $v^* = \theta/(c+1)$. If the model is correct, v^* and v should be close. Thus, using these relations, we can derive the robust estimate of parameter θ.

Also, there is a exponent term c, which can adjust weight and hence called tuning parameter. Windham (1995)[3] studied in how to select this parameter and called this method Robust Model Fitting. In this thesis, we propose another method of selection and do research with it.

1.3 Why need to do this question?

Past robust model, such as Huber estimate (1964)[2], Windham (1995)[3] estimate, and Akifumi Notsu, Shinto Eguchi (2016)[1] Gamma-clust robust estimate, there are tuning parameters. Tuning parameters will affect the robustness of the model (Figure 1.2). In Windham's model, parameter c determines the weight of each point. Selection of c is important when doing clustering analysis, since larger the c becomes, more robust the model will be, and number of clusts will closer to the number of data points; on the contrary, smaller the c becomes, less robust the model will be, and number of clusts will go to 1. In general, there is a reasonable interval to choose the parameter, and usually is selected

doi:10.6342/NTU201900783

subjectively. By using influence function and Fisher information, we give a objective selection criterion, and more details can be seen in chapter 3.

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

Chapter 2

Literature Review

2.1 Normal Robust Model

We use a simple example to elaborate Windham's iterative algorithm estimates. Assume there are 400 samples $\{x_1, x_2, ..., x_{400}\}$, and among of them, there are 340 samples are from $N(\mu, \sigma^2)$, $\mu = 0, \sigma^2 = 1$ and the rest are outliers around $\mu^* = 7$. We want to estimate μ and variance σ^2.

- First step : Find initials.

  Use sample mean and sample variance (MLE in normal distribution) as initials.

  $\hat{\mu} = \bar{x} = 0.99,\ \hat{\sigma}^2 = \frac{1}{400}\sum(x_i - \bar{x})^2 = 7.19$

- Second step : Weight data.

  $w_i = w(x_i, \hat{\mu}, \hat{\sigma}^2) = K\phi^c(x_i; \hat{\mu}, \hat{\sigma}^2)$, c is chosen to be 0.37, ϕ is normal density function, and K is a constant such that $\sum_{i=1}^{400} w_i$ is 1.

  Similarly, we use MLE on the weighted data, and get,

  $\hat{m} = \sum w_i x_i = 0.45$, $\hat{s}^2 = \sum w_i(x_i - \hat{m})^2 = 3.93$

- Third step : Find the non-weighted parameters such that we can get $\hat{m}$ and $\hat{s}^2$ in second step after weighting.

  That is, if the model is correct, which is satisfied normal distribution with density $\phi(x; \mu, \sigma^2)$ assumption, then the weighted density is $K^*\phi^{c+1}(x; \mu, \sigma^2)$, which

doi:10.6342/NTU201900783

Iteration	$\hat{\mu}$	$\hat{\sigma}^2$
1	0.993392	7.191821
2	0.4594806	5.380773
3	0.2499072	3.751013
4	0.09685292	2.415905
5	-0.01083447	1.445184
6	-0.05552865	1.044782
7	-0.06715192	0.9457484
8	-0.07040669	0.9195676
9	-0.07137085	0.912215

Table 2.1: Estimates for $\hat{\mu}$ and $\hat{\sigma}^2$.

is $\phi(x; \mu, \sigma^2/(c+1))$. Thus, we can revise the new parameters to $\hat{\mu}_+ = \hat{m} = 0.45$, $\hat{\sigma}^2_+ = 1.37\hat{s}^2 = 5.38$.

- Fourth step : Repeat step 1 to 3, until the estimates converge. In other words, replace $\hat{\mu}$ and $\hat{\sigma}^2$ by new estimates $\hat{\mu}_+$ and $\hat{\sigma}^2_+$ until they converge (Table 2.1).

2.2 General Description

We now use mathematical way to describe the example above. In the beginning, we observe that there are n one dimension data, $\{x_1, x_2, ..., x_n\}$ from the distribution F, and its empirical distribution is $\hat{F}$. $T(\hat{F})$ is a parameter estimator, and $T(\hat{F}) = (\bar{x}, \frac{1}{n}\sum(x_i - \bar{x})^2)$ is the maximum likelihood case. And note that $T(F)$ satisfies $E_F[\psi(x; T(F))] = 0$, where ψ is the score function of $T(F)$. Moreover, Windham defined weighted distribution as $dF_{c,t}(x) = w^*(x;t)dF(x)$, where $w^*(x;t) = K^* g^c(x;t)$, g is the density in the model, $K^* = \{E_F[g^c(x;t)]\}^{-1}$, and here we assume g is normal. If we plug $\hat{F}$ into F, then $d\hat{F}_{c,t}(x) = \hat{w}^*(x;t)d\hat{F}(x)$, where $\hat{w}^*(x;t) = \hat{K}^* g^c(x;t)$, and $\hat{K}^* = \{E_{\hat{F}}[g^c(x;t)]\}^{-1}$.

doi:10.6342/NTU201900783

Hence, the second step is,

$$\begin{aligned} T(\hat{F}_{c,t}) = (\hat{m}, \hat{s}^2) &= (\int x d\hat{F}_{c,t}, \int (x-\hat{m})^2 d\hat{F}_{c,t}) \\ &= (\int \hat{w}^*(x;t) x d\hat{F}, \int \hat{w}^*(x-\hat{m})^2 d\hat{F}) \\ &= (\frac{1}{n}\sum \frac{g^c(x_i;t)}{\frac{1}{n}\sum g^c(x_j;t)} x_i, \frac{1}{n}\sum \frac{g^c(x_i;t)}{\frac{1}{n}\sum g^c(x_j;t)} (x_i-\hat{m})^2) \\ &= (\sum w_i x_i, \sum w_i (x_i-\hat{m})^2), \end{aligned}$$

where $w_i = w(x_i;t) = \frac{g^c(x_i;t)}{\sum g^c(x_j;t)}$.

Next, we want to know the difference of the estimates between the weighted model and non-weighted one. Hence, assume that the normal density of F is $g(x;\mu,\sigma^2)$, We have,

$$\int w^*(x;\mu,\sigma^2) x dF = \frac{\int g^{c+1}(x;\mu,\sigma^2) x dx}{\int g^{c+1}(x;\mu,\sigma^2) dx} = \mu$$

$$\begin{aligned} \int w^*(x;\mu,\sigma^2)(x-\hat{m})^2 dF &= \int w^*(x;\mu,\sigma^2)(x-\mu)^2 dF \\ &= \frac{\int g^{c+1}(x;\mu,\sigma^2)(x-\mu)^2 dx}{\int g^{c+1}(x;\mu,\sigma^2) dx} \\ &= \frac{\sigma^2}{(c+1)} \end{aligned}$$

From the formula above, we know that mean is unchanged and the variance is multiplied by $\frac{1}{(c+1)}$ in the weighted model. Therefore, a function τ is defined to transform the estimator between weighted and non-weighted distribution,

$$(\hat{m}, \hat{s}^2) = \tau(\hat{\mu}_+, \hat{\sigma}^2_+) = (\hat{\mu}_+, \hat{\sigma}^2_+/(1+c))$$

$$(\hat{\mu}_+, \hat{\sigma}^2_+) = \tau^{-1}(\hat{m}, \hat{s}^2) = (\hat{m}, (1+c)\hat{s}^2)$$

For $N \geq 0$, $t^{(0)} = T(\hat{F})$,

$$t^{(N+1)} = \tau^{-1}\{T(\hat{F}_{c,t^{(N)}})\}$$

doi:10.6342/NTU201900783

is an iterative procedure. And the Windham robustified estimate (WRE) $T_c(\hat{F})$ is hence defined as $T_c(\hat{F}) = \hat{\theta}$, where $\hat{\theta}$ is the estimate from the last iteration.

2.3 Selection Criterion

T_c is a solution of $E_F[w^*(x;\theta)\psi(x;\tau(\theta))] = 0$, hence is an M-estimator, the corresponding score function is

$$\psi_c(x;\theta) = w^*(x;\theta)\psi(x,\tau(\theta)) \tag{2.1}$$

Let $h(t) = \tau^{-1}\{T(F_{c,t})\}$, if the process converges, we will have $h(\theta) = \theta$. Hence,

$$E_{F_{c,t}}[\psi[x;\tau\{h(t)\}]] = 0$$

Take derivative with respect to t, and derive the convergence rate $h'(t)$.

$$\begin{aligned}
&\frac{\partial}{\partial t}E_{F_{c,t}}[\psi[x;\tau\{h(t)\}]] = 0 \\
&\Rightarrow \frac{\partial}{\partial t}E_F[w^*(x;t)\psi[x;\tau\{h(t)\}]] = 0 \\
&\Rightarrow E_F[\left(\frac{\partial}{\partial t}w^*(x;t)\right)\psi[x;\tau\{h(t)\}] + w^*(x;t)\left(\frac{\partial}{\partial t}\psi[x;\tau\{h(t)\}]\right)] = 0 \\
&\Rightarrow E_F[cw^*(x;t)(\frac{\partial}{\partial t}\log g(x;t))\psi(x;\tau(h(t)))] + E_F[w^*(x;t)\frac{\partial\psi(x;\tau(h(t)))}{\partial\tau(h(t))}\frac{\partial\tau(h(t))}{\partial h(t)}\frac{\partial h(t)}{\partial t}] = 0
\end{aligned}$$

where $\frac{\partial}{\partial t}w^*(x;t) = cw^*(x;t)(\frac{\partial}{\partial t}\log g(x;t))$

When it is about to converge, i.e. $h(t) = t$,

$$\begin{aligned}
h'(t) &= -c\{E_{F_{c,t}}[\frac{\partial\psi(x;\tau(t))}{\partial\tau(t)}]\frac{\partial\tau(t)}{\partial t}\}^{-1}E_{F_{c,t}}[\psi(x;\tau(t))\frac{\partial}{\partial t}(\log g(x;t))] \\
&= -\frac{cE_F[\frac{\partial}{\partial t}(\log g(x;t))\psi_c(x;t)]}{E_F[w^*(x;t)\frac{\partial\psi(x;\tau(t))}{\partial\tau(t)}\frac{\partial\tau(t)}{\partial t}]} \\
&= \frac{-cE_F[\frac{\partial}{\partial t}(\log g(x;t))\psi_c(x;t)]}{E_F[\psi_c'(x;t)] - E_F[cw^*(x;t)\frac{\partial}{\partial t}(\log g(x;t))\psi(x;\tau(t))]} \\
&= \frac{-cE_F[\frac{\partial}{\partial t}(\log g(x;t))\psi_c(x;t)]E_F[\psi_c'(x;t)]^{-1}}{I - cE_F[\frac{\partial}{\partial t}(\log g(x;t))\psi_c(x;t)]E_F[\psi_c'(x;t)]^{-1}} \\
&= cB(t)\{I + cB(t)\}^{-1},
\end{aligned}$$

doi:10.6342/NTU201900783

Since

$$\psi_c'(x;t) = \frac{\partial}{\partial t}(\psi_c(x;t)) = cw^*(x;t)(\frac{\partial}{\partial t}\log g)\psi(x;\tau(t)) + w^*(x;t)\frac{\partial\psi(x;\tau(t))}{\partial\tau(t)}\tau', \quad (2.2)$$

where $B(t) = -E_F[\frac{\partial}{\partial t}(\log g(x;t))\psi_c(x;t)]E_F[\psi_c'(x;t)]^{-1}$, which is related to influence function. In fact,

$$B\{T_c(F)\} = E_F[-E_F[\psi_c'(x;t)|_{t=T_c(F)}]^{-1}\psi_c(x;T_c(F)) \cdot \frac{\partial}{\partial t}(\log g(x;t))|_{t=T_c(F)}]$$

$$= E_F[IF(x;T_c,F)s(x;T_c(F))],$$

where $IF(x;T_c,F) = -E_F[\psi_c'(x;t)|_{t=T_c(F)}]^{-1}\psi_c(x;T_c(F))$ is the influence function and $s(x;T_c(F)) = \frac{\partial}{\partial t}(\log g(x;t))|_{t=T_c(F)}$ is the score function.

According to Cauchy–Schwarz inequality,

$$\{E_F[s^2(x;T_c(F))]E_F[IF^2(x;T_c,F)]\}^{-1} \leq B^{-2}\{T_c(F)\}$$

And Windham noted that the left hand side is just the asymptotic relative efficiency, which is the reciprocal of the Fisher information times the asymptotic variance. Hence, he used B^{-2} as a criterion for choosing c, which is

$$\rho(c) = B^{-2}\{T_c(F)\}$$

By $h'(t) = cB(t)\{I + cB(t)\}^{-1}$,

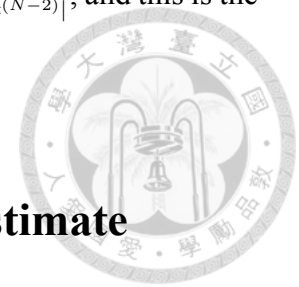
$$\rho(c) = (c/h'(t) - c)^2$$

Thus, the tuning parameter c is chosen by

$$\hat{c} = \arg\max_c \rho(c)$$

doi:10.6342/NTU201900783

In simulation, the convergence rate $h'(t)$ can be estimated by $\frac{|t^{(N)}-t^{(N-1)}|}{|t^{(N-1)}-t^{(N-2)}|}$, and this is the method Windham used.



2.4 γ-estimate and Weighted Robustified estimate

2.4.1 Some Notes about γ-clust

- γ-divergence

  Like K-L divergence, is a measure of the difference between two probability distributions.

- γ-cross entropy

  Suppose $\{x_1, x_2, ..., x_n\}$ are from the distribution with density f. And there is a density g_θ we assumed that it is the model density with θ unknown. The γ-cross entropy $d_\gamma(f, g_\theta)$ is defined as

$$d_\gamma(f, g_\theta) = -\frac{1}{\gamma}\log\{\int g(x;\theta)^\gamma f(x)dx\} + \frac{1}{1+\gamma}\log\int g(x;\theta)^{1+\gamma}dx$$

  $d_\gamma(f, g_\theta)$ can be empirically estimated by

$$d_\gamma(\hat{f}, g_\theta) = -\frac{1}{\gamma}\log\{\frac{1}{n}\sum_{i=1}^{n} g(x_i;\theta)^\gamma\} + \frac{1}{1+\gamma}\log\int g(x;\theta)^{1+\gamma}dx,$$

  where $\hat{f}$ is the empirical pdf.

  The small value of γ-cross entropy means two distributions are close, so the robust estimator $\hat{\theta}_\gamma$ can be defined by minimizing $d_\gamma(\hat{f}, g_\theta)$, i.e.

$$\hat{\theta}_\gamma = \arg\min_\theta d_\gamma(\hat{f}, g_\theta)$$

  We substitute g_θ to $g(x;\mu,\Sigma)$, where g is a p-variate normal density function.

$$d_\gamma(\hat{f}, g(x;\mu,\Sigma)) = -\frac{1}{\gamma}\log\{\frac{1}{n}\sum^{n} g(x_i;\mu,\Sigma)^\gamma\} + \frac{1}{1+\gamma}\log\int g(x;\mu,\Sigma)^{1+\gamma}dx$$

doi:10.6342/NTU201900783

$$\Rightarrow e^{-d_\gamma(\hat{f},g(x;\mu,\Sigma))} = \{\frac{1}{n}\sum^{n} g(x_i;\mu,\Sigma)^\gamma\}^{1/\gamma}\{\int g(x;\mu,\Sigma)^{1+\gamma}dx\}^{-\frac{1}{1+\gamma}}$$

Note that

$$\begin{aligned}
\int g(x;\mu,\Sigma)^{1+\gamma}dx &= \int\{(2\pi)^{-\frac{p}{2}}(det\Sigma)^{-\frac{1}{2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))\}^{1+\gamma}dx \\
&= (2\pi)^{-\frac{p\gamma}{2}}(1+\gamma)^{-\frac{1}{2}}(det\Sigma)^{-\frac{\gamma}{2}} \\
\Rightarrow e^{-d_\gamma(\hat{f},g(x;\mu,\Sigma))\gamma} &= \frac{1}{n}\sum g(x_i;\mu,\Sigma)^\gamma\{\int g(x;\mu,\Sigma)^{1+\gamma}dx\}^{-\frac{\gamma}{1+\gamma}} \\
&= \frac{1}{n}\sum g(x_i;\mu,\Sigma)^\gamma\{(2\pi)^{-\frac{p\gamma}{2}}(1+\gamma)^{-\frac{1}{2}}(det\Sigma)^{-\frac{\gamma}{2}}\}^{-\frac{\gamma}{1+\gamma}} \\
&\propto \sum g(x_i;\mu,\Sigma)^\gamma(det\Sigma)^{\frac{\gamma^2}{2(1+\gamma)}}
\end{aligned}$$

Thus, we derive a function of γ which is likelihood function in γ-divergence sense.

- γ-utility function

$$L_\gamma(\mu,\Sigma) = (det\Sigma)^{\frac{\gamma^2}{2(1+\gamma)}}\sum_{i=1}^{n} g(x_i;\mu,\Sigma)^\gamma$$

Hence, the robust estimator is to maximize $L_\gamma(\mu,\Sigma)$

2.4.2 Weighted Distribution

To maximize γ-utility function, we take derivatives with respect to μ and set it be 0. Here, $g(x;\mu,\sigma^2)$ is univariate normal density.

$$\frac{\partial}{\partial\mu}L_\gamma(\mu,\sigma^2) = \frac{\partial}{\partial\mu}\sum g^\gamma(x_i;\mu,\sigma^2)(\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}} = 0$$

doi:10.6342/NTU201900783

$$
\begin{aligned}
&\Rightarrow \frac{\partial \sum g^{\gamma}(x_i;\mu,\sigma^2)}{\partial \mu} = 0 \\
&\Rightarrow \gamma \sum g^{\gamma-1}(x_i;\mu,\sigma^2)\frac{\partial g(x_i;\mu,\sigma^2)}{\partial \mu} = 0 \\
&\Rightarrow \sum g^{\gamma}(x_i;\mu,\sigma^2)(x_i-\mu) = 0 \\
&\Rightarrow \hat{\mu} = \frac{\sum g^{\gamma}(x_i;\mu,\sigma^2)x_i}{\sum g^{\gamma}(x_i;\mu,\sigma^2)}
\end{aligned}
$$

Similarly, take derivatives w.r.t. σ^2,

$$
\begin{aligned}
&\frac{\partial}{\partial \sigma^2} L_{\gamma}(\mu,\sigma^2) = \frac{\partial}{\partial \sigma^2}\sum g^{\gamma}(x_i;\mu,\sigma^2)(\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}} = 0 \\
&\Rightarrow \frac{\gamma^2}{2(1+\gamma)}(\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}-1}\sum g^{\gamma}(x_i;\mu,\sigma^2) + (\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}}\sum \gamma g^{\gamma-1}(x_i;\mu,\sigma^2)\frac{\partial g(x_i;\mu,\sigma^2)}{\partial \sigma^2} = 0 \\
&\Rightarrow \frac{\gamma}{2(1+\gamma)}\sum g^{\gamma}(x_i;\mu,\sigma^2) + \sigma^2 \sum g^{\gamma-1}(x_i;\mu,\sigma^2)\frac{\partial g(x_i;\mu,\sigma^2)}{\partial \sigma^2} = 0
\end{aligned}
$$

and since

$$
\begin{aligned}
\frac{\partial g(x;\mu,\sigma^2)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}(2\pi\sigma^2)^{-\frac{1}{2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= (2\pi)^{-\frac{1}{2}}(-\frac{1}{2})(\sigma^2)^{-\frac{3}{2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} + (2\pi)^{-\frac{1}{2}}(\sigma^2)^{-\frac{1}{2}}\frac{(x-\mu)^2}{2(\sigma^2)^2}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= -\frac{1}{2}(\sigma^2)^{-1}g(x;\mu,\sigma^2) + g(x;\mu,\sigma^2)\frac{(x-\mu)^2}{2(\sigma^2)^2}
\end{aligned}
$$

We have,

$$
\begin{aligned}
&\frac{\gamma}{2(1+\gamma)}\sum g^{\gamma}(x_i;\mu,\sigma^2) + \sum g^{\gamma}(x_i;\mu,\sigma^2)(-\frac{1}{2} + \frac{(x_i-\mu)^2}{2\sigma^2}) = 0 \\
&\Rightarrow \frac{1}{1+\gamma}\sum g^{\gamma}(x_i;\mu,\sigma^2) = \sum g^{\gamma}(x_i;\mu,\sigma^2)\frac{(x_i-\mu)^2}{\sigma^2} \\
&\Rightarrow \hat{\sigma}^2 = (1+\gamma)\frac{\sum g^{\gamma}(x_i;\mu,\sigma^2)(x_i-\mu)^2}{\sum g^{\gamma}(x_i;\mu,\sigma^2)}
\end{aligned}
$$

We figure out that the estimator we derived from the aspect of γ-divergence is exactly a Windham robustified estimate.

doi:10.6342/NTU201900783

Chapter 3

Our Selection Criterion

In this chapter, we propose another criterion to choose c. From chapter two, we knew that the criterion B^{-2} is related to convergence rate of the estimates and hence can be easily calculated. However, if we look into it carefully, the true criterion is $\{E_F(s^2)E_F(IF^2)\}^{-1}$ rather than its upper bound B^{-2}. Therefore, our selection criterion of the tuning parameter c is,

$$\hat{c} = \arg\max_c \{E_F(s^2(x;T_c))E_F(IF^2(x;T_c,F))\}^{-1}$$

The following are some reasons that we prefer using $\{E_F(s^2E_F(IF^2))\}^{-1}$ (denoted sIF2) as criterion:

- The true criterion is $\{E_F(s^2)E_F(IF^2)\}^{-1}$ rather than its upper bound B^{-2}.
- In the same model, the selection of c is stable, since we calculated the criterion directly.
- The MSE of the estimator is smaller.
- If c is large, then $\rho(c) = (c/h'(t) - c)^2$ becomes large. That is, ρ tends to choose larger c.
- Although the convergence rate is easy to compute, it's unstable.

In each section we discuss and calculate the criterion sIF2 under various model distribution $g_\theta(x)$. The numerical performance is provided in chapter 4.

doi:10.6342/NTU201900783

3.1 g_θ is univariate normal, $\theta = \mu$

The distribution density function $g_\theta(x) = g(x;\mu)$ in this model is univariate normal with unknown mean μ and variance 1.

From eq. (1.1) $w^*(x;\mu) = K^* g^c(x;\tau(\mu))$ and from eq. (2.1) $\psi_c(x;\mu) = w^*(x;\mu)\psi(x;\tau(\mu))$, where $\tau(\mu) = \mu$ and $\psi(x;\mu) = x - \mu$

First, the score function is

$$s(x;\mu) = \frac{\partial}{\partial\mu}\{\log g(x;\mu)\}$$

$$= \frac{\partial}{\partial\mu}\{\log\frac{1}{\sqrt{2\pi}} - \frac{(x-\mu)^2}{2}\} = x - \mu$$

The Fisher information is

$$E_{\hat{F}}(s^2(x;\mu)) = \frac{1}{n}\sum(x_i - \mu)^2$$

Second, since $T_c(F)$ is an M-estimator, its influence function can be written as

$$\begin{aligned} IF(x;T_c,F) &= -E_F[\psi_c'(x;\mu)|_{\mu=T_c(F)}]^{-1}\psi_c(x;T_c(F)) \\ &= -E_F[cw^*(x;T_c(F))(x-T_c(F))^2 - w^*(x;T_c(F))]^{-1}w^*(x;T_c(F))(x-T_c(F)) \\ &= \frac{w^*(x;T_c(F))(x-T_c(F))}{1 - cE_F[w^*(x;T_c(F))(x-T_c(F))^2]} \\ \Rightarrow IF^2(x;T_c,F) &= \frac{w^{*2}(x;T_c(F))(x-T_c(F))^2}{(1 - cE_F[w^*(x;T_c(F))(x-T_c(F))^2])^2} \end{aligned}$$

From eq. (2.2), $\psi_c'(x;\mu) = cw^*(x;\mu)(x-\mu)^2 - w^*(x;\mu)$

Therefore, the asymptotic variance is

$$E_F(IF^2(x;T_c,F)) = \frac{E_F[w^{*2}(x;T_c(F))(x-T_c(F))^2]}{(1 - cE_F[w^*(x;T_c(F))(x-T_c(F))^2])^2}$$

$$E_{\hat{F}}(IF^2(x;T_c,\hat{F})) = \frac{n\sum w^2(x_i;T_c(\hat{F}))(x_i - T_c(\hat{F}))^2}{(1 - c\sum w(x_i;T_c(\hat{F}))(x_i - T_c(\hat{F}))^2)^2}$$

doi:10.6342/NTU201900783

And the asymptotic relative efficiency is

$$\{E_{\hat{F}}(s^2(x;T_c(\hat{F})))E_{\hat{F}}(IF^2(x;T_c,\hat{F}))\}^{-1}$$

$$= \{1-\sum w(x_i;T_c(\hat{F}))(x_i-T_c(\hat{F}))\}^2/\{\sum(x_i-T_c(\hat{F}))^2\sum w^2(x;T_c(\hat{F}))(x_i-T_c(\hat{F}))^2\}$$

Hence, we derived the criterion sIF2 in the model of distribution $g_\theta(x)$ is univariate normal with unknown mean μ and variance 1.

3.2 g_θ is bivariate normal, $\theta = \mu$

The distribution density function $g_\theta(x) = g(x,\mu)$ in this model is bivariate normal with unknown mean μ and variance $\mathcal{I}_2$, where $\mathcal{I}_2$ is a 2 by 2 identity matrix.

The Fisher information matrix is

$E_{\hat{F}}(s^2(x;\mu)) = E_{\hat{F}}[(x-\mu)(x-\mu)^T] = \frac{1}{n}\sum_{i=1}^n(x_i-\mu)(x_i-\mu)^T$

Robust Model Fitting part :

$E_F[\psi_c'(x;\mu)] = E_F[cw^*(x;\mu)(x-\mu)(x-\mu)^T + w^*(x;\mu)(-\mathcal{I}_2)]$ (from eq. (2.2))

$= cE_F[w^*(x;\mu)(x-\mu)(x-\mu)^T] - \mathcal{I}_2$

$E_{\hat{F}}[\psi_c'(x;\mu)] = c\sum w(x_i,\mu)(x_i-\mu)(x_i-\mu)^T - \mathcal{I}_2$

And the influence function is

$IF(x;T_c,F) = -E_F[\psi_c'(x;\mu)|_{\mu=T_c(F)}]^{-1}\psi_c(x;T_c(F))$

Hence,

$IF^2(x;T_c,F) = -E_F[\psi_c'(x;\mu)|_{\mu=T_c(F)}]^{-1}\psi_c(x;\mu)(-E_F[\psi_c'(x;\mu)|_{\mu=T_c(F)}]^{-1}\psi_c(x;T_c(F)))^T$

$= E_F[\psi_c'(x;\mu)|_{\mu=T_c(F)}]^{-1}\psi_c(x;T_c(F)\psi_c(x;T_c(F))^T E_F[\psi_c'(x;\mu)|_{\mu=T_c(F)}]^{-1}$

Note that

$E_F[\psi_c(x;T_c(F)\psi_c(x;T_c(F))^T] = E_F[w^*(x-T_c(F))(w^*(x-T_c(F)))^T]$

$E_{\hat{F}}[\psi_c(x;T_c(\hat{F})\psi_c(x;T_c(\hat{F}))^T] = \sum w_i^2(x_i-T_c(\hat{F}))(x_i-T_c(\hat{F}))^T$,

where w_i is the weight for the i-th data. We have

$E_{\hat{F}}[IF^2(x;T_c,\hat{F})]$

$= (c\sum w_i(x_i-T_c(\hat{F}))(x_i-T_c(\hat{F}))^T - \mathcal{I}_2)^{-1}(\sum w_i^2(x_i-T_c(\hat{F}))(x_i-T_c(\hat{F}))^T)$

doi:10.6342/NTU201900783

$(c\sum w_i(x_i - T_c(\hat{F}))(x_i - T_c(\hat{F}))^T - \mathcal{I}_2)^{-1}$

Finally, we derive the formula of $\{E_{\hat{F}}[s^2]E_{\hat{F}}[IF^2]\}^{-1}$ and we choose the reciprocal of the product of the largest eigenvalue of $E_{\hat{F}}[s^2(x;T_c(\hat{F}))]$ and $E_{\hat{F}}[IF^2(x;T_c,\hat{F})]$ as our criterion.

3.3 g_θ is univariate normal, $\theta = (\mu, \sigma^2)^T$

The distribution density function $g_\theta(x) = g(x,(\mu,\sigma^2)^T)$ in this model is univariate normal with unknown mean μ and variance σ^2.

The Fisher information matrix is

$E_F[s^2(x;\theta)]_{ij} = E_F[s(x;\theta)(s(x;\theta)^T)]_{ij} = E_F[(\frac{\partial}{\partial\theta_i}\log g(x;\theta))(\frac{\partial}{\partial\theta_j}\log g(x;\theta))]$

Here, $\theta = (\mu,\sigma^2)^T$

Hence,

$$E_F(s(x;\theta)(s(x;\theta))^T) = E_F\begin{bmatrix} \frac{(x-\mu)^2}{\sigma^4} & \frac{(x-\mu)}{\sigma^2}(-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4}) \\ \frac{(x-\mu)}{\sigma^2}(-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4}) & (-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4})^2 \end{bmatrix}$$

$$\psi(x;\theta) = \begin{bmatrix} x-\mu \\ (x-\mu)^2 - \sigma^2 \end{bmatrix}$$

Robust Model Fitting part :

$w^*(x;\theta) = \frac{1}{\sqrt{c+1}}\exp(-\frac{c}{2\sigma^2}(x-\mu)^2)$

$$\psi_c(x;\theta) = w^*(x;\theta)\psi(x;\tau_c(\theta)) = w^*(x;\theta)\begin{bmatrix} x-\mu \\ (x-\mu)^2 - \frac{\sigma^2}{c+1} \end{bmatrix}$$

$E_F[\psi_c'(x;\theta)]$

$$= E_F\begin{bmatrix} \frac{\partial}{\partial\mu}w^*(x;\theta)(x-\mu) & \frac{\partial}{\partial\sigma^2}w^*(x;\theta)(x-\mu) \\ \frac{\partial}{\partial\mu}w^*(x;\theta)[(x-\mu)^2-\frac{\sigma^2}{c+1}] & \frac{\partial}{\partial\sigma^2}w^*(x;\theta)[(x-\mu)^2-\frac{\sigma^2}{c+1}] \end{bmatrix}$$

$$= E_F\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

doi:10.6342/NTU201900783

We know

$$\frac{\partial}{\partial \mu} w^*(x;\theta) = c w^*(x;\theta) \frac{x-\mu}{\sigma^2}$$

and

$$\frac{\partial}{\partial \sigma^2} w^*(x;\theta) = c w^*(x;\theta) \frac{(x-\mu)^2}{2\sigma^4}$$

Thus,

$A = \frac{\partial}{\partial \mu} w^*(x;\theta)(x-\mu) = c w^*(x;\theta) \frac{(x-\mu)^2}{\sigma^2} - w^*(x;\theta)$

$B = \frac{\partial}{\partial \sigma^2} w^*(x;\theta)(x-\mu) = c w^*(x;\theta) \frac{(x-\mu)^3}{2\sigma^4}$

$C = \frac{\partial}{\partial \mu} w^*(x;\theta)[(x-\mu)^2 - \frac{\sigma^2}{c+1}] = c w^*(x;\theta)(\frac{(x-\mu)^3}{\sigma^2} - \frac{x-\mu}{c+1}) - 2w^*(x;\theta)(x-\mu)$

$D = \frac{\partial}{\partial \sigma^2} w^*(x;\theta)[(x-\mu)^2 - \frac{\sigma^2}{c+1}] = c w^*(x;\theta)(\frac{(x-\mu)^4}{2\sigma^4} - \frac{(x-\mu)^2}{2\sigma^2(c+1)}) - \frac{w^*(x;\theta)}{c+1}$

Hence, from the formula before

$IF^2(x;T_c,F) = -E_F[\psi_c'(x;\theta)|_{\theta=T_c(F)}]^{-1} \psi_c(x;T_c(F))(-E_F[\psi_c'(x;\theta)|_{\theta=T_c(F)}]^{-1} \psi_c(x;T_c(F)))^T$

$= E_F[\psi_c'(x;\theta)|_{\theta=T_c(F)}]^{-1} \psi_c(x;T_c(F)) \psi_c(x;T_c(F))^T E_F[\psi_c'(x;\theta)|_{\theta=T_c(F)}]^{-1}$

Thus, we derive the formula of $\{E_F[s^2]E_F[IF^2]\}^{-1}$ and choose the reciprocal of the product of the largest eigenvalue of $E_F[s^2(x;T_c(F))]$ and $E_F[IF^2(x;T_c,F)]$ as our criterion.

3.4 g_θ is qGaussian, $\theta = \mu$

3.4.1 Some Notes about qGaussian

- β-power model pdf (qGaussian):

$$f_\beta(x;\mu,\Sigma) = c_\beta (det 2\pi\Sigma)^{-\frac{1}{2}} \{1 - \frac{\beta}{2+(p+2)\beta}(x-\mu)^T\Sigma^{-1}(x-\mu)\}_+^{1/\beta}$$

,where $\beta > -2/(p+2)$, $x \in R^p$

As $\beta \to 0$, f_β converges to normal density.

$f_0(x;\mu,\Sigma) := \lim_{\beta\to 0} f_\beta(x;\mu,\Sigma)$

$= \lim_{\beta\to 0} c_\beta \|2\pi\Sigma\|^{-\frac{1}{2}} \{1 - \frac{1}{2/\beta+(p+2)}(x-\mu)^T\Sigma^{-1}(x-\mu)\}_+^{1/\beta}$

$= c_0 \|2\pi\Sigma\|^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$, which is $\mathcal{N}(\mu,\Sigma)$

And when $-2/(p+2) < \beta < 0$, f_β is t-distribution.

doi:10.6342/NTU201900783

- γ-cross entropy

  In section 2.4.1, we have given the form of empirical γ-cross entropy $d_\gamma(\hat{f}, g_\theta)$

$$d_\gamma(\hat{f}, g_\theta) = -\frac{1}{\gamma}\log\{\frac{1}{n}\sum_{i=1}^{n} g(x_i;\theta)^\gamma\} + \frac{1}{1+\gamma}\log\int g(x;\theta)^{1+\gamma}dx$$

  ,where $\hat{f}$ is the empirical pdf.

  Hence, we substitute g_θ to f_β,

$$\begin{aligned}
&d_\gamma(\hat{f}, f_\beta(x;\mu,\Sigma)) \\
&= -\frac{1}{\gamma}\log\{\frac{1}{n}\sum f_\beta(x_i;\mu,\Sigma)^\gamma\} + \frac{1}{1+\gamma}\log\int f_\beta(x;\mu,\Sigma)^{1+\gamma}dx \\
&\Rightarrow e^{-d_\gamma(\hat{f},f_\beta)} = \{\frac{1}{n}\sum f_\beta^\gamma(x;\mu,\Sigma)\}^{1/\gamma}\{\int f_\beta(x;\mu,\Sigma)^{1+\gamma}dx\}^{-\frac{1}{1+\gamma}}
\end{aligned}$$

  note that

$$\begin{aligned}
&\int f_\beta(x;\mu,\Sigma)^{1+\gamma}dx \\
&= \int c_\beta^{1+\gamma}(det2\pi\Sigma)^{-\frac{1+\gamma}{2}}\{1-\frac{\beta}{2+p+2\beta}(x-\mu)^T\Sigma^{-1}(x-\mu)\}_+^{\frac{1+\gamma}{\beta}}dx \\
&= \int \frac{c_\beta^{1+\gamma}}{c_{\beta/(1+\gamma)}}(det2\pi\Sigma)^{-\gamma/2}k^{p/2}c_{\beta/(1+\gamma)}(det2\pi k\Sigma)^{-1/2} \\
&\{1-\frac{\beta/(1+\gamma)}{2+(p+2)\beta/(1+\gamma)}(x-\mu)^T(k\Sigma)^{-1}(x-\mu)\}_+^{\frac{1+\gamma}{\beta}}dx \\
&= \frac{c_\beta^{1+\gamma}}{c_{\beta/(1+\gamma)}}(det2\pi\Sigma)^{-\gamma/2}k^{p/2}\int f_{\beta/(1+\gamma)}(x;\mu,k\Sigma)dx \\
&= \frac{c_\beta^{1+\gamma}}{c_{\beta/(1+\gamma)}}(det2\pi\Sigma)^{-\gamma/2}k^{p/2},
\end{aligned}$$

doi:10.6342/NTU201900783

where $k = \frac{\frac{\beta/(1+\gamma)}{2+(p+2)\beta/(1+\gamma)}}{\frac{\beta}{2+(p+2)\beta}} = \frac{2/\beta+(p+2)}{2(1+\gamma)/\beta+(p+2)}$

$$
\begin{aligned}
&\Rightarrow e^{-d_\gamma(\hat{f},f_\beta)\gamma} = \{\frac{1}{n}\sum f_\beta^\gamma(x;\mu,\Sigma)\}\{\int f_\beta(x;\mu,\Sigma)^{1+\gamma}dx\}^{-\frac{\gamma}{1+\gamma}} \\
&= \frac{1}{n}\sum f_\beta^\gamma\{\frac{c_\beta^{1+\gamma}}{c_{\beta/(1+\gamma)}}(det\Sigma)^{-\gamma/2}(\frac{k}{2\pi^\gamma})^{p/2}\}^{-\frac{\gamma}{1+\gamma}} \\
&\propto \sum f_\beta^\gamma(x_i;\mu,\Sigma)(det\Sigma)^{\frac{\gamma^2}{2(1+\gamma)}} \\
&= L_{\beta,\gamma}(\mu,\Sigma)
\end{aligned}
$$

Hence, the robust estimator is to maximize $L_{\beta,\gamma}(\mu,\Sigma)$

- γ-utility function (similar to likelihood function)

$$L_{\beta,\gamma}(\mu,\Sigma) = (det\Sigma)^{\frac{\gamma^2}{2(1+\gamma)}}\sum_{i=1}^{n} f_\beta(x_i;\mu,\Sigma)^\gamma$$

3.4.2 Weighted Distribution

To maximize this, take derivatives with respect to μ, and consider $x \in R^p$, where $p = 1$.

$$
\begin{aligned}
&\frac{\partial}{\partial\mu}\left((\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}}\sum f_\beta^\gamma(x_i;\mu,\sigma^2)(x_i;\mu,\sigma^2)\right) \\
&= (\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}}\gamma\sum f_\beta^{\gamma-1}(x_i;\mu,\sigma^2)\frac{\partial}{\partial\mu}f_\beta(x_i;\mu,\sigma^2) \\
&= (\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}}\gamma\sum f_\beta^{\gamma-1}(x_i;\mu,\sigma^2)(c_\beta\frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{\beta}\{1-\frac{\beta}{2+3\beta}\frac{(x_i-\mu)^2}{\sigma^2}\}_+^{\frac{1}{\beta}-1}\frac{\beta}{2+3\beta}\frac{2}{\sigma^2}(x_i-\mu)) \\
&= (\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}}\gamma\sum f_\beta^{\gamma-1}(x_i;\mu,\sigma^2)(c_\beta\frac{1}{\sqrt{2\pi\sigma^2}}\{1-\frac{\beta}{2+3\beta}\frac{(x_i-\mu)^2}{\sigma^2}\}_+^{\frac{1}{\beta}})^{1-\beta}c_\beta^\beta(\frac{1}{\sqrt{2\pi\sigma^2}})^\beta\frac{2}{2+3\beta}\frac{x_i-\mu}{\sigma^2} \\
&= (\sigma^2)^{\frac{\gamma^2}{2(1+\gamma)}}\gamma\sum f_\beta^{\gamma-\beta}(x_i;\mu,\sigma^2)(x_i-\mu)\frac{2c_\beta^\beta}{(2+3\beta)\sigma^2}(\frac{1}{\sqrt{2\pi\sigma^2}})^\beta
\end{aligned}
$$

Let the result equal to 0, we have

$$\hat{\mu} = \frac{\sum f_\beta^{\gamma-\beta}(x_i;\mu,\sigma^2)x_i}{\sum f_\beta^{\gamma-\beta}(x_i;\mu,\sigma^2)}$$

doi:10.6342/NTU201900783

Hence, we can define a weighted distribution

$dF_{\beta,\gamma,t}(x) = w^*_{\beta,\gamma}(x;t)dF(x)$ with $w^*_{\beta,\gamma}(x;t) = f_\beta^{\gamma-\beta}(x;t)/E_F[f_\beta^{\gamma-\beta}(x;t)]$

Then we calculate the weighted variance

$$
\begin{aligned}
&\int f_\beta^{\gamma-\beta}(x;\mu,\sigma^2)(x-\mu)^2 dF \\
&= \int (c_\beta \frac{1}{\sqrt{2\pi\sigma^2}}\{1 - \frac{\beta}{2+(2+p)\beta}\frac{(x-\mu)^2}{(\sigma)^2}\}_+^{1/\beta})^{\gamma-\beta+1} dx \\
&= \int (c_\beta \frac{1}{\sqrt{2\pi\sigma^2}})^{\gamma-\beta+1}\{1 - \frac{\beta}{2+(2+p)\beta}\frac{(x-\mu)^2}{(\sigma)^2}\}_+^{(\gamma-\beta+1)/\beta} dx \\
&= \int Constant\{1 - \frac{\beta}{2+(2+p)\beta}\frac{(x-\mu)^2}{(\sigma)^2}\}_+^{(\gamma-\beta+1)/\beta} dx \\
&(\frac{\beta}{2+(2+p)\beta} = \frac{1}{2/\beta+(2+p)} \to \frac{1}{2(\gamma-\beta+1)/\beta+(2+p)} = \frac{\beta}{2\gamma+2+p\beta}) \\
&= \frac{2+(2+p)\beta}{2\gamma+2+p\beta}\sigma^2
\end{aligned}
$$

$$
\Rightarrow \hat{\sigma} = \frac{2\gamma+2+p\beta}{2+(2+p)\beta}\frac{\sum f_\beta^{\gamma-\beta}(x;\mu,\sigma^2)(x-\mu)^2}{\sum f_\beta^{\gamma-\beta}(x;\mu,\sigma^2)}\}
$$

This result is same as in Local Fixed-Point Algortithm (Akifumi Notsu, Shinto Eguchi, 2016)[1].

3.4.3 Applied to gamma-clust

We applied the result to gamma-clust, and assume that the distribution density function $g_\theta(x) = f_\beta(x;\mu)$ in this model is univariate qGaussian with unknown mean μ and variance 1.

doi:10.6342/NTU201900783

First, the score function is,

$$\begin{aligned} s(x;\mu) &= \frac{\partial}{\partial\mu}\{\log f_\beta(x;\mu)\} \\ &= \frac{\partial}{\partial\mu}\{\log\frac{c_\beta}{\sqrt{2\pi}} + \frac{1}{\beta}\log\{1-\frac{\beta}{2+3\beta}(x-\mu)^2\}_+\} \\ &= \frac{1}{\beta}\frac{\frac{2\beta}{2+3\beta}(x-\mu)}{\{1-\frac{\beta}{2+3\beta}(x-\mu)^2)\}_+} \\ &= \frac{2}{2+3\beta}\frac{(x-\mu)}{\{1-\frac{\beta}{2+3\beta}(x-\mu)^2\}_+} \end{aligned}$$

Hence the Fisher information is,

$$E_{\hat{F}}(s^2(x;\mu)) = \frac{1}{n}(\frac{2}{2+3\beta})^2\sum_i\frac{(x_i-\mu)^2}{\{1-\frac{\beta}{2+3\beta}(x_i-\mu)^2\}_+^2}$$

And the influence function is,

$$\begin{aligned} IF(x;T_c,F) &= -E_F[\psi_c'(x;\mu)|_{\mu=T_c(F)}]^{-1}\psi_c(x;T_c(F)) \\ &= -E_F[((w^*(x;\mu))'\psi(x;\mu)+w^*(x;\mu)\psi'(x;\mu))|_{\mu=T_c(F)}]^{-1}w^*(x;T_c(F))\psi(x;T_c(F)) \\ &= -E_F[(cw^*(x;\mu)(\frac{\partial}{\partial\mu}\log f_\beta)(x-\mu)-w^*(x;\mu))|_{\mu=T_c(F)}]^{-1}w^*(x;T_c(F))(x-T_c(F)) \\ &= \frac{w^*(x;T_c(F))(x-T_c(F))}{1-\frac{2c}{2+3\beta}E_F[w^*(x;T_c(F))\frac{(x-T_c(F))^2}{\{1-\frac{\beta}{2+3\beta}(x-T_c(F))^2\}_+}]} \end{aligned}$$

Similar to Section 3.1, the empirical asymptotic relative efficiency is,

$$\{E_{\hat{F}}(s^2(x;T_c(\hat{F})))E_{\hat{F}}(IF^2(x;T_c,\hat{F}))\}^{-1}$$

$$= \{(\frac{2}{2+3\beta})^2\sum\frac{(x_i-T_c(\hat{F}))^2}{\{1-\frac{\beta}{2+3\beta}(x_i-T_c(\hat{F}))^2\}_+^2}(\frac{\sum w^2(x_i;T_c(\hat{F}))(x_i-T_c(\hat{F}))^2}{(1-\frac{2c}{2+3\beta}\frac{1}{n}\sum w(x_i;T_c(\hat{F}))\frac{(x_i-T_c(\hat{F}))^2}{\{1-\frac{\beta}{2+3\beta}(x_i-T_c(\hat{F}))^2\}_+})^2})\}^{-1}$$

Thus, we derived the criterion sIF2 in the model of distribution $g_\theta(x)$ is univariate qGaussian with unknown mean μ and variance 1.

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

Chapter 4

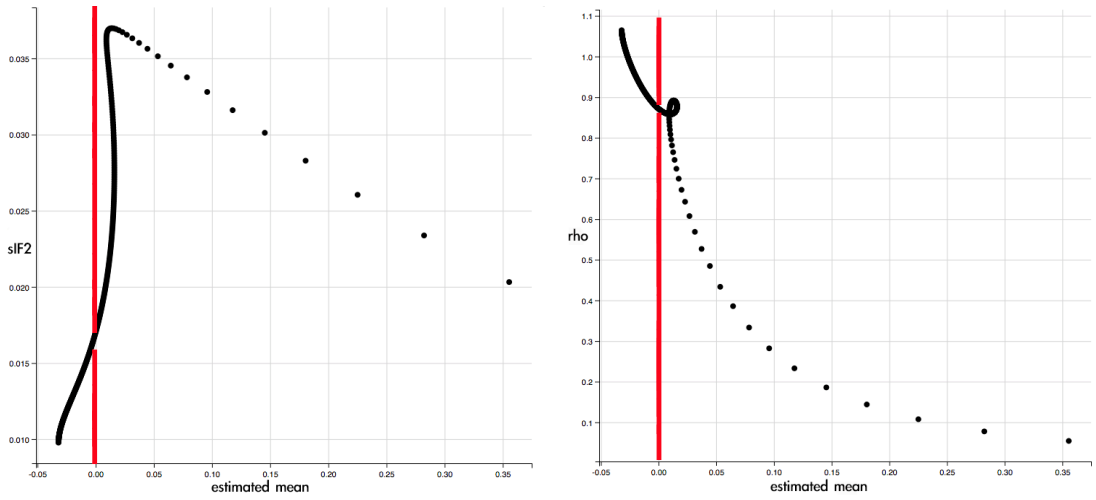
Numerical Examples

4.1 One Dimension Case

4.1.1 One component with outliers

In this subsection, we compare two criteria by presenting two examples.

For first example, assume there are 280 samples from $\mathcal{N}(0,1)$ which is the main distribution we are going to estimate and 120 samples from $\mathcal{N}(7,1)$ which is seemed to be outliers, and we already know that the variance of the main distribution is 1. The following is the result: The y-axis of Figure 4.1(a) means the value of $\{E_{\hat{F}}(s^2)E_{\hat{F}}(IF^2)\}^{-1}$, and



(a) sIF2 criterion (b) rho criterion

Figure 4.1:

doi:10.6342/NTU201900783

c	m	rho	sIF2	r	inverse of IF2
0.10	0.355485	0.054209	0.020318	0.300456	118.5937
0.20	0.044493	0.484657	0.035632	0.223171	224.6873
0.27	0.015466	0.723932	0.036972	0.240890	234.9016
0.28	0.013881	0.745662	**0.036986**	0.244859	235.0853
0.29	0.012623	0.764607	0.036973	0.249052	235.0815
0.30	0.011633	0.781335	0.036938	0.253393	234.9166
0.40	0.009357	0.867237	0.035837	0.300468	228.0527
0.50	0.011366	0.888907	0.034169	0.346544	217.3251
1.00	0.015714	0.870179	0.025893	0.517375	164.4990
2.00	-0.002410	0.877856	0.016394	0.680981	104.6453
3.00	-0.023187	0.973124	0.012213	0.752545	78.3862
4.00	-0.032000	**1.064666**	0.009778	0.794940	62.9035

Table 4.1: result of first example. r is the convergence rate in the 20th iteration.

the y-axis of Figure 4.1(b) means Windham's criterion ρ. And the x-axis of both figures (m) is the mean estimate each for a different c, from 0.1 to 4 with distance 0.01 and all estimates iterate for 20 times. Hence, the chosen c is 4.00 for the criterion ρ and 0.28 for the criterion sIF2 (Table 4.1).

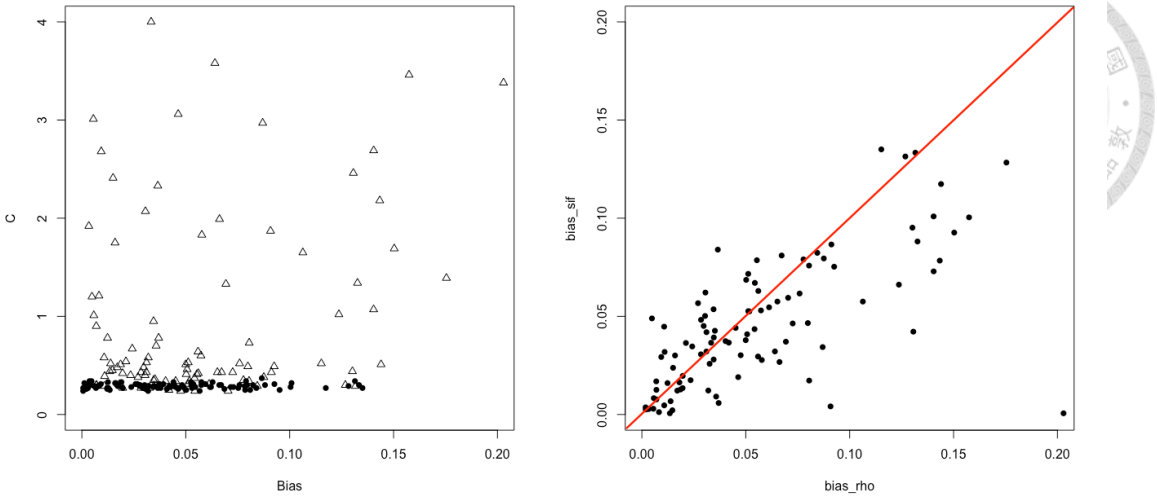
Next, we do this process for 100 different data with same distribution assumption, and compare their bias, which is the absolute value of the difference between 0 and the robustified mean estimates. In Figure 4.2(a), the y-axis is the chosen c, circle points stand for estimates from criterion sIF2 and triangle points for estimates from criterion ρ. We can see that the chosen parameter c by criterion sIF2 is much more stable for the parameter chosen by ρ. In Figure 4.2(b), it is clear that bias of sIF2 is smaller than of ρ, since there are fewer points in the up-left triangle of the figure than those in the down-right triangle part. We also compute their mean, variance, and mean squared error (Table 4.2).

	mean	variance	MSE
rho	0.064793	0.003796	0.007956
sIF2	0.056344	0.001902	0.005057

Table 4.2: Comparing data in first example

For second example, we assume there are 280 samples from $\mathcal{N}(-1, 1)$ which is the main distribution we are going to estimate and 120 samples from $\mathcal{N}(4, 1)$ which is seemed to be outliers, and we already know that the variance of the main distribution is 1. The

doi:10.6342/NTU201900783

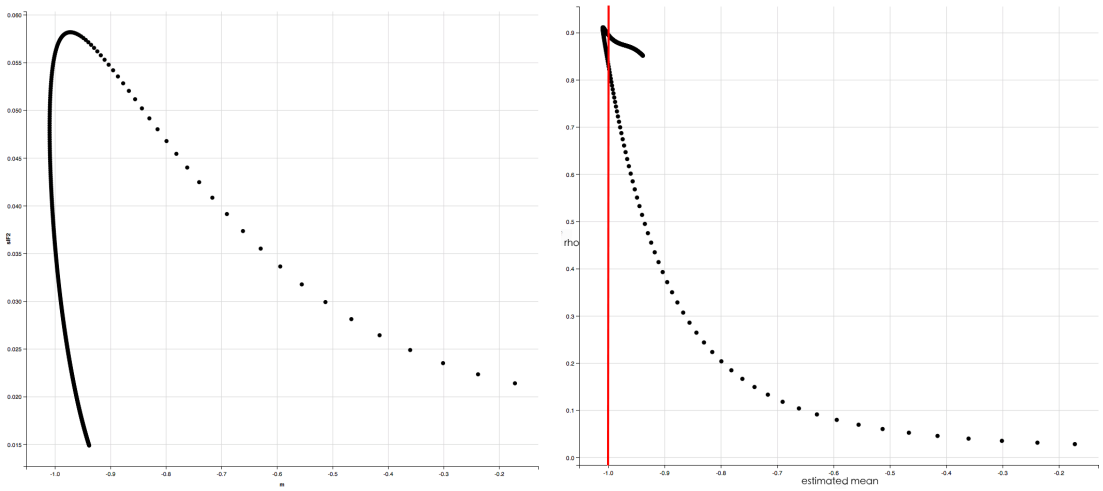


(a) (b)

Figure 4.2:

following is the result:

The y-axis of Figure 4.3(a) means the value of $\{E_{\hat{F}}(s^2)E_{\hat{F}}(IF^2)\}^{-1}$, and The y-axis of Figure 4.3(b) means Windham's criterion ρ. And the x-axis of both figures (m) is the mean estimate each for a different c, from 0.1 to 4 with distance 0.01 and all estimates iterate for 20 times. Hence, the chosen c is 1.12 for the criterion ρ and 0.5 for the criterion sIF2 (Table 4.3).



(a) sIF2 criterion (b) rho criterion

Figure 4.3:

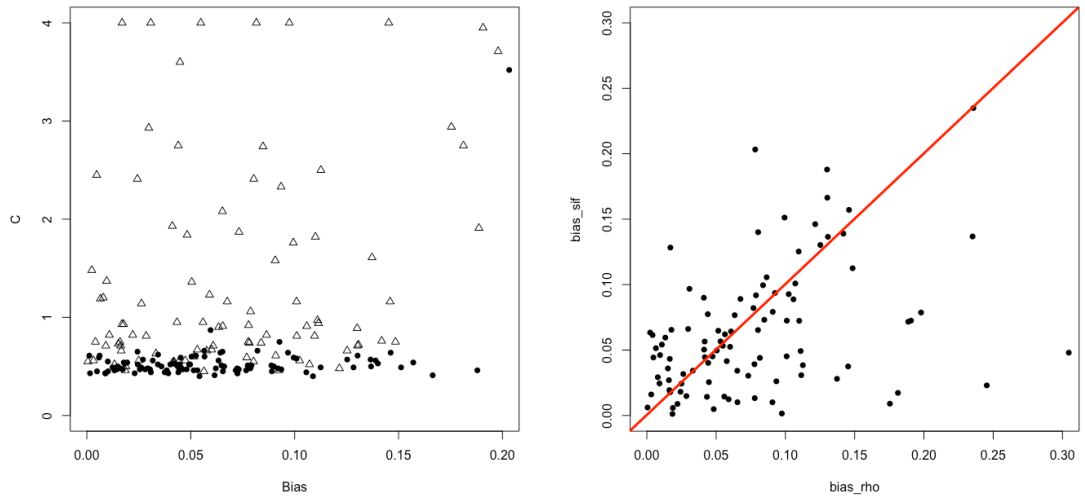
Next, we do this process for 100 different data with same distribution assumption,

doi:10.6342/NTU201900783

c	m	rho	sIF2	r	inverse of IF2
0.10	-0.171572	0.028217	0.021389	0.373163	58.6819
0.30	-0.855458	0.285477	0.051150	0.359583	168.9291
0.40	-0.934677	0.494710	0.056845	0.362530	192.7951
0.50	-0.971573	0.660704	**0.058169**	0.380855	199.7947
0.60	-0.990432	0.771304	0.057426	0.405888	198.5343
0.80	-1.005879	0.877478	0.053476	0.460634	185.8748
1.10	-1.009079	0.910806	0.046233	0.535446	160.8778
1.12	-1.008926	**0.910903**	0.045764	0.539912	159.2387
1.14	-1.008745	0.910869	0.045300	0.544310	157.6129
1.20	-1.008050	0.910123	0.043934	0.557102	152.8232
1.40	-1.004542	0.903702	0.039711	0.595584	137.9662
1.60	-0.999934	0.895686	0.036020	0.628336	124.9419

Table 4.3: result of second example. r is the convergence rate in the 20th iteration.

and compare their bias, which is the absolute value of the difference between -1 and the robustified mean estimates. In Figure 4.4(a), the y-axis is the chosen c, circle points stand for estimates from criterion sIF2 and triangle points for estimates from criterion ρ. We can see that the chosen parameter c by criterion sIF2 is much more stable by criterion ρ. In Figure 4.4(b), it is clear that bias of sIF2 is more smaller than of ρ, since there are fewer points in the up-left triangle of the figure than those in the down-right triangle part. We also compute their mean, variance, and mean squared error (Table 4.4).



(a) (b)

Figure 4.4:

doi:10.6342/NTU201900783

	mean	variance	MSE
rho	0.079325	0.004226	0.010477
sIF2	0.063453	0.002000	0.006006

Table 4.4: Comparing data in second example

From the simulations of those two types of mixture distribution data, we find out that Windham's robustified estimates can find the mean for the main distribution if two distributions are not too close, and even can be used to do clustering. And then we use the criterion sIF2 we suggested to choose the tuning parameter c. So, next subsection is the process of one dimension clustering.

4.1.2 Five components

Suppose there are 1000 samples, in which 200 from $\mathcal{N}(-14, 1)$, 200 from $\mathcal{N}(-10, 1)$, 200 from $\mathcal{N}(-4, 1)$, 200 from $\mathcal{N}(0, 1)$, 200 from $\mathcal{N}(5, 1)$. We only know that the variance of each component is 1.

First, we randomly choose 100 points from data, and let them be the initial points. After 20 times iteration, there are 100 mean estimates for one fixed c. Those estimates converge to several points. We let c be from 0.2 to 2 with width 0.05, and find out the most common number of converge points is 5. Then for those c with correct number of converge points, we use sIF2 criterion to choose c. The result is shown in Figure 4.5 and Table 4.5.

4.2 Two Dimension-One Component with Outliers

We extend data from one dimension to two dimension case.

Now, suppose there are 560 samples from $\mathcal{N}((0, 0)^T, \mathcal{I}_2)$ and 240 samples from $\mathcal{N}((3, 3)^T, \mathcal{I}_2)$, and we want to estimate the mean of the main distribution. We let c be from 0.3 to 3 with step 0.01. Figure 4.6 shows the result. The estimated mean is $(0.010235, 0.074160)$ and the chosen c is 0.55.

doi:10.6342/NTU201900783

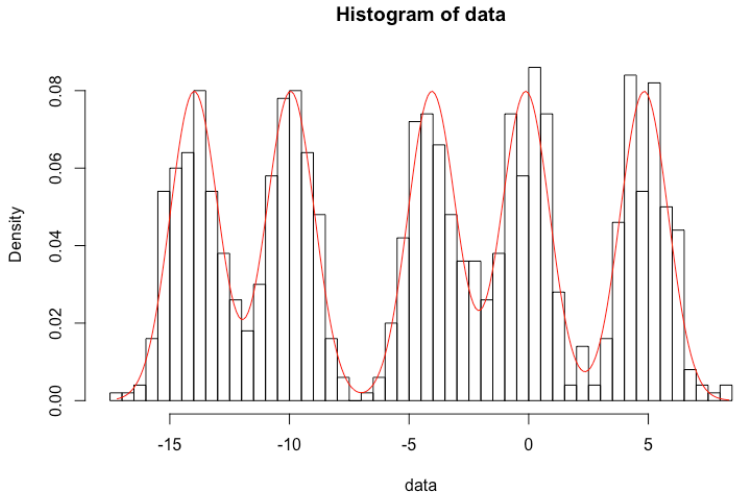


Figure 4.5: Red curve is the estimated density.

c	m_1	m_2	m_3	m_4	m_5	$sIF2_1$	$sIF2_2$	$sIF2_3$	$sIF2_4$	$sIF2_5$
0.60	-13.838	-10.092	-3.791	-0.313	**4.829**	.434	1.160	.859	.814	**.904**
0.65	-13.878	-10.057	-3.849	-0.272	4.835	.472	1.299	1.029	.930	.899
0.70	-13.908	-10.030	-3.893	-0.239	4.839	.497	1.409	1.171	1.025	.886
0.95	**-13.982**	-9.962	-4.012	-0.143	4.838	**.511**	1.652	1.523	1.247	.773
1.00	-13.989	-9.955	-4.025	-0.131	4.836	.501	1.663	1.547	1.256	.747
1.05	-13.994	**-9.949**	-4.036	**-0.121**	4.833	.490	**1.667**	1.563	**1.259**	.721
1.10	-13.999	-9.945	**-4.046**	-0.111	4.831	.478	1.665	**1.570**	1.255	.695
1.20	-14.006	-9.939	-4.063	-0.096	4.825	.451	1.648	1.570	1.234	.645
1.40	-14.012	-9.933	-4.086	-0.072	4.811	.395	1.584	1.525	1.156	.553
1.80	-14.009	-9.934	-4.111	-0.037	4.783	.300	1.417	1.374	.947	.402
2.00	-14.002	-9.937	-4.118	-0.025	4.769	.263	1.336	1.293	.841	.341

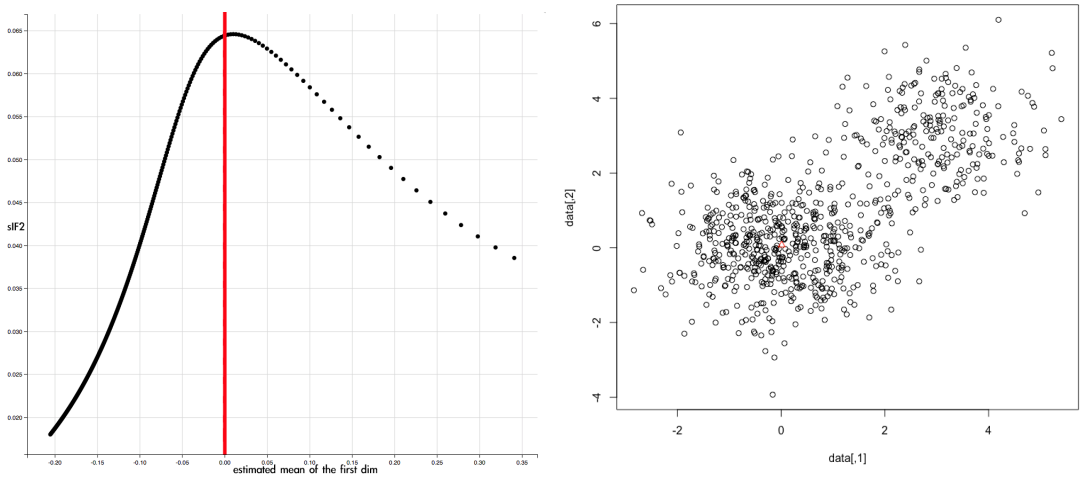
Table 4.5: Result of 1-dim clustering. m_i and $sIF2_i$ are estimated mean and criterion of group i

4.3 Variance Unknown

We now back to one dimension data but variance is unknown.

Now, suppose there are 340 samples from $\mathcal{N}(0, 1)$ and 60 samples from $\mathcal{N}(4, 0.5)$, and we want to estimate the mean and variance of the main distribution. We let c be from 0.3 to 3 with step 0.01. Figure 4.7 shows the result. The estimated mean is 0.027701, estimated variance is 0.988843 and the chosen c is 0.86.

doi:10.6342/NTU201900783



(a) (b)

Figure 4.6: Result of 2-dim data. (a) mean of the first dimension and sIF2. (b) scatter plot and the estimated mean (red triangle).

4.4 qGaussian Model

We give example of using qGaussian model to estimate mean.

$$\hat{\mu} = \sum f_\beta^{\gamma-\beta} x_i / \sum f_\beta^{\gamma-\beta}$$

From the formula above, it is similar to normal distribution case with exponent $c = \gamma - \beta$. Note that qGaussian density is a function of β, so if β is fixed, we can use the asymptotic efficiency criterion in 3.4.3 to choose γ.

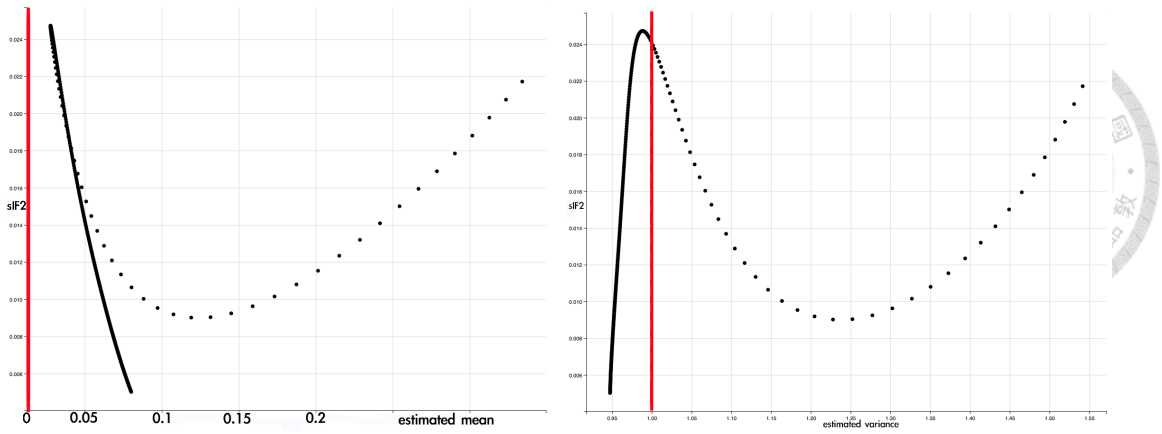
Suppose there are 400 samples, in which 280 samples from $\mathcal{N}(0, 1)$ and 120 samples from $\mathcal{N}(5, 1)$, and we already know the variance of the qGaussian model is 1. Table 4.6 shows the result:

β	0	0.2	0.3	0.4	0.5	0.6
mean	0.0363	-0.0083	-0.0293	-0.0710	0.0210	0.0107
chosen γ	0.52	1.08	1.62	2.96	0.94	1.21
chosen $\gamma - \beta$	0.52	0.88	1.32	2.56	0.44	0.61

Table 4.6: Result of qGaussain model. The sample mean of the main distribution is 0.0331.

Note that for different β, the γ we chosen $c = \gamma - \beta$ will be very different. And the

doi:10.6342/NTU201900783



(a) (b)

Figure 4.7: Variance unknown simulation. (a) estimated mean and sIF2. (b) estimated variance and sIF2.

tail of qGaussian density will decay to 0, which is different from normal density, may be an advantage on robustness or clustering.

As soon as we derive the qGaussain robustified estimate $\hat{\mu}$, there are points with 0 weight. For those of non-zero weight, we calculate the sample mean $\hat{\mu}_{trim}$, and compare to the qGaussain robustified estimate $\hat{\mu}$ we derived from the start. We run the estimated procedure for 100 times with $\beta = 0.5$, γ from 0.55 to 3.5 with step 0.01, and calculate the mean, variance, and mean squared error of these two robustified estimators.

	mean	variance	MSE
$\hat{\mu}$	0.15259	0.012274	0.035434
$\hat{\mu}_{trim}$	0.21052	0.007867	0.052107
distance is 3.5			
$\hat{\mu}$	0.05403	0.006732	0.009584
$\hat{\mu}_{trim}$	0.08278	0.005074	0.011876
distance is 4			
$\hat{\mu}$	0.02413	0.005352	0.005880
$\hat{\mu}_{trim}$	0.02483	0.004776	0.005344
distance is 5			
$\hat{\mu}$	-0.00173	0.006118	0.006060
$\hat{\mu}_{trim}$	-0.00099	0.005405	0.005351
distance is 6			

Table 4.7: Comparing $\hat{\mu}$ and $\hat{\mu}_{trim}$ with different distance of two populations.

We find out that $\hat{\mu}_{trim}$ is better, but if we change the data to 280 samples from $\mathcal{N}(0, 1)$ and 120 samples from $\mathcal{N}(4, 1)$, the result will be on the contrary. We simulate different

doi:10.6342/NTU201900783

cases with different distance of two populations in Table 4.7.

doi:10.6342/NTU201900783

doi:10.6342/NTU201900783

Bibliography

[1] Shinto Eguchi Akifumi Notsu. Robust clustering method in the presence of scattered observations. *Neural Computation*, 28(6):1141–1162, 2016.

[2] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[3] Michael P. Windham. Robustifying model fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):599–609, 1995.

doi:10.6342/NTU201900783