Robustness Property of Exponential Families and Student Distribution

Traineeship Report

Centrum Wiskunde & Informatica, Amsterdam, the Netherlands

Trainee: Evangelos Tsiamalos[1]
Supervisor: Prof. Peter D. Grünwald[2]

[1] University of Crete, Heraklion, Greece
[2] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2] Leiden University, Leiden, The Netherlands

[1]tsiamalos.e@gmail.com
[2]pdg@cwi.nl

July 29th, 2022

Contents

1 Introduction

This project focuses on the investigation of the robustness property of the exponential families on the Student distribution. The main motivation for that is that the Student distribution, which does not belong to the exponential families, approximates the Normal distribution. One could see this project as a part of a more general investigation of the behaviour of the likelihood ratio in the scope of hypothesis testing. For that reason, we can have in mind that the setting is that of a hypothesis test, and for simplicity, a one-sided hypothesis test of the form $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$.

First, I establish the setting and the definitions for this project. Although I provide the general setting for all Exponential Families, the main motivation is to investigate some of the properties in the very specific case of the Student Distribution, which we hope will have a similar behaviour to that of the Exponential Families. In terms of definitions and the structure of our setting, I follow Peter Grünwald's *The Minimum Description Length Principle* (Grunwald, 2007). The definitions and most of the material presented in this report are covered in Chapters 18 and 19 of Grunwald, 2007.

1.1 Exponential Families

Definition 1.1. *One Parameter Exponential Families - Cannonical Parametrization*
A family of distributions $\{P_\beta, \beta \in \Theta)\}$ is called One Parameter Exponential Family if they have densities of the form

$$f_\beta(x) = e^{\beta\phi(x)-\psi(\beta)}r(x) \ , \ x \in \mathbb{R} \tag{1}$$

This parametrization is called the Canonical Parametrization of the Exponential Family. We denote the Canonical Parameter Space as Θ_{can} and the Exponential Family as $\{f_\beta, \beta \in \Theta_{can}\}$

Remark. *We assume that Θ_{can} is an open, convex set.*

With this definition, along with certain properties we can obtain, we define the Mean-Value Parametrization.

Definition 1.2. *One Parameter Exponential Families - Mean Value Parametriation*
Each member of P_β of an exponential family can be identified by the expectation of its statistic $\mathbb{E}\phi(X) = \mu$. We denote $\Theta_{mean} = \{\mu \in \mathbb{R} : \exists\beta \in \Theta_{can}$ such that $\mathbb{E}_\beta\phi(X) = \mu\}$.

We are interested in some properties of the Mean-Value Parametrization:

1. Θ_{mean} is convex
2. For all $\mu_0 \in \Theta_{mean}$, the function $\mu \mapsto D(\mu \parallel \mu_0)$ is a strictly convex function of μ.

The second property, is going to be of importance for the rest of the project. For more details on the definitions and the basic properties see Grünwald 2007.

1.2 Robustness Property and Maximum Likelihood Estimator

Now I present a series of Information-Theoretic Propositions from Grünwald 2007 that are essential for our project.

For convenience, we define the extended KL-divergence as:

$$D_P(\mu_1 \parallel \mu_2) = \mathbb{E}_P\Big(-\log P_{\mu_2} - (-\log P_{\mu_1}) \Big) = D(P \parallel P_{\mu_2}) - D(P \parallel \mu_1)$$

Proposition 1.1 (Robustness Property)**.** *Let $\mu_0 \in \Theta_{mean}$ and let P be any distribution with $\mathbb{E}_P\phi(X) = \mu_0$. Then, for all $\mu \in \Theta_{mean}$,*

$$D_P(\mu_0 \parallel \mu) = D(\mu_0 \parallel \mu)$$

Now we are going to use this property specifically for the empirical distribution of the observed data $\underset{\sim}{x} =^t (x_1, \ldots, x_n) \in \mathbb{R}^n$. This distribution is defined as

$$\mathbb{P}(x) = \frac{|\{i \in \{1, \ldots, n\} : x_i = x\}|}{n} = \frac{1}{n}\sum_{j=1}^{n} \delta_{x_j}(x)$$

Proposition 1.2 (The Maximum Likelihood Estimator)**.** *The maximum likelihood estimator satisfies*

$$\hat{\mu}(\underset{\sim}{x}) = \mathbb{E}_{\hat{\mu}(\underset{\sim}{x})}\phi(X) = \frac{1}{n}\sum_{j=1}^{n} \phi(x_j)$$

and it exists whenever $\frac{1}{n}\sum_{j=1}^{n} \phi(x_j) \in \Theta_{mean}$.

Finally, combining these two propositions, we get

Proposition 1.3 (The log-likelihood ratio and the KL divergence)**.** *For any $\mu \in \Theta_{mean}$*

$$\log \frac{f_{\hat{\mu}(\underset{\sim}{x})}(\underset{\sim}{x})}{f_{\mu}(\underset{\sim}{x})} = n\mathbb{E}_{\hat{\mu}(\underset{\sim}{x})} \frac{f_{\hat{\mu}(\underset{\sim}{x})}(X)}{f_{\mu}(X)} = nD(\hat{\mu} \parallel \mu)$$

1.3 Main Example - Application

Now we will combine the properties presented in the sections above. This application was demonstrated by Professor Grunwald. We consider an 1-parameter exponential family with the mean value parametrization $\{f_\mu, \mu \in \Theta_{mean}\}$ and fix some μ_0. We denote $\beta = \beta(\mu)$ for any μ and $\beta_0 = \beta(\mu_0)$. Then we have the following properties:

1. For any $\mu > \mu_0$, and for any $\underset{\sim}{x} \in \mathbb{R}^n$, the log-likelihood ratio is a linear function of the Maximum Likelihood Estimator $\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} \phi(x_j)$

$$\log \frac{f_{\mu}(\underset{\sim}{x})}{f_{\mu_0}(\underset{\sim}{x})} = n(\beta - \beta_0)\hat{\mu} - n(\psi(\beta) - \psi(\beta_0)) = g_{\mu,\mu_0}(\hat{\mu})$$

2. For fixed μ_0, the function $\mu \longmapsto D(f_\mu || f_{\mu_0})$ is a convex function of μ.

3. Due to the robustness property in Section 1.2, we have that:

$$\frac{1}{n}\log \frac{f_{\hat{\mu}}(\underset{\sim}{x})}{f_{\mu_0}(\underset{\sim}{x})} = D(f_{\hat{\mu}} || f_{\mu_0})$$

4. For any $\mu > \mu_0$, we have the following:

$$g_{\mu,\mu_0}(\hat{\mu}) = \log \frac{f_\mu(\underset{\sim}{x})}{f_{\mu_0}(\underset{\sim}{x})} \leq \log \frac{f_{\hat{\mu}}(\underset{\sim}{x})}{f_{\mu_0}(\underset{\sim}{x})} = nD(f_{\hat{\mu}}||f_{\mu_0})$$

The 4^{th} one shows that the linear function $g_{\mu,\mu_0}(\hat{\mu})$ is always under the convex function $nD(f_{\hat{\mu}}||f_{\mu_0})$, and they are equal only when $\hat{\mu} = \mu$. In other words, the function $g_{\mu,\mu_0}(\hat{\mu})$ is the tangent of $nD(f_{\hat{\mu}}||f_{\mu_0})$ at point $\hat{\mu} = \mu$.

I present the example of the Normal Distribution with unknown mean μ and known variance. For simplicity, I choose the variance $\sigma^2 = 1$. Then the family $N(\mu, 1), \mu \in \mathbb{R}$ is an exponential family parametrized according to the mean-value parametrization, since:

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} = \frac{1}{\sqrt{2\pi}} e^{\mu x - \frac{1}{2}\mu^2 - \frac{1}{2}x^2}, x \in \mathbb{R}$$

which means that $\phi(X) = X$ and $\mathbb{E}X = \mu$.

Take $X_j \sim N(\mu, 1)$,$j = 1, \ldots, n$, with $\mu \in \mathbb{R}$, independent random variables. We denote $\underset{\sim}{X} =^t (X_1, \ldots, X_n)$ the vector of n-independent random variables. It is known that, for $\underset{\sim}{X} \in \mathbb{R}^n$, the maximum likelihood estimator for the unknown mean is

$$\hat{\mu} = \hat{\mu}(\underset{\sim}{x}) = \frac{1}{n} \sum_1^n x_j = \bar{X}_n$$

We fix some $\mu_0 \in \mathbb{R}$. For any $\mu > \mu_0$, the log-likelihood ratio is:

$$\log \frac{f_\mu(\underset{\sim}{x})}{f_{\mu_0}(\underset{\sim}{x})} = n\left(\bar{X}_n - \frac{1}{2}(\mu + \mu_0)\right)(\mu - \mu_0) = n\left(\hat{\mu} - \frac{1}{2}(\mu + \mu_0)\right)(\mu - \mu_0)$$

We can observe that for any choice of μ, μ_0, this is a linear function of $\hat{\mu}$:

$$\log \frac{f_\mu(\underset{\sim}{x})}{f_{\mu_0}(\underset{\sim}{x})} = g_{\mu,\mu_0}(\hat{\mu})$$

If we specifically choose $\mu = \hat{\mu}$, the log-likelihood ratio becomes:

$$\log \frac{f_{\hat{\mu}}(\underset{\sim}{x})}{f_{\mu_0}(\underset{\sim}{x})} = n\left(\hat{\mu} - \frac{\mu_0 + \hat{\mu}}{2}\right)(\hat{\mu} - \mu_0) = \frac{n}{2}(\hat{\mu} - \mu_0)^2$$

It is easy to calculate that the Kullback-Leibler divergence in the case of Normal Distribution is

$$D(f_\mu||f_{\mu_0}) = \frac{1}{2}(\mu - \mu_0)^2$$

We observe that, for fixed μ_0, the function $\mu \longmapsto D(f_\mu||f_{\mu_0})$ is a convex function of μ, and

$$\log \frac{f_{\hat{\mu}}(\underset{\sim}{x})}{f_{\mu_0}(\underset{\sim}{x})} = nD(f_{\hat{\mu}}||f_{\mu_0})$$

due to the robustness property.

The 4 properties described above are satisfied. Therefore, we have the image:

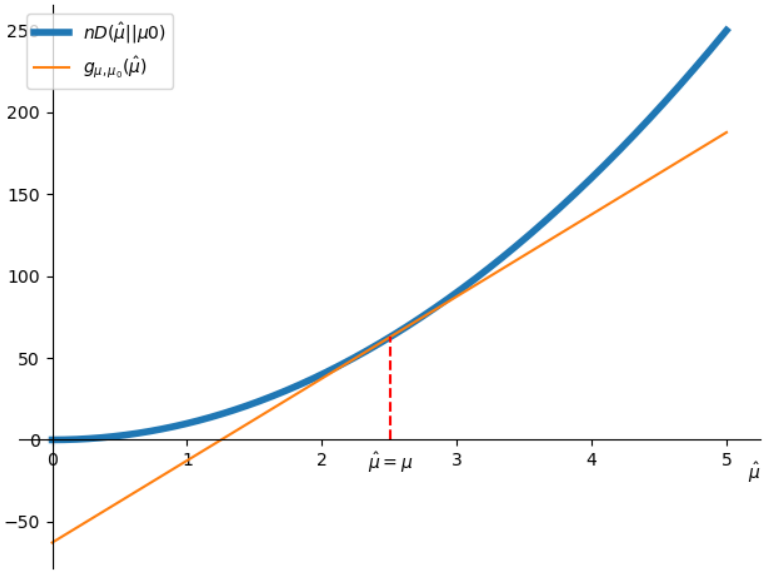


Figure 1: The application on the Normal Distribution

2 The Student Distributions

Like mentioned, the next step would be to investigate whether we can obtain a similar image to Figure 1.1 in the case of the Student Distributions.

We investigate the non-Standardized Student distribution. We say that random variable X follows the Non-Standardized Student distribution with mean θ and ν degrees of freedom if it has a probability density function of the form:

$$f_\theta(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}}\left(1+\frac{1}{\nu}(x-\theta)^2\right)^{-\frac{\nu+1}{2}}, \; x \in \mathbb{R}$$

We denote

$$X \sim Student_\nu(\theta)$$

For any $\nu \in \mathbb{N}$, the family $\{f_\theta(\cdot,\nu), \theta \in \mathbb{R}\}$ is not an exponential family. However, it is also known that for large ν, these distributions are approximating the Normal $N(\theta,1)$ distribution, and so it is natural to believe that they might have a better behaviour than other, non-exponential families.

First, we have to think about the key properties that enable the exponential families to create a setting like this. If we have some sufficiently regular single-parameter distribution with an unknown mean $\theta \in \mathbb{R}$ and if we take $\theta_0 \in \mathbb{R}$, then $D(f_\theta||f_{\theta_0}) = \frac{1}{2}I(\theta)(\theta-\theta_0)^2 + o((\theta-\theta_0)^2)$ Grünwald, 2007 . So in a neighbourhood of θ_0, we expect to have the same quadratic shape.

Also, by the Law of Large Numbers, we have that

$$\frac{1}{n}\log\frac{f_\theta(\underset{\sim}{x})}{f_{\theta_0}(\underset{\sim}{x})} \xrightarrow[n\to+\infty]{} \mathbb{E}_\theta \log\frac{f_\theta}{f_{\theta_0}} = D(f_\theta||f_{\theta_0}) \ , \ p_\theta\text{-a.s.}$$

which means that, for sufficiently large n, the log-likelihood ratio is going to be close to the Kullback Leibler Divergence.

Another thing that is not obvious, is the form of the function $g_{\theta,\theta_0}(\hat{\theta}) = \log\frac{f_\theta(\underset{\sim}{x})}{f_{\theta_o}(\underset{\sim}{x})}$. In the cases that we are going to examine later in this report, the maximum likelihood estimator cannot be analytically computed, and therefore has to be numerically estimated. For that reason, we use Newton's method to estimate the maximum likelihood estimator.

Estimating the Maximum Likelihood Estimator

If we have a family of distributions with densities $\{f_\theta, \theta \in \mathbb{R}\}$ and for $\underset{\sim}{x} \in \mathcal{X}^n$, we denote the log-likelihood function $l(\theta;\underset{\sim}{x}) = f_\theta(\underset{\sim}{x})$, $\theta \in \Theta$, then we employ the algorithm:

$$\theta_{k+1} = \theta_k - \frac{\frac{d}{d\theta}l(\theta;\underset{\sim}{x})}{\frac{d^2}{d\theta^2}l(\theta;\underset{\sim}{x})} \ , \ k \in \mathbb{N}$$

with some arbitrary starting value θ_0. In this case, since we try to estimate the mean, we use $\theta_0 = \bar{x}_n$. In the examples that we produce below, it is not true that $\hat{\theta} = \bar{x}_n$.

2.1 The Student Distribution for fixed ν

We follow the same tactic as the example in Section 1.3. For various values of the MLE we compute the log likelihood ratio $\log\frac{f_{\hat{\theta}}(\underset{\sim}{x})}{f_{\theta_o}(\underset{\sim}{x})}$ as well as the Kullback Leibler divergence $nD(f_{\hat{\theta}}||f_{\theta_0})$ and we make the plots. Again, for simplicity we take the setting of $H_0 : \theta = \theta_0$ vs $\theta > \theta_0$. If we try it for $\nu = 10$ and $n = 20$, the result looks something like this:

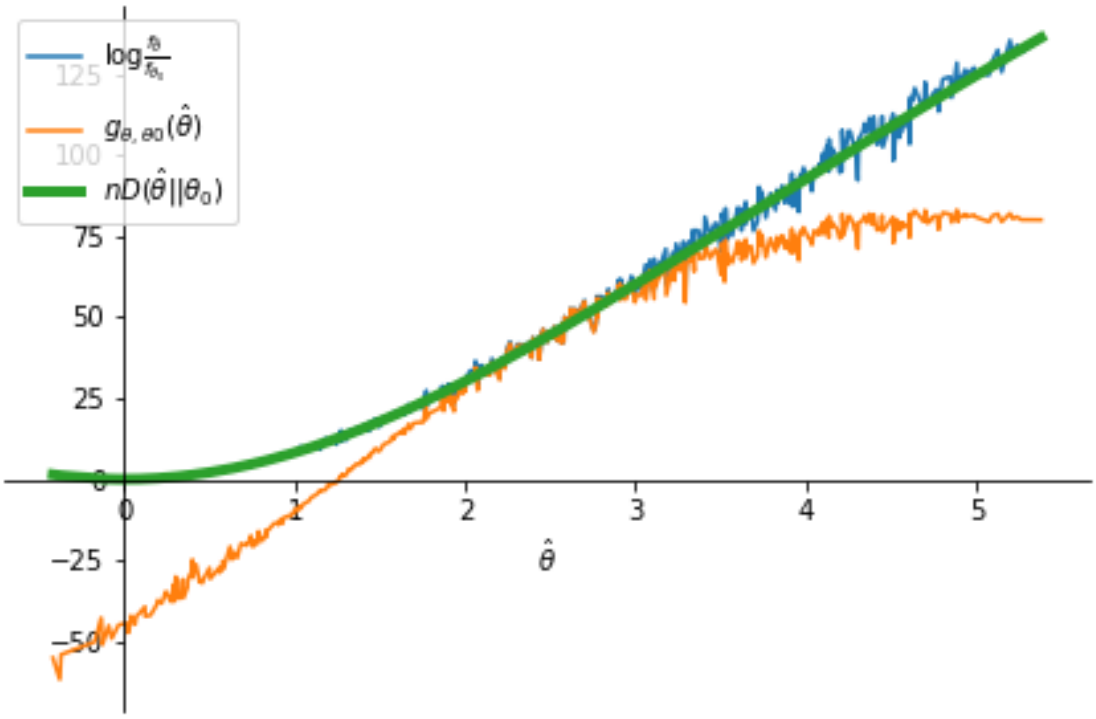


Figure 2: The application in the Student Distribution

Figure 2 is a typical image that we get and there are a few things to observe. First of all, the log-likelihood ratio (blue line) is not exactly equal to the Kullback Leibler (green line). This behaviour is not surprising though, because of the Law of Large Numbers. When we get a larger sample, the blue line starts to match the green one.

Another obvious observation is that the function $g_{\theta,\theta_0}(\hat{\theta})$ is no longer a linear function of $\hat{\theta}$ (orange line). In this case, the only thing that is common to the Normal Example is that it appears that the tangent point is at $\hat{\theta} = \theta$, but this was to be expected, because this is determined by the Maximum Likelihood property (property 4 in Section 1.3), which holds for any kind of distribution.

If we opt for a larger sample, the blue line is going to start matching the green line, but the shape of the orange one is not going to change, since it does not really depend on the sample size, but its noise will start to reduce. For example, if we take $n = 50$ and $n = 100$ we can see these changes.

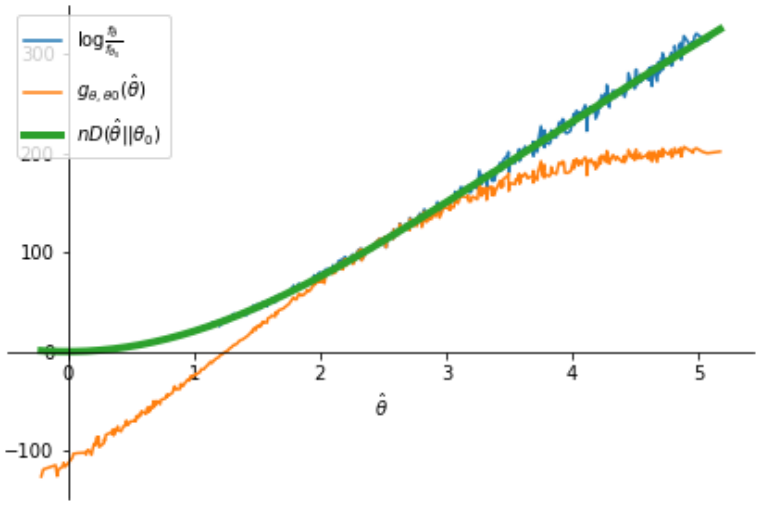


Figure 3: Application for $n = 50$

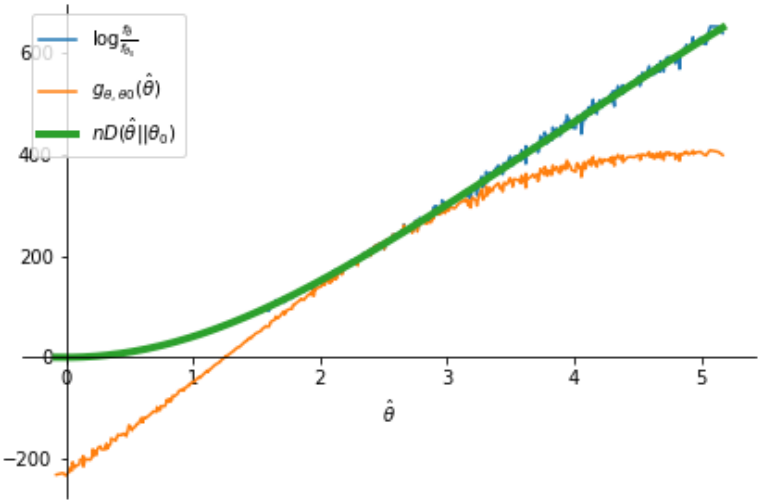


Figure 4: Application for $n = 100$

It is also worth mentioning that the KL divergence $D(f_{\hat{\theta}}||f_{\theta_0})$ of the Student distribution is a convex function of $\hat{\theta}$ only in a region around θ_0. However, one could describe it as quasi-convex when further away from θ_0.

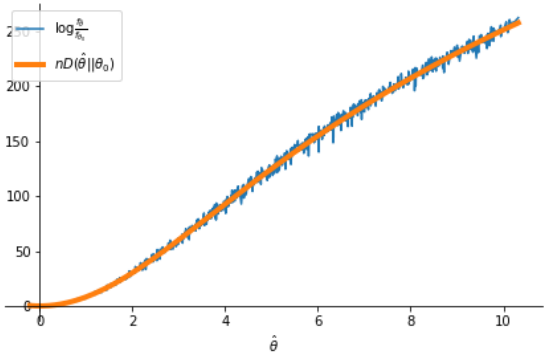
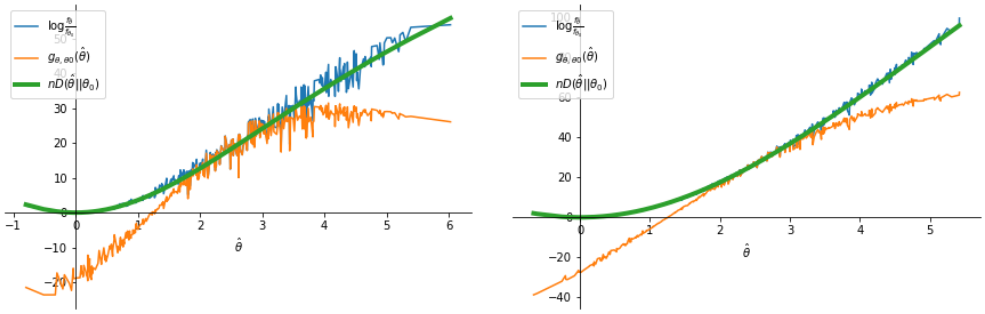


Figure 5: The KL divergence is not a convex function further from θ_0

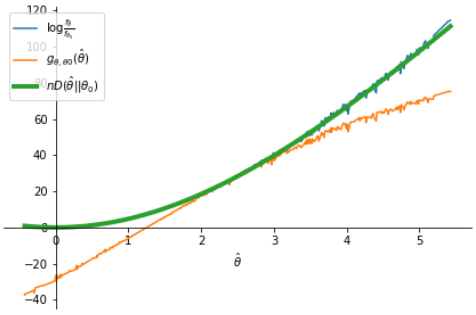
What is more interesting is what happens if ν starts getting larger and larger.

2.2 The Student Distribution for large ν

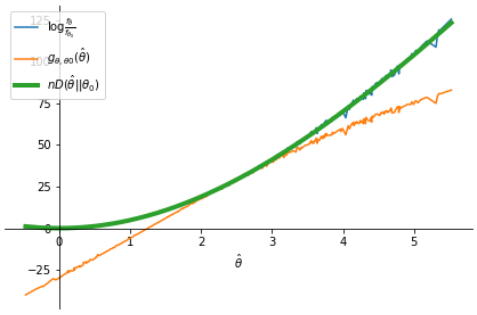
In this section, we are going to see the behaviour of this image as ν becomes bigger and bigger, for fixed sample size n. It is known that $Student_\nu(\theta) \xrightarrow[\nu\to+\infty]{distribution} N(\theta, 1)$. Therefore, we should expect that its behaviour will start to approximate that of the Normals. This means that we expect (a) the log-likelihood ratio to be closer to the KL-divergence, (b) the function $g_{\theta,\theta_0}(\hat{\theta})$ to be a closer to a linear function of $\hat{\theta}$ and (c) the KL divergence to have a shape that is more convex rather than quasi-convex.
Indeed, we produce the images for $n = 10$ and for variable ν.



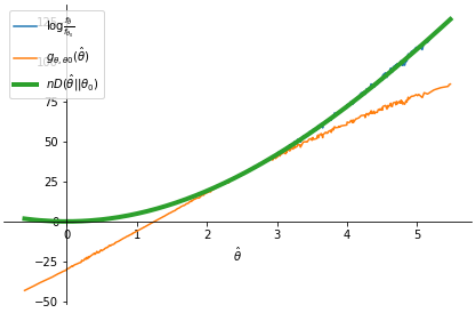
(a) $\nu = 5$ (b) $\nu = 25$



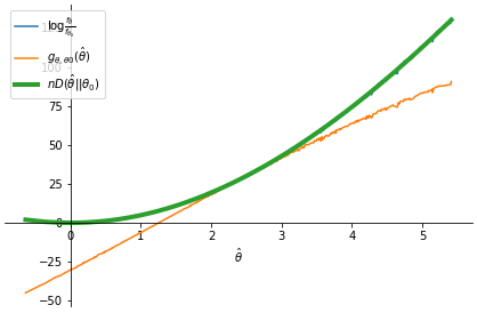
(c) $\nu = 45$



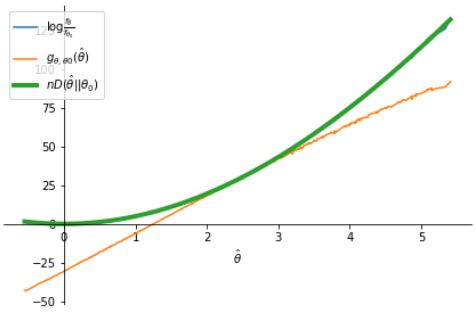
(d) $\nu = 65$



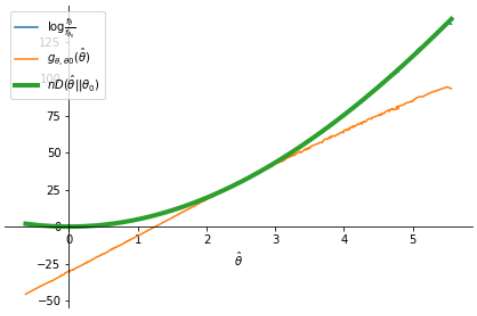
(e) $\nu = 85$



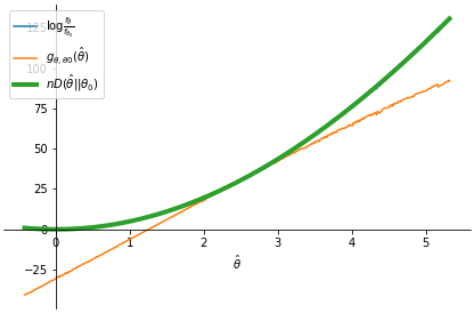
(f) $\nu = 125$



(g) $\nu = 145$



(h) $\nu = 165$



(i) ν~~10~~ 185

Figure 6: The example of the Student Distribution as ν gets larger

3 Scale Mixtures of Normals

So far, we have only mentioned that the Student distribution is expected to have similar behaviour to that of the Normal due to its asymptotic property. However, the Student distribution can be also written as a scale mixture of Normals. In particular,

$$f_T(x;\theta,\nu) = \int_0^{+\infty} f(x;\theta,\sigma^2)W(\sigma^2)d\sigma^2$$

where the $f_T(x;\theta,\nu)$ is the denity of $Student_\nu(\theta)$, $f(x;\theta,\sigma^2)$ is the density of $N(\theta,\sigma^2)$ and $W(\sigma^2)$ is the density of $Inverse-Gamma(a=\frac{\nu}{2}, b=\frac{\nu}{2})$.

For that reason, we could ask if there is any chance of detecting a similar behaviour on other scale mixtures of Normals. We tried the same application for the cases of mixtures with discrete priors, the Logistic Distribution and the Laplace Distribution, but there was no evidence of a behaviour that is better than that of the Student distribution (at least for small ν).

4 Some Observations

In this section, I would like to present some observations and remarks that I produced in this study.

4.1 The Student Distribution Density converges uniformly to the Normal Density

Proposition 4.1. *If $f_\nu(x)$ is the density of $Student_\nu$, $\nu \in \mathbb{N}$ and $f(x)$ is the density of $N(0,1)$, then $f_\nu \xrightarrow[\nu\to+\infty]{} f$ uniformly in $\mathbb{R}$. This means that:*

$$\sup_{x\in\mathbb{R}} \left| f_\nu(x) - f(x) \right| \xrightarrow[\nu\to+\infty]{} 0$$

Proof. Neal 2000 shows in his article that $f_\nu \xrightarrow[\nu\to+\infty]{} f$ uniformly in every interval $[a,b]$. So our job is to ensure the uniform convergence of the tails of the distributions.

Take $\epsilon > 0$

We have that

$$f_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{1}{\nu}x^2\right)^{-\frac{\nu+1}{2}} \ , \ x \in \mathbb{R}$$

and

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \ , \ x \in \mathbb{R}$$

We write $f_\nu(x) = C_\nu h_\nu(x)$ and $f(x) = Ch(x)$ where:

$$C_\nu = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \ , \ h_\nu(x) = \left(1 + \frac{1}{\nu}x^2\right)^{-\frac{\nu+1}{2}}$$

and

$$C = \frac{1}{\sqrt{2\pi}} \ , \ h(x) = e^{-\frac{1}{2}x^2}$$

Also we need the following lemmas.

Lemma 4.1. *For any $a > 0$, the sequence $h_n(a) = \left(1 + \frac{a}{n}\right)^{-\frac{n+1}{2}}$ is decreasing in n.*

Proof. It suffices to show that for any $a > 0$, the sequence $\left(1+\frac{a}{n}\right)^{n+1}$ is increasing in n. We write:

$$\left(1 + \frac{a}{n}\right)^{n+1} =$$

$$\sum_{k=0}^{n+1} \binom{n+1}{k} \left(\frac{a}{n}\right)^k =$$

$$1 + \sum_{k=1}^{n+1} \binom{n+1}{k} \frac{a^k}{n^k} =$$

$$1 + \sum_{k=1}^{n+1} \frac{(n+1)!}{k!(n+1-k)!} \frac{a^k}{n^k} =$$

$$1 + \sum_{k=1}^{n+1} \frac{1}{k!}(n+1-k+1)(n+1-k+2)\cdots(n+1-k+k)\frac{a^k}{n^k} =$$

$$1 + \sum_{k=1}^{n+1} \frac{a^k}{k!} \prod_{j=1}^{k} \frac{n+1-k+j}{n} =$$

$$1 + \sum_{k=1}^{n+1} \frac{a^k}{k!} \left(\prod_{j=1}^{k-3} \frac{n+1-k+j}{n} \right) \cdot \frac{n+1-k+k-2}{n} \cdot \frac{n+1-k+k-1}{n} \cdot \frac{n+1-k+k}{n} =$$

$$1 + \sum_{k=1}^{n+1} \frac{a^k}{k!} \left(\prod_{j=1}^{k-3} \frac{n+1-k+j}{n} \right) \cdot \frac{n-1}{n} \cdot 1 \cdot \frac{n+1}{n} =$$

$$1 + \sum_{k=1}^{n+1} \frac{a^k}{k!} \left(\prod_{j=1}^{k-3} (1 - \frac{k-1-j}{n}) \right) \cdot (1 - \frac{1}{n}) \cdot 1 \cdot (1 + \frac{1}{n}) =$$

$$1 + \sum_{k=1}^{n+1} \frac{a^k}{k!} \left(\prod_{j=1}^{k-3} (1 - \frac{k-1-j}{n}) \right) \cdot (1 - \frac{1}{n^2}) \overset{(1)}{\leq}$$

$$1+\sum_{k=1}^{n+1}\frac{a^k}{k!}\left(\prod_{j=1}^{k-3}(1-\frac{k-1-j}{n+1})\right)\cdot(1-\frac{1}{(n+1)^2})\overset{(2)}{\leq}$$

$$1+\sum_{k=1}^{n+2}\frac{a^k}{k!}\left(\prod_{j=1}^{k-3}(1-\frac{k-1-j}{n+1})\right)\cdot(1-\frac{1}{(n+1)^2})=$$

$$\left(1+\frac{a}{n+1}\right)^{n+2}$$

The inequality (1) is true because we replace n by $n+1$ in the denominator of each fraction. The inequality (2) is true because we simply add a non negative quantity in the sum (meaning the one for $k=n+2$). The final equality is straight forward and we can obtain it just like we did up to inequality (1).

□

Corollary 4.1.1. *For all $x \in \mathbb{R}$, the sequence $h_\nu(x)$, $\nu \in \mathbb{N}$ is a decreasing sequence.*

Lemma 4.2. *For any $\epsilon > 0$, there exists $\Delta > 0$ and $\nu_0 \in \mathbb{N}$ such that*

$$\sup_{|x|\geq\Delta_1}\left|h_\nu(x)-h(x)\right|<\epsilon$$

for all $\nu \geq \nu_0$

Proof. First of all, we observe that $h_\nu(x), h(x)$ are decreasing functions of x in $[0,+\infty)$. Also, $\lim_{x\leftarrow+\infty} h(x)=0$ and so, there exists $\Delta>0$ such that $h(x)<\frac{\epsilon}{4}$ for all $x>\Delta$.

Take $\Delta_1>\Delta$. Then, there exists $\nu_0\in\mathbb{N}$ such that $h_\nu(\Delta_1)-h(\Delta_1)<\frac{\epsilon}{4}$ for all $\nu\geq\nu_0$. That is because of the point-wise convergence of h_ν to h.

Essentially, we have that for any $x\geq\Delta_1$ and any $\nu\geq\nu_0$,

$$|h_\nu(x)-h(x)|\leq h_\nu(x)+h(x)\leq h_{\nu_1}(x)+\frac{\epsilon}{4}\leq h_{\nu_1}(\Delta_1)+\frac{\epsilon}{4}<h(\Delta_1)+\frac{\epsilon}{4}+\frac{\epsilon}{4}<\frac{\epsilon}{4}+\frac{\epsilon}{4}+\frac{\epsilon}{4}=\frac{3\epsilon}{4}<\epsilon$$

So we have shown that, for any $\epsilon>0$, there exists $\Delta_1>0$ and $\nu_0\in\mathbb{N}$ such that

$$\sup_{x\geq\Delta_1}\left|h_\nu(x)-h(x)\right|<\epsilon$$

for all $\nu\geq\nu_0$.

We can easily see that due to the symmetry of the functions, we have that

$$\sup_{|x|\geq\Delta_1}\left|h_\nu(x)-h(x)\right|<\epsilon$$

□

Remark. *This Δ_1 only depends on ϵ and not ν*

Remark. *The key observation is that the limit function has the same limit as each component of the sequence as $|x| \to +\infty$ and the sequence is decreasing in ν.*

Now we continue with the main proof.
We can show uniform convergence of f_ν to f.

Take $\epsilon > 0$.

Neal 2000 says that $f_\nu \xrightarrow[\nu \to +\infty]{} f$ uniformly in every interval $[a, b]$.

Also, he proves that $C_\nu \xrightarrow[\nu \to +\infty]{} C$, and so

1. There exists $M > 0$ such that $C_\nu < M$ for all $\nu \in \mathbb{N}$
2. There exists $\nu_1 \in \mathbb{N}$ such that $|C_\nu - C| < \frac{\epsilon}{3}$ for all $\nu \geq \nu_1$

From Lemma 4.2, we have that there exists $\Delta > 0$ and $\nu_2 \in \mathbb{N}$ such that, for all $\nu \geq \nu_2$,

$$\sup_{|x| \geq \Delta} \left| h_\nu(x) - h(x) \right| < \frac{\epsilon}{3M}$$

Finally, from Neal 2000, there exists $\nu_3 \in \mathbb{N}$ such that, for all $\nu \geq \nu_3$

$$\sup_{|x| \leq \Delta} \left| f_\nu(x) - f(x) \right| < \frac{\epsilon}{3}$$

Now we write:

$$\sup_{x \in \mathbb{R}} \left| f_\nu(x) - f(x) \right| \leq \sup_{|x| \leq \Delta} \left| f_\nu(x) - f(x) \right| + \sup_{|x| \geq \Delta} \left| f_\nu(x) - f(x) \right| =$$

$$\sup_{|x| \leq \Delta} \left| f_\nu(x) - f(x) \right| + \sup_{|x| \geq \Delta} \left| C_\nu h_\nu(x) - C h(x) \right| \leq$$

$$\sup_{|x| \leq \Delta} \left| f_\nu(x) - f(x) \right| + |C_\nu| \sup_{|x| \geq \Delta} \left| h_\nu(x) - h(x) \right| + \sup_{|x| \geq \Delta} |h(x)| \left| C_\nu - C \right| \leq$$

$$\sup_{|x| \leq \Delta} \left| f_\nu(x) - f(x) \right| + M \sup_{|x| \geq \Delta} \left| h_\nu(x) - h(x) \right| + \left| C_\nu - C \right|$$

And now, combining everything above, if we choose $\nu_0 = \max\{\nu_1, \nu_2, \nu_3\}$, we have that for all $\nu \geq \nu_0$:

$$\sup_{x \in \mathbb{R}} \left| f_\nu(x) - f(x) \right| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$$

□

Corollary 4.2.1. *For any $\epsilon > 0$, there exists $\nu_0 \in \mathbb{N}$ such that*

$$\sup_{x\in\mathbb{R}} \left| f_\nu(x;\theta) - f(x;\theta) \right| < \epsilon$$

for all $\nu \geq \nu_0$, for all $\theta \in \mathbb{R}$

Proof. Take $\theta_1, \theta_2 \in \mathbb{R}$. Then, there exists $\nu_0 \in \mathbb{N}$ such that

$$\sup_{x\in\mathbb{R}} \left| f_\nu(x;\theta_1) - f(x;\theta_1) \right| < \epsilon$$

and we have that

$$\sup_{x\in\mathbb{R}} \left| f_\nu(x;\theta_1) - f(x;\theta_1) \right| = \sup_{x\in\mathbb{R}} \left| f_\nu(x-\theta_2+\theta_1;\theta_1) - f(x-\theta_2+\theta_1;\theta_1) \right| = \sup_{x\in\mathbb{R}} \left| f_\nu(x;\theta_2) - f(x;\theta_2) \right|$$

□

Corollary 4.2.2. *For any $k \in \mathbb{N}$, the joint density*

$$f_\nu(\underset{\sim}{x};\theta) = \prod_{j=1}^{k} f_\nu(x_j;\theta) \xrightarrow[\nu\to+\infty]{} \prod_{j=1}^{k} f(x_j;\theta) = f(\underset{\sim}{x};\theta)$$

uniformly in $\mathbb{R}^k$.

Proof. Assume that we have two functions $f_n, g_n, f, g : \mathbb{R} \to \mathbb{R}_{>0}$ such that $f_n \xrightarrow[\nu\to+\infty]{} f$ and $g_n \xrightarrow[\nu\to+\infty]{} g$ uniformly in $\mathbb{R}$. Assume that f, g are bounded. Then, the function sequence $F_n : \mathbb{R}^2 \to \mathbb{R}_{>0}$ with $F_n(x,y) = f_n(x)g_n(y)$ converges uniformly on $\mathbb{R}^2$ to $F : \mathbb{R}^2 \to \mathbb{R}_{>0}$ with $F(x,y) = f(x)g(y)$. This is true because

$$\sup_{x,y\in\mathbb{R}^2} \left| F_n(x,y) - F(x,y) \right| = \sup_{x,y\in\mathbb{R}^2} \left| f_n(x)g_n(y) - f(x)g(y) \right| \leq$$

$$\sup_{x\in\mathbb{R}} |f_n(x)| \sup_{y\in\mathbb{R}} \left| g_n(y) - g(y) \right| + \sup_{y\in\mathbb{R}} |g(y)| \sup_{x\in\mathbb{R}} \left| f_n(x) - f(x) \right|$$

Due to boundedness of f, g and the uniform convergence of f_n to f, there exists $M > 0$ such that $\sup_{x\in\mathbb{R}} |f_n(x)| < M$ for all $n \in \mathbb{N}$ and $\sup_{y\in\mathbb{R}} |g(y)| < M$.

From that, using induction, we can expand the result to finite products.

□

4.2 The Maximum Likelihood Estimator of the Student Distribution

Fix $k \in \mathbb{N}$. Let $\nu \in \mathbb{N}$ and $\underset{\sim}{x} \in \mathbb{R}^k$. We denote $l_\nu(\theta; \underset{\sim}{x}), l(\theta; \underset{\sim}{x}) : \mathbb{R} \to \mathbb{R}_{>0}$ the likelihoods that correspond to the distributions $Student_\nu(\theta)$ and $N(\theta, 1)$. Essentially,

$$l_\nu(\theta; \underset{\sim}{x}) = \prod_{j=1}^{k} f_\nu(x_j; \theta)$$

and

$$l(\theta; \underset{\sim}{x}) = \prod_{j=1}^{k} f(x_j; \theta)$$

For any $\nu \in \mathbb{N}$, we denote $\hat{\theta}_\nu(\underset{\sim}{x}) = \underset{\theta \in \mathbb{R}}{\operatorname{argmax}}\, l_\nu(\theta; \underset{\sim}{x})$.

We consider any sequence of observations $\underset{\sim}{x}_\nu =^t (x_\nu(1), \ldots, x_\nu(k)) \in \mathbb{R}^k$ that come from any distribution.

Proposition 4.2. *For any $\epsilon > 0$, there exists $\nu_0 \in \mathbb{N}$ such that*

$$l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) - l_\nu(\bar{x}_\nu; \underset{\sim}{x}_\nu) < \epsilon$$

for any observations $\underset{\sim}{x}_\nu$, for all $\nu \geq \nu_0$

Proof. From Corollary 3.2.2, there exists $\nu_0 \in \mathbb{N}$ such that, for any $\theta \in \mathbb{R}$,

$$\sup_{\underset{\sim}{x} \in \mathbb{R}^k} \left| f_\nu(\underset{\sim}{x}; \theta) - f(\underset{\sim}{x}; \theta) \right| < \frac{\epsilon}{2}$$

and so,

$$|l_\nu(\theta; \underset{\sim}{x}_\nu) - l(\theta; \underset{\sim}{x}_\nu)| < \frac{\epsilon}{2} \text{ for all } \theta \in \mathbb{R}$$

Then, if we consider the fact that $\bar{x} = \underset{\theta \in \mathbb{R}}{\operatorname{argmax}}\, l(\theta; \underset{\sim}{x})$ for the Normal distribution,

$$l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) < l(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) + \frac{\epsilon}{2} \leq l(\bar{x}_\nu; \underset{\sim}{x}_\nu) + \frac{\epsilon}{2} < l_\nu(\bar{x}_\nu; \underset{\sim}{x}_\nu) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = l_\nu(\bar{x}; \underset{\sim}{x}_\nu) + \epsilon$$

□

Remark. *In the proof, we also showed that for any ϵ, there exists $\nu_0 \in \mathbb{N}$ such that*

$$|l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) - l(\bar{x}_\nu; \underset{\sim}{x}_\nu)| < \epsilon$$

for $\nu \geq \nu_0$, and any observations $\underset{\sim}{x}_\nu$ because

$$l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) < l(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) + \epsilon \leq l(\bar{x}_\nu; \underset{\sim}{x}_\nu) + \epsilon \Rightarrow$$

$$l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) - l(\bar{x}_\nu; \underset{\sim}{x}_\nu) < \epsilon$$

and

$$l(\bar{x}_\nu; \underset{\sim}{x}_\nu) \leq l_\nu(\bar{x}_\nu; \underset{\sim}{x}_\nu) + \epsilon \leq l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) + \epsilon \Rightarrow$$

$$l(\bar{x}_\nu; \underset{\sim}{x}_\nu) - l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) < \epsilon$$

Remark. *This proposition at first seems a bit obvious, but requires the uniform convergence of the Student Distributions. Also, this is true for any sample size. Essentially it says that for $\nu \gg$, the value of the Student likelihood at the sample mean can be very close to the actual maximum. However, this does not provide any information about how close these two estimators are. Most importantly though, this is true for any sequence of observations.*

Problem. *Can we show that for large ν we have $\hat{\theta}_\nu(\underset{\sim}{x}_\nu) \approx \bar{x}_\nu$?*

We try to simulate it by taking multiple samples for multiple θs, and take the $\max_{\underset{\sim}{x}_\nu, \theta} |\hat{\theta}_\nu(\underset{\sim}{x}_\nu) - \bar{x}_\nu|$ as a sequence of ν we can get Figure 7 that shows that indeed those two quantities come asymptotically close, no matter the sample or θ.

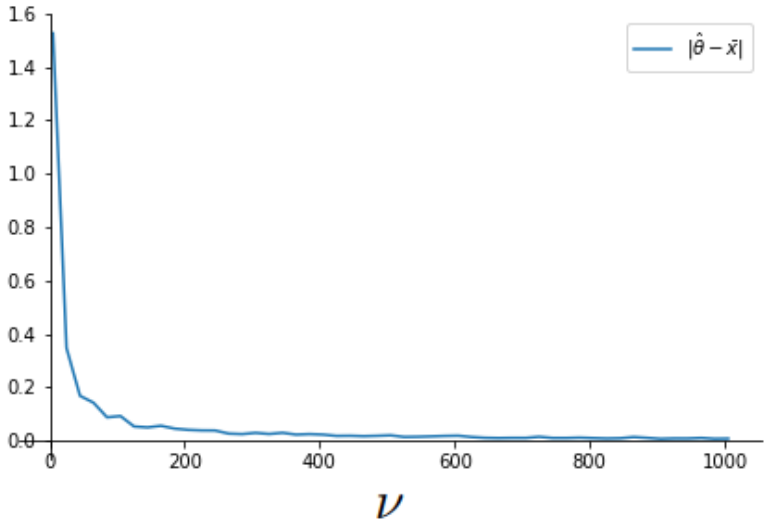


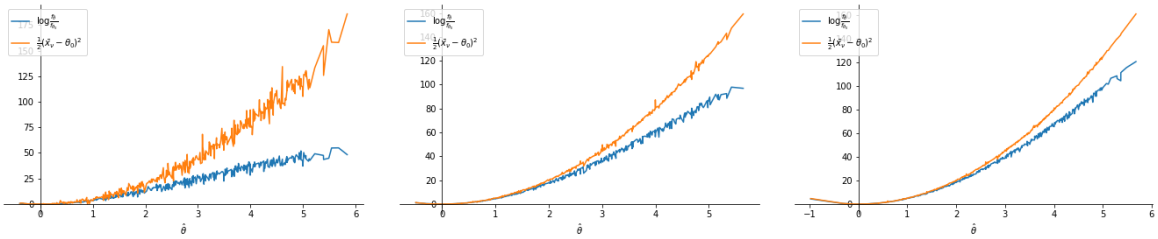
Figure 7: The distance of the Student MLE and the sample mean for large ν

4.3 The Log-Likelihood Ratios

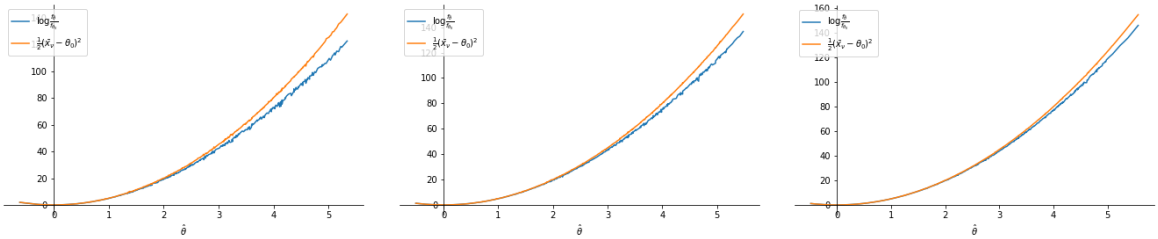
Since for large ν the densities of the student distributions are very close to the normal ones, it is only natural to expect that $\log \frac{l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}); \underset{\sim}{x}_\nu)}{l_\nu(\theta_0; \underset{\sim}{x}_\nu)}$ and $\log \frac{l(\bar{x}_\nu; \underset{\sim}{x}_\nu)}{l(\theta_0; \underset{\sim}{x}_\nu)}$ should be close for large ν. This would mean that

$$\log \frac{l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}); \underset{\sim}{x}_\nu)}{l_\nu(\theta_0; \underset{\sim}{x}_\nu)} \approx \frac{k}{2}(\bar{x}_\nu - \theta_0)^2 \text{ for large } \nu$$

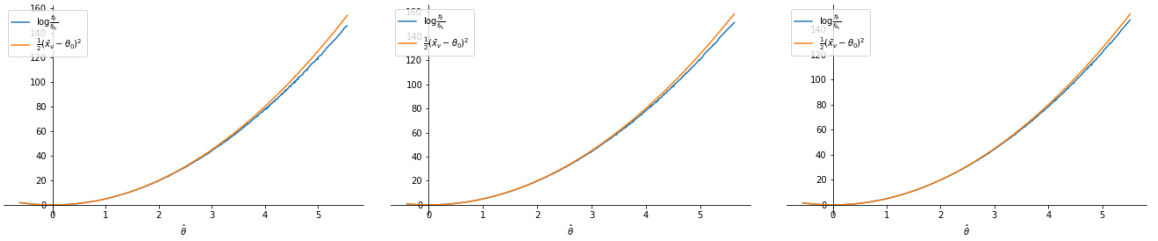
For example, in Figure 8 we see some plots that illustrate this:



(a) $\nu = 5$ (b) $\nu = 25$ (c) $\nu = 45$



(d) $\nu = 85$ (e) $\nu = 145$ (f) $\nu = 245$



(g) $\nu = 305$ (h) $\nu = 405$ (i) $\nu = 505$

Figure 8: Demonstration of $\log \frac{l_\nu(\hat{\theta}_\nu)}{l_\nu(\theta_0)}$, $\frac{k}{2}(\bar{x}_\nu - \theta_0)^2$ for large ν

One would be confident about it due to Remark 1 of Proposition 4.2, but there is a problem. The logarithm is not a uniformly continuous function in $[0, +\infty)$. Therefore, if two quantities are close, we cannot say that the difference of their logarithms are going to be close as well. However, the logarithm is uniformly continuous in $[c, +\infty)$ for any $c > 0$, because its derivative is bounded in $[c, +\infty)$, and so it is Lipsitz-continuous in $[c, +\infty)$, ensuring uniform continuity.

We can write, for any ν and for any observation $\underset{\sim}{x}_\nu$,

$$\left| \log \frac{l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu)}{l_\nu(\theta_0; \underset{\sim}{x}_\nu)} - \log \frac{l(\bar{x}_\nu; \underset{\sim}{x}_\nu)}{l(\theta_0; \underset{\sim}{x}_\nu)} \right| =$$

$$\left|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l_\nu(\theta_0;\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu) + \log l(\theta_0;\underset{\sim}{x}_\nu)\right| \leq$$

$$\left|l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu)\right| + \left|\log l_\nu(\theta_0;\underset{\sim}{x}_\nu) - \log l(\theta_0;\underset{\sim}{x}_\nu)\right|$$

So, if we can somehow ensure that all the quantities in the logarithm become smaller than some threshold $c > 0$ with small probability, we can use the uniform convergence and Remark 1 from Proposition 4.2.

Proposition 4.3. *Take $\theta \geq \theta_0$. Then, for any $\epsilon > 0$ and for any $\delta > 0$, there exists $\nu_0 \in \mathbb{N}$ such that*

$$P_\theta\left(|\log \frac{l_\nu(\hat{\theta}_\nu(\underset{\sim}{x});\underset{\sim}{x}_\nu)}{l_\nu(\theta_0;\underset{\sim}{x}_\nu)} - \frac{k}{2}(\bar{x}_\nu - \theta_0)^2| \geq \delta\right) < \epsilon$$

for all $\nu \geq \nu_0$.

Proof. We are going to need the following lemmas

Lemma 4.3. *Take $\theta \geq \theta_0$. For any $\epsilon > 0$ and for any $\delta > 0$, there exists $\nu_0 \in \mathbb{N}$ such that*

$$P_\theta\left(|\log l_\nu(\theta_0;\underset{\sim}{x}_\nu) - \log l(\theta_0;\underset{\sim}{x}_\nu)| \geq \delta\right) < \epsilon$$

for all $\nu \geq \nu_0$.

Proof. There exists $c > 0$ such that

$$P_\theta\left(x \in \mathbb{R} : f(x;\theta_0) < (\frac{3c}{2})^{\frac{1}{k}}\right) < \frac{\epsilon}{k}$$

This is true because can always find a region that is close to infinity the probability of observing such extreme data is less arbitrary small, and so the densities can become very small.

For this c, we take the set

$$A = \prod_{j=1}^{k}\left\{x_j \in \mathbb{R} : f(x_j;\theta_0) \geq (\frac{3c}{2})^{\frac{1}{k}}\right\} \subset \mathbb{R}^k$$

Then $P_\theta(A^c) < \epsilon$, because:

$$P_\theta(A^c) = P_\theta\left(\bigcup_{j=1}^{k}\left\{x_j \in \mathbb{R} : f(x_j;\theta_0) \leq (\frac{3c}{2})^{\frac{1}{k}}\right\}\right) \leq \sum_{j=1}^{k} P_\theta\left(\left\{x_j \in \mathbb{R} : f(x_j;\theta_0) \leq (\frac{3c}{2})^{\frac{1}{k}}\right\}\right) < k\frac{\epsilon}{k} = \epsilon$$

For $\underset{\sim}{x} \in A$ we have that

$$f(\underset{\sim}{x};\theta_0) = \prod_{j=1}^{k} f(x_j;\theta_0) \geq \frac{3c}{2} \geq c$$

Due to uniform convergence, there exists $\nu_1 \in \mathbb{N}$ such that

$$|f_\nu(\underset{\sim}{x};\theta_0) - f(\underset{\sim}{x};\theta_0)| < \frac{c}{2}$$

for all $\underset{\sim}{x} \in \mathbb{R}^k$ and for all $\nu \geq \nu_1$.

So for $\nu \geq \nu_1$ and for $\underset{\sim}{x} \in A$, we have

$$f_\nu(\underset{\sim}{x};\theta_0) - f(\underset{\sim}{x};\theta_0) < \frac{c}{2} \Rightarrow f_\nu(\underset{\sim}{x};\theta_0) > f(\underset{\sim}{x};\theta_0) - \frac{c}{2} > \frac{3c}{2} - \frac{c}{2} = c$$

So, for all $\underset{\sim}{x} \in A$, $\nu \geq \nu_1$,

$$f_\nu(\underset{\sim}{x};\theta_0), f(\underset{\sim}{x};\theta_0) \geq c$$

Now the logarithm is a uniformly continuous function in $[c, +\infty)$, so there exists $\nu_2 \in \mathbb{N}$ such that

$$|\log f_\nu(\underset{\sim}{x};\theta_0) - \log f(\underset{\sim}{x};\theta_0)| < \delta$$

for all $\underset{\sim}{x} \in A$ and for all $\nu \geq \nu_2$.
We choose $\nu_0 = \max\{\nu_1, \nu_2\}$. Then for $\nu \geq \nu_0$,

$$P_\theta\left(|\log l_\nu(\theta_0;\underset{\sim}{x}_\nu) - \log l(\theta_0;\underset{\sim}{x}_\nu)| \geq \delta\right) =$$

$$P_\theta\left(|\log l_\nu(\theta_0;\underset{\sim}{x}_\nu) - \log l(\theta_0;\underset{\sim}{x}_\nu)| \geq \delta, \underset{\sim}{x}_\nu \in A\right) + P_\theta\left(|\log l_\nu(\theta_0;\underset{\sim}{x}_\nu) - \log l(\theta_0;\underset{\sim}{x}_\nu)| \geq \delta, \underset{\sim}{x}_\nu \in A^c\right) \leq$$

$$0 + P_\theta(A^c) < \epsilon$$

□

Lemma 4.4. *Take $\theta \geq \theta_0$. For any $\epsilon > 0$ and for any $\delta > 0$, there exists $\nu_0 \in \mathbb{N}$ such that*

$$P_\theta\left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu)| \geq \delta\right) < \epsilon$$

for all $\nu \geq \nu_0$.

Proof. We are going to do something similar, but in order to do so, we have to ensure that the estimators can be in a specific interval with small probability.
First of all, it s very easy to show that if we have random variables X_ν, X such that the densities $f_\nu \xrightarrow[\nu\to\infty]{} f$ uniformly in $\mathbb{R}$, then we have that $X_\nu \xrightarrow[\nu\to\infty]{} X$ in distribution. We can use the dominated convergence theorem to prove this.

Now if we have $X_\nu(j), j = 1, \ldots, k$ independent and $X_\nu(j) \xrightarrow[\nu\to\infty]{d} X(j), j = 1 \ldots k$ and $X(j), j = 1 \ldots k$ are independent, then $\sum_{j=1}^k X_\nu(j) \xrightarrow[\nu\to\infty]{d} \sum_{j=1}^k X(j)$. This is very easy to prove using characteristic functions.

Finally, if $X_\nu \xrightarrow[\nu\to\infty]{d} X$, then for $c > 0$ we have $cX_\nu \xrightarrow[\nu\to\infty]{d} cX$. This comes directly from the definition of the convergence in distribution.

Combining these for our case, we have that

$$\frac{1}{k}\sum_{j=1}^{k} X_\nu(j) \xrightarrow[\nu\to\infty]{d} \frac{1}{k}\sum_{j=1}^{k} X(j)$$

where X_j independent following $N(\theta, 1)$ So

$$\frac{1}{k}\sum_{j=1}^{k} X_\nu(j) \xrightarrow[\nu\to\infty]{d} N(\theta, \frac{1}{k})$$

Another important point is that, because the distribution function $F_{N(\theta,\frac{1}{k})}$ is continuous, the convergence of the distribution functions is uniform.

Now we can start controlling the estimators.
Just like before, there exists $\theta_2 > 0$ such that

$$P_\theta(|Y_{\sim N(\theta,\frac{1}{k})}| > \theta_2) < \frac{\epsilon}{4}$$

This θ_2 can be thought as the most extreme values that the estimator $\bar{X}_\nu$ can take.

Because of convergence in distribution

$$\sup_{x\in\mathbb{R}} |P_\theta(\bar{X}_\nu \le x) - P(Y_{\sim N(\theta,\frac{1}{k})} \le x)| \xrightarrow[\nu\to+\infty]{} 0$$

The uniformity in $\mathbb{R}$ is true because the limit distribution function is continuous.

There exists $\nu_1 \in \mathbb{N}$ such that

$$\sup_{x\in\mathbb{R}} |P_\theta(|\bar{X}_\nu| > x) - P(|Y_{\sim N(\theta,\frac{1}{k})}| > x)| < \frac{\epsilon}{4}$$

for all $\nu \ge \nu_1$, and so

$$|P_\theta(|\bar{X}_\nu| > \theta_2) - P(|Y_{\sim N(\theta,\frac{1}{k})}| > \theta_2)| < \frac{\epsilon}{4} \Rightarrow$$

$$P_\theta(|\bar{X}_\nu| > \theta_2) < P(|Y_{\sim N(\theta,\frac{1}{k})}| > \theta_2) + \frac{\epsilon}{4} < \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}$$

Now we continue exactly like the proof of the previous lemma, but we work thinking that the estimator satisfies $-\theta_2 \le \bar{x} \le \theta_2$.

There exists $c > 0$ such that

$$P_\theta\left(x \in \mathbb{R} : f(x;\bar{\theta}) < (\frac{3c}{2})^{\frac{1}{k}}\right) < \frac{\epsilon}{2k}$$

for all $\bar{\theta} \in [-\theta_2, \theta_2]$. Again, these are the most extreme values that the densities of this set of normal distributions can get.
For this c, we take the set

$$A_\theta = \prod_{j=1}^{k} \left\{ x_j \in \mathbb{R} : f(x_j; \bar{\theta}) \geq (\frac{3c}{2})^{\frac{1}{k}} \text{ for all } \bar{\theta} \in [-\theta_2, \theta_2] \right\} \subset \mathbb{R}^k$$

Then $P_\theta(A_\theta^c) < \epsilon$, because:

$$P_{\bar{\theta}}(A_\theta^c) = P_\theta \left(\bigcup_{j=1}^{k} \left\{ x_j \in \mathbb{R} : f(x_j; \theta) < (\frac{3c}{2})^{\frac{1}{k}} \text{ for some } \bar{\theta} \in [-\theta_2, \theta_2] \right\} \right) \leq$$

$$\sum_{j=1}^{k} P_\theta \left(\left\{ x_j \in \mathbb{R} : f(x_j; \bar{\theta}) < (\frac{3c}{2})^{\frac{1}{k}} \text{ for some } \bar{\theta} \in [-\theta_2, \theta_2] \right\} \right) < k \frac{\epsilon}{2k} = \frac{\epsilon}{2}$$

For $\underset{\sim}{x} \in A_\theta$ we have that

$$f(\underset{\sim}{x}; \theta) = \prod_{j=1}^{k} f(x_j; \bar{\theta}) \geq \frac{3c}{2} \geq c \text{ for all } \bar{\theta} \in [-\theta_2, \theta_2]$$

Due to Remark 1 of Proposition 4.2, there exists $\nu_2 \in \mathbb{N}$ such that

$$|f_\nu(\underset{\sim}{x}_\nu; \hat{\theta}_\nu(\underset{\sim}{x}_\nu)) - f(\underset{\sim}{x}; \bar{x}_\nu)| < \frac{c}{2}$$

for all $\underset{\sim}{x} \in \mathbb{R}^k$ and for all $\nu \geq \nu_2$. So, for $\underset{\sim}{x}_\nu \in A_\theta$ and for $\bar{x}_\nu \in [-\theta_2, \theta_2]$,

$$f_\nu(\underset{\sim}{x}_\nu; \hat{\theta}_\nu(\underset{\sim}{x}_\nu)) - f(\underset{\sim}{x}; \bar{x}_\nu) < \frac{c}{2} \Rightarrow f_\nu(\underset{\sim}{x}_\nu; \hat{\theta}_\nu(\underset{\sim}{x}_\nu)) > f(\underset{\sim}{x}_\nu; \bar{x}_\nu) - \frac{c}{2} > \frac{3c}{2} - \frac{c}{2} = c$$

The logarithm is a uniformly continuous function in $[c, +\infty)$, so there exists $\nu_3 \in \mathbb{N}$ such that

$$|\log f_\nu(\underset{\sim}{x}_\nu; \hat{\theta}_\nu(\underset{\sim}{x}_\nu)) - \log f(\underset{\sim}{x}; \bar{x}_\nu)| < \delta$$

for $\bar{x}_\nu \in [-\theta_2, \theta_2]$,for all $\underset{\sim}{x}_\nu \in A$ and for all $\nu \geq \nu_3$.

We choose $\nu_0 = \max\{\nu_1, \nu_2, \nu_3\}$. Then for $\nu \geq \nu_0$,

$$P_\theta \left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu; \underset{\sim}{x}_\nu)| \geq \delta \right) =$$

$$P_\theta \left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu; \underset{\sim}{x}_\nu)| \geq \delta, \{\bar{x}_\nu \in [-\theta_2, \theta_2]\} \right) +$$

$$P_\theta \left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu); \underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu; \underset{\sim}{x}_\nu)| \geq \delta, \{\bar{x}_\nu \notin [-\theta_2, \theta_2]\} \right) \leq$$

$$P_\theta\left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu)| \geq \delta, \{\bar{x}_\nu \in [-\theta_2,\theta_2]\}\right) + P_\theta(|\bar{X}_\nu| > \theta_2) <$$

$$P_\theta\left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu)| \geq \delta, \{\bar{x}_\nu \in [-\theta_2,\theta_2]\}\right) + \frac{\epsilon}{2} =$$

$$P_\theta\left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu)| \geq \delta, \{\bar{x}_\nu \in [-\theta_2,\theta_2]\}, A\right) +$$

$$P_\theta\left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu)| \geq \delta, \{\bar{x}_\nu \in [-\theta_2,\theta_2]\}, A^c\right) + \frac{\epsilon}{2} \leq$$

$$0 + P_\theta(A^c) + \frac{\epsilon}{2} < 0 + \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

□

Now to conclude with the main proof, using the lemmas, there exist $\nu_1, \nu_2 \in \mathbb{N}$ such that

$$P_\theta\left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu)| \geq \frac{\delta}{2}\right) < \frac{\epsilon}{2} \text{ for all } \nu \geq \nu_1$$

and

$$P_\theta\left(|\log l_\nu(\theta_0;\underset{\sim}{x}_\nu) - \log l(\theta_0;\underset{\sim}{x}_\nu)| \geq \frac{\delta}{2}\right) < \frac{\epsilon}{2} \text{ for all } \nu \geq \nu_2$$

So, if we set $\nu_0 = \max\{\nu_1, \nu_2\}$, we have that for $\nu \geq \nu_0$

$$P_\theta\left(|\log \frac{l_\nu(\hat{\theta}_\nu(\underset{\sim}{x});\underset{\sim}{x}_\nu)}{l_\nu(\theta_0;\underset{\sim}{x}_\nu)} - \frac{k}{2}(\bar{x}_\nu - \theta_0)^2| \geq \delta\right) =$$

$$P_\theta\left(\left|\log \frac{l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu)}{l_\nu(\theta_0;\underset{\sim}{x}_\nu)} - \log \frac{l(\bar{x}_\nu;\underset{\sim}{x}_\nu)}{l(\theta_0;\underset{\sim}{x}_\nu)}\right| \geq \delta\right) \overset{(1)}{\leq}$$

$$P_\theta\left(|\log l_\nu(\hat{\theta}_\nu(\underset{\sim}{x}_\nu);\underset{\sim}{x}_\nu) - \log l(\bar{x}_\nu;\underset{\sim}{x}_\nu)| \geq \frac{\delta}{2}\right) + P_\theta\left(\log l_\nu(\theta_0;\underset{\sim}{x}_\nu) - \log l(\theta_0;\underset{\sim}{x}_\nu)| \geq \frac{\delta}{2}\right) <$$

$$\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Inequality (1) is true because, if h, g any functions, then $\{|f+g| \geq \delta\} \subset \{|f| \geq \frac{\delta}{2}\} \cup \{|g| \geq \frac{\delta}{2}\}$. This is simply because, $x \in \{|f+g| \geq \delta\}$ and $|f(x)|, |g(x)| < \frac{\delta}{2}$, then

$$\delta \leq |f(x) + g(x)| \leq |f(x)| + |g(x)| < \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

which is a contradiction. □

5 Software Used

In the scope of this project, all the calculations and plots where done using Python. In particular, for the simulation of random variable data we used the scipy.stats package.

References

Grünwald P.D. 2007. *The Minimum Description Length Principle.* MIT Press.

Neal D.K.,2000. *Uniform Convergence of the T-Distributions.* MISSOURI JOURNAL OF MATHEMATICAL SCIENCES.

END OF REPORT