JMLR: Workshop and Conference Proceedings 29:1–12, 2013 ACML 2013

Minimum Message Length Inference of Laplace Distribution

Author Name1 ABC@SAMPLE.COM
Address 1

Author Name2 XYZ@SAMPLE.COM
Address 2

Editor: Cheng Soon Ong and Tu Bao Ho

Abstract

This paper aims at bringing out the merits of using the Laplace distribution over the Normal distribution. The choice of the best distribution is objectively made using the minimum message length (MML) principle. The message length for transmitting data using a Laplace distribution is derived and its parameters are estimated. This method of transmission is compared with that of transmistting the data using the Normal distribution. This is explored in the context of superposition of protein structures. The optimal superposition of protein structures (described using a Normal distribution) minimizing the L2 norm is computed using the Kearsley's method and the superposition minimizing the L1 norm (described using a Laplace model) is approximated using Monte Carlo simulation. These two are compared with respect to their model complexity and the overall fit to the data using MML.

Keywords: Laplace, Normal, MML, Kearsley, Monte Carlo simulation

1. Introduction

Normal distribution is widely used in modelling a set of data whose true distribution is unknown. In many problems, the objective function is formulated as a sum of squares, (the L2 norm) and this function is minimized or maximized depending on the application. Normal distribution has a huge impact on the cost function because of the squared nature of the individual terms. If there are outliers in the dataset, the final inference might be skewed to accommodate the outliers in the model description. The Laplace distribution, however, is robust to outliers as the objective function involves the sum of the absolute values of the difference of the individual terms (L1 norm). The choice of selecting a Laplace over Normal is investigated in this paper. This selection is made by formulating the objective function using Minimum Message Length (MML). The distribution which results in the best compression of data is chosen to be the best model.

Normal distribution is expressed in terms of squared difference from the mean μ (1),

$$\text{pdf}(x) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

and hence objective functions based on minimizing the total least squares result in a closed form analytical solution. On the contrary, because of the mathematical nature of the Laplace

(2) which is expressed as the absolute difference from the mean, does not offer a closed form solution.

$$\mathrm{pdf}(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{2}$$

That being said, we however, believe that it has potential benefits. Laplace distribution is a model of choice in areas as diverse as signal processing (Eltoft et al., 2006), image denoising (Rabbani et al., 2006), gene expression studies (Bhowmick et al., 2006), market risk prediction (Haas et al., 2005), and machine learning (Cord et al., 2006). In most of these applications, a mixture model of Laplace distributions is used.

The main contribution of this work is the derivation of the message length expression and estimation of the MML parameters for the Laplace distribution. The general procedure to formulate the message length expression for transmitting data using some statistical model is outlined in Wallace and Freeman (1987). The MML method of estimating parameters for a number of distributions has been well established (Wallace, 2005). This general approach is followed in the derivation of the message length expression for the Laplace distribution.

We demonstrate its use in two cases. The first scenario involves data being randomly generated from an underlying Laplace distribution. This data is then fitted using both Normal and Laplace models. We use the derived formulation of MML for modelling the data using Laplace distribution. It is observed that the compression in message length is better when compared to transmitting it over a Normal distribution. Since the original distribution from which the data was generated happens to be Laplace, it is reasonable to guess that a Laplace model fits better. In general, if the data is generated from an underlying true distribution, modelling that data using the same statistical model results in a better fit. This is done as a validation check to ensure that the MML formulation for the Laplace case is consistent with this observation.

Further, we demonstrate its use by applying to the problem of optimal superposition. Examples of protein structures are considered and they are superposed using Kearsley's transformation. This results in an orientation of the proteins which minimizes the total least squares deviations. This optimal superposition is then encoded using a Normal distribution. A related superposition which minimizes the total absolute value of the deviations is then computed using a Monte Carlo simulation. This optimal superposition is encoded using a Laplace distribution.

2. The Minimum Message Length (MML) Framework

2.1. Inductive Inference

Wallace and Boulton (1968) developed the first practical criterion for model selection using information theory. MML provides an elegant framework to compare any two competing hypotheses that model some observed data. The hypothesis that results in the shortest overall message length is chosen as the best one, in line with traditional statistical inference using the Bayesian method.[1]

Using Bayes' theorem to explain some observed data D by hypothesis H, we get:

$$\Pr(H\&D) = \Pr(H) \times \Pr(D|H) = \Pr(D) \times \Pr(H|D)$$

where $\Pr(H\&D)$ is the joint probability of data D and hypothesis H. $\Pr(H)$ is the prior probability of hypothesis H, $\Pr(D)$ is the prior probability of probability of data D, $\Pr(H|D)$ is the posterior probability of H given D, and $\Pr(D|H)$ is the likelihood. MML uses the following result from information theory: given an event E with a probability $\Pr(E)$, the message length $I(E)$ for an optimal code is given by $I(E) = -\log_2(\Pr(E))$ bits (Shannon, 1948). Applying this insight to the Bayes' theorem, we get the following relationship between conditional probabilities in terms of optimal message lengths:

$$I(H\&D) = I(H) + I(D|H) = I(D) + I(H|D)$$

In the traditional Bayesian framework, the hypothesis H with the largest posterior probability $\Pr(H|D)$ is often preferred. Among the terms in the above equation, $\Pr(H)$ (and hence $I(H)$) can usually be estimated well for *some* reasonable prior(s) on hypotheses. Given the data D and a chosen prior H, the likelihood $\Pr(D|H)$ can also be estimated. Whilst comparing two competing hypotheses, the prior of observed data $\Pr(D)$ can be ignored as it is a common factor. Hence, for two competing hypotheses, H and H', we have:

$$I(H|D) - I(H'|D) = I(H) + I(D|H) - I(H') - I(D|H')$$

The message, therefore, comprises of two parts:

1. Statement of the hypothesis H (given by $I(H)$)
2. Statement of the data D using the hypothesis (given by $I(D|H)$)

2.2. Parameter Estimation using MML

The hypothesis is a statistical model which is characterized by a set of parameters. MML treats the parameters and data as entities which need to be passed on as information by a transmitter to a receiver. There is a message length associated with encoding both the parameters and the data. The main difference between MML and other Bayesian methods of inference is in the importance they give to the parameters of the model. As an example, most methods based on manimum likelihood (ML) ignore the prior distribution of the parameters. In the MML approach, both the parameters and the data are stated to some precision. The message length varies with the precision. If the parameters are stated more accurately than required, the message length might be longer although this might lead to a better fit to the data. MML works by identifying the precision to which these parameters need to be stated. Wallace and Freeman (1987) provide an approximation to infer these precision values which balance the tradeoff of precision of parameters and the fit to the data. Such an elegant way allows us to compare two hypotheses which model the same data but using different sets of parameters.

1. http://allisons.org/ll/MML/

3. Message Lengths of Normal & Laplace distributions

Wallace and Freeman (1987) derived the code length of the two part message as

$$\begin{aligned} I(\bar{\theta}, D) &= I(\bar{\theta}) + I(D|\bar{\theta}) \\ &\approx \underbrace{\frac{d}{2}\log\kappa_d - \log h(\bar{\theta}) + \frac{1}{2}\log(det\, F(\bar{\theta}))}_{\text{part1}} + \underbrace{L(\bar{\theta}) + \frac{d}{2}}_{\text{part2}} \end{aligned} \tag{3}$$

where $\bar{\theta}$ is the set of model parameters, d is the number of parameters, κ_d is the d-dimensional lattice quantization constant (Conway and Sloane, 1984), $h(\bar{\theta})$ is the prior probability of the parameters, $det(\mathrm{F}(\bar{\theta}))$ is the determinant of the expected Fisher matrix, and $L(\bar{\theta})$ is the negative log likelihood of observed data. The MML estimates $\hat{\bar{\theta}}_{\text{MML}}$ of the parameters are determined by minimizing (3).

3.1. Normal distribution

The parameters describing the Normal distribution (1) are the mean μ and the standard deviation σ. Let $D = \{x_1, x_2, \ldots, x_N\}$ be the observed data containing N samples, ϵ be the precision to which each datum is stated. Let R_μ, R_σ be the range of μ and $\log\sigma$ respectively. ϵ, R_μ, R_σ are hyperparameters which are introduced in Wallace (2005). The derivation of the MML estimates for a Normal distribution is presented in Wallace (2005). The MML estimates are:

$$\begin{aligned} \hat{\mu}_{\text{MML}} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\ \hat{\sigma}_{\text{MML}} &= \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu}_{\text{MML}})^2} \end{aligned}$$

The corresponding minimized message length is given as

$$\begin{aligned} \therefore I_{min} &= I(\hat{\mu}_{\text{MML}}, \hat{\sigma}_{\text{MML}}) \\ &= 1 + \log\kappa_2 + \log(\mathrm{R}_\mu \mathrm{R}_\sigma) + \frac{1}{2}\log(2N^2) \\ &\quad + \frac{N}{2}\log\left(\frac{2\pi}{\epsilon^2}\right) + \frac{N-1}{2}\log\left(\frac{\sum_{n=1}^{N}(x_n - \hat{\mu}_{\text{MML}})^2}{N-1}\right) + \frac{N-1}{2} \end{aligned} \tag{4}$$

3.2. Laplace distribution

The novelty of this paper lies in the derivation of the MML estimates of the parameters of the Laplace distribution. The parameters describing a Laplace distribution are the μ and b. μ is the mean of the distribution and b is the scale parameter. The probability density function is given in (2). To derive these estimates, we use the Wallace-Freeman approximation (Wallace and Freeman, 1987). This requires:

- a likelihood function

- the Fisher information matrix
- prior distributions on the parameters

Using (2), the *likelihood function* is

$$f(D|\bar{\theta}) = \prod_{n=1}^{N} \frac{\epsilon}{2b} e^{-\frac{|x_n - \mu|}{b}}$$

and hence the *negative log-likelihood* is computed as

$$\begin{aligned} L(\bar{\theta}) &= -\log f(D|\bar{\theta}) \\ &= N \log\left(\frac{2}{\epsilon}\right) + N \log b + \frac{1}{b}\sum_{n=1}^{N} |x_n - \mu| \end{aligned} \tag{5}$$

The maximum likelihood (ML) estimates for μ and b are given by

$$\begin{aligned} \hat{\mu}_{\mathrm{ML}} &= \mathrm{median}\{x_n\} \\ \hat{b}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} |x_n - \hat{\mu}_{\mathrm{ML}}| \end{aligned}$$

Computation of Fisher information $\mathbf{F}(\bar{\boldsymbol{\theta}})$

The Fisher information matrix is given by

$$\mathrm{F}(\mu, b) = \begin{pmatrix} \mathrm{E}\left[\frac{\partial^2 L}{\partial \mu^2}\right] & \mathrm{E}\left[\frac{\partial^2 L}{\partial \mu \partial b}\right] \\ \mathrm{E}\left[\frac{\partial^2 L}{\partial b \partial \mu}\right] & \mathrm{E}\left[\frac{\partial^2 L}{\partial b^2}\right] \end{pmatrix}$$

where E[.] is the expected value of that quantity. Using (5),

$$\begin{aligned} \frac{\partial^2 L}{\partial b^2} &= -\frac{N}{b^2} + \frac{2}{b^3}\sum_{n=1}^{N} |x_n - \mu| \\ \mathrm{E}\left[\frac{\partial^2 L}{\partial b^2}\right] &= -\frac{N}{b^2} + \frac{2}{b^3}\mathrm{E}\left[\sum_{n=1}^{N} |x_n - \mu|\right] \end{aligned}$$

$$\begin{aligned} \mathrm{E}\left[|x - \mu|\right] &= \int_{-\infty}^{\infty} |x - \mu| . \frac{1}{2b} . e^{-\frac{|x-\mu|}{b}} dx \\ &= \frac{1}{2b} \int_{-\infty}^{\mu} -(x - \mu) e^{\frac{(x-\mu)}{b}} dx + \int_{\mu}^{\infty} (x - \mu) e^{-\frac{(x-\mu)}{b}} dx \\ &= b \end{aligned}$$

$$\therefore \mathrm{E}\left[\frac{\partial^2 L}{\partial b^2}\right] = -\frac{N}{b^2} + \frac{2}{b^3}(Nb) = \frac{N}{b^2}$$

The computation of $\mathrm{E}\left[\frac{\partial^2 L}{\partial \mu^2}\right]$ and $\mathrm{E}\left[\frac{\partial^2 L}{\partial \mu \partial b}\right]$ is involved and hence, tucked away in section A. Using those results, we have

$$\mathrm{F}(\mu, b) = \begin{pmatrix} \frac{N}{b^2} & 0 \\ 0 & \frac{N}{b^2} \end{pmatrix}$$

$$\therefore det(\mathrm{F}(\mu, b)) = \frac{N^2}{b^4} \tag{6}$$

Priors on the parameters

A prior probability on μ and b is assumed in accordance with the prior assumed in the case of Normal. The ranges from which μ and $\log b$ are drawn are prespecified as R_μ and R_b respectively (Wallace, 2005).

$$\begin{aligned} \therefore h(\bar{\theta}) = h(\mu, b) &= h(\mu)h(b) \\ &= \frac{1}{\mathrm{R}_\mu}.\frac{1}{b\mathrm{R}_b} \end{aligned} \tag{7}$$

Using (3), (6), (7),

$$I(\mu, b) = (\log \kappa_2 + \log(\mathrm{R}_\mu \mathrm{R}_\sigma) + \log N - \log b) + \left(N \log\left(\frac{2}{\epsilon}\right) + N \log b + \frac{1}{b}\sum_{n=1}^{N} |x_n - \mu| + 1\right)$$

To obtain the MML estimates $\hat{\mu}_{\mathrm{MML}}$ and $\hat{b}_{\mathrm{MML}}$ which results in minimum I, $\frac{\partial I}{\partial \mu} = 0$ and $\frac{\partial I}{\partial b} = 0$. The MML estimates are therefore, given by

$$\begin{aligned} \hat{\mu}_{\mathrm{MML}} &= \mathrm{median}\{x_n\} \\ \hat{b}_{\mathrm{MML}} &= \frac{1}{N-1}\sum_{n=1}^{N} |x_n - \hat{\mu}_{\mathrm{MML}}| \end{aligned} \tag{8}$$

The corresponding minimized message length is given as

$$\begin{aligned} \therefore I_{min} &= I(\hat{\mu}_{\mathrm{MML}}, \hat{b}_{\mathrm{MML}}) \\ &= 1 + \log \kappa_2 + \log(\mathrm{R}_\mu \mathrm{R}_\sigma) + \log N + N \log\left(\frac{2}{\epsilon}\right) \\ &\quad + (N-1)\log\left(\frac{\sum_{n=1}^{N} |x_n - \hat{\mu}_{\mathrm{MML}}|}{N-1}\right) + (N-1) \end{aligned} \tag{9}$$

4. Experiments

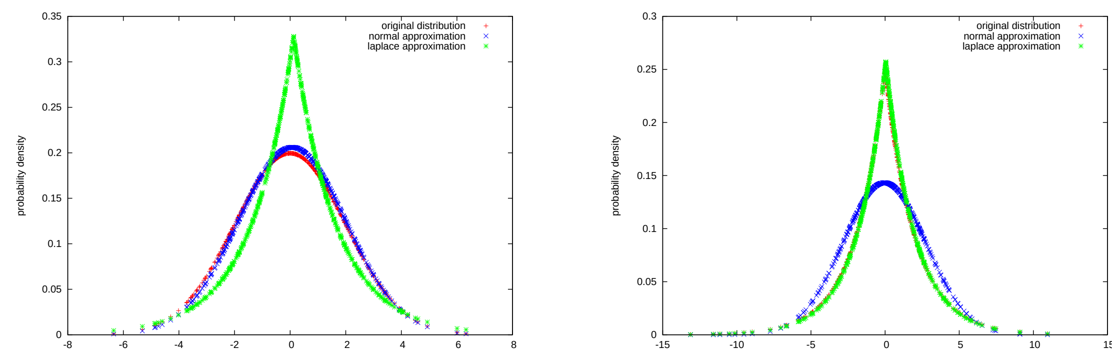
We demonstrate the use of Laplace estimates in two scenarios.

4.1. Data generation and modelling

In the first case, data is generated randomly from a distribution (Normal & Laplace) separately. This data is then modelled using the two distributions. It is observed that if the true distribution is a Laplace, then it is better modelled by a Laplace. (9) is used to determine the code length when the data is modelled using the Laplace distribution. (4) is used to encode the data using a Normal distribution.

As an example, 500 random samples are generated from each of the distributions. The mean of the true distribution is taken to be 0 and the spread (standard deviation σ for a Normal and scale parameter b for a Laplace) is chosen to be 2. Figure 1 shows the original distributions and the corresponding Normal and Laplace approximations. In 1(a), the true distribution is normal (red curve). The Normal approximation (blue curve) overlaps almost entirely with the red curve which is an indication of a good fit. The Laplace approximation (green curve) significantly deviates from the original distribution. The same argument holds for 1(b) where the underlying distribution of the data is Laplace, and hence, in this case, the Laplace seems to be a good fit.

Table 1 provides a comparison of the estimates of the two distributions. The message length (msglen) is computed in bits. It can be seen when the true distribution is Laplace, the message length corresponding to the Laplace estimate (6690.91 bits) is smaller compared to that of the Normal estimate (6755.68 bits).



(a) Original distribution: Normal

(b) Original distribution: Laplace

Figure 1: Approximation of data using Normal & Laplace distributions

Table 1: Comparison of the estimates

True distribution	True mean	True spread	Normal estimates			Laplace estimates		
			mean	spread	msglen	mean	spread	msglen
Normal	0	2	0.0722611	1.93638	6493.89	0.127191	1.52389	6518
Laplace	0	2	-0.0455104	2.78563	6755.68	0.0471194	1.9376	6690.91

4.2. Superposition of protein structures

Given any two vector sets $U = \{u_1, u_2, \ldots, u_m\}$ and $V = \{v_1, v_2, \ldots, v_m\}$ where each u_i and v_i, $(i \in \{1, 2, \ldots, m\})$ is a vector in 3D space, the superpositioning problem refers to finding a suitable transformation on V to align it with U such that the sum of squares of the deviations of each corresponding vector is minimum. If a transformation on V results in an altered vector set $V' = \{v'_1, v'_2, \ldots, v'_m\}$, the objective function corresponding to the sum of squares (L2 norm) of all deviations:

$$\sum_{i=1}^{m} \|v'_i - u_i\|^2 \tag{10}$$

where $\|.\|$ denotes the vector norm, needs to be minimized. Kearsley (1989) provides a solution to this by resolving the transformation into translation and rotation. The centres of mass of the two vector sets are translated to the origin and the problem then reduces to finding the rotation matrix which minimizes the total least squares. This is done by representing the rotation matrix using a quaternion and then solving the resultant eigen value decomposition problem. As such, Kearsley (1989) offers an analytical way to solve the *least squares* superposition problem. One specific application of this is in the superposition of two protein structures.

If we consider only the sum of absolute deviations (L1 norm), the objective function that needs to be minimized would then be

$$\sum_{i=1}^{m} \|v'_i - u_i\| \tag{11}$$

The superposition problem can be formulated in the MML framework as finding the orientation of two proteins such that the deviations of each corresponding point are communicated in an effective manner. Superposition based on minimizing total least squares correspond to stating the deviations using a Normal distribution. Superposition based on minimizing the absolute value of the deviations correspond to transmitting the deviations using a Laplace distribution. Keynes (1911) showed that the Laplace distribution minimized the absolute deviation from the median (which is also corroborated by the MML estimate of Laplace parameters (8)) and is, hence, pertinent for our current discussion.

Minimizing (11) does not yield a closed form solution. As such, one needs to adopt approximate methods to find the best superposition. The one used in this paper is a version that uses Monte Carlo simulation. It is described below:

1. Apply Kearsley's transformation and find the superposition that corresponds to least sum of squares of the deviations. In this state, the value of the objective function (11) is computed.

2. From this orientation, the protein is perturbed randomly. If the new orientation results in a better value of (11), the new orientation is accepted. If however, the value of the objective function is less than the previous value, the new orientation is accepted with a small probability.

3. This is repeated for many iterations. The process is expected to converge to the global minimum. As such, this would correspond to the optimal superposition which minimizes the sum of absolute deviations.

Results

In our experiment, the initial least squares superposition using Kearsley's method is done using SUPER (Collier et al., 2012).

TODO

Acknowledgments

Acknowledgements should go at the end, before appendices and references.

References

Debjani Bhowmick, AC Davison, Darlene R. Goldstein, and Yann Ruffieux. A laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics*, 7(4):630–641, 2006.

James H. Collier, Arthur M. Lesk, Maria Garcia de la Banda, and Arun Siddharth Konagurthu. Super: a web server to rapidly screen superposable oligopeptide fragments from the protein data bank. *Nucleic Acids Research*, 40(Web-Server-Issue):334–339, 2012.

J. H. Conway and N. J. A. Sloane. On the voronoi regions of certain lattices. *SIAM Journal on Algebraic and Discrete Methods*, 5:294–305, 1984.

Aurélien Cord, Christophe Ambroise, and Jean-Pierre Cocquerez. Feature selection in robust clustering based on laplace mixture. *Pattern Recogn. Lett.*, 27(6):627–635, April 2006.

T. Eltoft, Taesu Kim, and Te-Won Lee. On the multivariate laplace distribution. *Signal Processing Letters, IEEE*, 13(5), 2006.

Markus Haas, Stefan Mittnik, and Marc Paolella. Modeling and predicting market risk with laplace-gaussian mixture distributions. 2005.

Simon K. Kearsley. On the orthogonal transformation used for structural comparisons. *Acta. Cryst.*, A45:208–210, 1989.

J. M. Keynes. The principal averages and the laws of error which lead to them. *Journal of the Royal Statistical Society*, 74(3):pp. 322–331, 1911.

H. Rabbani, M. Vafadust, and S. Gazor. Image denoising based on a mixture of laplace distributions with local parameters in complex wavelet domain. In *Image Processing, 2006 IEEE International Conference on*, pages 2597–2600, 2006.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–, july, october 1948.

C. S Wallace. *Statistical and Inductive Inference using Minimum Message Length.* Information Science and Statistics. SpringerVerlag, 2005.

C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.

C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(3):pp. 240–265, 1987.

Appendix A. Derivations involved in the computation of Laplace Fisher

$$\frac{\partial L}{\partial \mu} = -\frac{1}{b}\sum_{n=1}^{N}\frac{(x_n-\mu)}{|x_n-\mu|} \qquad \left(\text{using}\frac{d}{dx}|x| = \frac{d}{dx}\sqrt{x^2} = \frac{x}{|x|}\right)$$

This is discontinuous as it is piecewise constant. Hence to calculate $\frac{\partial^2 L}{\partial \mu^2}$, the following approach is adopted:

Assume that the actual distribution has parameters m and b. The receiver, however, decodes the mean as μ to an accuracy of parameter value δ. As such μ is a random variable and it is fair to reason out that $\mu \in \left[m-\frac{\delta}{2}, m+\frac{\delta}{2}\right]$. I˙1t is assumed that μ follows a uniform distribution in this range. Using this assumption, now we compute the $\mathrm{E}\left[\frac{\partial L}{\partial \mu}\right]$ and subsequent calculations. From our assumptions, $\mathrm{pdf}(\mu) = \frac{1}{\delta}$.

$$\therefore \frac{\partial L}{\partial \mu} \approx \mathrm{E}\left[\frac{\partial L}{\partial \mu}\right] = -\frac{1}{b}\mathrm{E}\left[\sum_{n=1}^{N}\frac{x_n-\mu}{|x_n-\mu|}\right]$$

$$\begin{aligned}\mathrm{E}\left[\frac{x-\mu}{|x-\mu|}\right] &= \int_{-\infty}^{\infty}\frac{x-\mu}{|x-\mu|}.\frac{1}{2b}.e^{\frac{|x-m|}{b}}dx\\ &= \int_{-\infty}^{\mu}-\frac{1}{2b}e^{-\frac{|x-m|}{b}}dx + \int_{\mu}^{\infty}\frac{1}{2b}e^{-\frac{|x-m|}{b}}dx\end{aligned}$$

(i) Let $\mu < m$

$$\begin{aligned}\therefore \mathrm{E}\left[\frac{x-\mu}{|x-\mu|}\right] &= \int_{-\infty}^{\mu}-\frac{1}{2b}e^{\frac{x-m}{b}}dx + \int_{\mu}^{m}\frac{1}{2b}e^{\frac{x-m}{b}}dx + \int_{m}^{\infty}\frac{1}{2b}e^{-\frac{x-m}{b}}dx\\ &= 1-e^{\frac{\mu-m}{b}}\end{aligned}$$

(ii) Let $\mu > m$

$$\begin{aligned}\therefore \mathrm{E}\left[\frac{x-\mu}{|x-\mu|}\right] &= \int_{-\infty}^{m}-\frac{1}{2b}e^{\frac{x-m}{b}}dx + \int_{m}^{\mu}-\frac{1}{2b}e^{-\frac{x-m}{b}}dx + \int_{\mu}^{\infty}\frac{1}{2b}e^{-\frac{x-m}{b}}dx\\ &= -(1-e^{-\frac{\mu-m}{b}})\end{aligned}$$

(i) and (ii) can be merged and hence, $\mathrm{E}\left[\frac{x-\mu}{|x-\mu|}\right] = -sgn(\mu - m)(1 - e^{-\frac{|\mu-m|}{b}})$. From the argument above,

$$\frac{\partial L}{\partial \mu} \approx \mathrm{E}\left[\frac{\partial L}{\partial \mu}\right] = -\frac{1}{b}\mathrm{E}\left[\sum_{n=1}^{N}\frac{x_n-\mu}{|x_n-\mu|}\right]$$
$$= \frac{N}{b} sgn(\mu - m)(1 - e^{-\frac{|\mu-m|}{b}}) \tag{12}$$
$$\therefore \frac{\partial^2 L}{\partial \mu^2} = \frac{N}{b^2} e^{\frac{-|\mu-m|}{b}}$$
$$\mathrm{E}\left[\frac{\partial^2 L}{\partial \mu^2}\right] = \frac{N}{b^2}\mathrm{E}\left[e^{\frac{-|\mu-m|}{b}}\right]$$

$$\mathrm{E}\left[e^{\frac{-|\mu-m|}{b}}\right] = \int_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} e^{\frac{-|\mu-m|}{b}}.\frac{1}{\delta}.d\mu$$
$$= \frac{1}{\delta}\int_{m-\frac{\delta}{2}}^{m} e^{\frac{\mu-m}{b}} d\mu + \int_{m}^{m+\frac{\delta}{2}} e^{-\frac{\mu-m}{b}} d\mu$$
$$= 2b\left(\frac{1-e^{-\frac{\delta}{2b}}}{\delta}\right)$$
$$= 2b\left(\frac{1}{2b} - \frac{1}{2b}\mathcal{O}\left(\frac{\delta}{2b}\right)\right) \qquad (\text{assuming} \quad \delta \ll 2b)$$
$$\approx 1$$

$$\therefore \mathrm{E}\left[\frac{\partial^2 L}{\partial \mu^2}\right] = \frac{N}{b^2}(1) = \frac{N}{b^2}$$

Using (12),

$$\frac{\partial^2 L}{\partial b \partial \mu} = N sgn(\mu - m)\left[-\frac{1}{b^2} - \left(e^{\frac{-|\mu-m|}{b}}\left(-\frac{1}{b^2}\right) + \frac{1}{b}e^{\frac{-|\mu-m|}{b}}\frac{|\mu-m|}{b^2}\right)\right]$$
$$= -\frac{N}{b^2} sgn(\mu - m)(1 - e^{-\frac{|\mu-m|}{b}}) - \frac{N}{b}.\frac{(\mu-m)}{b^2}.e^{-\frac{|\mu-m|}{b}}$$
$$\therefore \mathrm{E}\left[\frac{\partial^2 L}{\partial b \partial \mu}\right] = -\frac{N}{b^2}(\mathrm{E}_1 - \mathrm{E}_2) - \frac{N}{b^3}\mathrm{E}_3, \qquad \text{where}$$

$$E_1 = E[sgn(\mu - m)] = \int\limits_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} sgn(\mu - m).\frac{1}{\delta}.d\mu = \frac{1}{\delta}\int\limits_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} \frac{\mu - m}{|\mu - m|}d\mu$$

$$= \frac{1}{\delta}\int\limits_{-\frac{\delta}{2}}^{\frac{\delta}{2}} \frac{t}{|t|}dt = 0 \qquad \text{(as the integrand is an odd function)}$$

$$E_2 = E[sgn(\mu - m)e^{-\frac{|\mu-m|}{b}}] = \frac{1}{\delta}\int\limits_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} \frac{\mu - m}{|\mu - m|}e^{-\frac{|\mu-m|}{b}}d\mu$$

$$= \frac{b}{\delta}\int\limits_{-\frac{\delta}{2}}^{\frac{\delta}{2}} \frac{t}{|t|}e^{-|t|}dt = 0 \qquad \text{(as the integrand is an odd function)}$$

$$E_3 = E[(\mu - m)e^{-\frac{|\mu-m|}{b}}] = \frac{1}{\delta}\int\limits_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} (\mu - m)e^{-\frac{|\mu-m|}{b}}d\mu$$

$$= \frac{b^2}{\delta}\int\limits_{-\frac{\delta}{2}}^{\frac{\delta}{2}} te^{-|t|}dt = 0 \qquad \text{(as the integrand is an odd function)}$$

$$\therefore E\left[\frac{\partial^2 L}{\partial b \partial \mu}\right] = 0$$