JMLR: Workshop and Conference Proceedings 29:1–16, 2013 ACML 2013

Minimum Message Length Estimate of Parameters of Laplace Distribution

Parthan Kasarapu PARTHAN.KASARAPU@MONASH.EDU
Faculty of IT, Monash University, Clayton, VIC 3800, Australia

Lloyd Allison LLOYD.ALLISON@MONASH.EDU
Faculty of IT, Monash University, Clayton, VIC 3800, Australia

Enes Makalic EMAKALIC@UNIMELB.EDU.AU
Centre for MEGA Epidemiology, The University of Melbourne, Carlton, VIC 3053, Australia

Editor: Cheng Soon Ong and Tu Bao Ho

Abstract

The Laplace distribution offers a number of uses in statistical inference and modelling of symmetric data with long tails. We report here for the first time the derivation of the minimum message length (MML) estimates of location (μ) and scale (b) parameters of the Laplace distribution for any observed data. We demonstrate an application of this work to compare and contrast the quality of orthogonal superposition of two (spatial) vector sets under ℓ_1 and ℓ_2 norms.

Keywords: Laplace distribution, Normal distribution, minimum length encoding

1. Introduction

The Laplace distribution is a continuous probability density function that is dependent on the absolute value of the difference between the mean and the data point *i.e.,* the negative log probability varies linearly with the absolute difference. Because of this characteristic, the inference of parameters of this distribution (based on some empirical data) is not severely affected by outliers. The Laplace distribution is the 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 normal distribution, on the other hand, is sensitive to outliers because of the quadratic nature of the contributions of individual terms. In many problems, formulating an objective function (that needs to be optimized) as the sum of squared deviations (the ℓ_2 norm) results in a closed form solution. Its ℓ_1 equivalent, however, does not have an analytical solution (for data with more than one dimension). As such, one needs to adopt approximate methods to find the best superposition corresponding to the ℓ_1 norm. We show this in the context of the superposition of vector sets and use a Monte Carlo simulation to optimize the objective function when it is formulated using the ℓ_1 norm. To compare the superpositions obtained when using these two objective functions, we use the minimum

© 2013 P. Kasarapu, L. Allison & E. Makalic.

message length (MML) criterion. MML is used to adjudicate which is better on individual test cases.

MML is an information theoretic method of Bayesian inference (Wallace and Boulton, 1968). MML is best understood through a communication process between an imaginary transmitter and receiver connected over a Shannon channel. The method involves encoding the model and its fit to the data. A model that results in the smallest total message length is inferred as the most suitable under this criterion. Indeed the message length paradigm provides an objective means to differentiate between competing models and select the best one. In this paper, we use this technique to evaluate the overall fit when the data is modelled using a Normal and a Laplace distribution.

While the MML expression for the Normal distribution has been worked out previously (Wallace and Boulton, 1968; Wallace, 2005), the MML estimates for the Laplace distribution have not been characterized. The main contribution of our 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 Boulton (1975) and is referred to as Strict Message Length (SMML) Inference. This is computationally infeasible and is shown to be NP-hard for most models (Farr and Wallace, 2002). Wallace and Freeman (1987) provide a quadratic approximation which is often easy to compute and is commonly referred to as the *Wallace-Freeman* approximation. The MML method of estimating parameters for a number of distributions using this approximation has been well established (Wallace, 2005). In this paper, we use the Wallace–Freeman quadratic approximation to derive the message length expression for the Laplace distribution.

The results demonstrate the use of Laplace estimates in two cases. The first scenario involves data being randomly generated from a Laplace distribution. This data is then fitted using both Normal and Laplace models. We use the derived formulation of Laplace MML to compute the Laplace fit. This is then compared against the fit using a Normal distribution.

As a demonstration of a practical real world application, we consider the problem of superposition of vector sets. The objective function for the superposition problem can be formulated where the deviations of the corresponding vectors are calculated using the ℓ_1 norm or the ℓ_2 norm. In MML parlance, the optimal superposition would correspond to stating the deviations concisely. We encode the deviations using both Normal and Laplace distributions. The distribution which results in the best compression is chosen to be the best model for those two vector sets.

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 competing hypotheses that model some observed data. As mentioned before, 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] which maximizes the posterior distribution of the parameters given the data. Using Bayes's 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 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's 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. When comparing 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 discriminative ability of MML lies in its consideration of the model complexity and the error of the fit. In the MML setting, the total message length, therefore, involves a statement cost of the hypothesis (given by $I(H)$) and a statement cost of the data given the hypothesis (given by $I(D|H)$).

2.2. Parameter Estimation using MML

The hypothesis is a statistical model which is characterized by its parameters. MML accounts for both the model complexity and its explanatory power of the data observed. MML works by maximizing the expectation of the posterior probability and this is where it differs from the traditional Bayesian methods. This involves determining the *accuracy of parameter values (AOPV)* for continuous parameters. AOPV is a measure of the uncertainty in stating a parameter and it is this region over which the expectation is maximized. As per MML, a parameter need not be stated very accurately. The optimal precision to which it needs to be stated is computed as part of the MML inference. A good description of the procedure is outlined in Oliver and Baxter (1994).

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

3. Message Length of Laplace distribution

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

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

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 $\ell(\bar{\theta})$ is the negative log likelihood of observed data. The MML estimates $\hat{\bar{\theta}}_{\mathrm{MML}}$ of the parameters are determined by minimizing expression (1).

3.1. Laplace distribution

The contribution of this paper is in the derivation of the MML estimates of the parameters of the Laplace distribution which have not been characterized previously. The parameters describing a Laplace distribution are the location(μ) and the scale (b). The probability density function (pdf) is given in (2).

$$\mathrm{pdf}(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \tag{2}$$

To derive the MML estimates, we use the Wallace–Freeman approximation. This estimation as per (1) requires the likelihood function, the Fisher information matrix, and the prior distributions on the parameters. Let $D = \{x_1, x_2, \ldots, x_N\}$ be the observed data containing N samples and ϵ be the precision to which each datum is stated. Let R_μ, R_b be the range of μ and $\log b$ respectively. ϵ, R_μ, R_b are hyperparameters introduced in Wallace (2005). These are fixed and will not be inferred from the observed data. Using Equation 2, the *likelihood function* is given by

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

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

$$\begin{aligned} L(\bar{\theta}) &= -\log f(D|\bar{\theta}) \\ &= N\log\left(\frac{2}{\epsilon}\right) + N\log b + \frac{1}{b}\sum_{n=1}^{N}|x_n - \mu| \end{aligned} \tag{3}$$

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

$$\begin{aligned} \hat{\mu}_{\mathrm{ML}} &= \mathrm{median}\{x_1, x_2, \ldots, x_n\} \\ \hat{b}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}|x_n - \hat{\mu}_{\mathrm{ML}}| \end{aligned} \tag{4}$$

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 Equation 3, we have:

$$\frac{\partial^2 L}{\partial b^2} = -\frac{N}{b^2} + \frac{2}{b^3}\sum_{n=1}^{N}|x_n - \mu|$$

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

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

$$\text{Therefore,}\quad \mathrm{E}\left[\frac{\partial^2 L}{\partial b^2}\right] = -\frac{N}{b^2} + \frac{2}{b^3}(Nb) = \frac{N}{b^2} \tag{5}$$

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, discussed in section A. Using those results, we have

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

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

Priors on the parameters

A prior probability on μ and b is assumed in accordance with the prior assumed in the case of Normal. It is assumed that μ and b have independent priors, that μ has a uniform prior density (location invariance), and that $\log b$ has a uniform prior density (scale invariance) (Wallace, 2005). The ranges from which μ and $\log b$ are drawn are assumed to be R_μ and R_b respectively. The joint prior probability of the parameters is then given by

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

From (1), (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 $\hat{\mu}_{\text{MML}}$ and $\hat{b}_{\text{MML}}$ 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

$$\hat{\mu}_{\text{MML}} = \text{median}\{x_1, x_2, \ldots, x_n\}$$

$$\frac{\partial I}{\partial b} = 0 \Rightarrow \frac{N-1}{\hat{b}} - \frac{1}{\hat{b}^2}\sum_{n=1}^{N} |x_n - \hat{\mu}_{\text{MML}}| = 0$$

$$\text{Therefore,} \quad \hat{b}_{\text{MML}} = \frac{1}{N-1}\sum_{n=1}^{N} |x_n - \hat{\mu}_{\text{MML}}| \tag{8}$$

The ML estimator (4) has N in the denominator which differs from the MML estimator (8) which has $(N-1)$ in the denominator. Hence, the ML estimate tends to underestimate b slightly. Using these estimates, the minimum message length becomes

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

4. MML Inference for a Normal distribution

A Normal distribution whose density function (pdf) is given by (10)

$$\text{pdf}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{10}$$

is characterized by the mean μ and the standard deviation σ. The MML estimates for a normal distribution are presented in Wallace (2005); Wallace and Freeman (1987) and are:

$$\hat{\mu}_{\text{MML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

The corresponding minimized message length is given as

$$\begin{aligned} I_{min} &= I(\hat{\mu}_{\text{MML}}, \hat{\sigma}_{\text{MML}}) \\ &= 1 + \log \kappa_2 + \log(\mathrm{R}_\mu \mathrm{R}_\sigma) + \frac{1}{2}\log(2N^2) \\ &\quad + \frac{N}{2}\log\left(\frac{2\pi}{\epsilon^2}\right) + \frac{N-1}{2}\log\left(\frac{\sum_{n=1}^{N}(x_n - \hat{\mu}_{\text{MML}})^2}{N-1}\right) + \frac{N-1}{2} \end{aligned} \tag{11}$$

5. Experiments

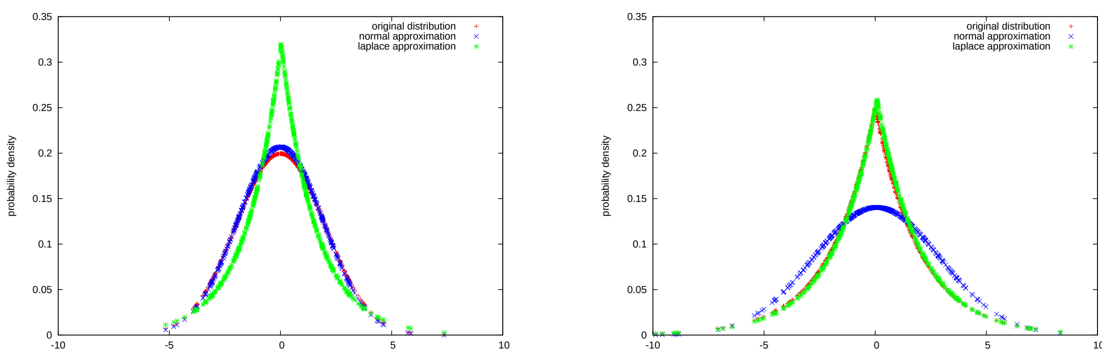
We demonstrate the use of Laplace estimates in two scenarios.

5.1. 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/Normal, then the compression in message length is better when it is modelled using a Laplace/Normal distribution respectively. This is indeed expected and is done as a validation check to ensure that the derived MML formulation for the Laplace is consistent with the observation. Equations (9) and (11) are used to determine the code length when the data is modelled using the Laplace and Normal distributions.

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 indicative 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 (6688.78 bits) is smaller compared to that of the Normal estimate (6770.24 bits).



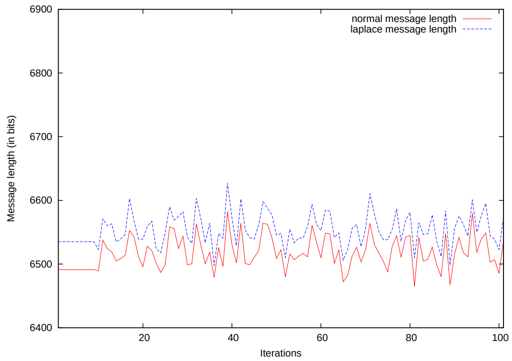
(a) Original distribution: Normal (b) Original distribution: Laplace

Figure 1: Approximation of data using Normal & Laplace distributions

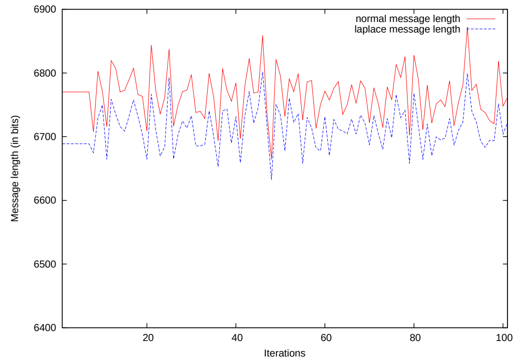
Figure 2 compares the message lengths over 100 iterations. Each iteration involves generating 500 random data samples and modelling using both distributions. For each iteration, there will be a message length corresponding to each of the distributions. In 2(a), the original distribution is Normal and it is observed that over all the iterations, the message length for the Normal (red) is consistently less than that of the Laplace (blue). A

Table 1: Comparison of the estimates (for a single iteration)

True distribution	True mean	True spread	Normal estimates			Laplace estimates		
			mean	spread	msglen	mean	spread	msglen
Normal	0	2	-0.00687	1.92919	**6491.21**	0.0269173	1.56037	6535.04
Laplace	0	2	0.06303	2.84256	6770.24	0.0789925	1.93187	**6688.78**



(a) Original distribution: Normal



(b) Original distribution: Laplace

Figure 2: Comparison of message lengths over 100 iterations

similar but reversed behaviour is observed when the data was generated using the Laplace distribution as can be seen in 2(b).

5.2. Superposition of vector sets

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 U to align it with V such that the deviations of each vector in V with its counterpart in U is minimized. Let the transformation be effected by a translation vector t and a rotation matrix R. Let it result in an altered vector set $U' = \{u'_1, u'_2, \ldots, u'_m\}$, where $u'_i = R(u_i - t)$. Both the objective functions corresponding to the sum of squares (ℓ_2 norm) of all deviations

$$\sum_{i=1}^{m} \|v_i - u'_i\|^2 = \sum_{i=1}^{m} \|v_i - R(u_i - t)\|^2 \tag{12}$$

and the objective function corresponding to the sum of absolute deviations (ℓ_1 norm)

$$\sum_{i=1}^{m} \|v_i - u'_i\| = \sum_{i=1}^{m} \|v_i - R(u_i - t)\| \tag{13}$$

(where $\|.\|$ denotes the vector norm) need to be minimized. 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 encoded in an effective manner. Superposition based on minimizing total least squares corresponds 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.

Minimization of (12) yields $t = \left(\frac{\sum_{i=1}^m u_i}{m} - \frac{R\sum_{i=1}^m v_i}{m}\right)$. Substituting this value of t in (12) results in the modified objective function:

$$\sum_{i=1}^{m} \left|\left| \left(v_i - \frac{\sum_{i=1}^m v_i}{m}\right) - R\left(u_i - \frac{\sum_{i=1}^m u_i}{m}\right) \right|\right|^2 \tag{14}$$

Kearsley (1989) provides a solution to (12) by resolving the transformation into translation and rotation. The centres of mass of the two vector sets are translated to the origin (14) and the problem then reduces to finding the rotation matrix which minimizes the total least squares. This involves respresenting the rotation matrix using a quaternion and then solving the resultant eigenvalue decomposition problem. As such, Kearsley (1989) offers an analytical way to solve the *least squares* superposition problem.

Minimizing (13), however, does not yield a closed form solution. Differentiating (13) with respect to t and setting it to 0 yields

$$\sum_{i=1}^{m} \frac{Rv_i - (u_i - t)}{\|Rv_i - (u_i - t)\|} = 0 \tag{15}$$

In this case, R and t cannot be separated and, hence, an analytic solution does not exist. As such, one needs to resort to approximate methods to find the best ℓ_1 superposition. The one used in this paper is based on 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 (13) is computed.

2. From this orientation, the protein is perturbed randomly. If the new orientation results in a better value of the L1 norm (13), 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 minute probability.

3. This is repeated for a certain number of 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.

The two vector sets are first superposed using the Kearsley's method and the message length (I_N) computed through MML inference using a Normal distribution. Monte Carlo simulation is performed (as discussed above) from this stage and the final orientation is obtained. At this point, the message length (I_L) is computed through MML inference using a Laplace distribution. Two cases arise:

- If $I_L < I_N$, then there exists a superposition that is better than the one resulting from minimizing the sum of squared deviations (12).

- If $I_N < I_L$, then the superposition obtained by minimizing (12) is better. Since the minimal ℓ_1 superposition is obtained using a Monte Carlo simulation (which is terminated after a certain number of iterations), it could also be possible that the optimal solution wasn't found.

The aim of this exercise is to show that not all vector sets have their optimal superpositions dictated by minimizing sum of squared deviations. It also drives home the use of MML estimators in determining the kind of superposition to be considered.

Results

We apply the problem of superposition to protein structures. In this context, the vector sets corresponds to the three dimensional coordinates of the α-carbon atoms of amino acid residues constituting the proteins' backbone. We use SUPER (Collier et al., 2012) (an implementation of Kearsley's orthogonal superposition) to get all protein segments from the Protein Data Bank which fit to a Root Mean Squared Deviation (RMSD) of 5 Angstroms (Å) or better. The protein segment corresponding to PDB ID 2IC7, chain A and residues 132-162. This protein segment was randomly considered and there were ~23000 segments which fit within a RMSD of 5 Å. The message length corresponding to this optimal ℓ_2 superposition is calculated using (11). This orientation will also have a certain value for the sum of absolute deviations. The message length to encode these absolute differences is computed using (9). For these 23000 segments across several proteins, we determine the optimal ℓ_1 superposition using our Monte Carlo simulation. The superposition resulting after 1000 iterations is regarded as the optimal ℓ_1 superposition. The message length to encode the absolute differences is computed. There were about ~1700 instances where the optimal ℓ_1 superposition is better than the ℓ_2 equivalent. The results of the Monte Carlo simulation are shown in Table 2.

Each time we superimpose some protein segment with that of 2IC7, the other segment is treated as a 'fixed structure' and 2IC7 segment is randomly perturbed from its optimal ℓ_2 superposition. Each segment in the other protein is uniquely identified by its chain ID and residue span. These are obtained using SUPER. RMSD is the error of fit of the ℓ_2 superposition. The initial/final ℓ_1 deviations correspond to the average of the absolute deviations in both these orientations; msglen(ℓ_2) is the message length using a Normal distribution to encode the deviations in the optimal ℓ_2 superposition; msglen (initial ℓ_1) corresponds to the message length of encoding the absolute deviations before perturbations and msglen (final ℓ_1) is the message length to encode the deviations after the Monte Carlo simulation. As expected, the final ℓ_1 deviation is less than the initial one. It is observed

Table 2: Comparison of the minimum message lengths

Fixed structure	RMSD (in Å)	Initial ℓ_1 deviation (in Å)	Final ℓ_1 deviation (in Å)	msglen (ℓ_2)	msglen (initial ℓ_1)	msglen (final ℓ_1)
3IGJ [A:139-169]	1.834	1.910	1.872	1134.82	1102.03	**1100.28**
4G87 [A:291-321]	2.112	2.555	2.465	1153.54	1165.38	**1138.59**
3R2W [B:435-465]	3.437	4.637	4.593	**1218.17**	1228.74	1222.79

this happens for most of the cases which suggest the presence of the ℓ_1 optimum somewhere close to the initial position.

For the discussion below, in Figure 3, 4(a), 4(b), the red curve corresponds to the fixed protein; the blue curve is the optimal ℓ_2 superposition (of the 2IC7 segment), and the green curve corresponds to its optimal ℓ_1 superposition.

The first row in Table 2 corresponds to Figure 3. Its msglen(initial ℓ_1) < msglen(ℓ_2). This suggests that ℓ_1 superposition supercedes the ℓ_2 superposition, and it only gets better as the structure is perturbed as is evidenced by msglen(final ℓ_1) – the message length after the Monte Carlo simulation. Here, it can be seen that the red curve (the fixed protein) is closer to the green curve (ℓ_1 superposition) than to the blue curve (ℓ_2 superposition) in the majority of the protein structure.

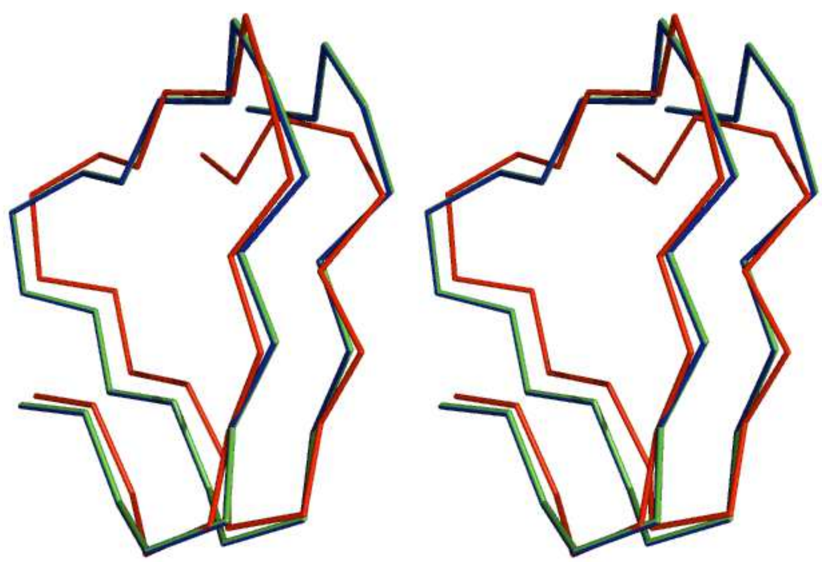
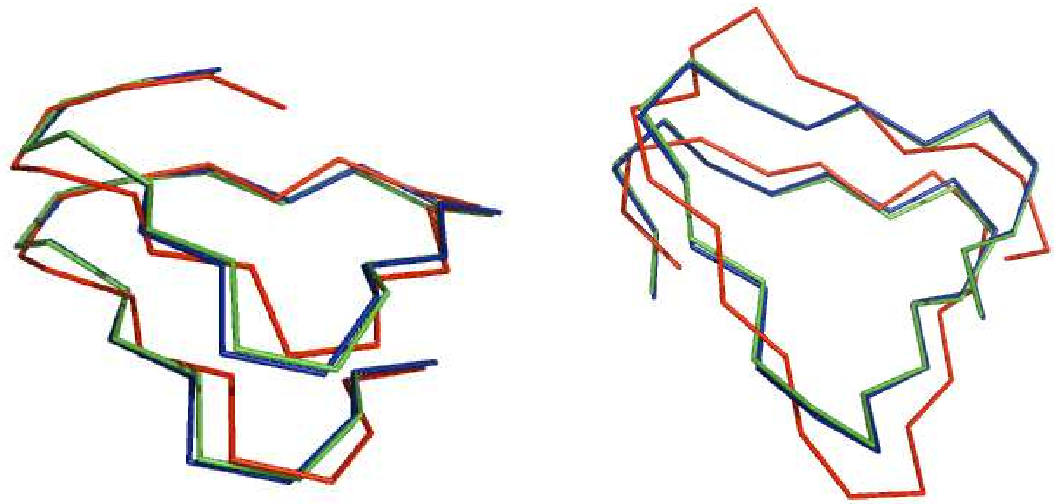


Figure 3: Stereo images of the superposition of segments from 2IC7 (red) & 3IGJ (fixed). Initial & final ℓ_1 superposition (green) better than ℓ_2 superposition (blue).

The second row in Table 2 corresponds to the case where the final ℓ_1 superposition (1138.59 bits) is better than the ℓ_2 superposition (1153.54 bits). The initial ℓ_1 superposition (1165.38 bits) is not the optimal one and hence, the Monte Carlo simulation goes through a series of iterations to reach the local optimum. In 4(a), the red curve is closer to the green curve (final ℓ_1) than to the blue curve (ℓ_2 superposition). The third row in Table 2 demonstrates the case where the ℓ_2 superposition (1218.17 bits) is optimal compared to the ℓ_1 superposition (1222.79 bits). On careful inspection, one can see that the red curve (in 4(b)) is closer to the blue curve corresponding to the ℓ_2 superposition than to the green curve (the optimal ℓ_1 superposition).



(*a*) Final ℓ_1 superposition (green) better than ℓ_2 superposition (blue)

(*b*) ℓ_2 superposition (blue) better than ℓ_1 superposition (green)

Figure 4: Comparing the protein superpositions with respect to ℓ_1 and ℓ_2 norm

6. Conclusion

We have derived the MML estimators for the Laplace distribution and applied them in the context of encoding the superposition of vector sets. This is used to distinguish the quality of superposition for any two vector sets. In general, if the objective function is formulated as the sum of absolute differences, we have a framework using MML to encode these values using a Laplace distribution. The quality of the overall fit can be evaluated using the MML inference technique and can be compared with other models. For the specific problem of superposition of two vector sets, the choice of ℓ_1 or ℓ_2 superposition is made by comparing the message lengths obtained by encoding deviations using ℓ_1 and ℓ_2 norm.

Acknowledgments

We thank Arun Konagurthu for the numerous discussions we had regarding the superposition application. We also thank Maria Garcia de la Banda and James Collier for proof reading the paper.

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.

G. E. Farr and C. S. Wallace. The complexity of strict minimum message length inference. *The Computer Journal*, 45(3):285–292, 2002.

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.

J.J. Oliver and R.A. Baxter. MML and Bayesianism: Similarities and differences. *Dept. Comput. Sci. Monash Univ., Clayton, Victoria, Australia, Tech. Rep*, 206, 1994.

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 D. M. Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.

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 assume that $\mu \in \left[m-\frac{\delta}{2}, m+\frac{\delta}{2}\right]$. İ1t 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, $\mathrm{pdf}(\mu) = \frac{1}{\delta}$.

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

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

(i) Let $\mu < m$

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

(ii) Let $\mu > m$

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

(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,

$$\begin{aligned}
\frac{\partial L}{\partial \mu} \approx \mathrm{E}\left[\frac{\partial L}{\partial \mu}\right] &= -\frac{1}{b}\mathrm{E}\left[\sum_{n=1}^{N}\frac{x_n-\mu}{|x_n-\mu|}\right] \\
&= \frac{N}{b} sgn(\mu-m)(1-e^{-\frac{|\mu-m|}{b}}) \qquad (16)\\
\text{Therefore,} \quad \frac{\partial^2 L}{\partial \mu^2} &= \frac{N}{b^2} e^{\frac{-|\mu-m|}{b}} \\
\mathrm{E}\left[\frac{\partial^2 L}{\partial \mu^2}\right] &= \frac{N}{b^2}\mathrm{E}\left[e^{\frac{-|\mu-m|}{b}}\right]
\end{aligned}$$

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

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

Using (16),

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

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

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

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

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

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

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

$$\text{Therefore,} \quad E\left[\frac{\partial^2 L}{\partial b \partial \mu}\right] = 0$$