JMLR: Workshop and Conference Proceedings 29:1–14, 2013 ACML 2013

Minimum Message Length Estimate of Parameters of Laplace Distribution

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

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

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

Editor: Cheng Soon Ong and Tu Bao Ho

Abstract

The Laplace distribution offers a number of uses in statistical inference and modelling on 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 norm.

Keywords: Laplace distribution, Normal distribution, MML, 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 quadratic nature of the contributions of individual terms to the objective function. 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 relatively more robust to outliers as the objective function involves the sum of the absolute values of the difference of the individual terms (L1 norm). The outliers are not penalised as drastically when compared with the Normal distribution. The use of MML to discriminate between a Laplace and a Normal is explored in this paper. This selection is made by formulating the objective function using minimum message length (MML) principle. We use the derived minimum message length expression for the Laplace. The minimum message length expression for the Normal has been worked out previously. The message lengths resulting from these two distributions are compared against each other. 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. In contrast, because of the mathematical nature of the Laplace (2) which is expressed as the absolute difference from the mean, there does not exist a closed form solution (for data with more than one dimension).

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

In spite of the lack of an analytical solution, 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) and is referred to as Strict Message Length (SMML) Inference. This is computationally infeasible and is shown to be NP-hard (Farr and Wallace, 2002). Wallace and Freeman (1987) provide an approximate form which 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 approximation to derive the the message length expression for the Laplace distribution.

MML is an information theoretic way of Bayesian inference. The method involves encoding the model and its fit to the data. This encoded message needs to be communicated between a hypothetical transmitter and a receiver. A model with the smallest message length directly relates to a better fit. This provides a mechanism to select the best model from a set of models. In this paper, we use this intuition to evaluate the overall fit when the data is modelled using a Normal and a Laplace distribution.

We 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.

We show a practical application of this research on 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 L1 norm or the L2 norm. An optimal superposition would effectively minimize the sum of all the deviations.

In MML parlance, the optimal superposition would correspond to stating the deviations concisely. We encode the deviations using both Normal and Laplace distributions and compare them. In particular, we experiment with protein structures, in which case, the vector sets correspond to the three dimensional coordinates of the atoms.

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'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. 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 $I(D|H)$)

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

2.2. Parameter Estimation using MML

The hypothesis is a statistical model which is characterized by its 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 given the parameters. To calculate the message length to encode parameters, the prior density on the parameters is partitioned into cells each with a distinct index. It is this unique identifier that is transmitted across and the receiver interprets the parameters as the mid-point of the corresponding cell. If there are d parameters describing the model, each cell has an uncertainty in volume which is called the *accuracy of parameter value (AOPV)*. The parameter estimation process requires the determination of this precision to which the parameters need to be stated. A good description of the procedure is outlined in Oliver and Baxter (1994).

The message length varies with the precision to which the parameters are stated. 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. The optimal precision is expressed in terms of the determinant of the Fisher information matrix as shown in Oliver and Baxter (1994).

The overall message length consists of the contribution due to the parameters (model complexity) and the data given the parameters (error of fit). MML allows us to balance this tradeoff by choosing the parameters that minimize the overall message length. Such an elegant way allows us to compare two hypotheses which model the same data but using different sets of parameters.

3. Message Lengths of Normal & Laplace distributions

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

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

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 $\hat{\bar{\theta}}_{\mathrm{MML}}$ 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); Wallace and Boulton (1968). The MML estimates are:

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

The corresponding minimized message length is given as

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

3.2. Laplace distribution

The contributions 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 μ and b. μ is the *location* parameter 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

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

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),

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

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

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

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

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

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

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

As established in Oliver and Baxter (1994), the optimal precision to which the parameters need to be stated is given by: $AOPV \propto \frac{1}{\sqrt{det(F(\bar{\theta}))}}$. The Normal Fisher is $\frac{2N^2}{\sigma^4}$ (Wallace, 2005) and the Laplace Fisher as derived above is $\frac{N^2}{b^4}$. There is a marked difference in the optimal AOPVs computed in the two cases. For the same value of spread ($\sigma = b$), the uncertainty in the parameter values for a Laplace is $\sqrt{2}$ times that of Normal, which means that the parameters for a Laplace need to be stated less precisely when compared with the Normal.

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.

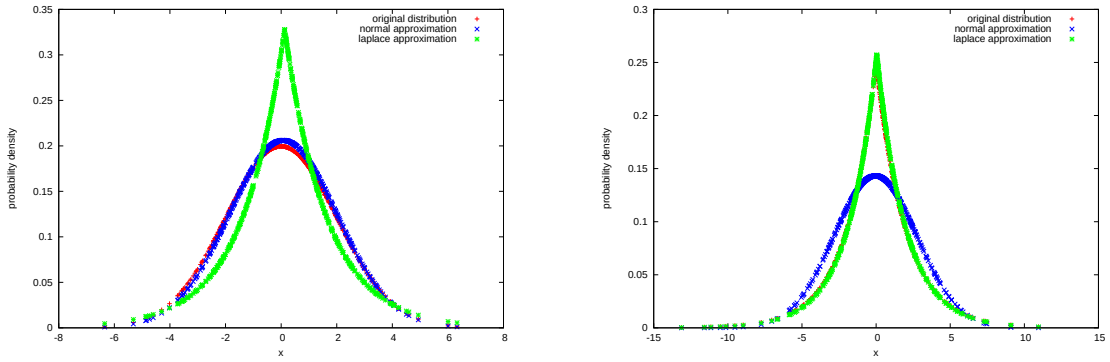
4.1. Data generation and modelling

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/Normal, then the compression in message length is better when it is modelled using a Laplace/Normal distribution. 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. (9) and (4) are used to determine the code length when the data is modelled using the Laplace and Normal distributions respectively.

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

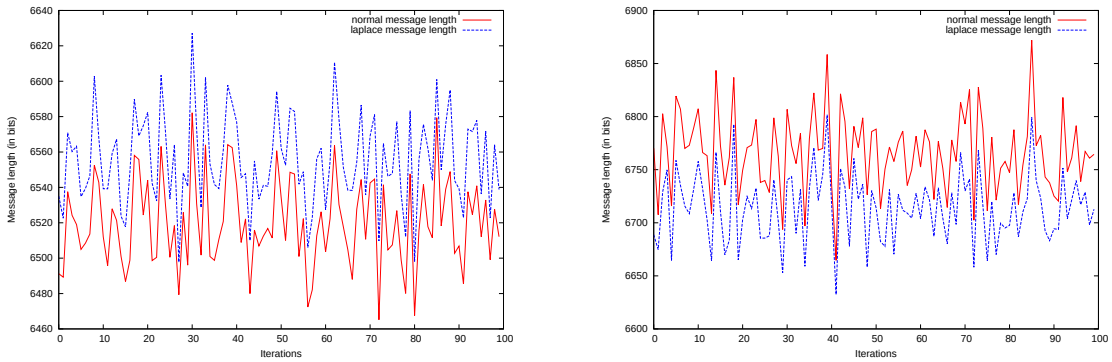
Table 1: Comparison of the estimates

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

Figure 2 compares the message lengths over 100 iterations. 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). In 2(b), the original distribution is Laplace and it is observed that over all the iterations, the message length for the Laplace (blue) is consistently less than that of the Normal (red).

4.2. 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 V to align it with U such that the deviations of each vector in V with its counterpart in U is minimum. If a transformation on V results in an



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

Figure 2: Comaprison of message lengths over many iterations

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:

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

or the objective function corresponding to the sum of absolute deviations (L1 norm)

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

(where $\|.\|$ denotes the vector norm) needs 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 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.

Kearsley (1989) provides a solution to (10) 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 involves respresenting 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.

Minimizing (11), however, does not yield a closed form solution. As such, one needs to adopt approximate methods to find the best superposition corresponding to the L1 norm. 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 the L1 norm (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 minute 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.

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 which is optimal than the one resulting from minimizing the sum of squared deviations (10).

- If $I_N < I_L$, then the superposition obtained by minimizing (10) is better. Since the minimal L1 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 point 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. The vector sets would correspond to the three dimensional coordinates of the α-carbon atoms of amino acid residues constituting the proteins' backbone. In our experiment, the initial least squares superposition using Kearsley's method is done using SUPER (Collier et al., 2012).

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.

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 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,

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

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

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

Using (12),

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

$$\mathrm{E}_1 = \mathrm{E}[sgn(\mu - m)] = \int_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} sgn(\mu - m).\frac{1}{\delta}.d\mu = \frac{1}{\delta}\int_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} \frac{\mu - m}{|\mu - m|}d\mu$$

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

$$\mathrm{E}_2 = \mathrm{E}[sgn(\mu - m)e^{-\frac{|\mu-m|}{b}}] = \frac{1}{\delta}\int_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} \frac{\mu - m}{|\mu - m|}e^{-\frac{|\mu-m|}{b}}d\mu$$

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

$$\mathrm{E}_3 = \mathrm{E}[(\mu - m)e^{-\frac{|\mu-m|}{b}}] = \frac{1}{\delta}\int_{m-\frac{\delta}{2}}^{m+\frac{\delta}{2}} (\mu - m)e^{-\frac{|\mu-m|}{b}}d\mu$$

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

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