Universal Models for the Exponential Distribution

Daniel F. Schmidt and Enes Makalic

***Abstract*—This note considers the problem of constructing information theoretic universal models for data distributed according to the exponential distribution. The universal models examined include the sequential Normalised Maximum Likelihood (SNML) code, conditional Normalised Maximum Likelihood (CNML) code, the Minimum Message Length (MML) code and the Bayes mixture code (BMC). The CNML code yields a codelength identical to the Bayesian mixture code, and within $O(1)$ of the MML codelength, with suitable data driven priors.**

***Index Terms*—MDL, MML, Universal Models**

I. Introduction

Consider n i.i.d. data points $\mathbf{x}^n = (x_1, \ldots, x_n)$ distributed according to the exponential distribution

$$p(\mathbf{x}^n|\theta) = \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) \tag{1}$$

This paper is concerned with derivation of universal models [1] for the exponential distribution. An important universal model in the Minimum Description Length (MDL) literature is the Normalised Maximum Likelihood (NML) density [1], [2], [3]

$$p_{\mathrm{NML}}(\mathbf{x}^n) = \frac{p(\mathbf{x}^n|\hat{\theta}_{\mathrm{ML}}(\mathbf{x}^n))}{\int_{\mathbb{R}^n} p(\mathbf{y}^n|\hat{\theta}_{\mathrm{ML}}(\mathbf{y}^n)) d\mathbf{y}^n} \tag{2}$$

which, when it exists, minimises the maximum coding regret relative to a family of source distributions.

The NML distribution is not readily computable for the exponential distribution since the integral in the denominator of (2) (i.e. the parametric complexity) diverges over $\mathbb{R}_+^n$. In the case of the linear-Gaussian regression model an ingenious bounding of the dataspace followed by renormalisation circumvents the problem of infinite parametric complexity [4]. In the case of the exponential distribution, placing a lower bound on the sufficient statistic does not alleviate the problem, hence several alternative universal models are considered. The universal models examined in this paper are based on sequential Normalised Maximum Likelihood (SNML) [5], conditional Normalised Maximum Likelihood (CNML) [6], the Minimum Message Length code (MML87) [7], [8], [9], [10] and the Bayes mixture code (BMC) [11].

II. Universal Models for the Exponential Distribution

A. Sequential NML

The sequential NML procedure (SNML) was recently introduced in [5], [6] as an alternative code to Predictive MDL [1], [12]. SNML allows one to perform prediction over new data which was previously not possible with the conventional NML distribution. The basic idea is to transmit data sequentially, such that the code for the new data is conditioned on all the previous data and attains minimax regret. Consider a set $\mathbf{x}^n \in \mathbb{R}_+^n$ of n i.i.d. data samples distributed as per an exponential distribution with unknown parameter $\theta \in \mathbb{R}_+$; the Maximum Likelihood estimate for θ is given by $\hat{\theta}_{\mathrm{ML}}(\mathbf{x}^n) = n / \sum_{i=1}^n x_i$. The sequential NML procedure transmits each data point $(x_{m+1}, \ldots, x_n)$ sequentially, such that the code for the x_t datapoint is based on previous data $\mathbf{x}^{t-1} = (x_{m+1}, \ldots, x_{t-1})$. The value $m > 0$ is the number of datapoints required for the Maximum Likelihood estimate to be computable; in the case of the exponential distribution $m = 1$. It is assumed that some 'base' coding distribution $p_1(\cdot)$ exists for the first m datapoints, and consequently their transmission and codelength are omitted in the following discussion.

The sequential predictive density for a new data point, x_t, is defined by the CNML distribution

$$p(x_t|\mathbf{x}^{t-1}) = \frac{p\left(\mathbf{x}^{t-1}, x_t|\hat{\theta}_{\mathrm{ML}}(\mathbf{x}^{t-1}, x_t)\right)}{\int_0^\infty p\left(\mathbf{x}^{t-1}, y|\hat{\theta}_{\mathrm{ML}}(\mathbf{x}^{t-1}, y)\right) dy} \tag{3}$$

The joint likelihood of x_t and $\mathbf{x}^{t-1}$ at the Maximum Likelihood estimate is

$$p\left(\mathbf{x}^{t-1}, x_t|\hat{\theta}_{\mathrm{ML}}(\mathbf{x}^{t-1}, x_t)\right) = \left(\frac{t}{e\left(\sum_{i=1}^{t-1} x_i + x_t\right)}\right)^t \tag{4}$$

The normalising constant in (3) can be evaluated analytically

$$\int_0^\infty \left(\frac{t}{e\left(\sum_{i=1}^{t-1} x_i + y\right)}\right)^t dy = \frac{t^t}{e^t(t-1)} \left(\sum_{i=1}^{t-1} x_i\right)^{1-t} \tag{5}$$

Finally, substituting (4) and (5) into (3) yields the one-step-ahead sequential NML distribution for the data point x_t conditioned on $\mathbf{x}^{t-1}$:

$$p(x_t|\mathbf{x}^{t-1}) = (t-1) \left(\sum_{i=1}^{t-1} x_i\right)^{t-1} \left(\sum_{i=1}^{t} x_i\right)^{-t} \tag{6}$$

For finite n the sequential predictive NML distribution is clearly not an exponential distribution. The total sequential NML codelength for data $\mathbf{x}^n$ is therefore $I_{\mathrm{SNML}}(\mathbf{x}^n) = -\sum_{t=2}^n \log p(x_t|\mathbf{x}^{t-1})$. As in the case of the predictive MDL universal model, this codelength is dependent on the *order* of the data. Ideally, one would try all possible permutations of $\mathbf{x}^n$ and use some statistic (such as the minimum) as a representative codelength. However, the number of permutations $n!$ renders this exhaustive procedure infeasible even for moderate sizes of n.

Daniel F. Schmidt and Enes Makalic are with Monash University
Clayton School of Information Technology
Clayton Campus Victoria 3800, Australia.
Telephone:+61 3 9905 9555, Fax: +61 3 9905 9422
Email: {Daniel.Schmidt,Enes.Makalic}@infotech.monash.edu.au

B. Conditional NML Distribution

An alternative approach that helps to reduce the arbitrary aspects of the SNML code is to construct a conditional NML distribution for $(n-1)$ datapoints conditioned on a single data point, say x_j. This reduces the number of permutations to n possible choices of the conditioning datum, and allows one to compute the CNML codelength for all $j = 1, \ldots, n$ and average as suggested by [13]. In the following, the notation $\mathbf{x}^{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$ denotes the dataset comprising all x_i with the j-th datapoint $(i \neq j)$ omitted. The CNML density for $\mathbf{x}^{-j}$ conditional on x_j is

$$p(\mathbf{x}^{-j}|x_j) = \frac{p\left(x_j, \mathbf{x}^{-j}|\hat{\theta}_{\mathrm{ML}}(x_j, \mathbf{x}^{-j})\right)}{\int_{\mathbb{R}_+^{n-1}} p\left(x_j, \mathbf{y}^{n-1}|\hat{\theta}_{\mathrm{ML}}(x_j, \mathbf{y}^{n-1})\right) \mathbf{dy}^{n-1}}$$

In the case of the exponential model (1), the conditional NML distribution for $\mathbf{x}^{-j}$ is

$$p(\mathbf{x}^{-j}|x_j) = \Gamma(n) \left(\sum_{i=1}^{n} x_i \right)^{-n} x_j \qquad (7)$$

In order to remove the dependency on the particular choice of x_j, one may average the negative logarithm of (7) over n possible choices of j leading to an 'average' codelength for $\mathbf{x}^n$. The problem with this approach is that the resulting codelength is defined for $(n-1)$ datapoints rather than n; this has little effect when n is large, but for small n it can cause problems for model selection. An alternative method is to augment the dataset by one extra artificial datapoint $\mathbf{x}^{n+1} = (x_0, \mathbf{x}^n)$ and compute the conditional NML distribution conditioned on this new datum. The new datapoint is chosen such that $\hat{\theta}_{\mathrm{ML}}(\mathbf{x}^{n+1}) = \hat{\theta}_{\mathrm{ML}}(\mathbf{x}^n)$, i.e. the introduction of x_0 leaves the Maximum Likelihood estimate unaltered. The choice $x_0 = (1/n \sum_{i=1}^n x_i)$ satisfies this restriction, and conditioning on x_0 yields a code, $I_{\mathrm{CNML}}(\mathbf{x}^n)$, for the remaining n datapoints

$$n \log \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) + (n+1)\log(n+1) - \log\Gamma(n+1) \qquad (8)$$

The approach of using an augmented dataset $\mathbf{x}^{n+1}$ as described above may also be used to find an SNML code for n datapoints; i.e. transmit the new dataset $\mathbf{x}^{n+1}$ using the SNML code $I_{\mathrm{SNML}}(\mathbf{x}^{n+1})$.

C. Bayes Mixture Code

The Bayes mixture code (BMC) [11] is based on the Bayesian universal model given by $p_{\mathrm{BAYES}}(\mathbf{x}^n) = \int_\Theta p(\mathbf{x}^n|\theta)\pi(\theta)d\theta$, where $\pi(\cdot)$ is a suitable prior distribution. The Bayesian universal model is clearly the marginal distribution of the data given the mixing distribution $\pi(\cdot)$; this type of universal model produces 'one-part' codes that are similar to what Wallace calls the 'non-explanation' code ([10], pp. 154). A possible prior over θ is the conjugate exponential prior $\pi(\theta|\alpha) = \alpha \exp(-\alpha\theta)$. This prior is equivalent to a posterior obtained from the Jeffreys' prior with one 'observation' equal to α. The marginal distribution of the data given α is

$$p_{\mathrm{BAYES}}(\mathbf{x}^n) = \alpha \left(\sum_{i=1}^{n} x_i + \alpha \right)^{(-n-1)} \Gamma(n+1) \qquad (9)$$

Following the idea of [14], the codelength (9) is minimised by choosing $\hat{\alpha} = 1/n \sum_{i=1}^n x_i$, which yields the complete codelength $I_{\mathrm{BAYES}}(\mathbf{x}^n)$:

$$n \log \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) + (n+1)\log(n+1) - \log\Gamma(n+1) \qquad (10)$$

It is known that the Bayes code and the conditional NML code coincide asymptotically (see Theorem 11.3 in [11]). Interestingly, the Bayes code with this choice of empirical prior is equivalent to the conditional NML code with a suitable choice of 'extra data' x_0 even in the finite sample case. This is also similar to the procedure derived in [15]; however, we sacrifice the artificial data point x_0 rather than any of the real data yielding a code for n data points.

D. Minimum Message Length Code

The Minimum Message Length (MML) universal model differs from the NML and BMC models in that it is a *two part* code that first transmits a fully specified model (i.e. an estimate of θ) with codelength $I_{87}(\theta)$, and then transmits the data $\mathbf{x}^n$, assuming the transmitted θ is the true generating model, with codelength $I_{87}(\mathbf{x}^n|\theta)$. In contrast to the NML/BMC codes one obtains an explicit estimate of parameters in addition to a measure of complexity. The Wallace-Freeman MML87 approximation, $I_{87}(\mathbf{x}^n, \theta)$, is used to obtain the MML codelengths in this paper; for univariate models this is

$$-\log \pi(\theta) + \frac{1}{2}\log J(\theta) - \frac{1}{2}\log 12 + \frac{1}{2} - \log p(\mathbf{x}^n|\theta) \qquad (11)$$

where $J(\theta)$ is the Fisher information and $\pi(\cdot)$ is a prior density over the parameter space. For the exponential distribution, the Fisher information of θ is $J(\theta) = n/\theta^2$. Using the conjugate exponential prior yields the MML87 estimate $\hat{\theta}_{87}(\mathbf{x}^n|\alpha) = (n+1)/(\sum_{i=1}^n x_i + \alpha)$. The MML87 estimator has the effect of augmenting the data $\mathbf{x}^n$ by one datapoint with a value of α; thus, by choosing $\hat{\alpha} = 1/n \sum_{i=1}^n x_i$ one arrives at essentially the same 'dataset', $\mathbf{x}^{n+1}$, used by the CNML code. Empirical priors such as the one suggested here 'cheat' by peeking at the data and have previously been used when applying MML to mixture modelling and shrinkage estimation [10], [16].

The resulting MML codelength at the estimate $\hat{\theta}_{87}(\mathbf{x}^n|\alpha)$ is

$$I_{87}\left(\mathbf{x}^n, \hat{\theta}_{87}(\mathbf{x}^n|\alpha)\right) = (n+1)\log\left(\sum_{i=1}^{n} x_i + \alpha \right) + n$$
$$+ \frac{1}{2}\log\left(\frac{n}{\alpha^2}\right) - (n+1)\log(n+1) + O(1) \qquad (12)$$

Minimising the message length (12) for α also yields the choice $\hat{\alpha}(\mathbf{x}^n) = 1/n \sum_{i=1}^n x_i$. It is interesting to note that this choice of α renders the MML87 estimator equivalent to the Maximum Likelihood estimator. The final joint message length, $I_{87}\left(\mathbf{x}^n, \hat{\theta}_{\mathrm{ML}}(\mathbf{x}^n)\right)$, of $\mathbf{x}^n$ and $\hat{\theta}_{87}(\mathbf{x}^n|\alpha)$ using $\alpha = \hat{\alpha}(\mathbf{x}^n)$ is

$$n \log \left(\sum_{i=1}^{n} x_i \right) - \left(n - \frac{1}{2} \right) \log n + n + \frac{1}{2}(3 - \log(12))$$

Note that the MML87 codelength is asymptotically equivalent to the CNML code as seen by applying Stirling's approximation to (8) which yields

$$\begin{aligned} I_{\mathrm{CNML}}(\mathbf{x}^n) &\approx n\log\left(\sum_{i=1}^n x_i\right) + (n+1)\log(n+1) \\ &\quad -(2n+1/2)\log n + n - (1/2)\log(2\pi) \end{aligned}$$

The difference $I_{87}(\mathbf{x}^n, \hat{\theta}_{87}(\mathbf{x}^n|\hat{\alpha}(\mathbf{x}^n))) - I_{\mathrm{CNML}}(\mathbf{x}^n)$ approaches $(1/2)(1+\log(\pi/6))$ as $n \to \infty$ which closely matches the approximate bound derived in [10] (pp. 238).

III. Prediction with the sequential NML distribution

The predictive NML density is interesting in its own right, given that it is the predictive distribution of new data x_{n+1} given observed data $\mathbf{x}^n$ that minimises maximum coding regret. The order of the data $\mathbf{x}^n$ has no effect on the predictive distribution (assuming the generating model is i.i.d.), and for the exponential distribution the mean of the predictive NML distribution over x_{n+1} is $\mathrm{E}\left[x_{n+1}\right] = (\sum_{i=1}^n x_i)/(n-1)$. The mean of the predictive distribution obtained by Maximum Likelihood and the CNML predictive distribution are both biased with a tendency to overestimate the true mean. The KL divergence [17] between the predictive NML distribution, for some new data point x_{n+1} having observed data $\mathbf{x}^n$, and a true generating exponential distribution with parameter θ is

$$\begin{aligned} \Delta\left(\theta || p(x_{n+1}|\mathbf{x}^n)\right) &= \log\left(\theta\sum_{i=1}^n x_i\right) - \log n - 1 \\ &\quad +(n+1)\exp\left(\theta\sum_{i=1}^n x_i\right)\mathrm{Ei}\left(\theta\sum_{i=1}^n x_i\right) \qquad (13) \end{aligned}$$

where $\mathrm{Ei}(\cdot)$ denotes the exponential integral. The KL risk for the maximum likelihood plug-in distribution and the SNML predictive distribution is:

$$\begin{aligned} \mathrm{E}\left[\Delta\left(\theta||\hat{\theta}_{\mathrm{ML}}(\mathbf{x}^n)\right)\right] &= \psi(n) + 1/(n-1) - \log n \\ \mathrm{E}\left[\Delta\left(\theta||p(x_{n+1}|\mathbf{x}^n)\right)\right] &= \psi(n) + 1/n - \log n \end{aligned}$$

where $\psi(\cdot)$ is the digamma function. The SNML predictive distribution dominates the ML plug-in distribution in terms of KL risk since $\mathrm{E}[\Delta(\theta||\hat{\theta}_{\mathrm{ML}}(\mathbf{x}^n))] - \mathrm{E}\left[\Delta\left(\theta||p(x_{n+1}|\mathbf{x}^n)\right)\right] \geq 0$ for all n. The SNML predictive distribution is KL consistent as

$$\lim_{n\to\infty}\left\{\psi(n) + 1/n - \log n\right\} = 0$$

It is straightforward to show that (13) asymptotically coincides with the KL divergence of the MML87/ML predictive density.

A. Bayesian Interpretation

It is possible to interpret the predictive SNML distribution within a Bayesian framework. Here, the distribution (6) may be expressed as a Bayesian predictive distribution of x_{n+1} given $\mathbf{x}^n$ using a weighted mixture of exponentials with a gamma mixing function

$$p(x_{n+1}|\mathbf{x}^n) = \int_0^\infty \mathrm{Exp}\left(x_{n+1}|\theta\right)\mathrm{Gam}\left(\theta|n, \sum_{i=1}^n x_i\right)d\theta$$

The gamma mixing function corresponds to a posterior obtained from an exponential likelihood and the Jeffreys' prior $\pi_J(\theta) \propto 1/\theta$. Thus, the predictive SNML distribution is seen to imply a gamma posterior density over the parameter θ.

IV. Conclusion

This paper has derived several information theoretic criteria for the exponential distribution. Under suitable choices of 'hyperparameters' all four criteria yield codelengths within $O(1)$ of each other. The ordering problem inherent in the complete SNML code was removed by constructing a CNML code for n data points conditioned on a single artificial data point x_0. The same approach should be possible for other exponential models that possess sufficient statistics. The CNML code yielded a codelength identical to the BMC, and within $O(1)$ of the MML codelength, with suitable data driven priors. This raises the interesting question whether other statistical models exist for which such a correspondence holds?

Acknowledgements

The authors would like to thank Teemu Roos, the anonymous reviewers and the editor for their helpful comments which substantially improved the paper.

References

[1] J. Rissanen, *Information and Complexity in Statistical Modeling*, 1st ed., ser. Information Science and Statistics. Springer, 2007.

[2] ——, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[3] ——, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1712–1717, 2001.

[4] ——, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[5] T. Roos and J. Rissanen, "On sequentially Normalized Maximum Likelihood models," in *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, Tampere, Finland, 2008.

[6] J. Rissanen and T. Roos, "Conditional NML universal models," in *Proc. 2007 Info. Theory and Applications Workshop*, 2007, pp. 337–341.

[7] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.

[8] C. Wallace and D. Boulton, "An invariant Bayes method for point estimation," *Class. Soc. Bulletin*, vol. 3, no. 3, pp. 11–34, 1975.

[9] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *Journal of the Royal Statistical Society (Series B)*, vol. 49, no. 3, pp. 240–252, 1987.

[10] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, 1st ed., ser. Information Science and Statistics. Springer, 2005.

[11] P. D. Grünwald, *The Minimum Description Length Principle*, ser. Adaptive Communication and Machine Learning. The MIT Press, 2007.

[12] J. Rissanen, "A predictive-least squares principle," *IMA J. Math. Contr. Inform.*, vol. 3, pp. 211–222, 1986.

[13] J. O. Berger and L. R. Pericchi, "Objective Bayesian methods for model selection: introduction and comparison," *Institute of Mathematical Statistics Lecture Notes (Monograph series)*, vol. 38, pp. 135–207, 1997.

[14] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[15] F. Liang and A. Barron, "Exact minimax strategies for predictive density estimation, data compression, and model selection," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2708–2726, November 2004.

[16] E. Makalic and D. F. Schmidt, "Minimum message length shrinkage estimation," *Statistics & Probability Letters*, 2009, (In Press, Corrected Proof), DOI: 10.1016/j.spl.2008.12.021.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.