Minimum Message Length Inference and Mixture Modelling of Inverse Gaussian Distributions

Daniel F. Schmidt and Enes Makalic

Centre for MEGA Epidemiology, The University of Melbourne
Carlton, VIC 3053, Australia
{dschmidt,emakalic}@unimelb.edu.au

Abstract. This paper examines the problem of modelling continuous, positive data by finite mixtures of inverse Gaussian distributions using the minimum message length (MML) principle. We derive a message length expression for the inverse Gaussian distribution, and prove that the parameter estimator obtained by minimising this message length is superior to the regular maximum likelihood estimator in terms of Kullback–Leibler divergence. Experiments on real data demonstrate the potential benefits of using inverse Gaussian mixture models for modelling continuous, positive data, particularly when the data is concentrated close to the origin or exhibits a strong positive skew.

1 Introduction

A common approach to learning structure in complex data is through clustering, or more generally, finite mixture modelling. A finite mixture model with K classes models a probability distribution as

$$p(\mathbf{y}_i; \boldsymbol{\alpha}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) = \sum_{k=1}^{K} \alpha_k p_k(\mathbf{y}_i; \boldsymbol{\theta}_k), \tag{1}$$

where $\boldsymbol{\alpha}$ are the K mixing weights and $\boldsymbol{\theta}_i$ are the vectors of parameters that define the component distributions $p_i(\cdot)$. In general, $\boldsymbol{\alpha}$, K and $\boldsymbol{\theta}_i$ are all unknown and we are required to learn an appropriate mixture model using only the observed data. Unsupervised learning of finite mixture models has been one of the most successful applications of the information theoretic minimum message length (MML) [1–3] principle of inductive inference. Over the past forty years this work has been continuously improved, with refinements to the coding schemes and the addition of new distributions. This paper extends MML mixture modelling further by the inclusion of the inverse Gaussian distribution for positive, continuous data. We say that a variable $Y_i \sim IG(\mu, \lambda)$ if the probability density function for $Y_i = y$ is given by

$$p(y_i; \mu, \lambda) = \left(\frac{1}{2\pi\lambda y_i^3}\right)^{\frac{1}{2}} \exp\left(-\frac{(y_i - \mu)^2}{2\mu^2\lambda y_i}\right), \tag{2}$$

where $\mu > 0$, $\lambda > 0$. The inverse Gaussian is a flexible distribution for modelling continuous, positive data. This work therefore helps to fill a hole in the current MML mixture modelling literature. Previous work [4] has examined modelling this type of data using the gamma distribution. Unfortunately, the treatment of the prior distributions used in [4] is somewhat superficial, and the resulting criterion depends on arbitrarily chosen hyperparameters, which have a crucial effect on the estimation of the number of classes. This paper offers an alternative distribution modelling for continuous, positive data, and uses sensible, data driven choices for the necessary prior distributions. These ideas could easily be further adapted to alternative distributions such as the gamma and Weibull.

2 The Minimum Message Length Principle

Minimum message length (MML) [1, 3] is an information theoretic Bayesian principle for inductive inference. The fundamental idea is that compressing data equates to learning the structure in the data. Theoretical results support the argument that if we can substantially compress the data, then there is a high probability we have learned something about the underlying process that produced the data [5]. In contrast to more traditional statistical procedures for learning, such as those based on hypothesis testing, the MML principle generalises in a straightforward manner to cover estimation of both conventional continuous model parameters in addition to structural parameters, such as the number of components in a mixture model [3].

To learn a model from data $\mathbf{y}$ using MML, we must first posit a countable set of candidate models $\gamma \in \Gamma$, each with associated parameters $\boldsymbol{\theta}_\gamma \in \Theta_\gamma$. We then compare the models in terms of their ability to compress the data. To do this, we view the compressed data as a message composed of two components. The first component, or *assertion*, encodes the details of the model, such as the structural and continuous parameters; the length of the assertion, in base-e digits, is $I(\gamma) + I(\boldsymbol{\theta}_\gamma|\gamma)$. The second component, or *detail*, encodes the data with the aid of the previously stated model, and is of length $I(\mathbf{y}|\gamma, \boldsymbol{\theta}_\gamma)$. The total message length may then be used as a measure of quality of fit of a model to the data, which automatically takes into account the complexity of the model as well as its ability to explain the data. To estimate a model, including structural parameters, from observed data using MML, we solve

$$\left\{\hat{\gamma}_{\mathrm{MML}}(\mathbf{y}), \hat{\boldsymbol{\theta}}_{\mathrm{MML}}(\mathbf{y})\right\} = \underset{\gamma\in\Gamma, \boldsymbol{\theta}\in\Theta_\gamma}{\arg\min} \left\{I(\gamma) + I(\boldsymbol{\theta}_\gamma|\gamma) + I(\mathbf{y}|\boldsymbol{\theta}_\gamma, \gamma)\right\}. \qquad (3)$$

Coding of the structural parameters is straightforward due to the equivalence of discrete codewords and probability mass functions, i.e., $I(\gamma) = -\log \pi_\gamma(\gamma)$, where $\pi_\gamma(\cdot)$ is a suitable prior distribution over Γ. The coding of the continuous parameters assertion is more problematic, as any single point of a probability density function has measure zero. It is therefore necessary to quantise the continuous parameters, rendering them essentially discrete. While there are a variety of ways in which the resulting codelengths can be computed [6–8], if the

model is sufficiently regular it is most convenient to use the Wallace–Freeman (MML87) codelength approximation [7] for models with continuous parameters. For a model with k continuous parameters $\boldsymbol{\theta}_\gamma$, the MML87 codelength for $I_{87}(\mathbf{y}, \boldsymbol{\theta}_\gamma, \gamma) \approx I(\gamma) + I(\boldsymbol{\theta}_\gamma|\gamma) + I(\mathbf{y}|\boldsymbol{\theta}_\gamma, \gamma)$ is

$$-\log \pi_\gamma(\gamma) - \log p(\mathbf{y}|\boldsymbol{\theta}_\gamma, \gamma) - \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\gamma) + \frac{1}{2}\log |\mathbf{J}_n(\boldsymbol{\theta}; \gamma)| + c(k), \qquad (4)$$

with

$$c(k) = -\frac{k}{2}\log 2\pi + \frac{1}{2}\log k\pi + \psi(1),$$

where $p(\cdot)$ is the probability density function for the model, $\pi_{\boldsymbol{\theta}}(\cdot)$ is the prior distribution for $\boldsymbol{\theta}_\gamma \in \Theta_\gamma$, $\mathbf{J}_n(\cdot)$ is the Fisher information matrix for n samples, and $\psi(\cdot)$ denotes the digamma function. Under suitable regularity conditions, the MML87 approximation is within a term of order $o_n(1)$ of the exact message length. An extensive discussion of the MML principle, along with the associated coding procedures, can be found in the excellent book by C. S. Wallace [3].

2.1 Message Lengths of Mixture Models

This section summarises the message length expressions for general finite mixture models. For the purposes of simplicity, we restrict our discussion to the case of univariate data, though the ideas extend in a straightforward manner to the multivariate case. The treatment is necessarily brief, and for a much more complete discussion of the message lengths of mixture models in general, the reader is referred to [3], pp. 275–297.

From (1) it is clear that a mixture model consists of K classes, and models the probability density function of the observed data as a weighted sum of these K classes. We first require some notation. Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the observed data, let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ denote the parameters of the distributions associated with each of the K classes, and recall that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ denotes the mixing weights. An important quantity is the *degree of membership* of each datum to each class. Let $\mathbf{R} \in (0,1)^{n \times K}$ denote the matrix of class memberships. The entries of this matrix are given by

$$r_{i,k} = \frac{\alpha_k\, p(y_i; \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} \alpha_j\, p(y_i; \boldsymbol{\theta}_j)}, \qquad (5)$$

which can be interpreted as the posterior probability of data y_i belonging to class k, treating the mixing weights as *a priori* probabilities of belonging to the K classes. From this quantity we can derive the effective sample sizes associated with each class as

$$n_k = \sum_{i=1}^{n} r_{i,k}. \qquad (6)$$

The totality of parameters for a mixture model of inverse Gaussian distributions is then $\boldsymbol{\Phi} = \{K, \boldsymbol{\alpha}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$. The length of a message that encodes both the data given the mixture model parameters $\boldsymbol{\Phi}$, and the mixture model parameters themselves may be found using the lengths of the following message components:

1. A codeword for K. We chose a uniform distribution over $\{1, \ldots, K_1\}$ so that $I(K) = \log K_1$, with the choice of K_1 being essentially irrelevant.
2. A statement of the parameters $\boldsymbol{\alpha}$. This is done by treating these as the parameters of a multinomial distribution with cell counts n_k, yielding a codeword of length

$$I(\boldsymbol{\alpha}) = \left(\frac{K-1}{2}\right) \log n - \frac{1}{2} \sum_{k=1}^{K} \log \alpha_k - \log \Gamma(K).$$

3. A statement of the model parameters $\boldsymbol{\theta}_k$ for each class and variable. Appealing to the independence arguments in [3] (pp. 291–293) we can decompose the statement of this parameters into a sum of components of length $I(\boldsymbol{\theta}_k)$, the details of which depend on the particular distribution in question. For the inverse Gaussian, these are detailed in Section 3.2.
4. The data, given the previously stated mixture model parameters, which is given by

$$I(\mathbf{y}|\boldsymbol{\Phi}) = -\sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k p(y_i; \boldsymbol{\theta}_k). \tag{7}$$

   One of the most interesting aspects of MML mixture modelling is that the above codelength of the data can be itself be broken down into two parts: a first part, stating which class each data point belongs to, and a second part in which the data is coded using that particular class. Due to the clever way in which the assignment to classes is coded the joint codelength for these two components reduces to (7).

Using these components, the total codelength for the mixture model with parameters $\boldsymbol{\Phi}$ is given by

$$I(\mathbf{y}, \boldsymbol{\Phi}) = I(K) + I(\boldsymbol{\alpha}) + I(\mathbf{y}|\boldsymbol{\Phi}) + \sum_{k=1}^{K} I(\boldsymbol{\theta}_k) + c(d), \tag{8}$$

where $d = (K-1) + \sum_{k=1}^{K} |\boldsymbol{\theta}_k|$ is the total number of continuous parameters in the mixture model. To estimate a mixture model using MML we seek the values of $\boldsymbol{\Phi}$ that minimise (8), which is usually done using the expectation-maximisation algorithm coupled with a suitable non-linear search for the structural components, the details of which lie outside the scope of this paper.

3 MML Inference of Inverse Gaussian Models

To compute message lengths for inverse Gaussian models, and therefore find MML estimates for the parameters μ and λ, we use the MML87 approximation (4). This requires the following ingredients: (i) a likelihood function, (ii) the Fisher information matrix and (iii) appropriate prior distributions. From (2)

it is straightforward to see that the negative log-likelihood of data under an $IG(\mu, \lambda)$ model can be compactly written as

$$-\log p(\mathbf{y}; \mu, \lambda) = \frac{n}{2}\log(2\pi\lambda) + \frac{3}{2}\sum_{i=1}^{n}\log y_i - \frac{n}{\mu\lambda} + \frac{S_1}{2\mu^2\lambda} + \frac{S_2}{2\lambda} \tag{9}$$

where

$$S_1 = \sum_{i=1}^{n} y_i, \quad S_2 = \sum_{i=1}^{n}\frac{1}{y_i},$$

are minimal sufficient statistics for the inverse Gaussian distribution. The maximum likelihood estimates for μ and λ are given by

$$\hat{\mu}_{\mathrm{ML}}(\mathbf{y}) = \frac{S_1}{n}$$
$$\hat{\lambda}_{\mathrm{ML}}(\mathbf{y}) = \frac{S_1 S_2 - n^2}{nS_1}$$

The entries of the Fisher information can be found by noting that $\mathrm{E}[S_1] = n\mu$ and $\mathrm{E}[S_2] = n(1/\mu + \lambda)$. We then have

$$\mathbf{J}_n(\mu, \lambda) = \begin{pmatrix} \frac{n}{\mu^3\lambda} & 0 \\ 0 & \frac{n}{2\lambda^2} \end{pmatrix},$$

and thus

$$|\mathbf{J}_n(\mu, \lambda)| = \frac{n^2}{2\mu^3\lambda^3}. \tag{10}$$

To perform MML inference we need priors on μ and λ. We assume the two parameters are *a priori* independent. We could use the conjugate priors (half-normal for μ, inverse-gamma for λ) [9], but instead choose to use simpler component-wise Jeffreys' priors (i.e., Jeffreys' priors for each parameter, assuming that all other parameters are known). This is the same procedure as is done in the MML treatment of the standard univariate Gaussian distribution. We then have

$$\pi_{\mu,\lambda}(\mu, \lambda) = \pi_\mu(\mu)\pi_\lambda(\lambda) \tag{11}$$
$$\pi_\mu(\mu) = \frac{\sqrt{\mu_0}}{2\mu^{\frac{3}{2}}}, \qquad \mu \in (\mu_0, \infty),$$
$$\pi_\lambda(\lambda) = \frac{1}{\log(\lambda_1/\lambda_0)\lambda}, \quad \lambda \in (\lambda_0, \lambda_1)$$

where sensible, data-driven choices for μ_0, λ_0 and λ_1 will be discussed later.

Substituting (9), (10) and (11) into (4), and minimising for μ and λ yields the MML87 estimates

$$\hat{\mu}_{87}(\mathbf{y}) = \frac{S_1}{n}, \tag{12}$$
$$\hat{\lambda}_{87}(\mathbf{y}) = \frac{S_1 S_2 - n^2}{(n-1)S_1}.$$

It is clear that $\hat{\mu}_{87}(\mathbf{y}) = \hat{\mu}_{\mathrm{ML}}(\mathbf{y})$, and $\hat{\lambda}_{87}(\mathbf{y}) = (n/(n-1))\hat{\lambda}_{\mathrm{ML}}(\mathbf{y})$. Substituting these estimates into the message length yields the minimised message length, $I_{87}(\mathbf{y}, \hat{\mu}_{87}(\mathbf{y}), \hat{\lambda}_{87}(\mathbf{y}))$:

$$\left(\frac{n-1}{2}\right)\log\left(2\pi e\hat{\lambda}_{87}(\mathbf{y})\right) + \frac{3}{2}\sum_{i=1}^{n}\log y_i + \log\left(\frac{\sqrt{2}\,n\log(\lambda_1/\lambda_0)}{\sqrt{\mu_0}}\right) + \psi(1). \quad (13)$$

It is clear that the choice of the prior hyperparameters $(\lambda_0, \lambda_1, \mu_0)$ has no effect on the MML estimators of the μ and λ parameters. However, in the setting of mixture modelling, in which a model can potentially comprise several inverse Gaussian distributions, the choice of these hyperparameters will have a crucial effect on the message length. Section 3.2 addresses the use of inverse Gaussian distributions in the mixture modelling setting, and details a data driven way of selecting these hyperparameters.

3.1 Behaviour of the MML Estimates

Let μ^* and λ^* denote the true parameter values, i.e., $y_1, \ldots, y_n \sim IG(\mu^*, \lambda^*)$. It is well known that

$$\hat{\mu}_{\mathrm{ML}}(\mathbf{y}) \sim IG\left(\mu^*, \frac{n}{\lambda^*}\right), \quad (14)$$

and it follows immediately that $\mathrm{E}\left[\hat{\mu}_{\mathrm{ML}}(\mathbf{y})\right] = \mathrm{E}\left[\hat{\mu}_{87}(\mathbf{y})\right] = \mu^*$, i.e., both ML and MML87 yield unbiased estimates of μ^*. To explore the behaviour of estimates of λ^* we use the fact that

$$\left(\frac{n}{\lambda^*}\right)\hat{\lambda}_{\mathrm{ML}}(\mathbf{y}) \sim \chi^2_{n-1}, \quad (15)$$

where χ^2_ν denotes a chi-squared distribution with ν degrees of freedom. Using this result, along with the fact that $\hat{\lambda}_{87}(\mathbf{y}) = (n/(n-1))\hat{\lambda}_{\mathrm{ML}}(\mathbf{y})$, it is straightforward to show that

$$\mathrm{E}\left[\hat{\lambda}_{\mathrm{ML}}(\mathbf{y})\right] = \left(\frac{n-1}{n}\right)\lambda^*, \quad \mathrm{E}\left[\hat{\lambda}_{87}(\mathbf{y})\right] = \lambda^*. \quad (16)$$

The maximum likelihood estimator exhibits a downward bias, while the MML87 estimator is unbiased. These results closely parallel those found in the case of the usual univariate Gaussian distribution.

Measures of estimator quality such as bias and expected squared error suffer from the fact that they are parameterisation dependent. This issue can be circumvented by examining the behaviour of the estimators in terms of measures that are invariant under reparameterisations. A common choice is the Kullback–Leibler (KL) [10] divergence. The KL divergence from the true, generating $IG(\mu^*, \lambda^*)$ to an approximating $IG(\hat{\mu}, \hat{\lambda})$ is

$$\Delta(\mu^*, \lambda^* || \hat{\mu}, \hat{\lambda}) = \frac{1}{2}\log\left(\frac{\hat{\lambda}}{\lambda^*}\right) + \left(\frac{1}{\hat{\lambda}}\right)\left(\frac{\lambda^*}{2} + \frac{1}{2\mu^*} + \frac{\mu^*}{2\hat{\mu}^2} - \frac{1}{\hat{\mu}}\right) - \frac{1}{2}. \quad (17)$$

Estimators may be assessed in terms of their expected KL divergence, or *KL risk*, for a particular sample size n. Let $R_n(\hat{\mu}, \hat{\lambda}) \equiv \mathrm{E}[\Delta(\mu^*, \lambda^* || \hat{\mu}(\mathbf{y}), \hat{\lambda}(\mathbf{y}))]$ denote the KL risk for sample size n, the expectation being taken with respect to the generating model $IG(\mu^*, \lambda^*)$. It is of interest to compare the KL risks of the ML and MML87 estimators. This is done by examining the difference in KL risks. The risk difference is given by

$$R_n(\hat{\mu}_{\mathrm{ML}}, \hat{\lambda}_{\mathrm{ML}}) - R_n(\hat{\mu}_{87}, \hat{\lambda}_{87}) = \left(\frac{3\mu^*\lambda^* + n^2 + n}{2(n-3)n^2} \right) + \frac{1}{2} \log \left(\frac{n-1}{n} \right),$$

which is strictly greater than zero for all $n > 3$. This shows that for all $n > 3$, *the MML87 estimator has strictly lower KL risk than the maximum likelihood estimator*, irrespective of the model (μ^*, λ^*) that generated the data. In the case that the sample size $n \leq 3$, it turns out that both MML87 and ML estimators have infinite KL risk, and neither is demonstrably more accurate in terms of KL divergence.

3.2 Inverse Gaussian Distributions in MML Mixture Models

The minimised message length the for inverse Gaussian models given by (13) is not exactly appropriate for the mixture model case, as it is based on complete membership of all the data to a single inverse Gaussian model. It is, however, straightforward to adapt the message length expressions to the mixture setting by appealing to the independence arguments outlined by Wallace in [3]. It can be shown that the appropriate negative log-likelihood for the k-th inverse Gaussian component in a mixture model is

$$\frac{n_k}{2} \log(2\pi\lambda) + \frac{3}{2} \sum_{i=1}^{n} r_{i,k} \log y_i - \frac{n_k}{\mu\lambda} + \frac{S_{k,1}}{2\mu^2\lambda} + \frac{S_{k,2}}{2\lambda} \tag{18}$$

where

$$S_{k,1} = \sum_{i=1}^{n} r_{i,k}\, y_i, \quad S_{k,2} = \sum_{i=1}^{n} \frac{r_{i,k}}{y_i},$$

are the appropriately weighted sufficient statistics. Due to the form of (18), the Fisher information is simply given by $\mathbf{J}_{n_k}(\mu_k, \lambda_k)$, with n_k being the effective sample size for class k given by (6). The prior distributions remained unchanged, and the MML estimates become

$$\hat{\mu}_k(\mathbf{y}) = \frac{S_{k,1}}{n_k},$$
$$\hat{\lambda}_k(\mathbf{y}) = \frac{S_{k,1}S_{k,2} - n_k^2}{(n_k - 1)S_{k,1}}.$$

As was previously noted, the choice of the prior hyperparameters μ_0, λ_0 and λ_1 becomes an issue in the mixture model setting, as the particular values chosen will have a crucial effect on the estimate of the number of classes unless the

sample size is very large. To solve this problem, we use two simple, data driven choices for the hyperparameters. The μ_0 hyperparameter sets the lower-bound on the prior density for μ. From the form of the MML estimator $\hat{\mu}_k(\mathbf{y})$ we can easily determine that the smallest value the estimate may assume is equal to the smallest data point in $\mathbf{y}$. Therefore, we have

$$\hat{\mu}_0 = \min_i \{y_i\}.$$

The form of the estimate for $\hat{\lambda}_k(\mathbf{y})$ is complex, and determination of the smallest and largest values it may assume, given a particular dataset $\mathbf{y}$ is computationally intensive. To avoid this problem, we instead use the simple idea proposed by Rissanen to deal with the similar problem of infinite parametric complexity [11]. This involves setting $\lambda_0 = e^{-a}$, $\lambda_1 = e^a$, with $a \in \{1, 2, \ldots\}$ the smallest positive integer such that

$$e^{-a} \leq \lambda_k \leq e^a, \quad k = 1, \ldots, K.$$

Using this choice of priors, the quantity needed for mixture modelling is

$$I(\mu_k, \lambda_k) = \log n_k - \frac{1}{2} \log \hat{\lambda}_k(\mathbf{y}) + \log \left(\frac{2\sqrt{2}a}{\sqrt{\hat{\mu}_0}} \right). \tag{19}$$

For the resulting codelength of the entire mixture model to be valid it must also include the length of the codewords needed to state the hyperparameters $\hat{\mu}_0$ and a. The total codelength of a mixture model for inverse Gaussian distributions, using these empirical priors, then becomes

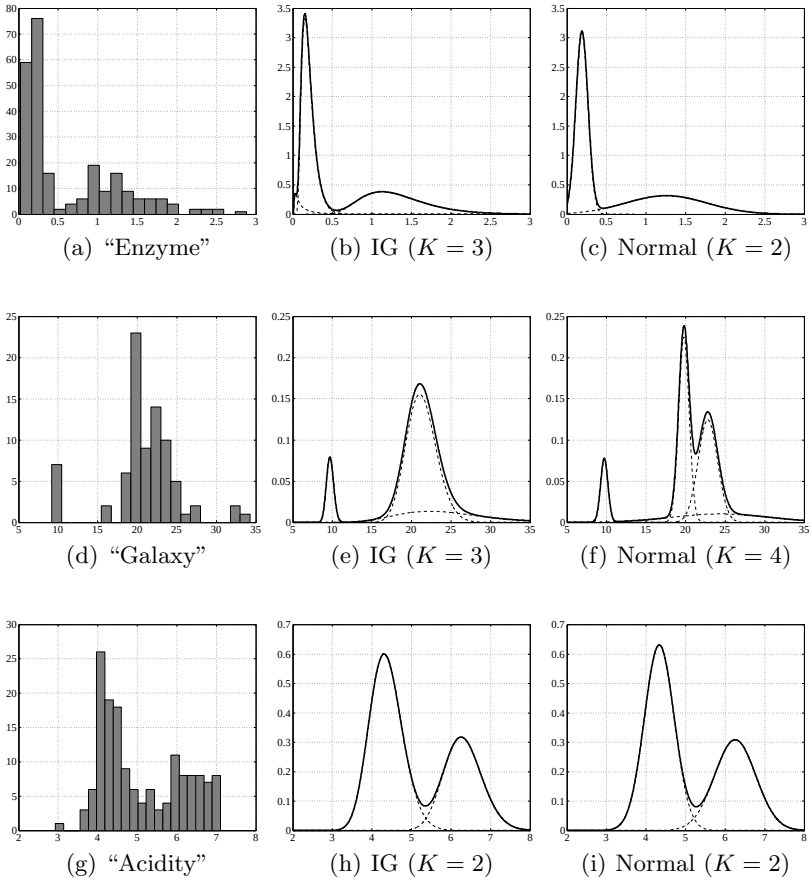
$$I(\mathbf{y}, \boldsymbol{\Phi}, a, \hat{\mu}_0) = I(\mathbf{y}, \boldsymbol{\Phi}) + I(\hat{\mu}_0) + I(a).$$

where $I(\mathbf{y}, \boldsymbol{\Phi})$ is given by (8). We note that $\hat{\mu}_0$ is a continuous parameter, and may be stated with a codelength of $I(\hat{\mu}_0) \approx (1/2) \log n$. The hyperparameter a is a positive integer, and following [11], we code this using the log-star code for integers, yielding a codelength of $I(a) \approx \log_*(a) + 2.86$, where $\log_*(x) = \log x + \log \log x + \ldots$, the logarithms iterating until they become negative.

4 Experiments

There have been a large number of previous simulation studies conducted demonstrating that MML is, in general, superior to commonly used asymptotic techniques such as Akaike's information criterion (AIC) [12] or the Bayesian information criterion (BIC) [13] in the context of estimating a finite mixture model (for example, [4, 14]). Given that the inverse Gaussian model satisfies the conditions for the MML87 approximation, and therefore yields sensible codelengths, there is no compelling reason to expect any significant departure from this trend.

Therefore, we conclude the paper by comparing MML inverse Gaussian mixture modelling against regular MML univariate Gaussian mixture modelling on the three real datasets: (i) "Enzyme", (ii) "Acidity", and (iii) "Galaxy" (see [15]).



(a) "Enzyme" (b) IG ($K = 3$) (c) Normal ($K = 2$)

(d) "Galaxy" (e) IG ($K = 3$) (f) Normal ($K = 4$)

(g) "Acidity" (h) IG ($K = 2$) (i) Normal ($K = 2$)

Fig. 1. Mixture Modelling using Inverse Gaussian (IG) and Normal Distributions

All three datasets are composed of non-negative data, and their histograms suggest that multiple modes are a possibility. For each dataset, two mixture models were estimated, one using inverse Gaussian distributions, and a second using univariate Gaussian distributions, univariate Gaussian mixture modelling being the standard approach to clustering of continuous variables used by most software packages. The histograms of the datasets, along with plots of the estimated Gaussian and inverse Gaussian mixture models are presented in Figure 1. The differences between the estimated models for each of the datasets are summarised below:

– **"Enzyme"**: For this dataset the advantage obtained by using a positively skewed distribution for positive, continuous data was substantial. The estimated Gaussian mixture model was composed of two classes and had a total message length of 86.19 nits (base-e digits), while the inverse Gaussian

mixture model was composed of three classes with a total message length of 69.34 nits. The large difference in message lengths suggests that the inverse Gaussian model offers a significantly better fit to the data than the regular Gaussian model; this is primarily due to the clustering of data near the origin, as well as the positively skewed nature of the data further away from the origin, both of which cause problems for Gaussian distributions.

– **"Galaxy"**: The estimated Gaussian mixture model contained four classes with a total message length of 236.16 nits, and the inverse Gaussian model was composed of 3 classes, with a message length of 235.5 nits. Both mixture models identify the small peak around $y = 10$ as a separate class, but differ in the way that they model the large cluster from $y = 15$ to $y = 30$. The inverse Gaussian mixture model has identified this cluster as unimodal, while the Gaussian mixture model has split the cluster into two separate classes. The difference in message lengths indicates a slight preference for the inverse Gaussian explanation, but it is not great enough to make any conclusive decision.

– **"Acidity"**: The estimated Gaussian mixture model was composed of two classes with a total message length of 209.4 nits, and the estimated inverse Gaussian mixture model also contained two classes, with a total message length of 210.57 nits. From Figure 1 we can see that the data is not close to the origin and appears to be reasonably tightly clustered around $y = 3$ through $y = 7$. Both mixture models are very similar, the primary difference being the height and width of the first peak. This similarity is also mirrored in the message lengths of the two models, which are very close, the Gaussian mixture model being slightly preferred.

The above analyses suggest that the mixture modelling with inverse Gaussian distributions can lead to big improvements over regular Gaussian mixture modelling if the data exhibit positive skewness or are clustered close to the origin. In both of these cases, the regular Gaussian distribution, being symmetric and defined over the entire real line, will be unable to provide an excellent fit to the data. Of course, the strength of MML mixture modelling is that the message length is comparable between mixture models of different distributions. This highlights an important property of MML mixture modelling: for a given dataset, we may use the message length to not only to select the number of classes, but also to select an an appropriate distribution for the classes themselves.

References

1. Wallace, C.S., Boulton, D.M.: An information measure for classification. Computer Journal **11**(2) (August 1968) 185–194
2. Wallace, C.S., Dowe, D.L.: MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. Statistics and Computing **10**(1) (January 2000) 73–83

3. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. First edn. Information Science and Statistics. Springer (2005)
4. Agusta, Y., Dowe, D.: Unsupervised learning of gamma mixture models using minimum message length. In Hamza, M., ed.: Proceedings of the 3rd IASTED conference on Artificial Intelligence and Applications, Benalmadena, Spain, ACTA Press (September 2003) 457–462
5. Grünwald, P.D.: The Minimum Description Length Principle. Adaptive Communication and Machine Learning. The MIT Press (2007)
6. Wallace, C., Boulton, D.: An invariant Bayes method for point estimation. Classification Society Bulletin **3**(3) (1975) 11–34
7. Wallace, C.S., Freeman, P.R.: Estimation and inference by compact coding. Journal of the Royal Statistical Society (Series B) **49**(3) (1987) 240–252
8. Schmidt, D.F.: A new message length formula for parameter estimation and model selection. In: Proc. 5th Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-11). (2011)
9. Banerjee, A.K., Bhattacharyya, G.K.: Bayesian results for the inverse gaussian distribution with an application. Technometrics, **21**(2) (1979) 247–251
10. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1) (March 1951) 79–86
11. Rissanen, J.: Fisher information and stochastic complexity. IEEE Transactions on Information Theory **42**(1) (January 1996) 40–47
12. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6) (December 1974) 716–723
13. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2) (1978) 461–464
14. Bouguila, N., Ziou, D.: High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(10) (2007) 1716–1731
15. Richardson, S., Green, P.J.: On bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society. Series B (Methodological) **59**(4) (1997) 731–792