Lecture Notes in Artificial Intelligence (LNAI) 1394, pp 96-109, Springer (1998).

Single Factor Analysis in MML Mixture Modelling

Russell T. Edwards[1,2] and David L. Dowe[2]

[1] Supercomputing and Astrophysics Research Group, Swinburne University of Technology, John St, Hawthorn, Victoria 3122, Australia
redwards@pulsar.physics.swin.edu.au,
[2] School of Computer Science and Software Engineering, Monash University, Wellington Rd, Clayton, Victoria 3168, Australia
dld@cs.monash.edu.au

Abstract. Mixture modelling concerns the unsupervised discovery of clusters within data. Most current clustering algorithms assume that variables within classes are uncorrelated. We present a method for producing and evaluating models which account for inter-attribute correlation within classes with a single Gaussian linear factor. The method used is Minimum Message Length (MML), an invariant, information-theoretic Bayesian hypothesis evaluation criterion. Our work extends and unifies that of Wallace and Boulton (1968) and Wallace and Freeman (1992), concerned respectively with MML mixture modelling and MML single factor analysis. Results on simulated data are comparable to those of Wallace and Freeman (1992), outperforming Maximum Likelihood. We include an application of mixture modelling with single factors on spectral data from the Infrared Astronomical Satellite. Our model shows fewer unnecessary classes than that produced by AutoClass (Goebel et. al. 1989) due to the use of factors in modelling correlation.

Keywords: Minimum Message Length, MML, Statistical and Machine Learning, Noise Handling, Induction in KDD.

1 Introduction

This paper introduces a method for inferring models of data based on the combination of two kinds of model frameworks: mixture models [13, 16, 14, 18] and factor models [7, 20, 15]. Mixture modelling, variously also known as a form of clustering, intrinsic classification, and numerical taxonomy, is the modelling of a statistical distribution by a mixture of other distributions, known as components or classes. Mixture modelling is prone to overfitting since the best fit to any body of data, in terms of maximum likelihood, is typically achieved by having many classes, each fitting a small number things [1]. The Minimum Message Length (MML) principle, presented by Wallace and Boulton [16], is a Bayesian method

[1] where a 'thing' is a datum of observed attribute values

which penalises overly complex hypotheses, and for this reason it has been used for the evaluation of mixture model hypotheses [16, 14, 18][2].

Factor analysis concerns the modelling of inter-attribute correlation by the assertion of the existence of some attribute of things which was not measured, but which has an effect on the attributes that were observed. The MML estimator for factors [20, 15] produces results that are superior on average to maximum likelihood estimators.

Real-world data often exhibits correlation structure which is well suited to modelling by a mixture of distributions with factors. We have developed an MML method for the estimation of the parameters and structure of this class of models. We present an analysis of the results of this estimator on simulated data and on spectral data from the Infrared Astronomical Satellite (IRAS) [12], which shows the value of the inclusion of factors in mixture modelling. Spectral data contains a large amount of correlation, much of which corresponds to the variation in strength of continuous parameters of sources, such as their temperature or the abundance of certain ions and molecules. This kind of variation may be handled by mixture models (Goebel et. al. [6] have published such a model for the IRAS data), but we believe it is better modelled by mixtures of factors, resulting in a much reduced number of classes.

2 Parameter Estimation by Minimum Message Length

The minimum message length (MML) principle as stated by Wallace and Boulton [16] asserts that the "best" hypothesis about data is that which minimizes the length of a two part message conveying the hypothesis and encoding the data given this hypothesis. Elementary information theory results tell us that an event of probability p may be encoded in $-\log_2 p$ bits. Therefore, assuming a prior distribution on hypotheses, we may encode our hypothesis H in length $-\log_2 \Pr(H)$ bits, or $-\log_e \Pr(H)$ nits, a unit of convenience since we are often taking the logarithm of exponentials of base e. Given the hypothesis, which somehow conveys a probability distribution for the data (by, for example, encoding the mean and standard deviation for a Normal distribution), the data are encoded in $-\log \Pr(D|H)$ nits. Hence, the entire message is conveyed in $-\log \Pr(H) - \log \Pr(D|H) = -\log \Pr(H).\Pr(D|H)$ nits, and the MML estimate minimizes this quantity, or equivalently, maximizes $\Pr(H).\Pr(D|H)$. By Bayes' rule this corresponds to maximizing $\Pr(D).\Pr(H|D)$, and since $\Pr(D)$ is independent of H, the MML estimate therefore maximizes the Bayesian posterior probability $\Pr(H|D)$.

Unlike the usual Bayesian method of maximizing the posterior density, MML works with probability masses by recognizing that all data is necessarily measured to a finite precision, and that point estimates should also be recorded to a finite precision. Thus the precision of all distribution parameters, as well as their expected values under quantization are estimated. The precision of distribution

[2] More information regarding mixture modelling with MML and other methods is available at `http://www.cs.monash.edu.au/~dld/mixture.modelling.page.html`.

parameters is determined as a trade-off between the cost of encoding with extra precision, and the cost of encoding the data with a sub-optimal hypothesis due to rounding. For continuous-valued parameters, with the optimum quantizing volume (to a second order quadratic Taylor expansion approximation), the hypothesis is encoded in $-\log \frac{h(\boldsymbol{z})}{\sqrt{k_D^D F(\boldsymbol{z})}}$ nits, where $h(\boldsymbol{z})$ is the prior density at the D-dimensional parameter vector estimate $\boldsymbol{z}$, k_D is the optimal D-dimensional lattice constant and $F(\boldsymbol{z})$ is the determinant of the matrix of expected second partial derivatives of the negative log-likelihood with respect to the elements of $\boldsymbol{z}$ (known as the Fisher information). Given the hypothesis $\boldsymbol{z}$ encoded to this precision, the second part of the message incurs an overhead due to rounding of the parameters of $D/2$ nits, giving a length of $-\log \Pr(D|\boldsymbol{z}) + D/2$. The MML estimate of $\boldsymbol{z}$ is that which minimizes the length of this two part message. Unlike the Bayesian maximum a posteriori estimate which maximizes the posterior density, the MML estimator is invariant under re-parameterisation. For a more detailed treatment see [19] or [17]. For estimators of the parameters of specific probability distributions, see [18].

3 Mixture Modelling by Minimum Message Length

3.1 Fundamentals

The original application of MML by Wallace and Boulton [16] was in mixture modelling. Recall that MML involves the evaluation of the length of a hypothetical message describing a hypothesis and encoding the data based on this hypothesis. For mixture models, we assume a message composed of the following parts:

1a. The number of components. (All numbers are considered equally likely a priori, although this could easily be modified.)

1b. The relative abundance of each component. (Creating names or labels for each component of length $-\log_2$ of the relative abundance, via a Huffman code, gives us a way of referring to components later when, e.g., we wish to say which component a particular data thing belongs to.)

1c. For each component, the distribution parameters of the component

1d. For each thing, the component to which it is estimated to belong. (This can be done using the Huffman code referred to in 1b above.)

2. The attribute values of each thing in turn, encoded using parameters of the hypothesis.

Given an assignment of things to classes, this message format allows the independent estimation of the distribution parameters of each class and the relative abundances of each class to minimize parts 1a-c of the message. Likewise, given distribution parameters and relative abundances, it allows the independent assignment of each thing to a class to minimize parts 1d and 2 of the message.

3.2 Snob and the EM Algorithm "Adjust Cycle"

The Snob [16, 14, 18] program for MML mixture modelling uses a version of the Expectation Minimization (EM) algorithm [10] to find a model with locally minimum message length. Given some initial number of classes and assignment of things to classes, the *first step* is to estimate the distribution parameters which minimize the length of the parts of the hypothesis and the data which pertain to each attribute of each class. Given these estimates for the parameters of each class, the *second step* is to re-assign things to classes in a manner which minimizes the combined length of parts 1d and 2. These two steps are repeated until the message length has reached convergence, at which point a local minimum of message length must have been found.

The topology of the mixture model is evolved by considering at each step of iteration the effect of either promoting the iterated *sub-classes* of a class, or *combining* two classes into one, and performing the alteration if it would decrease the message length.

3.3 Partial Assignment

Part 1d of the message described in the previous section implicitly restricts us to hypotheses, H, which assert with 100% definiteness which component each thing belongs to. Given that the population that we might encounter could consist of two different but highly over-lapping distributions, forcing us to state definitely which component each thing belongs to is bound to cause us to mis-classify outliers from one distribution as belonging to another. In the case of two over-lapping (but distinguishable) 1-dimensional Normal distributions, this would cause us to over-estimate the difference in the component means and under-estimate the component standard deviations. Since what we seek is a message which enables us to encode the attribute values of each thing as concisely as possible, we note that a shorter message can be obtained by a probabilistic (or partial) assignment of things to components. The reason for this is that[14, p77] if $p(j, x), j = 1, ..., J$, is the probability of component j generating datum x, then the total assignment of x to its best component results in a message length of $-\log(\max_j p(j, x))$ to encode x whereas, letting $P(x) = \sum_j p(j, x)$, a partial assignment of x having probability $p(j, x)/P(x)$ of being in component j results in a shorter message length of $-\log(P(x))$ to encode x. As shown in [14, p77], this shorter length is achievable by a message which asserts definite membership of each thing by use of a special coding trick.

4 Single Factor Analysis by Minimum Message Length

Classification is a way of modelling inter-attribute correlation. Whilst mixtures of uncorrelated multivariate distributions are able to model any form of inter-attribute correlation, they do so at a cost if the correlation does not arise due to actual homogeneous unknown sub-populations in the things to be modelled.

Typically in real-world data we might expect to find correlation structure of a continuous kind rather than just the disjoint style of mixture models. One way of modelling this is to hypothesize the existence of some continuous-valued attribute of things, which was not measured, and which represents a common "factor" affecting the attribute values that *were* measured. Hence, the attribute values of a thing, $\boldsymbol{x_n}$ where $x_{nk} \in \mathbb{R}$ are modelled by

$$x_{nk} = \mu_k + v_n a_k + \sigma_k r_{nk} \tag{1}$$

and v_n and r_{nk} are independently distributed from $N(0,1)$ for $1 \leq n \leq N$ and $1 \leq k \leq K$, where N is the number of things and K is the number of attributes per thing. The r_{nk} term represents the usual model of uncorrelated Gaussian variance, whilst $v_n a_k$ models the effect of the common factor: v_n is the *factor score* of thing n and a_k is the *factor load* for attribute k. Thus the effect of the factor on the value of attribute k of thing n is characterised by the factor score a_k with which it affects the kth attribute values of all things, and the factor load v_n with which all the attribute values of thing n are affected by the factor.

One example of modelling by factors is the concept of the Intelligence Quotient (IQ). IQs by definition are Normally distributed among the world's population (or, at least, among the people who have had IQ tests!) with a standard mean and variance. A person's IQ is estimated from their results on one or more IQ tests, for it is assumed that intelligence is a monotonic function of score on such tests. In terms of factors, a person's IQ is the factor score, and the factor loads embody what each intelligence test tells us about IQ, whilst the person's results at different tests would comprise their attributes. A high factor score would indicate high IQ, and one would expect that people with high IQ would perform well on most intelligence tests. A large positive factor load would indicate a test which is highly sensitive to intelligence in terms of the variation in its results, whilst a zero factor load would indicate a test which produces results that are totally uncorrelated to the intelligence of the person taking it.

For this kind of data, a mixture model must become overly complex in order to achieve a good fit — the usual result is that for each true class in the data (there may be just one), a mixture model produces several components in order to cover it fully. This effect is illustrated in figure 1, which depicts some two-dimensional data generated with a single class containing a very strong factor ($|\boldsymbol{a}|/|\boldsymbol{\sigma}| = 5$), and the probability density assigned to the data space by the MML mixture model of this data. The mean of the data appears around the middle of the plane, with a factor load vector of the same sign and (as plotted) equal magnitude in both attributes, as evidenced by the spread of data along the diagonal. The sign of $\boldsymbol{a}$ is arbitrary, but assuming a_1 and a_2 are positive, data points toward the upper right side of the plane would have large positive factor scores. Note that the optimal mixture model of this data produced three classes dotted along the factor to account for the correlation, where clearly there is only one intrinsic population in the data. This process of assigning multiple classes to account for continuous variation is similar to the way some human-designed classification schemes cope with correlation: a number of sub-classes

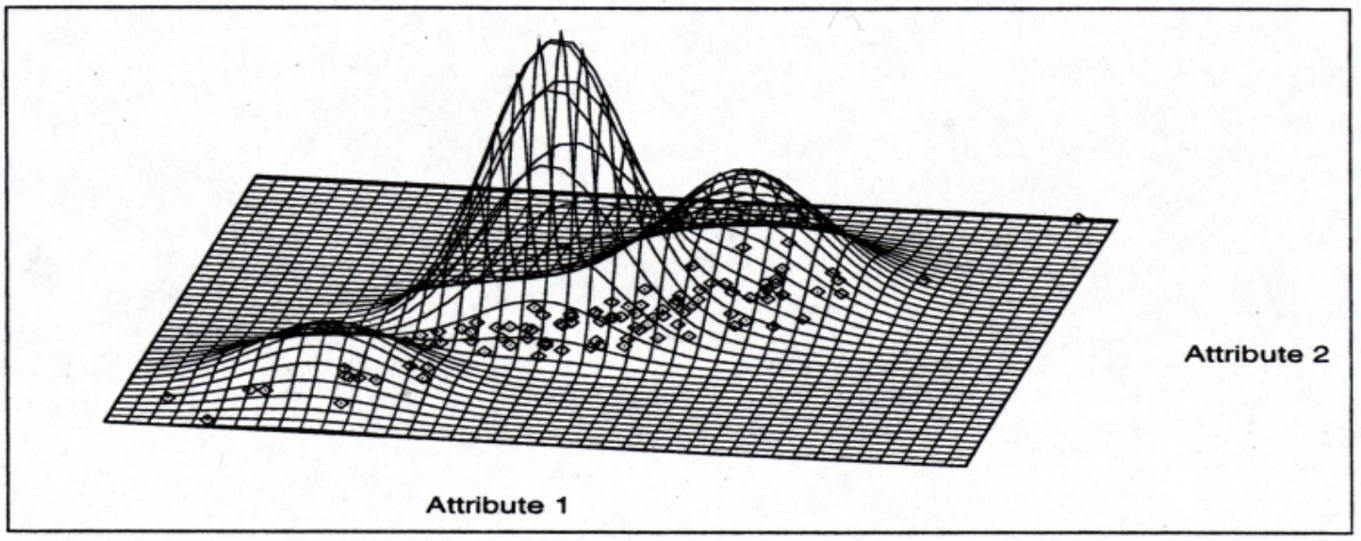

Fig. 1. Pseudo-randomly generated data with a strong Gaussian factor. The point s (which are on the $z = 0$ plane) show the data points, whilst the surface shows the probability density over attribute values, assigned by the MML mixture of three Gaussian classes.

are produced for every real class, with each sub-class (e.g. low, medium, high) spanning a certain range of data corresponding to different strengths of the factor (factor scores). Since we are interested in inferring something about the true class structure of the data (which in this case means only a single class, not three), it is better to model linear correlation directly.

An MML estimator for single factor models of the form in equation 1 was developed by Wallace and Freeman [20]. This estimator allows simultaneous estimation of the factor loads and scores, unlike maximum likelihood estimators, which cannot do so in a consistent manner. In addition, where there was evidence for a common factor this estimator is more accurate than the Maximum Likelihood estimators, and where there is insufficient evidence for a factor, the uncorrelated multivariate Gaussian model is reverted to by virtue of its smaller message length. Wallace [15] has also derived an MML estimator for models involving multiple factors.

5 MML Mixture Modelling with Single Factors

It is the nature of many mixture modelling applications that the data would be well modelled by a set of classes with the presence of factors with each class, or perhaps, shared across classes. We have developed a method for MML estimation of such a model.

Recall the mixture model message format described in section 3.1. The amendment of this format to allow factor models is conceptually simple: for part 1c we first encode the distribution parameters of any attributes not to be modelled as Gaussian (for example, discrete values [16], angles [17], non-negative reals). We then specify whether or not the remaining attributes are modelled with a factor. If the two choices are equally likely a priori, this specification is of

constant message length and may be ignored for the purposes of model discrimination within this framework. If a factor is not used, the means and standard deviations are encoded as usual, otherwise the factor parameters, consisting of the means, standard deviations, loads and scores, are encoded. Other parts of the message remain in the same format as before.

5.1 Modifications for Mixture Models Incorporating Factors

The search strategy used by Snob and described in section 3.1 works on the assumption that step two of the adjust cycle (section 3.2), the re-assignment of things to classes, will not affect the length of parts 1a-1c of the message. This appears not to be strictly true since in order to be able to send the second part of the message in a length equal to the negative log likelihood plus the small constant rounding costs (section 2), the estimates must be the MML estimates, encoded to the correct precision. Just as in section 2 it was assumed that the hypothesis cost varies little with the rounding of estimates, here it is assumed that the length of the hypothesis is a slowly varying function of the data and thus that any corrections that ought to be made to allow costing of the second part of the message as described, would have negligible effect on the length of parts 1a-1c. However, a small change in the number of things to be modelled by a factor (as occurs in step two of the adjust cycle) results in a large change in the cost of the hypothesis, as there will be a different (possibly larger or smaller) set of factor scores to encode.

When re-assigning a thing to a different class, we must do so based on a knowledge of not only how this will affect the length of parts 1d and 2 of the message, but seemingly also on how this will affect the length of the new part 1c which must arise as a result of re-assignment. Mainly, we must take into account the difference between the cost of describing the factor score of this thing in its new class compared to its old class, which may be large, especially if one of the classes is not modelled with a factor, meaning no factor score need be specified. If this cost is not properly taken into account, the message length can actually increase over successive iterations of the adjust cycle.

The Fisher information matrix for factor models is not diagonal in the elements involving the factor scores, which means that in transformation of the parameter space to allow optimal independently quantised coding [19], the factor scores are not separable from some other parameters. This means that we cannot simply attribute a single cost for the specification of a factor score, for use in re-assignment as described above. Recall from section 3.2 that step two of the adjust cycle for mixtures of uncorrelated distributions, re-assignment of things to classes in a manner which minimizes the lengths of parts 1d and 2 (and now, 1c) may by done optimally by considering the class assignment of each thing independently from that of every other thing. This is not true for mixtures of factors, because of the impossibility of apportioning the class-assignment-dependent parts of 1c (the specification of factor scores) to the culprit things. As a result of this, it seems the only way to perform step 2 of the adjust cycle optimally whilst taking into account the effects re-assignment has on the length of part 1c is to