find an overall assignment of all things to classes at once which minimizes the new lengths of parts 1c, 1d and 2, given that all estimates in 1c are to remain the same, apart from the factor scores.

As a result of these implications of the factor model, optimal re-assignment in step 2 of the adjust cycle appears to become an infeasible combinatorial search problem of exponential complexity, as compared to the usual Snob adjust cycle with its independent assignment of things in linear time. Considering the intractability of such a search, we have devised an iterative hill-climbing scheme. This scheme assigns one thing at a time by considering the effect the assignment would have on parts 1c, 1d and 2. This process is applied repetitively to random things, meaning that the assignment of things is continually updated in random order (to avoid bias), until a locally minimal assignment is found. For all moderate to large datasets this process becomes too slow and instead each thing is re-assigned only once, which is equivalent to considering what effect the assignment would have if the thing was the only thing whose assignment changes. This is not optimal except when the Snob adjust cycle has converged, but in practice it appears to always reduce the message length.

5.2 Partial Assignment with Factors

The method of re-assignment described in the previous section assumes total assignment, however it is desirable to use partial assignment to produce a consistent estimator (see section 3.3). The coding trick for partial assignment does not appear to be compatible with the scheme described in the previous section, for it requires that we be able to cost, or estimate the cost, of the parts of 1c, 1d and 2 that are attributable to each thing. In the previous section we calculated only the overall change in cost of 1c for a simultaneous assignment of all things to classes.

As noted in section 5.1, there appears to be no way of separating the costs of individual factor scores from the other factor parameters, and hence there is no way to distribute the cost as this scheme would require. Wallace [15] uses a polar representation for the factor score vectors in models involving multiple factors, which may allow independent coding.

5.3 Relationship to Previous Work

Hinton et. al [8] present an interesting method for fitting mixtures of single factors, based on a type of neural network called an "autoencoder" [9]. The network attempts to duplicate the input data at its output. A model is fitted essentially by minimizing the description length of a two-part message consisting of the factor scores and the differences between the input and output. Since it ignores the cost of all other parameters (e.g. the distribution parameters of each class, which should penalize mixtures of too many components), and encodes the output errors according to a fixed, empirical prior, this method appears equivalent to Maximum Likelihood, marginalized over factor scores [20] and assuming a prior value for σ. Maximum Likelihood is known to over-fit in problems where

the number of parameters to be estimated grows rapidly [11, 5, 20, 15], in fact it is the conjecture of Dowe [4] that no non-Bayesian method can be always invariant and statistically consistent while providing internally consistent parameter estimates. As a result of the reliance on Maximum likelihood, we suspect that simulation results would demonstrate over-fitting in the method of Hinton et. al.

6 Tests and IRAS LRS Spectral Classification Results

As a first step we have implemented mixture models with a single factor per class using total assignment, as described in section 5.1. The search process was based on that of Snob, with the addition of new heuristics for formation of initial models and sub-classes[3].

6.1 Tests and Results on Simulated Data

We generated data based on the parameters used by Wallace and Freeman [20]. There were three trials of 1000 data sets of 200 observations per dataset, with five attributes per observation. Each data set consisted of 100 observations from a class with means μ_k all equal to 4 and 100 observations from a class with means μ_k all zero. All standard deviations σ_k were unity , and all factor load vectors $\boldsymbol{a}$ were parallel to $(2, 3, 4, 5, 6)$. In all data in each of the three trials, the load vector lengths $|\boldsymbol{a}|$ were 1.5, 1.25 and 1.0 respectively, representing strong, weak, and barely detectable factors. These classes are moderately well separated, their variances in the fifth attribute being approximately 2 in the case of $|\boldsymbol{a}| = 1.5$. The means of various statistics and their standard errors are presented in table 1. In all cases the MML model had class structure very close to the true values; tabulated results for classes 1 and 2 refer to the estimated classes corresponding to true classes with means of 4 and 0 respectively. These results only apply to classes of datasets where MML preferred a factor model to the uncorrelated model.

With one exception, noted below, the results of the factor analysis concur with those of Wallace and Freeman [20], in terms of the mean squared length of $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{\beta}}$ (where $\beta_k = a_k/\sigma_k$), and the factor load error vector $\hat{\boldsymbol{a}} - \boldsymbol{a}$, and the mean of the square of the sine of the angle between the true and estimated load vectors. As noted in Wallace and Freeman [20], these results outperform maximum likelihood techniques. For the barely detectable factor $|\boldsymbol{a}| = 1.0$, the factor load length is overestimated. This is presumed to be a selection effect, since for about one third of dataset classes an uncorrelated model was preferred, due to statistical variation obscuring the presence of a factor, with remaining statistical variation in the data included for analysis producing apparently stronger factors.

There is a discrepancy in the statistics of the $\sum_k \log \hat{\sigma}_k$, included to detect any bias in the estimation of the standard deviations (the true values of

[3] See forthcoming Monash University School of Computer Science and Software Engineering technical report for details of these heuristics.

	$\|\boldsymbol{a}\| = 1.5$		$\|\boldsymbol{a}\| = 1.25$		$\|\boldsymbol{a}\| = 1.0$	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
$\hat{\boldsymbol{a}}^2$	2.243	2.226	1.583	1.580	1.129	1.120
$\pm$	0.016	0.016	0.013	0.014	0.010	0.010
$\hat{\boldsymbol{\beta}}^2$	2.333	2.310	1.674	1.745	1.214	1.197
$\pm$	0.020	0.019	0.016	0.086	0.013	0.012
$(\hat{\boldsymbol{a}} - \boldsymbol{a})^2$	0.097	0.098	0.102	0.109	0.102	0.111
$\pm$	0.002	0.002	0.002	0.007	0.003	0.006
$\sin^2$ error angle in $\hat{\boldsymbol{a}}$	0.031	0.032	0.050	0.052	0.078	0.084
$\pm$	0.001	0.001	0.001	0.001	0.002	0.002
$\sum_k \log \hat{\sigma}_k$	-0.027	-0.025	-0.033	-0.043	-0.070	-0.063
$\pm$	0.006	0.006	0.006	0.010	0.007	0.006
$\hat{\mu}_1$	3.9986	-0.0016	3.9933	-0.0011	3.9975	0.0020
$\pm$	0.0034	0.0033	0.0053	0.0035	0.0040	0.0038
$\hat{\mu}_2$	4.0030	-0.0025	3.9977	-0.0002	4.0000	0.0012
$\pm$	0.0035	0.0036	0.0054	0.0035	0.0040	0.0042
$\hat{\mu}_3$	3.9907	-0.0017	3.9874	-0.0029	3.9865	-0.0006
$\pm$	0.0039	0.0039	0.0055	0.0050	0.0042	0.0042
$\hat{\mu}_4$	3.9851	-0.0004	3.9818	0.0007	3.9830	-0.0006
$\pm$	0.0041	0.0042	0.0056	0.0040	0.0043	0.0043
$\hat{\mu}_5$	3.9968	-0.0019	3.9936	-0.0015	3.9973	0.0019
$\pm$	0.0045	0.0045	0.0058	0.0045	0.0046	0.0048
Number of data sets estimated to have factors	998	998	964	957	669	690
Relative abundance (class 1)	0.4999		0.4994		0.5000	
$\pm$	0.0002		0.0005		0.0001	

Table 1. Estimates and errors on simulated data

which were all 1). Unlike MML estimation of single factors without mixture modelling, there appears to be a small but significant bias towards underestimating the standard deviations. Moreover, there may also be a very slight bias towards underestimating the means of class 1 (the class with 'true' means of 4). These effects are presumed to be a result of the inconsistency of total assignment (section 3.3), however it should be noted that the mean value of $\sum_k \log \hat{\sigma}_k$ for our estimator is considerably closer to zero than that obtained when using the maximum likelihood estimator out of the context of mixture modelling. This estimator shows means of approximately -0.113, -0.148 and -0.20 for $|\boldsymbol{a}| = 1.5, 1.25$ and 1.0 respectively [20].

6.2 Application: IRAS LRS Spectral Classification

We have also applied our program to astronomical point-source infrared spectra from the Infrared Astronomical Satellite (IRAS) Low Resolution Spectrometer (LRS). This data consists of spectra of 5425 point sources, in the wavelength range 7μm - 23 μm. Measurements were made with two instruments with overlapping wavelength coverage. The spectra as provided to the public by NASA have been corrected in a number of ways[1] to take into account calibration details and inconsistencies between the various components of the spectrometer, however many of the spectra still exhibit a lower reading from the red (long

wavelength) instrument than the blue (short wavelength) in the region of overlap despite measures to minimize this. For input to our modelling software, spectra were normalized to a mean attribute value of 1 to remove the very large variance in intensities from spectrum to spectrum.

A Bayesian maximum a posteriori classification of this data has been published by Goebel et. al [6]. This classification was produced by the AutoClass program [2] which assumes no correlation within classes. A total of 77 classes were found, with the distribution parameters of these classes being clustered into 'metaclasses'. As noted by Goebel et. al., there is considerable serial correlation in the attribute values (i.e. intensity values for given wavelengths).

In addition, there is large amount of correlation of the continuous kind which in our opinion is well modelled by factors. The AutoClass classification contains many classes which are identical to other classes apart from the presence of say a slightly larger broad peak at a certain wavelength. Rather than modelling the variance from spectral features of different strengths with large numbers of classes, and then trying to cluster them according to their distribution parameters, it is better to model them by a factor, where the factor scores correspond to the strength of the feature in a particular spectrum, and the factor loads indicate the effect the feature has on spectra.

Table 2 depicts the best classification yet found by our program for the IRAS LRS data. In each class, the points lie on the attribute means, the error bars extend one standard deviation above and below the means, and the factor loads are plotted as a line which usually crosses the horizontal axis around the middle. In addition to the parameters displayed as above, there are two lines representing $\mu_k + a_k$ and $\mu_k - a_k$, which can be identified as lines approximately following the shape of the mean points. These are included to give an indication of the range effects the factor can have for spectra with factor scores $\pm$ one standard deviation from the mean of 0. There are a total of twelve classes.

The major difference between our classification and that of AutoClass is of course the large reduction in the number of classes. We believe this indicates that our classes are closer to the true class structure of the data. By and large, the factor in each class accounts for variation in the strength of some spectral feature : the 10μm and 10μm silicate emission bands in classes 2, 6 and 7, the 8μm band in classes 8, 9, and 10, and the colour temperature (which affects the overall slope of the curve) for the approximately blackbody spectra in classes 4, 5, 11 and 12. In many classes the factor accounts for the variation in the amount by which the intensities from the red and blue instruments miss in the overlap region : classes 4, 2 and 12 are strong examples of this. Also, in some classes the factor appears to account for variation in the *shape* of spectral features : in classes 6 and 7 in particular, the (negative) peak in the factor loads occurs at a lower wavelength than the peak in the means, and hence the factor score models the variation in peak emission wavelength as well as in the size of the peak.

For many classes the factor tries to model correlation structure which arises from more than one of the causes above. The strongest example of this is class 3, which would appear to simultaneously attempt to model features at 8μm,