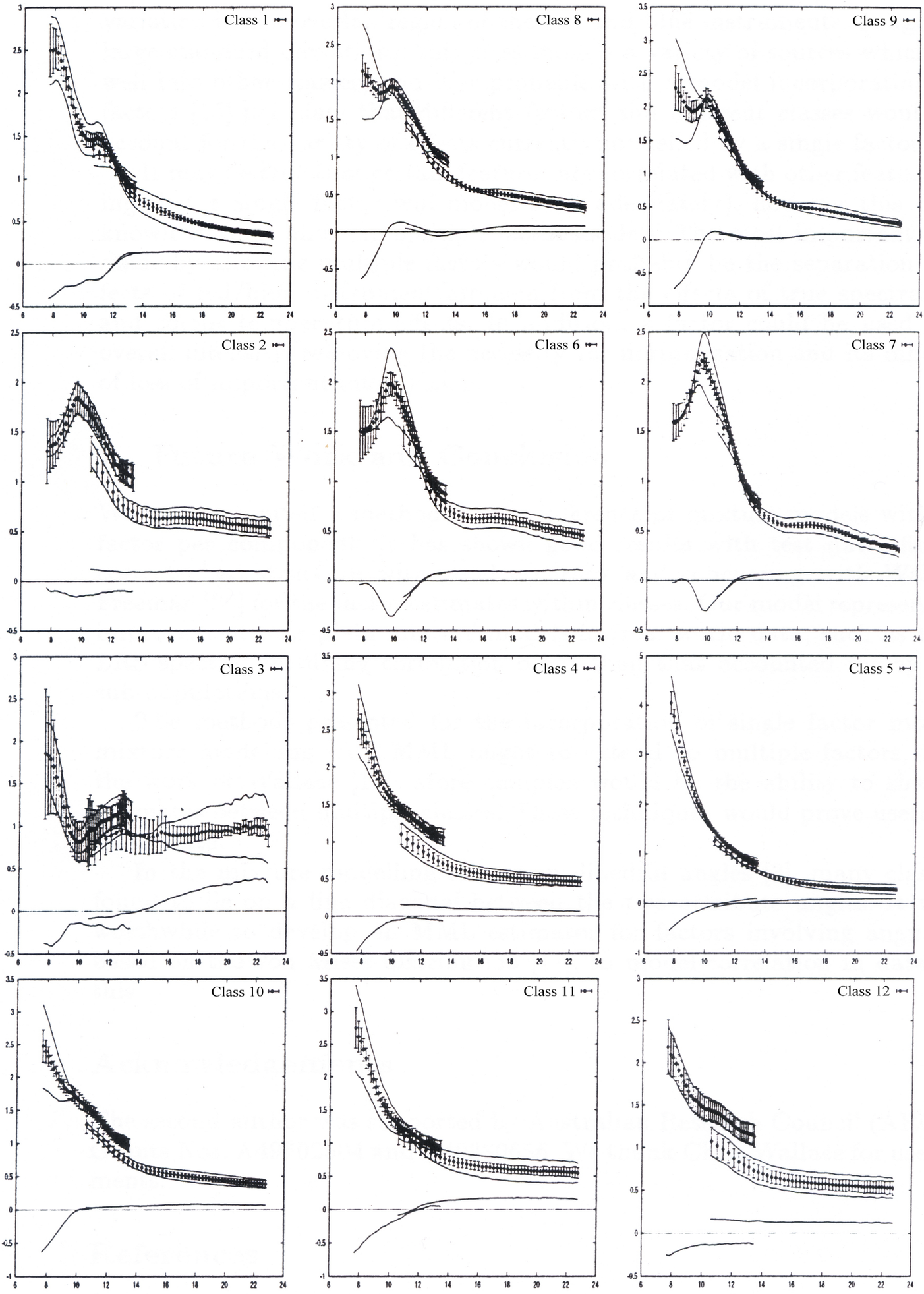

Table 2. IRAS LRS classes. The horizontal scale is in microns.

10μm, 11.5μm, 13μm, as well as the 'plateau' feature at 16μm-22μm, and the variation in the overlap region of the red and blue instruments. Judging by the large standard deviations, this class models a variety of sources which don't fit well into other classes, and it is probable with a model incorporating multiple factors [15] per class that different factors and different classes would arise to account for the variety of effects currently modelled by a single factor.

It may be true that certain features are correlated with other features (meaning that a single factor will model both adequately), however this cannot be known without allowing multiple factor models. The most important difference made by allowing multiple factors would probably be the separation of the effects of red/blue instrument problem from the effects of true spectral features and colour temperature effects. In addition, a factor could be used to model overall intensity, removing the necessity for normalization and its inherent risk of loss of important information.

7 Future Work and Conclusion

We have presented a method for the inference of mixture models with a single factor per component. It has shown good results with test data, finding the correct class structure with good accuracy and concurring with Wallace and Freeman [20] for the factor estimates within classes. Our model represents a large improvement over plain mixture models for real world data, such as the IRAS LRS spectra, involving correlation other than that accounted for by separate sub-populations.

The methods presented for the incorporation of single factor models into mixture modelling with MML ought to extend to multiple factors, based on the work of Wallace [15]. More complex would be the ability to share factor estimates between multiple classes. These techniques would prove useful for the IRAS data.

In the mixture modelling of protein dihedral angles [3], many classes were found to lie on a line diagonal between the two attribute angles. It would be worthwhile to develop an MML estimator for factors involving angular data, based on the von Mises distribution [17], to model correlation in data such as this.

Acknowledgements

The second author was supported by Australian Research Council (ARC) Large Grants Nos. A49602504 and A49330656. We thank Chris Wallace for useful comments.

References

1. G. Beichman, H.J. Neugebauer, P.E. Clegg, and Y.J. Chester. IRAS catalogs and atlases : Explanatory supplement, 1988.

2. P. Cheeseman, M. Self, J. Kelly, W. Taylor, D. Freeman, and J. Stutz. Bayesian classification. In *Seventh National Conference on Artificial Intelligence*, pages 607–611, Saint Paul, Minnesota, 1988.
3. D.L. Dowe, L. Allison, T.I. Dix, L. Hunter, C.S. Wallace, and T. Edgoose. Circular clustering of protein dihedral angles by Minimum Message Length. In *Proc. 1st Pacific Symp. Biocomp.*, pages 242–255, HI, U.S.A., 1996.
4. D.L. Dowe, R.A. Baxter, J.J. Oliver, and C.S. Wallace. Point Estimation using the Kullback-Leibler Loss Function and MML. In *Proc. 2nd Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, Melbourne, Australia, April 1998. Springer Verlag. accepted, to appear.
5. D.L. Dowe and C.S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. In *Proc. Computing Science and Statistics - 28th Symposium on the interface*, volume 28, pages 614–618, 1997.
6. J. Goebel, K. Volk, H. Walker, F. Gerbault, P. Cheeseman, M. Self, J. Stutz, and W. Taylor. A bayesian classification of the IRAS LRS Atlas. *Astron. Astrophys.*, (225):L5–L8, 1989.
7. H.H. Harman. *Modern Factor Analysis*. University of Chicago Press, USA, 2nd edition, 1967.
8. Hinton, Dayan, and Revow. Modelling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, 8(1):65–74, 1997.
9. G.E. Hinton and R. Zemel. Autoencoders, minimum description length and Helmholz free energy. In Cowan et. al, editor, *Advances in Neural Information Processing Systems*, 1994.
10. G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1996.
11. J. Neyman and E.L. Scott. Consistent estimates based on partially consistent observations. *Econometrika*, 16:1–32, 1948.
12. F.M. Olnon and E. Raimond. IRAS catalogs and atlases : Atlas of low resolution spectra, 1986.
13. D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, Inc., 1985.
14. C.S. Wallace. Classification by Minimum Message Length inference. In G. Goos and J. Hartmanis, editors, *Advances in Computing and Information – ICCI '90*, pages 72–81. Springer-Verlag, Berlin, 1990.
15. C.S. Wallace. Multiple Factor Analysis by MML Estimation. Technical Report 95/218, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1995. submitted to J. Multiv. Analysis.
16. C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.
17. C.S. Wallace and D.L. Dowe. MML estimation of the von Mises concentration parameter. Technical report TR 93/193, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1993. prov. accepted, Aust. J. Stat.
18. C.S. Wallace and D.L. Dowe. MML mixture modelling of Multi-state, Poisson, von Mises circular and Gaussian distributions. In *Proc. 6th Int. Workshop on Artif. Intelligence and Stat.*, pages 529–536, 1997. An abstract is in Proc. Sydney Int. Stat. Congr. (SISC-96, 1996), p197; and also in IMS Bulletin (1996), 25 (4), p410.
19. C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society (Series B)*, 49:240–252, 1987.
20. C.S. Wallace and P.R. Freeman. Single factor analysis by MML estimation. *Journal of the Royal Statistical Society (Series B)*, 54:195–209, 1992.