Cross-Validated Likelihood for Model Selection in Unsupervised Learning

Padhraic Smyth*
Department of Information and Computer Science
University of California, Irvine
CA 92697-3425
smyth@ics.uci.edu

1 Introduction

Cross-validation is a well-known technique in supervised learning to select a model from a family of candidate models. Examples include selecting the best classification tree using cross-validated classification error (Breiman et al., 1984) and variable selection in linear regression using cross-validated predictive squared error (Hjort, 1995).

Cross-validation is less seldomly used in *unsupervised* learning such as clustering. It is popular in kernel density estimation for choosing the smoothing parameter (the kernel bandwidth). However, it does not appear to have been used for the problem of determining cluster structure in clustering problems, i.e., solving the problem of how many clusters to fit to a given data set. This is the problem addressed in this paper.

Cross-validated likelihood can be viewed as an appropriate metric for model selection in probabilistic clustering, in particular for finite mixture models. In this paper, the use of cross-validated likelihood for clustering is investigated and the method is applied to a real problem where "truth" is unknown, i.e., determining the number of intrinsic "regimes" in records of upper atmosphere pressure taken daily since 1948 over the Northern Hemisphere.

2 Clustering with Mixture Models

The probabilistic mixture modelling approach to clustering is well-known: one assumes that the data from each cluster can be described in parametric form (usually Gaussian) and the overall marginal density of the data is a finite mixture model. The clustering problem becomes one of finding the weights and parameters for the component densities. Parameter estimation is often carried via iterative local likelihood maximization using the EM algorithm, given that k, the number of components is fixed.

So far so good. The main difficulties arise when one also wants to estimate k. Likelihood alone is of no use, since the likelihood can always be increased by increasing k irrespective of the true model. Bayesian and penalized likelihood methods provide alternative approaches for "honest" estimates of the number of components. The fully Bayesian approach is to treat the number of components k as a parameter and obtain a posterior distribution on

*Also with the Jet Propulsion Laboratory 525-3660, California Institute of Technology, Pasadena, CA 91109.

k given the data and the models. Even for the relatively simple Gaussian mixture model, this posterior cannot be calculated in closed form and must either be approximated or estimated via sampling techniques (for example, see Richardson and Green (1996) for a recent fully Bayesian treatment of mixture modelling with an unknown number of components). Penalized likelihood methods (such as AIC, BIC, MDL etc.) offer a simpler alternative, but as pointed out by Titterington, Makov, and Smith (1986), there are significant limitations on the applicability of these methods to mixture problems.

In this context cross-validated likelihood is a potentially interesting alternative as a model selection criterion. It can be readily be shown that cross-validated likelihood is an "honest" assessor of the "best" model in the sense that on average it will tend to choose the model from the candidate set which is closest to the true model generating the data. Distance here is the Kullback-Leibler distance (or cross-entropy) between the true model and the candidate model.

3 Choosing a Cross Validation Method

Consider that we have N data points from which to select a model. The fundamental idea of cross validation (CV) is to repeatedly partition the data into two sets, one of which is used to build the model and the other is used to evaluate the statistic of interest. Let M be the number of partitions, S_i be the ith subset used for evaluation, and $D \setminus S_i$ be the remainder of the data used for building the model. Thus, the cross-validated estimate of the kth model is defined as:

$$L_k^{cv} = \frac{1}{M} \sum_{i=1}^{M} L(S_i | \theta_k(D \setminus S_i))$$

where $\theta_k(D \setminus S_i)$ denotes the parameters for the kth model estimated from the ith training subset and $1 \leq k \leq K$.

There are a number of different cross-validation methodologies and they largely differ in how the partitions are chosen. "v-fold" cross validation uses v disjoint test partitions $\{S_1, \ldots, S_v\}$ each of size N/v. Well known examples are $v = N$ ("leave-one-out") and $v = 10$ which is used in CART (Breiman et al, 1984).

There are two primary motivations in practice for generating a cross-validated estimate: they are related, but different. The first is getting an unbiased ("honest") estimate of generalization performance for a particular model. The second motivation is comparing generalization performance across multiple models for the purpose of model selection. Much work in CV focuses on the former, while we are interested here in the latter. The same cross-validation estimator may not be optimal for both purposes. For example, an estimator with a constant bias across different models could be optimal for model selection but very sub-optimal for estimating generalization performance.

For model selection in linear regression, Burman (1989), Shao (1993), and Zhang (1993) have each investigated a particular CV procedure where M partitions are generated independently with a fixed fraction β being used as test samples, and $1 - \beta$ being used for parameter estimation in each case. (Burman calls it "repeated-learning-testing" or RLT, and Shao calls it "Monte Carlo cross validation" or MCCV—we will adopt the latter acronym). The main difference between this and the v-fold method is that each data point may be used as a test point more than once. Shao and Zhang each made the interesting observation that for model selection in linear regression, the theoretically optimal value of β (from a

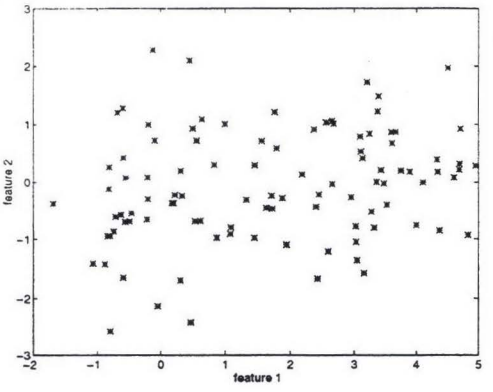

Figure 1: Scatter plot of a typical simulated two-cluster data set.

model selection viewpoint, rather than a performance estimation viewpoint) could be quite large. For example, Zhang discusses results where *testing* on 75% of the data yields a 97% or higher chance of selecting the correct model compared to an 85% or lower chance for *training* on 75% of the data. This is somewhat counter-intuitive to the standard CV practice where most of the data is used to fit the model and only a small fraction is used for evaluation. Kearns (1996), based on theoretical analysis of particular supervised learning problems, shows that the optimal fraction to be set aside for testing in cross validation decreases as the target function becomes more complex relative to the sample size. MCCV can also be viewed as a type of bootstrap method but without replacement. The relation to the bootstrap warrants further theoretical and empirical investigation.

4 The Performance of MCCV in Mixture Model Selection

4.1 Results on Simulated Data

We have investigated the application of MCCV to the problem of selecting the best number of components in a finite Gaussian mixture model. A non-trivial two cluster problem is shown in Figure 1. 2-dimensional data are generated from two equiprobable Gaussians with $\Sigma_1 = \Sigma_2 = I$ and with means 3 standard deviations apart. 50 data sets were generated from this model for each sample size: sample sizes ranged from 100 to 400. The MCCV procedure (with varying values of β), a penalized likelihood criterion (BIC), and 10-fold CV were each used to select a Gaussian mixture model among a family of three mixture models: $k = 1, 2$, and 3. The number of separate random partitions in each run of the MCCV procedure was set to $M = 20$ (this value was used in all analyses described in this paper). The BIC criterion is the maximum over k of the log-likelihood minus a penalty term of $p_k/2 \log(N)$, where p_k is the number of independent parameters in the k component model and N is the sample size.

The Gaussian mixture models were fitted via the Expectation-Maximization (EM) procedure. The EM procedure was initialized by selecting the highest likelihood solution obtained from 10 different runs of the k-means algorithm, where each k-means run was initialized randomly. This helped to avoid bad local minima in parameter space.

Sample Size	BIC	10-fold CV	MCCV $\beta = 0.7$	MCCV $\beta = 0.5$	MCCV $\beta = 0.3$
100	0.12	0.32	0.00	0.16	0.42
200	0.42	0.36	0.22	0.74	0.80
300	0.76	0.36	0.74	0.98	0.88
400	0.94	0.22	0.98	0.94	0.64

Table 1: Fraction of times the correct model of size 2 was selected.

Table 1 shows the fraction of times the correct model ($k = 2$) was chosen out of the 50 experiments, for each of the criteria and each of the sample sizes. It is clear that 10-fold CV is relatively poor for this problem (in fact it often *over-estimated* the number of components). It is also evident that for each of the sample sizes there is some setting of β (the fraction set aside for testing) for which MCCV outperforms BIC. It does not appear that there is a value of β which is universally best (as a function of sample size). As more data is available, it seems plausible that the optimal β increases. This is borne out in Table 1 where $\beta = 0.7$ is best for larger sample sizes and worst for smaller sample sizes. At this time there is no systematic method to automatically determine the best β to use for a particular problem when the true structure is unknown: this is an interesting open research question. However, we have found that the choice of $\beta = 0.5$ appears to be reasonably robust across a variety of problems and this value is used for the rest of the MCCV results reported in this paper. Smyth (1996) contains further results comparing MCCV, 10-fold CV, BIC, and the Autoclass algorithm (which is an approximation to the full Bayesian solution): in that work, Autoclass and MCCV (with $\beta = 0.5$ were determined to be roughly equally accurate in terms of model selection, BIC tended to under-estimate the true number of components, and 10-fold cross-validation was largely inaccurate.

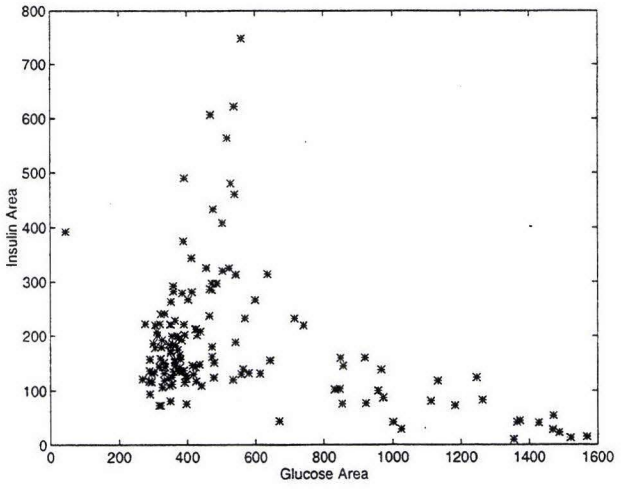

Figure 2: Scatter plot in 2 dimensions of 3-dimensional diabetes data

4.2 Analysis of Reaven and Miller's Diabetes Data

Reaven and Miller (1979) analyzed 3-dimensional plasma measurement data for 145 subjects who were clinically diagnosed into three groups: normal, chemically diabetic, or overtly

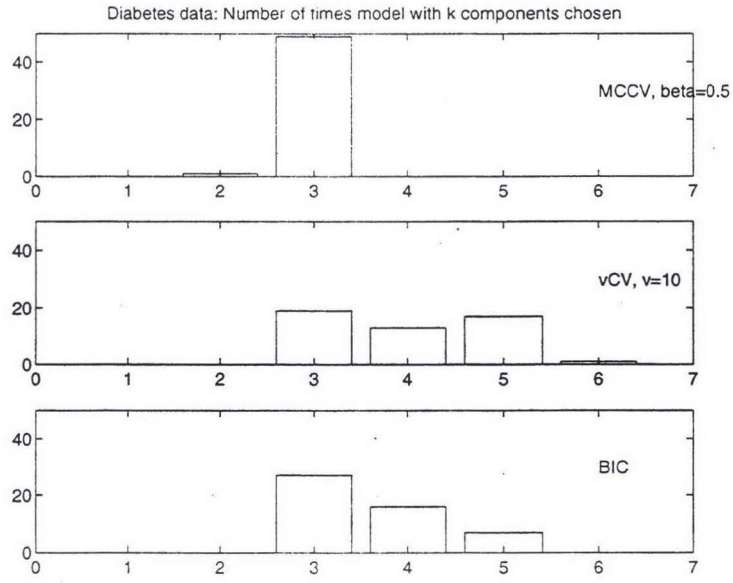

Figure 3: Results from multiple runs on diabetes data.

diabetic. This data set has since been analyzed in the statistical clustering literature by Symons (1981) and Banfield and Raftery (1993). When viewed in any of the 2-dimensional projections along the measurement axes, the data are not separated into obvious groupings: however, some structure is discernible (Figure 2).

The MCCV, 10-fold CV, and BIC criteria were each run 50 times on this data set. The only difference between each run for the BIC criterion was the fact that the Expectation Maximization (EM) procedure was started from (potentially) different randomly chosen initial conditions (as chosen from the 10 k-means runs) and thus could (and did) converge to different solutions. The MCCV and 10-fold CV methods had additional variability across runs in that the train/test partitions of the data were randomly chosen for each of the 50 runs.

Figure 3 summarizes the results. The MCCV method selected the model with 3 components 49 out of 50 times. The model selected by 10-fold CV ranged from 3 to 6 components. BIC chose 3 components about half of the time, and otherwise chose 4 or 5: these results for BIC are consistent with those of Banfield and Raftery (1993) whose "BIC-like" criterion indicated evidence for between 3 and 6 clusters and was maximized at 4. For real data sets such as this, one does not necessarily know what the real "truth" is. Nonetheless it is encouraging that the MCCV procedure agrees almost entirely with the clinical diagnosis for this data: for all cases for which the location of the clusters with $k = 3$ were examined, they matched the location of the clinical classes.

5 Application to Atmospheric Regime Detection

Records of the 700mb geopotential height (the height in meters at which the Earth's atmosphere registers a pressure of 700mb) have been collected daily in the Northern hemisphere since the 1948. A "map" for a given day corresponds to a spatial pressure pattern on a

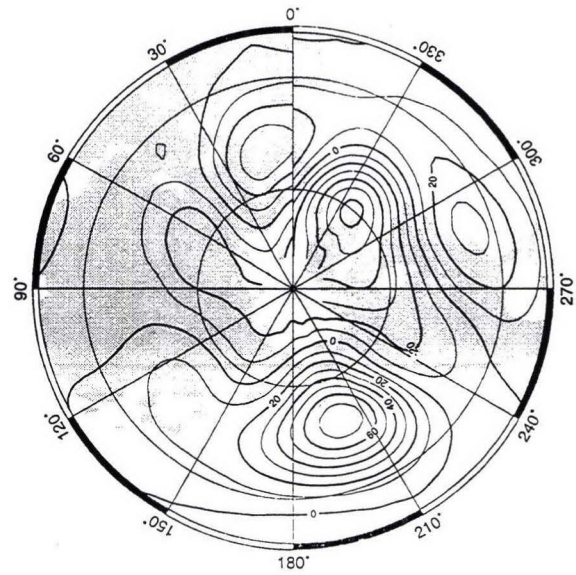

Figure 4: The Gulf of Alaska ridge regime which corresponds to the mean of one of the three components in a Gaussian mixture model for the geopotential height data.

grid. The historical record is a temporal sequence (time series) of maps. A variety of researchers have analysed the structure of these records to determine if there exist "regimes" of behavior, namely particular spatial pressure patterns which recur at regular intervals (Mo and Ghil, 1987; Cheng and Wallace 1993). Existence of these regimes has significant implications for understanding the low-frequency variability of the Earth's atmosphere. The spatial dimension of the data is removed by projection into the leading principal component (PC) directions. The "regime-identification" problem can to first-order be treated as that of finding clusters in the projected PC space.

Figure 4 shows the mean of one of three Gaussian mixture components fitted to the historical record projected into the first 2 principal component directions, and subsequently mapped back to the spatial grid. This particular pattern has been found in other studies and corresponds to the well known Gulf of Alaska ridge regime. Prior work has tried to answer the question of how many of these patterns recur in a reliable manner in the historical record. A variety of non-probabilistic clustering schemes have been tried (such as hierarchical clustering and bump-hunting). However, it is difficult to formulate objective approaches to the "how many clusters" question in a non-probabilistic clustering context and typically this question has been addressed in an ad hoc manner.

We fitted Gaussian mixture models to the data in the first 2 PC directions, varying the number of components k from 1 to 6, with a view to answering in a more objective manner the question of how many stable regimes exist. The log-likelihoods as estimated by MCCV are shown in Figure 5(a). Translating these into posterior probabilities on the k's (assuming uniform model priors) shows that there is clear evidence for 3 regimes (Figure 5(b)). Note that in general the posterior probability distribution can indicate whether the

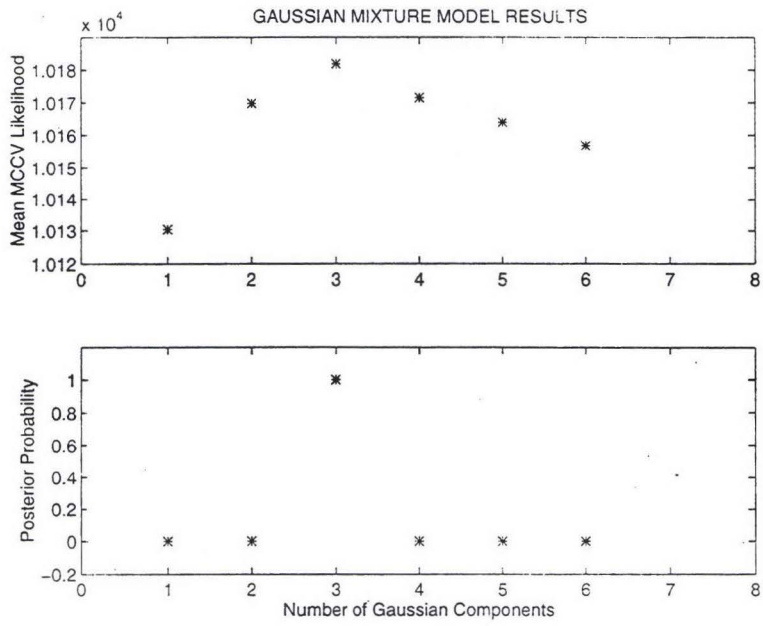

Figure 5: MCCV criteria for different numbers of components (k) in a Gaussian mixture model for the 700mb Northern hemisphere geopotential height data.

degree of evidence for a single value of k (the degree of "peaking" about any k).

These results are the first truly objective evidence of *multi-modality* in geopotential height in the Northern hemisphere. Furthermore the choice of 3 components is intriguing since this is also the same number of components arrived at by Cheng and Wallace (1993) using qualitative arguments based on an entirely different clustering methodology. They used hierarchical clustering (Ward's algorithm) based directly on correlation distances between spatial grid maps (without any principal component projection). Using sub-sampling techniques to estimate the robustness of the clusters, they concluded (based on a subjective analysis of the results) that there was strong evidence for 3 clusters. We have compared the maps corresponding to the particular 3 cluster centres in their paper with the maps corresponding to the means of the three Gaussian components found by the EM algorithm: they are qualitatively identical. This is a remarkable fact given that the methodologies were quite different.

In summary, cross-validated clustering has contributed to the first objective confirmation of Cheng and Wallace's hypothesis that there exist 3 stable regimes in the upper atmosphere geopotential height variability of the Northern Hemisphere. An important aspect of the cross-validated method is that a scientist or domain expert can easily interpret and understand the methodology once they grasp the basic concept of likelihood, i.e., the selected model is that which is estimated to have the highest likelihood on out-of-sample data. In contrast, and somewhat unfortunately, Bayesian and penalized likelihood model selection methods are likely to not be fully trusted by experts not versed in statistics.

6 Conclusion

Cross-validation as a tool for model selection is not limited to supervised learning problems such as regression or classification. For probabilistic models in unsupervised learning, cross-validated likelihood is a principled way to select between models of varying complexity. It provides a data-driven alternative to the more well-known penalized likelihood and Bayesian techniques. In this paper the utility of the method was demonstrated on a number of clustering problems, and in particular the methodology was able to provide the first objective evidence of a well-known scientific conjecture concerning the Earth's atmosphere.

7 Acknowledgements

The author would like to acknowledge Joe Roden (JPL), Professor Michael Ghil and Dr. Kayo Ide (UCLA) and Professor Andy Fraser (Portland State University) for their contributions to the atmospheric data analysis portion of the work described in this paper.

References

Banfield, J. D., and A. E. Raftery, Model-based Gaussian and non-Gaussian clustering, *Biometrics*, 49, 803–821, Sept. 1993.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J., *Classification and Regression Trees*, Wadsworth, 1984.

Burman, P., 'A comparative study of ordinary cross-validation, v-fold cross-validation, and the repeated learning-testing methods,' *Biometrika*, 76(3), 503–514, 1989.

Cheng, X., and Wallace, J. M. 'Cluster analysis of the Northern hemisphere wintertime 500-hPa height field: spatial patterns,' *J. Atmos. Sci.*, 50(16), 2674–2696, 1993.

Chow, Y. S., S. Geman, and L. D. Wu, 'Consistent cross-validated density estimation,' *The Annals of Statistics*, 11(1), 25–38, 1983.

Hjorth, J. S. U., *Computer Intensive Statistical Methods: Validation model selection and bootstrap*, Chapman and Hall, UK, 1994.

Kearns, M., 'A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split,' in *Advances in Neural Information Processing 8*, Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E. (eds.), Cambridge, MA: The MIT Press, 183–189, 1996.

Mo, K. C., and M. Ghil, 'Cluster analysis of multiple planetary flow regimes,' *J. Geophys. Res.*, 93, 10927–10951, 1987.

Reaven, G. M., and R. G. Miller, 'An attempt to define the nature of chemical diabetes using a multi-dimensional analysis,' *Diabetologia*, 16, 17–24, 1979.

Richardson, S. and Green, P. J., 'On Bayesian analysis of mixtures with an unknown number of components,' preprint, 1996.

Shao, J., 'Linear model selection by cross-validation,' *J. Am. Stat. Assoc.*, 88(422), 486–494, 1993.

Smyth, P., 'Clustering using Monte Carlo cross-validation,' *Proceedings of the Second International Conference on Knowledge Discovery*, AAAI Press, 1996: also available as `ftp://ftp.ics.uci.edu/pub/smyth/papers/kdd96.ps.Z`.

Symons, M., 'Clustering criteria and multivariate normal mixtures,' *Biometrics*, 37, 35–43, 1981.

Titterington, D. M., A. F. M. Smith, U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Chichester, UK: John Wiley and Sons, 1985.

Zhang, P., 'Model selection via multifold cross validation,' *Ann. Statist.*, 21(1), 299–313, 1993.