Model Selection for Probabilistic Clustering using Cross-Validated Likelihood

Padhraic Smyth*
Information and Computer Science
University of California, Irvine
CA 92697-3425

February 20, 1998

Technical Report UCI-ICS 98–09
Information and Computer Science
University of California, Irvine

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

Abstract

Cross-validated likelihood is investigated as a tool for automatically determining the appropriate number of components (given the data) in finite mixture modelling, particularly in the context of model-based probabilistic clustering. The conceptual framework for the cross-validation approach to model selection is direct in the sense that models are judged directly on their out-of-sample predictive performance. The method is applied to a well-known clustering problem in the atmospheric science literature using historical records of upper atmosphere geopotential height in the Northern hemisphere. Cross-validated likelihood provides strong evidence for three clusters in the data set, providing an objective confirmation of earlier results derived using non-probabilistic clustering techniques.

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 has also been used in *unsupervised* learning in the context of kernel density estimation for automatically choosing smoothing parameters (e.g., Silverman, 1986). However, it has not been applied to 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 may be due in part to the fact that for many clustering techniques there is no obvious score-function (for the number of clusters) to cross-validate. However, probabilistic model-based clustering (using finite mixture densities) is an exception, in that any score function which measures the quality of fit of the density also provides a candidate function for model selection.

In this paper cross-validated likelihood is investigated as an appropriate score function for model selection in probabilistic clustering, in particular for choosing the number of component densities in finite mixture models. Section 2 briefly reviews the application of mixture models to clustering. Section 3 discusses the use of cross-validated likelihood for choosing the number of mixture components. In Section 4 the method is applied to a well-known problem in atmospheric science, namely determining the number of "regimes" (or clusters) in records of upper atmosphere pressure taken daily since 1947 over the Northern Hemisphere. The cross-validation methodology provides an objective validation of earlier results from non-probabilistic clustering studies in the atmospheric science literature.

2 Clustering using Mixture Models

There is a long tradition in the statistical literature of using mixture models to perform probabilistic clustering (e.g, see Everitt and Hand, 1980; Titterington, Smith and Makov, 1986; and McLachlan and Basford, 1988). A key feature of the mixture approach to clustering is the ability to handle *uncertainty* about cluster membership in a probabilistic manner by allowing overlap of the clusters. Furthermore, the probabilistic model provides a framework for finding the optimal weights, locations, and shapes of the component clusters in a principled manner.

Let $\underline{X}$ be a d-dimensional random variable and let $\underline{x}$ represent a particular value of $\underline{X}$, e.g., an observed data vector with d components. A finite mixture probability density function for $\underline{X}$ can be written as

$$f^{(k)}(\underline{x}|\Phi^{(k)}) = \sum_{j=1}^{k} \alpha_j g_j(\underline{x}|\underline{\theta}_j) \tag{1}$$

where k is the number of components in the model and each of the g_j are the component density functions. The $\underline{\theta}_j$ are the parameters associated with density component g_j and the α_j are the relative "weights" for each component j, where $\sum_j \alpha_j = 1$ and $\alpha_j > 0, 1 \leq j \leq k$. $\Phi^{(k)} = \{\alpha_1, \ldots, \alpha_k, \underline{\theta}_1, \ldots, \underline{\theta}_k\}$ denotes the set of parameters for the overall mixture model with k components.

Let $D^{\text{train}} = \{\underline{x}_1, \ldots, \underline{x}_N\}$ denote the training data from which the model parameters are estimated. Assuming independent observations from an underlying true density $f(\underline{x})$,

the log-likelihood of $\Phi^{(k)}$ is defined as

$$l(\Phi^{(k)}|D^{\text{train}}) = \log p(D^{\text{train}}|\Phi^{(k)}) = \sum_{i=1}^{N} \log\Big(\sum_{j=1}^{k} \alpha_j g_j(\underline{x}|\underline{\theta}_j)\Big). \tag{2}$$

(Note that there are alternative objective functions which can be maximized in the clustering context, e.g., see Bensmail et al. (1997) for clustering using the "classification likelihood" function). Direct maximization of the mixture log-likelihood expression in Equation 2 is difficult except in trivial special cases. Thus, much of the popularity of mixture models in recent years is due to the existence of efficient iterative estimation techniques for obtaining maxima of this likelihood. In particular, the expectation-maximization (EM) procedure (Dempster et al. 1977; McLachlan and Krishnan, 1997) is a general technique for obtaining maximum-likelihood parameter estimates in the presence of missing data. In the mixture model context, the "missing data" are interpreted as the unknown or hidden labels that identify which data points originated from which mixture component. The EM procedure typically converges in parameter space to a local maximum of the log-likelihood function, but there is no guarantee of global convergence. Hence, the procedure is often initialized from multiple randomly chosen initial estimates and the largest of the resulting set of maxima is chosen as the final solution. It is well-known that there are various singular solutions to the maximum likelihood equations with infinitely large likelihood (such as having a cluster containing only one data point). Thus, in practice, the search for parameters is usually constrained to interior regions of parameter space which do not contain such spurious solutions (Hathaway, 1985).

The application of parameter estimation techniques (such as EM) assume that k (the number of components) is fixed. In practice, it is frequently the case that k is unknown, and, thus, one would like to also be able to infer some information about k from the data. Likelihood (as defined in Equation (2)) is of no direct use, since the likelihood on the training data can always be increased by increasing k irrespective of the true model. It is also well known that standard hypothesis testing methods (such as testing the hypothesis $k = 2$ against $k = 1$) fail in this context due to the breakdown of the standard assumptions on the asymptotic properties of the estimators (Feng and McCulloch, 1996). Alternatives, such as the bootstrap likelihood ratio (Aitkin et al., 1981; McLachlan, 1987) have been proposed, but not extensively investigated.

Bayesian and penalized likelihood methods provide more general approaches for "honest" estimates of the number of components. Penalized likelihood methods (such as AIC, BIC, MDL etc.) are typically based on approximations based on asymptotic arguments. They have the advantage of being relatively simple to implement since one simply penalizes the log-likelihood by an additive factor. However, as pointed out by Titterington, Makov, and Smith (1986), there are significant limitations on the applicability of these standard methods to mixture problems. Baxter and Oliver (1997) discuss the use of penalized likelihood methods which are not based on asymptotic arguments; however, their results are obtained only for univariate problems with one or two mixture components.

The fully Bayesian approach is to treat the number of components k as a parameter and obtain a posterior distribution on 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 analytically or estimated via sampling techniques. For example, Richardson and Green (1996) provide a Bayesian treatment of mixture modelling with an

unknown number of components using Monte Carlo Markov Chain (MCMC) methods.

The Bayesian and penalized likelihood approaches can be viewed from a single perspective by noting that the penalized likelihood methods can each be derived as different approximations to the full Bayesian solution (see Chickering and Heckerman (1997) for a full discussion of this viewpoint). Thus, in practice, existing model selection methods for mixture densities largely rely on approximations of one form or another. For any of these approximations (whether it be penalized likelihood, direct approximations to Bayes, or Monte Carlo sampling of the Bayesian solution) the results obtained can be dependent in a non-transparent manner on the quality of the underlying approximations or simulations. In the next section we discuss the use of cross-validation as an alternative approach.

3 Cross-validated Likelihood

Let $f(\underline{x})$ be the "true" probability density function for $\underline{x}$ and let $D^{\text{train}} = \{\underline{x}_1, \ldots, \underline{x}_N\}$ be a random sample from f as before. A set of finite mixture models with k components are fitted to D, where k ranges from 1 to $k_{\max}$. Thus, we have an indexed set of estimated models, $f^{(k)}(\underline{x}|\hat{\Phi}^{(k)}), 1 \leq k \leq k_{\max}$, where each $f^{(k)}(\underline{x}|\hat{\Phi}^{(k)})$ has been fitted to the same data set D^{train}.

Let $l_k^{\text{train}} = l(\Phi_k(D_{\text{train}})|D_{\text{train}})$ denote the log-likelihood of the fitted model with k components, where the parameters Φ_k have been fit to the training data D_{train} and the log-likelihood has been evaluated on the same data (as in Equation 2). l_k^{train} is a non-decreasing function of k since the increased flexibility of more mixture components allows better fit to the data (increased likelihood) as k is increased. Thus, l_k^{train} cannot directly provide any clues as to the *true* mixture structure in the data, if such structure exists.

Imagine instead that one has a large test data set D^{test} which was not used in fitting any of the models. Let $l_k^{\text{test}} = l_k(\Phi_k(D^{\text{train}})|D^{\text{test}})$ be the log-likelihood, in a manner analogous to Equation 2, where the models are fit to the training data D^{train} but the log-likelihood is evaluated on data D^{test} with N_{test} data points. One can interpret this "test log-likelihood" as a function of the "parameter" k, keeping all other parameters and D_{train} fixed. Intuitively, this test likelihood l_k^{test} should be a more useful estimator (than the training data likelihood l_k^{train}) for comparing mixture models with different numbers of components. (This test log-likelihood is also known as the log predictive score in the Bayesian model selection literature (e.g., see Good, 1952)).

For convenience of notation, let $f_k(\underline{x})$ denote the model with k components with parameters $\Phi_k(D^{\text{train}})$ fitted using D^{train}, and let

$$i_k = -\frac{l_k^{\text{test}}}{N_{\text{test}}} = -\frac{1}{N_{\text{test}}} l(\Phi_k(D^{\text{train}})|D^{\text{test}})$$

be the negative test log-likelihood per sample. Taking the expectation of i_k with respect to all training data sets of size N drawn from $f(\underline{x})$,

$$\begin{aligned} E[i_k] &= -\frac{1}{N_{\text{test}}} E\left[l(\Phi_k(D^{\text{train}})|D^{\text{test}})\right] \\ &= -\frac{1}{N_{\text{test}}} \sum_{j=1}^{N_{\text{test}}} E[\log f_k(\underline{x}_j)] \\ &= \int f(\underline{x}) \log \frac{1}{f_k(\underline{x})} d\underline{x} \end{aligned}$$

$$= \int f(\underline{x}) \log \frac{f(\underline{x})}{f_k(\underline{x})} d\underline{x} - \int f(\underline{x}) \log \frac{1}{f(\underline{x})} d\underline{x} \tag{3}$$

i.e., the expected value of i_k is the Kullback-Leibler (KL) distance between $f(\underline{x})$ and $f_k(\underline{x})$, minus a constant which is independent of k. Thus, the test log-likelihood l_k^{test} (scaled appropriately) is an unbiased estimator (within a constant) of this KL distance. The KL distance in turn defines how far the model $f_k(\underline{x})$ is from the true f and is strictly positive unless $f_k(\underline{x}) = f(\underline{x})$. Thus, the test log-likelihood is an unbiased estimator of the KL distance between truth and the models under consideration, and this motivates its use as a model selection criterion in this context.

Of course one typically does not have a large independent test data set such as D^{test} available. Thus, a practical alternative is to use l_k^{cv} for model selection instead, namely, a cross-validated estimate of l_k^{test}. In cross-validation the data are repeatedly partitioned 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. For the ith partition let S_i be the data subset used for evaluation of the log-likelihood and $D \setminus S_i$ be the remainder of the data used for building the model. Thus, the cross-validated estimate of the test log-likelihood for the kth model is defined as:

$$l_k^{cv} = \frac{1}{M} \sum_{i=1}^{M} l(\Phi_k(D \setminus S_i)|S_i) \tag{4}$$

where $\Phi_k(D \setminus S_i)$ denotes the parameters for the kth model estimated from the ith training subset, and $l(\Phi_k(D \setminus S_i)|S_i)$ is the log-likelihood evaluated on the data in S_i using the parameters estimated from the data $D \setminus S_i$.

It is worth noting that cross-validation will necessarily be less efficient in its use of the data compared to a fully Bayesian approach, i.e., it estimates l_k^{test} for models trained on some fraction β of the data (the training partition size), rather than on the full data. Thus, in this sense, the fully Bayesian approach is in principle more efficient in its use of the available data. Of course, as mentioned earlier, implementing the Bayesian approach in practice involves approximation of one form or another and, indeed, cross-validation itself can be viewed as a different type of approximation in the Bayesian context (Dawid, 1984).

In general, consider the case when the model family under consideration includes the true data generating distribution $f(\underline{x})$; let this particular model have k_{true} components. Both the Bayesian and cross-validation methodologies will tend to converge to k_{true} (as a function of k, from below) as the sample size is increased, i.e., for very small data sets there are only enough data to support the $k = 1$ hypothesis, but gradually as the sample size N is increased the selected model $\hat{k}$ increases until it "locks-on" to k_{true}. For cases where truth is not within the model family, the Bayesian approach encounters some conceptual difficulties (see for example the discussion of Raftery, Madigan, and Volinsky (1996), pages 346–347). In contrast, it is clear from the KL distance equations above, that the cross-validation methodology will directly seek that model from within the model family which is closest to truth.

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 for example (Breiman et al, 1984)). 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—the latter acronym will be used in this paper). 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.

Smyth (1996) compares the performance of a variety of model selection methods, including MCCV, 10-fold cross-validation, BIC, and the Autoclass algorithm (Cheeseman and Stutz, 1996) which is an analytic 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 often unreliable. In general, there appears to be no obvious systematic method to automatically determine the best β to use for a particular problem when the true structure is unknown, although the choice of $\beta = 0.5$ appears to be reasonably robust across a variety of problems (Smyth, 1996). In terms of choosing the number of different partitions M, the larger the value the less the variability in the log-likelihood estimates. In practice, values of M between 20 and 50 appear adequate for most applications.

Finally, it is worth noting that there is an extra computational cost incurred by repeated cross-validation, namely $k_{\max}$ different models are to be estimated and evaluated M different times. Compared to the simpler penalized likelihood methods (such as AIC or BIC) this is an increase in computation by a factor M. It is not clear how cross-validation compares to simulation-based (Bayesian) MCMC methods in terms of computational cost. Each method is somewhat "open-ended" in that the more computation one is willing to expend, the better the results on average. It seems likely that the typical cross-validation methodology will be closer in computational cost to a typical MCMC implementation than to the penalized likelihood methods.

4 Application of the Cross-Validated Clustering Method to Atmospheric Geopotential Height Data

4.1 Problem Background

Detection and identification of "regime-like" behavior in atmospheric circulation patterns is a problem which has attracted a significant amount of attention in atmospheric science. (As defined in the atmospheric science literature, *regimes* are recurrent and persistent spatial patterns which can be identified from atmospheric data sets (Cheng and Wallace, 1993; Kimoto and Ghil, 1993)). The most widely-used data set for these studies consists of daily measurements since 1947 of *geopotential height* on a spatial grid of over 500 points in the Northern Hemisphere (NH). Geopotential height is the height in meters at which the atmosphere attains a certain pressure (e.g., one has 500mb height data, 700mb height data, etc.). It can loosely be considered analogous to atmospheric pressure, particularly since one can visualize the data using contour maps with "lows," "highs", "ridges," and so forth.

Research on low-frequency atmospheric variability using geopotential heights during the past decade has demonstrated that on time scales longer than about a week, large-scale atmospheric flow fields appear to exhibit recurrent and persistent regimes. Direct identifi-

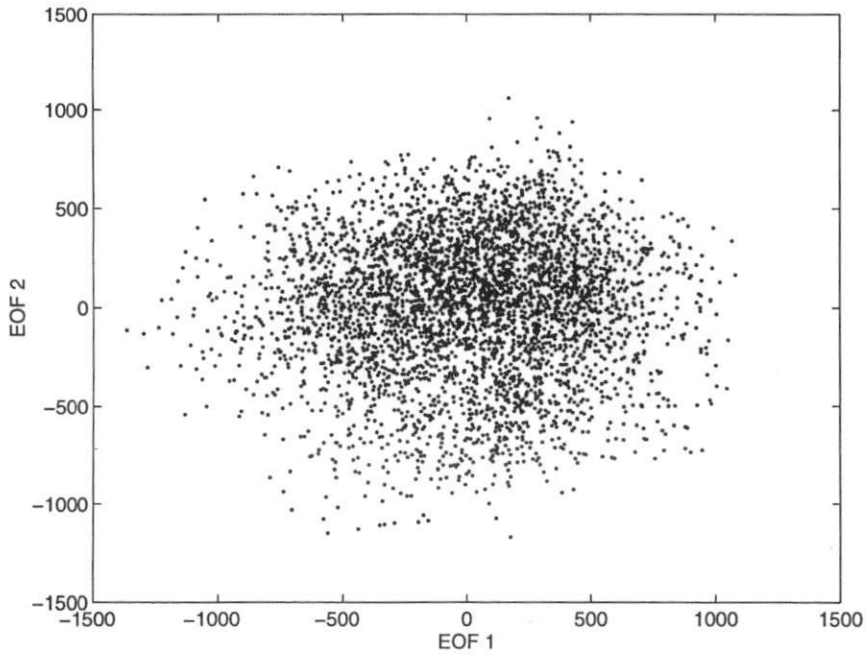

Figure 1: Scatter plot of NH winter data projected into first 2 EOF directions.

cation of these regimes in observed flow fields is difficult. This has motivated the use of a variety of cluster analysis algorithms to objectively classify observed geophysical fields into a small set of preferred regimes or categories, e.g., fuzzy clustering (Mo and Ghil, 1988), kernel density estimation and "bump hunting" (Kimoto and Ghil, 1993), hierarchical clustering (Cheng and Wallace, 1993), and least-squares (or k-means) clustering (Michelangeli, Vautard, and Legras (1995)).

While these approaches have produced useful and repeatable results (in terms of significant cluster patterns), there is nonetheless a degree of subjectivity in the application of these clustering techniques which is undesirable. In particular, none of these methods have provided a fully objective answer to the question of how many clusters exist. Thus, among the different studies, it is not clear how many different regimes can be reliably identified.

We analyzed the same data as has been used in almost all of the other clustering studies on this topic (e.g., Kimoto and Ghil (1993)), namely, daily observations of the NH 700-mb geopotential heights on a $10^\circ \times 10^\circ$ diamond grid (with 541 grid points), compiled at NOAA's Climate Analysis Center. The data are subject to a number of specific preprocessing steps (full details are provided in Smyth, Ghil and Ide (1998)). For the purposes of this paper it is sufficient to know that the daily 541 spatial grid points (or maps) are treated as 541-dimensional data vectors and then projected into a subspace defined by a few leading principal component directions for this 541-dimensional space. We will use the atmospheric science terminology of "empirical orthogonal functions" (or EOFs; Preisendorfer (1988)) to refer to the principal component directions in the rest of the paper. Projections used in the results described here range from the first 2 to the first 12 EOFs. Figure 1 shows data from the 3960 days defined as "winter" projected onto the first two EOFs. This projected winter data set is the "standard" data set which has been typically used in clustering studies in

Table 1: Cross-validated log-likelihood and estimated posterior probabilities, as a function of k, from 20 random partitions of 44 winters.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Cross-validated log-likelihood	-29164	-29153	-29137	-29148	-29156	-29165
Posterior probability	0.0	0.0	1.0	0.0	0.0	0.0

the past and it is on this data set that the application of cross-validation for model selection is investigated below.

4.2 Application of Mixture Model Clustering

We applied the mixture model cross-validation methodology to the data described in Section 4.1, using Gaussian components with unconstrained (full) covariance matrices. (These results, and various extensions, are described in more detail in Smyth, Ghil, and Ide (1998)). In all experiments the number of cross-validation partitions was $M = 20$ and the fraction of data β contained in each test partition was set to 0.5. The number of clusters (mixture components) was varied from $k = 1, \ldots 15$. The log-likelihoods for $k > 6$ were invariably much lower than those for $k \leq 6$ so for clarity only the results for $k = 1, \ldots, 6$ are presented. The estimated cross-validated log-likelihoods and approximate posterior probabilities on k are tabulated in Table 1 (the posterior probabilities are simply the exponentiated and normalized log-likelihoods). There is clear evidence for 3 clusters, i.e., the cross-validation estimate of the posterior probability for 3 clusters is effectively 1 and all others are effectively zero. Figure 2 shows the three-cluster solution (means and covariance shapes) in the two-dimensional EOF-space.

Note that the absolute values of the log-likelihoods are irrelevant—strictly speaking, likelihood is only defined within an arbitrary constant. Figure 3 shows the test log-likelihoods on the 20 different cross-validation partitions, relative to the log-likelihood on each partition of the $k = 3$ model (dotted line equal to zero). $k = 3$ clearly dominates. Note that for any particular partition $k = 3$ (the dotted line with value 0) is not necessarily always the highest likelihood model, but on average across the partitions it is significantly better than the other possible values for k.

4.3 Robustness of the Results

Numerous runs on the same data with the same parameters but with different randomly-chosen winter partitions ($M = 20$) always provided the same result, namely, an estimated posterior probability of $p(k = 3) \geq 0.999$ in all cases. The relative cross-validated likelihoods over 10 different runs are shown in Figure 4.

We also investigated the robustness of the method to the dimensionality of the EOF-space. The maps were projected into different subspaces, namely the first d EOF dimensions, with $d = 2, \ldots, 12$. As a function of the dimensionality d, the posterior probability mass was concentrated at $k = 3$ (i.e., $p(k = 3) \approx 1$) until $d = 6$, at which point the mass "switched" to become concentrated at $k = 1$ (i.e., $p(k = 1) \approx 1$)). Thus, as the dimensionality increases beyond $d = 6$, the cross-validation method does not provide any evidence to support a model more complex than a single Gaussian bump. This is to be expected since the number

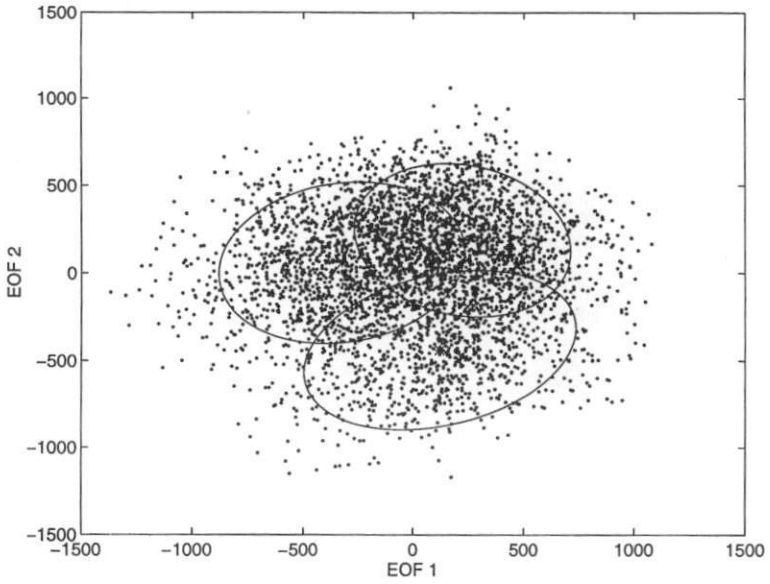

Figure 2: Scatter plot of NH winter data projected into the first 2 EOF directions with estimated means and covariance matrix shapes (ellipses) superposed as fitted by the EM procedure with a 3-component Gaussian mixture model. The ellipses represent contours of the density function which are 3-sigma from the means.

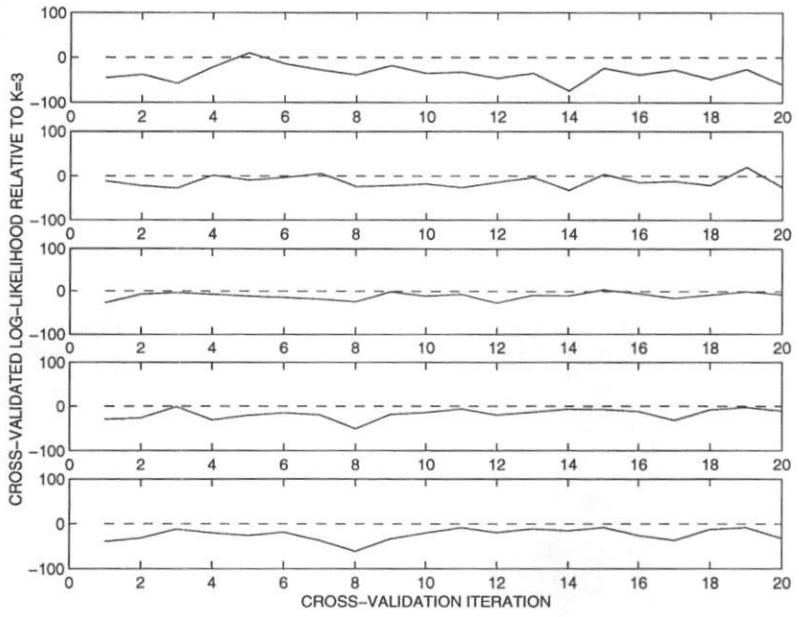

Figure 3: Log-likelihood of the test partition data on each cross-validation iteration (from 1 to 20) relative to the log-likelihood of the $k = 3$ model for (from top) (a) $k = 1$, (b) $k = 2$, (c) $k = 4$, (d), $k = 5$, and (e) $k = 6$.

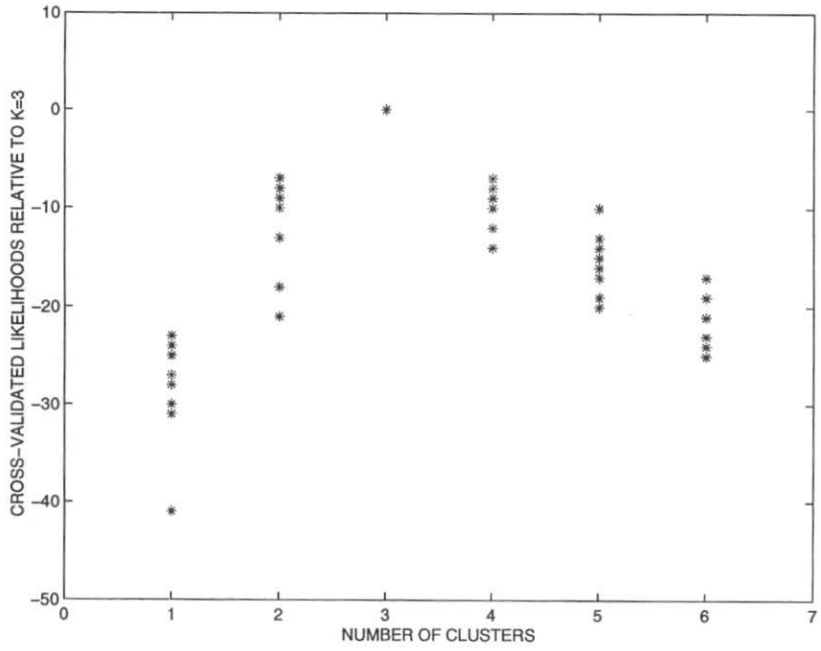

Figure 4: Cross-validated log-likelihoods for $k = 1, \ldots, 6$ relative to the cross-validated log-likelihood of the $k = 3$ model for 10 such different randomly-chosen cross-validation partitions

of parameters in a k-component Gaussian mixture model grows as kd^2. Since the total amount of data to fit the models is fixed, as the dimensionality d increases the estimates of the more complex models are less reliable and cannot be justified by the data.

For the three-component Gaussian model we investigated the variability in the physical grid maps obtained across different numbers of EOF dimensions. Note that each data point in a projected EOF space can be represented as a pressure map on the original grid (since each point is a linear combination of EOF vectors, and each EOF vector is a map). Thus, cluster centers in the EOF space can be "mapped back" to equivalent grid points in the original spatial grid to create spatial contour maps. Using the first d EOF dimensions, $d = 3, \ldots, 12$, a Gaussian mixture model with 3 components was fit to the data for each d. For each value of d, 3 physical maps were obtained from the centers of the 3 Gaussians. The pattern correlations (as defined in Wallace and Cheng (1993), page 2676) were then calculated between each of these maps (from d dimensions) and the corresponding maps obtained from 2 EOF dimensions. The results are shown in Table 2. It is clear that here is a very high correlation between the 2d EOF maps and each of the maps obtained in up to 12 EOF dimensions. This indicates that as the dimensionality of the E0F space grows beyond $d = 2$, the clusters in any of these dimensional spaces are essentially the same as for the the two-dimensional sub-space.

4.4 Comparisons with Bayesian and Penalized Likelihood Techniques

Autoclass 2.0 (Cheeseman and Stutz, 1996) was applied to the same data as described above. The default version of Autoclass (full covariance matrices) returned $k = 3$ as by far

Table 2: Pattern correlation coefficients between maps fitted using d EOF dimensions, $3 \leq d \leq 12$, and maps fitted using 2 EOF dimensions.

EOF Dimensionality d	r_1	r_2	r_3
3	0.978	0.961	0.998
4	0.974	0.960	0.999
5	0.947	0.957	0.976
6	0.946	0.946	0.957
7	0.945	0.951	0.945
8	0.931	0.946	0.938
9	0.938	0.953	0.941
10	0.946	0.951	0.949
11	0.927	0.943	0.934
12	0.945	0.946	0.935

the most likely choice for the number of clusters, i.e., no other k values had any significant posterior probability. For the same data we also calculated the BIC criterion which penalizes the training log-likelihood by an additive factor of $-k/2 \log N$. The BIC criterion was maximized at $k = 1$ (by a substantial margin). Thus, the cross-validation and Bayesian methods are in agreement (with $k = 3$), while BIC is more conservative.

4.5 Interpretation and Discussion of the Cluster Results

An important aspect of this problem is the scientific interpretation of the clusters obtained. The scientific interpretation is obtained by projecting the cluster centers (the Gaussian means) "back" to the grid-space as described earlier, and then directly interpreting the physical significance of the resulting spatial patterns.

Figure 5 shows the three maps corresponding to the three Gaussian centers on the left and the three maps corresponding to the "most distinct clusters of the wintertime 500mb field" on the right (Cheng and Wallace, 1993; also in Wallace, 1996). These two sets of maps have a high degree of qualitative similarity to each other. The upper maps (a) and (b) both clearly possess a distinctive ridge over the Gulf of Alaska. The middle maps (c) and (d) are characterized by a very distinctive blocking pattern over southern Greenland. The bottom maps (e) and (f) have a more complex pattern described as the "Rockies ridge" in Cheng and Wallace (1993, p.2680). The Cheng and Wallace results are considered among the most authoritative on this topic, and these particular three spatial patterns (or regimes) are frequently discussed in the atmospheric science literature.

Cheng and Wallace's methodology for arriving at three clusters was based on a combination of ad hoc resampling techniques and subjective judgement of the hierarchical clustering results. In their own words, "the more reproducible clusters are strung out along three well-defined branches of the family tree" (Cheng and Wallace, 1993). It is interesting to note that the cross-validation results described in this paper were obtained completely independently, i.e., the cross-validation data analysis was carried out without knowledge at that time of the Cheng and Wallace result. Thus, the cross-validation results provided an objective and independent validation of the earlier work. For further discussion of the physical interpretation of the results see Smyth, Ghil, and Ide (1998).

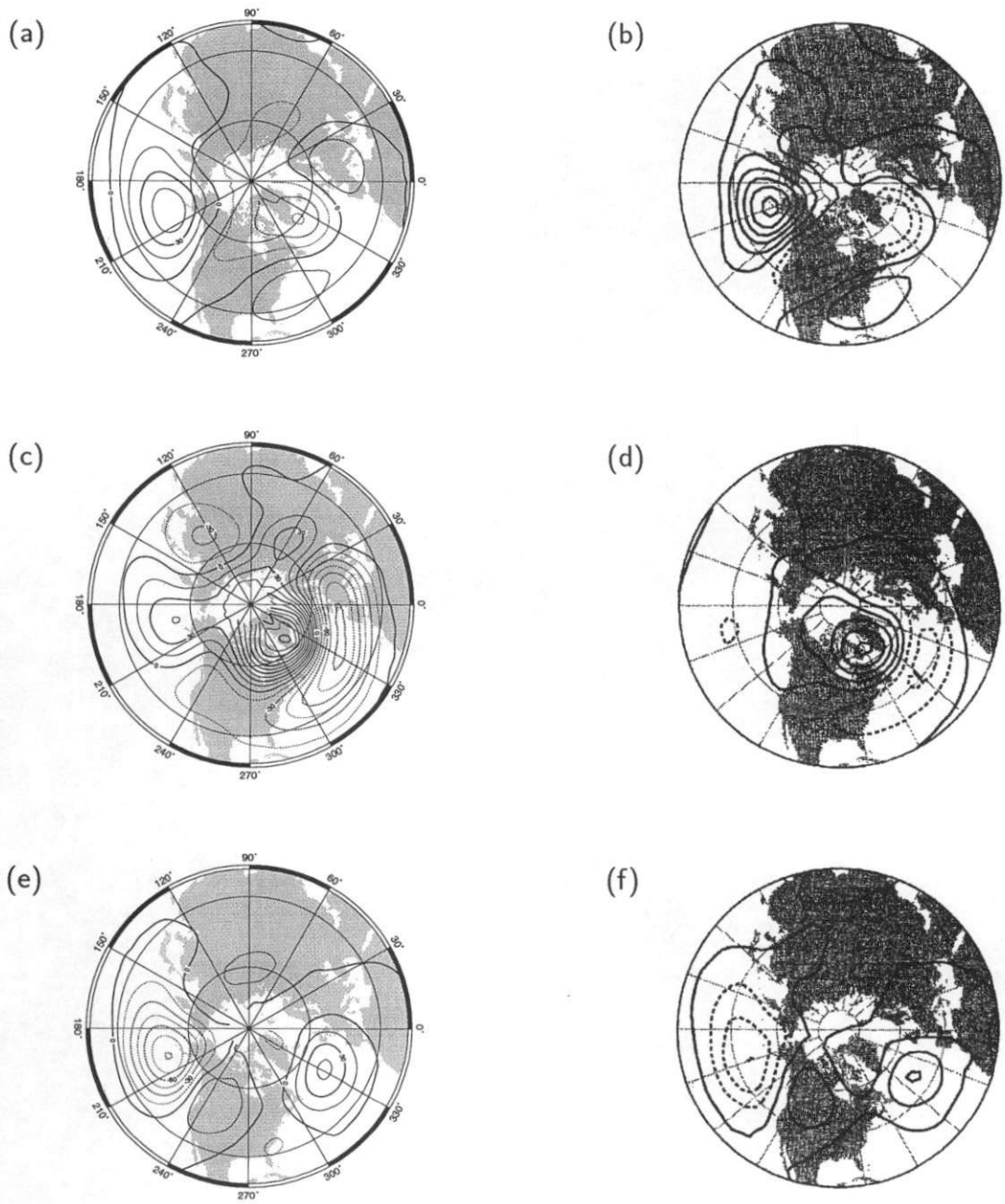

Figure 5: Geopotential height maps for the 3 cluster centers of the mixture model (left: panels a, c and e) and of Cheng and Wallace's (1993) hierarchical cluster model (right: panels b, d and f) which are reproduced in Wallace (1996) (panels b,d, and f reproduced by permission).

An obvious question is whether or not the results are sensitive to the projection methodology being used, i.e., would projection pursuit for example lead to different clusters? The answer would appear to be no. The similarity of the maps in Figure 5 (where one set is obtained in EOF-space and the other set by directly clustering the grid patterns) indicates that the EOF projection does not impact the resulting clusters. Cheng and Wallace (1993) also reached the same conclusion, by finding that hierarchical clustering in EOF space produced essentially the same clusters as the clusters obtained with no EOF projection.

5 Discussion and Conclusion

Cross-validated likelihood can play a useful practical role in model selection among different mixture density models. The conceptual framework is simpler than the typical penalized likelihood or Bayesian approach in that models are directly judged on their out-of-sample predictive ability, as estimated in a cross-validated fashion. The simplicity of the framework makes it directly applicable to a wide variety of practical problems. In this paper, only the problem of finding the correct numbers of components for Gaussian mixture models was discussed. However, one can in principle easily apply the methodology to a much broader class of mixture problems, such as selecting among different independence structures (e.g., see Bensmail et al (1997) and Thiesson et al (1998)) or model selection in the context of Markov models (e.g., see Smyth (1997) for an application to hidden Markov models).

Directions for further work on cross-validated likelihood include a bias-variance characterization for better understanding of the trade-offs involved in choosing β (see for example the work of Shao (1993) and Zhang (1993) in a regression context and Kearns (1996) in a classification context), and comparative studies between penalized likelihood, Bayesian, and cross-validation methodologies. In related work, Smyth and Wolpert (1998) extend the framework in this paper to *model averaging*, again using cross-validation to empirically determine the model weighting coefficients rather than using posterior probabilities on the models obtained from a Bayesian analysis.

A final point concerns the acceptance of any model selection methodology by domain experts (in this case, atmospheric scientists). The scientists participating in this work indicated a much greater willingness to trust a methodology based on cross-validation than a Bayesian analysis. This trust was due in large part to the direct interpretation which can be given to the cross-validation result (i.e., one seeks the model which predicts best on out-of-sample data). In contrast, the Bayesian formulation of the problem was perceived as indirect and less appealing. To put it another way, the scientists were far more willing to defend a cross-validation model selection procedure among their peers than they would be willing to defend a Bayesian model selection procedure. This is an important point. It is suggestive that while *in theory* a fully Bayesian analysis can be viewed as the optimal approach, *in practice* a cross-validation methodology can be a practical alternative, particularly when data and computational resources are relatively plentiful.

Acknowledgements

The author would like to acknowledge the considerable assistance of collaborators Michael Ghil, Kayo Ide, Joe Roden, and Andy Fraser on the atmospheric data analysis and also acknowledge Dr. Masahide Kimoto for providing the atmospheric data set described in this

paper. The author would also like to gratefully acknowledge valuable discussions on model selection with David Heckerman, David Madigan, Usama Fayyad, and Michael Turmon. The research described in this paper was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration and was supported in part by NSF CAREER award IRI-9703120.

6 References

Aitkin, M., Anderson, D., and Hinde, J. 'Statistical modelling of data on teaching styles (with discussion),' *J. R. Statist. Soc. A*, 144, 419-461, 1981.

Baxter, R. A. and Oliver J. J., 'Finding overlapping distributions using MML' in *Preliminary Papers of the Sixth International Workshop on AI and Statistics*, Fort Lauderdale, FL, January 1997.

Bensmail, H., Celeux, G., Raftery, A., and Robert, C. P., 'Inference in model-based cluster analysis,' *Statistics and Computing*, 7, 1–10, 1997.

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.

Cheeseman, P. and Stutz. J., 'Bayesian classification (AutoClass): theory and results,' in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.), Cambridge, MA: AAAI/MIT Press, pp. 153–180, 1996.

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.

Chickering, D. M., and Heckerman, D., 'Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables,' *Machine Learning*, 29(2/3), 181–244.

Dawid, A. P., 'Present position and potential developments: some personal views. Statistical theory: the prequential approach,' *J. R. Statist. Soc. A*, 147, 278–292 (with discussion), 1984.

Everitt, B. S., and Hand, D. J., *Finite Mixture Distributions*, London: Chapman and Hall, 1981.

Feng, Z. D., and McCulloch, C. E., 'Using bootstrap likelihood ratios in finite mixture models,' *J. R. Statist. Soc. B*, 58(3), 609–617, 1996.

Good, I. J., 'Rational decisions,' *J. R. Statist. Soc. B*, 14, 107–114, 1952.

Hathaway, R. J., 'A constrained formulation of maximum-likelihood estimation for normal mixture distributions,' *Ann. Stat.*, 13, 795-800, 1985.

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.

Kimoto, M., and Ghil, M., 1993 'Multiple flow regimes in the Northern hemisphere winter: Part I: methodology and hemispheric regimes,' *J. Atmos. Sci.*, 50(16), pp.2625–2643.

McLachlan, G. J., 'On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture,' Appl. Statist., 36, 318–324, 1987.

McLachlan, G. J. and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker, 1988.

McLachlan, G. J., and Krishnan, T., *The EM Algorithm and Extensions*, New York: John Wiley and Sons, 1997.

Michelangeli, P-A., Vautard, R., and Legras, B., 1995, 'Weather regimes: recurrence and quasi-stationarity,' *J. Atmos. Sci.*, 52(8), 1237–1256.

Mo, K., and Ghil, M., 'Cluster analysis of multiple planetary flow regimes,' *J. Geophys. Res.*, 93, D9, 10927–10952, 1988.

Preisendorfer, R.W., 1988: *Principal Component Analysis in Meteorology and Oceanography.* C.D. Mobley (Ed.), Elsevier, Amsterdam..

Raftery, A. E., Madigan, D., and Volinsky, C. 'Accounting for model uncertainty in survival analysis improves predictive performance,' in *Bayesian Statistics 5*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), Oxford University Press, 323–349, 1996.

Richardson, S. and Green, P. J., 'On Bayesian analysis of mixtures with an unknown number of components,' Mathematics Research Report S–96–01, University of Bristol, 1996.

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

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

Smyth, P., 'Clustering using Monte-Carlo cross validation,' in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, pp.126–133, 1996.

Smyth, P., 'Clustering sequences using hidden Markov models,' in *Advances in Neural Information Processing 9*, M. C. Mozer, M. I. Jordan and T. Petsche (eds.), Cambridge, MA: MIT Press, 648–654, 1997.

Smyth, P. and D. Wolpert, 'Stacked density estimation,' in *Advances in Neural Information Processing Systems 10*, M. Kearns, M. I. Jordan and S. Solla (eds), to appear, 1998.

Smyth, P., M. Ghil, and K. Ide, 'Multiple regimes in Northern hemisphere height fields via mixture model clustering,' Technical Report UCI-ICS 98–08, Information and Computer Science, University of California, Irvine, 1998.

Thiesson, B, Meek, C., Chickering, D. M., and Heckerman, D., 'Learning mixtures of Bayesian networks,' Technical Report MSR-TR-97-30, Microsoft Research, Redmond, WA, December 1997.

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

Wallace. J. M., 'Observed Climatic Variability: Spatial Structure,' in *Decadal Climate Variability*, D. L. T. Anderson and J. Willebrand (eds.), NATO ASI Series, Springer Verlag, 1996.

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