Statistics and Computing (2000) **9**, 63–72

Model selection for probabilistic clustering using cross-validated likelihood

PADHRAIC SMYTH

Information and Computer Science, University of California, Irvine, CA 92697-3425 (Also with the Jet Propulsion Laboratory 126-347, California Institute of Technology, Pasadena, CA 91109)

Submitted February 1998 and accepted September 1998

Cross-validated likelihood is investigated as a tool for automatically determining the appropriate number of components (given the data) in finite mixture modeling, particularly in the context of model-based probabilistic clustering. The conceptual framework for the cross-validation approach to model selection is straightforward in the sense that models are judged directly on their estimated out-of-sample predictive performance. The cross-validation approach, as well as penalized likelihood and McLachlan's bootstrap method, are applied to two data sets and the results from all three methods are in close agreement. The second data set involves a well-known clustering problem from the atmospheric science literature using historical records of upper atmosphere geopotential height in the Northern hemisphere. Cross-validated likelihood provides an interpretable and objective solution to the atmospheric clustering problem. The clusters found are in agreement with prior analyses of the same data based on 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 compared to penalized likelihood and bootstrap techniques on two real data sets, including 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 1985, 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

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

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

$$\begin{aligned} l\big(\Phi^{(k)} \mid D^{\text{train}}\big) &= \log p\big(D^{\text{train}} \mid \Phi^{(k)}\big) \\ &= \sum_{i=1}^{N} \log\left(\sum_{j=1}^{k} \alpha_j g_j(\underline{x}_i \mid \underline{\theta}_j)\right). \end{aligned} \tag{2}$$

(Note that there are alternative objective functions which can be maximized in the clustering context, e.g., see Celeux and Govaert (1995) 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 convergence to a global maximum. 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 datapoint). Thus, in practice, only maxima in the interior of the parameter space are considered as solutions to the maximum likelihood estimation problem (see Section 4.1 for further details and also McLachlan and Peel (1998) for discussion). The parameters found in this manner using a data set D will be denoted by $\hat{\Phi}^{(k)}(D)$.

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.

Consider the problem of testing the hypothesis of k components versus the hypothesis of $k+1$ components. Tests based on the likelihood ratio test statistic,

$$\lambda = \frac{p\big(D^{\text{train}} \mid \hat{\Phi}^{(k)}\big)}{p\big(D^{\text{train}} \mid \hat{\Phi}^{(k+1)}\big)}, \tag{3}$$

cannot be directly applied in the mixture context due to the breakdown of the standard assumptions on the asymptotic properties of the estimators (Feng and McCulloch 1996).

In the bootstrap approach of McLachlan (1987) (see also Aitkin, Anderson and Hinde 1981) the distribution of $-2\log\lambda$ is estimated by generating B bootstrap samples under the null hypothesis (i.e., from a model fitted to D^{train} with k components), and then estimating $-2\log\lambda$ on each bootstrap sample after fitting models with k and $k+1$ components to each of the B samples. The value of $-2\log\lambda^{\text{train}}$ (obtained from fitting models with k and $k+1$ components to D^{train}) is then compared with the bootstrap values of the ratio statistic to create a bootstrap version of a conventional likelihood-ratio test. Results in McLachlan (1987) and McLachlan and Peel (1997) demonstrate the utility of the bootstrap method on small problems with relatively few datapoints and two or three clusters. The simulation studies in McLachlan and Peel (1997) illustrate a small bias in the bootstrap approach, leading to a slight tendency to be biased in favor of the null hypothesis of k components. Nonetheless, the method appears to be a quite useful framework for the general mixture model model selection problem.

Bayesian and penalized likelihood methods also provide general frameworks for "honest" estimates of the number of components. Penalized likelihood methods (sch as AIC, BIC, MDL, etc.) are typically derived from approximations based on asymptotic arguments as the training data size N approaches ∞ (Schwarz 1978, Kass and Raftery 1995). 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, Smith, and Makov (1985), there are significant theoretical limitations on the applicability of these standard methods to mixture problems. Nonetheless, despite these theoretical reservations, penalized likelihood methods have often been found to work quite well for model selection in mixture problems (for example see the recent results of Fraley and Raftery (1998) using the BIC score function).

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 such as Markov Chain Monte Carlo method (MCMC). Lavine and West (1992), Diebolt and Robert (1994) and Bensmail *et al.* (1997) provide examples of the application of sampling techniques to Bayesian inference for mixture models.

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 approximation-based methods (whether it be penalized likelihood, closed-form approximations to the Bayesian solution, 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 to those discussed above. The cross-validation framework is closest in spirit to the bootstrap method of McLachlan (1987) in the sense that it is a likelihood-based method (rather than fully Bayesian) and has similar computational complexity.

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^{train}, where k ranges from 1 to $k_{\max}$. Thus, we have an indexed set of estimated models, $f^{(k)}(\underline{x} \mid \hat{\Phi}^{(k)})$, $1 \leq k \leq k_{\max}$, where each $f^{(k)}(\underline{x} \mid \hat{\Phi}^{(k)})$ has been fitted to the same data set D^{train}.

Let $l_k^{\text{train}} = l(\hat{\Phi}^k(D^{\text{train}}) \mid D^{\text{train}})$ denote the usual log-likelihood of the fitted model with k components, where the parameters $\hat{\Phi}^{(k)}$ have been determined from 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 $l_k^{\text{test}} = l_k(\hat{\Phi}^{(k)}(D^{\text{train}}) \mid D^{\text{test}})$ 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} datapoints. 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 (Good 1952)).

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

$$i_k = -\frac{l_k^{\text{test}}}{N_{\text{test}}} = -\frac{1}{N_{\text{test}}} l\big(\hat{\Phi}^{(k)}(D^{\text{train}}) \bigm| D^{\text{test}}\big)$$

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 $f(\underline{x})$,

$$\begin{aligned} E[i_k] &= -\frac{1}{N_{\text{test}}} E\big[l\big(\hat{\Phi}^{(k)}(D^{\text{train}}) \bigm| D^{\text{test}}\big)\big] \\ &= -\frac{1}{N_{\text{test}}} \sum_{j=1}^{N_{\text{test}}} E[\log f_k(\underline{x}_j)] \\ &= \int f(\underline{x}) \log \frac{1}{f_k(\underline{x})}\, d\underline{x} \\ &= \int f(\underline{x}) \log \frac{f(\underline{x})}{f_k(\underline{x})}\, d\underline{x} + \int f(\underline{x}) \log \frac{1}{f(\underline{x})}\, d\underline{x} \qquad (4) \end{aligned}$$

i.e., the expected value of i_k is the Kullback-Leibler (KL) distance (Cover and Thomas 1991) between $f(\underline{x})$ and $f_k(\underline{x})$ (the first term on the right in equation (4)), plus a constant which is independent of k (namely, the entropy of the true density function $f(\underline{x})$, the second term on the right above). Thus, the test log-likelihood l_k^{test} (scaled appropriately) is an unbiased estimator (within a constant) of the 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 $D \backslash S_i$ 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:

$$l_k^{\text{cv}} = \frac{1}{M} \sum_{i=1}^{M} l\big(\hat{\Phi}^{(k)}(D \backslash S_i) \bigm| S_i\big) \qquad (5)$$

where $\hat{\Phi}^{(k)}(D \backslash S_i)$ denotes the parameters for the kth model estimated from the ith training subset, and $l(\hat{\Phi}^{(k)}(D \backslash S_i) \mid S_i)$ is the log-likelihood evaluated on the data in S_i using the parameters estimated from the data $D \backslash S_i$.

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 of the data, 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, 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). This main difference between this and the v-fold method is that each datapoint may be used as a test point more than once. 10-fold cross-validation was found to be much less reliable than MCCV ($\beta = 0.5$) in terms of choosing the correct number of components on a set of both simulated and real mixture modeling problems (Smyth 1996).

In general, there appears to be no obvious systematic method for automatically determining the best value of 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 of M 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 roughly a factor M. The bootstrap approach also increases the computational burden by a factor of B (the number of bootstrap samples) over the penalized likelihood methods. Thus, if B and M are of the same order, the computational cost of the bootstrap and cross-validation will be comparable. An important point to note is that the computational cost of both the bootstrap and cross-validation approaches are *linear* in the number of resampling runs B or M. In addition, both methods could be easily and efficiently implemented on parallel computing hardware, e.g., using B or M parallel processors.

In this paper we limit our attention to experiments with the non-Bayesian approaches (penalized likelihood, bootstrap, and cross-validation) but clearly there is room for further study comparing these methods to the Bayesian MCMC methodologies.

4. Applications of cross-validated clustering

4.1. *Experimental methods*

For the cross-validation and BIC results below the EM algorithm was implemented as follows. The algorithm was started from 3 different randomly chosen initial partitions of the data as well as from 3 different partitions generated by the k-means algorithm. The algorithm was stopped in each case either when (a) 500 iterations of the algorithm were complete, or (b) when the increase in log-likelihood from the most recent two iterations of EM was below 10^{-4} of the increase in log-likelihood from the first two iterations of EM. The parameter solution consisting of the largest likelihood among these solutions was then chosen as the maximum likelihood solution, excluding any solution where $\hat{\sigma}_l^j < 0.01\sigma_l$, where $\hat{\sigma}_l^j$ is the estimated standard deviation for the lth variable in the Gaussian model of the jth component $(1 \leq j \leq k)$ and σ_l is the population standard deviation for the lth variable, $1 \leq l \leq d$. This latter step eliminates spurious local maxima on the edges of parameter space.

For the BIC method, the above procedure is run once on all of the training data and the penalty term $(p_k/2) \log n$ is subtracted from the maximum likelihood value. For the cross-validation method, for each partition of the data, models with $k = 1$ to $k = k_{\max}$ components are fitted on the training portion of the data using exactly the method above, and the log-score (or test set log-likelihood) is then calculated (for each k) on the out-of-sample test portion of the data. This procedure is repeated on M randomly-chosen partitions of the data and the resulting log-likelihood scores are averaged over M (equation (5)). In the results reported here the data are partitioned repeatedly into two disjoint subsets of equal size (i.e., MCCV with $\beta = 0.5$).

The MIXFIT software (courtesy of G. McLachlan) was also used to test the bootstrap method on the data sets described below. The software was configured to run the EM algorithm using the best initial condition found from among (a) 3 randomly chosen partitions of the data and (b) 3 partitions generated from the k-means algorithm. The option to initialize the MIXFIT algorithm using hierarchical clustering was disabled due to its $O(N^2)$ complexity ($N = 3960$ for one of the data sets below). Monitoring of the maxima found by EM in parameter space provided no evidence that poor local maxima were being found, i.e., for these particular data sets the random/k-means method of initializing EM appeared quite adequate. The EM algorithm was halted using the default method in MIXFIT, i.e., after either (a) a maximum of 500 iterations or (b) when the change in likelihood between the current iteration and the likelihood ten iterations previous is less than 10^{-6}, whichever of (a) or (b) occurs earlier.

99 bootstrap replications ($B = 99$) were run for each data set, testing $k = 1$ vs $k = 2$, $k = 2$ vs. $k = 3$, and so forth. $B = (100/\alpha) - 1$ permits significance testing at the α% level (McLachlan and Peel 1997, 1998). For $B = 99$ for example, the bootstrap algorithm roughly matches the same number of computational steps as the cross-validation approach with $M = 100$ and allows significance testing at the 1% level.

4.2. *Clustering of diabetes patients*

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 diabetic. This data set has since been analyzed in the clustering literature by Symons (1981), Banfield and Raftery (1993), and Fraley and Raftery (1998). Here we analyze the unlabeled data, i.e., the 3-dimensional measurements without any class labels.

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 (Fig. 1). For example in the plot of *sspg* versus *glucose* there is a cluster of points in the lower left corner as well as two "wings" to the data in roughly orthogonal directions.

Table 1 summarizes the results for BIC and cross-validated likelihood ($M = 100$). Both BIC and cross-validated likelihood choose $k = 3$ as the most likely value for k. The bootstrap method (as implemented in the MIXFIT software described earlier, using $B = 99$) indicated that the null hypothesis $k = 2$ is rejected (vs. $k = 3$) at the 1% level but that the null hypothesis $k = 3$ is not rejected at the 1% level when compared to $k = 4$. All three methods (cross-validation, bootstrap, and BIC) are in agreement on this data set, finding the same number of classes ($k = 3$) as that of the original clinical classification. Table 1 also shows the BIC and cross-validated scores per datapoint (original scores divided by N and $N/2$ respectively) indicating fairly close agreement between these normalized scores.

Table 1. *Comparison of scores for Gaussian mixture models with components from $k = 1$ to $k = 4$ on the diabetes patient data*

Scoring method	$k = 1$	$k = 2$	$k = 3$	$k = 4$
p_k	9	19	29	39
Likelihood	−2545.8	−2355.9	−2303.5	−2279.8
BIC	−2583.2	−2433.0	−2420.4	−2436.6
CV-Likelihood	−1287.5	−1219.6	−1207.8	−1229.5
BIC/datapoint	−17.82	−16.78	−16.69	−16.80
CV-Likelihood/ datapoint	−17.88	−16.94	−16.78	−17.08

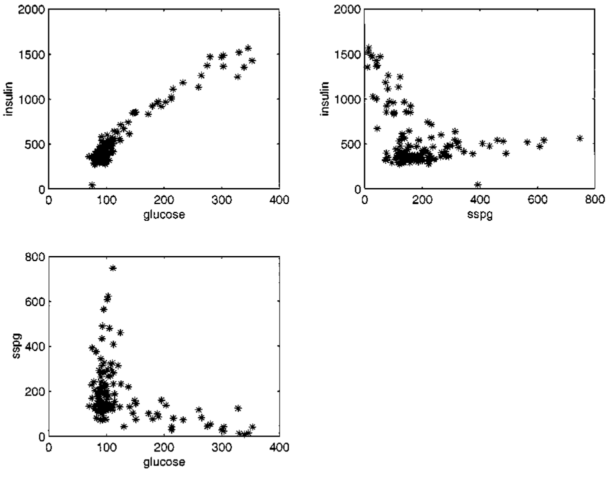

Fig. 1. *Scatter plots of diabetes data, without the class labels.*

4.3. *Application of cross-validated clustering to atmospheric geopotential height data*

4.3.1. 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 500 mb height data, 700 mb 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 identification 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.

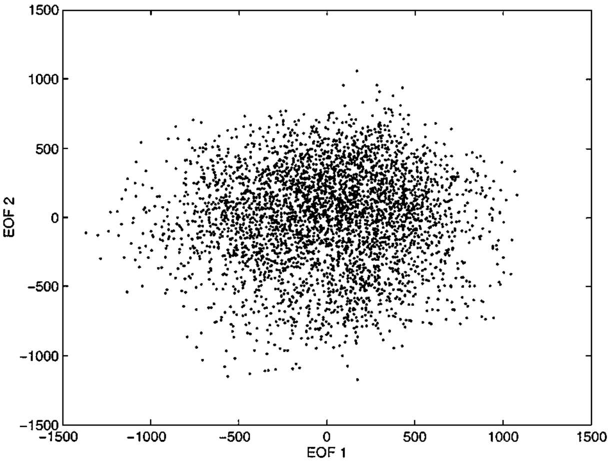

Fig. 2. *Scatter plot of NH winter data projected into first 2 EOF directions*

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° × 10° 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, Ide and Ghil (1999)). 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. The 12-dimensional data are publically available online by anonymous ftp from `ftp.ics.uci.edu/pub/smyth/data/atmos/`.

Figure 2 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 the past and it is on this data set that the application of cross-validation for model selection is investigated below.

4.3.2. 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, Ide, and Ghil (1999)). In all experiments the number of cross-validation partitions was $M = 100$ 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 $k = 1$ to $k = 15$. 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 are tabulated in Table 2 in addition to the BIC scores. The cross-validation score is maximized at $k = 3$ as is the BIC score, providing evidence of three clusters under the Gaussian mixture model assumption. The normalized scores for BIC and cross-validation also appear reasonably correlated. Approximate posterior probabilities on k (assuming equal priors on different values of k) can be calculated by exponentiating and normalizing the cross-validated log-likelihoods. For the figures in Table 2, the posterior probability estimated in this fashion is effectively 1 for $k = 3$ and 0 for other k values (within 4 decimal places).

The bootstrap method was also applied to the same data with $B = 99$ bootstrap replications. The null hypotheses $k = 2$ (versus $k = 3$) was rejected at the 1% level, while the null hypothesis $k = 3$ was not rejected at the 1% level, when compared to $k = 4$.

Thus, the cross-validation, bootstrap, and BIC methods all point to $k = 3$ as the most likely number of components, assuming a Gaussian mixture model. On checking the maximum likelihood parameter values for each method (cross-validation, bootstrap, BIC) there was no evidence of any problems with poor local maxima, i.e., all methods appeared to be finding the same maxima consistently for each value of k. Figure 3 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 4 shows the test log-likelihoods on the first 20 (from 100 total) 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

Table 2. *Comparison of scores for Gaussian mixture models with components from $k = 1$ to $k = 6$ on the atmospheric data*

Scoring method	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
p_k	5	11	17	23	29	35
Likelihood	−1890.4	−1799.3	−1745.6	−1733.7	−1724.3	−1711.7
BIC	−1923.6	−1869.7	−1853.3	−1878.7	−1906.6	−1931.2
CV-Likelihood	−982.7	−967.7	−949.1	−963.2	−967.8	−972.8
BIC/datapoint	−0.486	−0.472	−0.468	−0.474	−0.481	−0.489
CV-Likelihood/datapoint	−0.496	−0.489	−0.479	−0.486	−0.489	−0.491

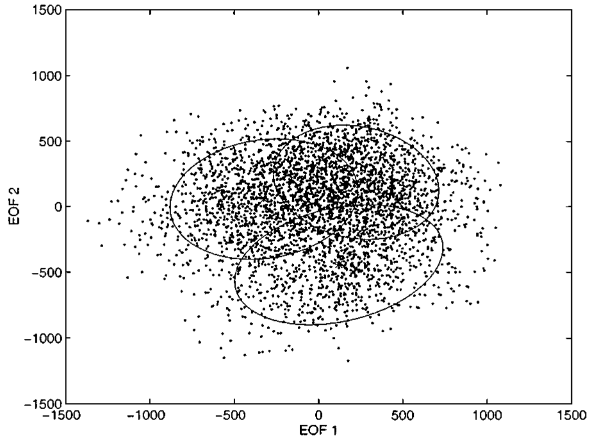

Fig. 3. *Scatter plot of NH winter data projected into the first 2 EOF directions with estimated means and covariance matrix shapes (ellipses) superposed as fitted by the EM procedure with a 3-component Gaussian mixture model. The ellipses represent contours of the density function which are 3-sigma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

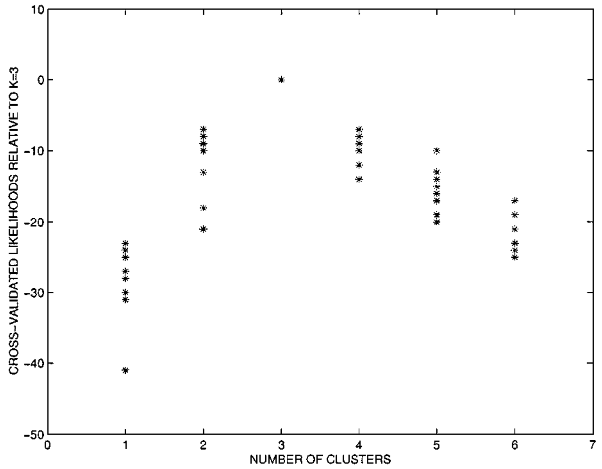

Fig. 5. *Cross-validated log-likelihoods for $k = 1, \ldots, 6$ relative to the cross-validated log-likelihood of the $k = 3$ model for 10 such different randomly-chosen cross-validation partitions*

average across the partitions it is significantly better than the other possible values for k.

4.3.3. 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 Fig. 5 where the likelihoods for each k are plotted relative to $k = 3$.

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 of parameters in a k-component Gaussian mixture model grows as kd^2. Since the total amount of data

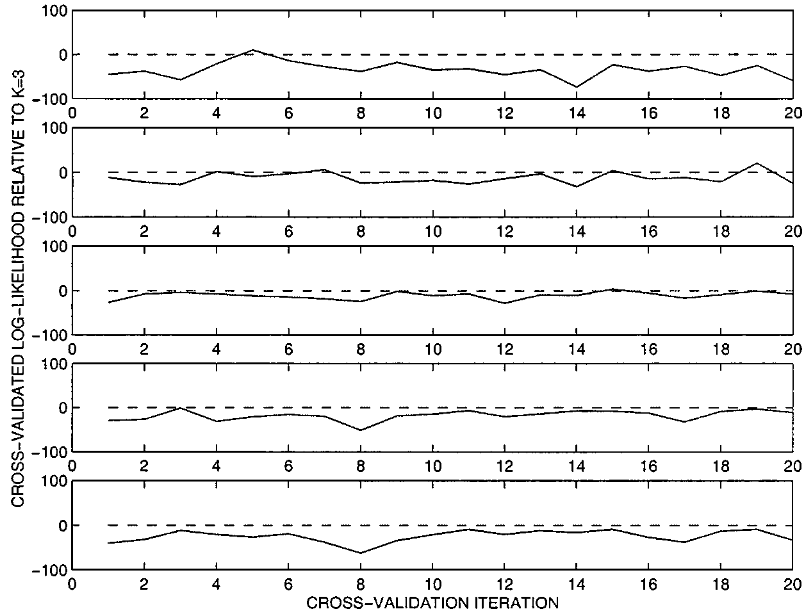

Fig. 4. *Log-likelihood of the test partition data on for the first 20 cross-validation runs relative to the log-likelihood of the $k = 3$ model for (from top) (a) $k = 1$, (b) $k = 2$, (c) $k = 4$, (d) $k = 5$, and (e) $k = 6$*

Table 3. *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

to fit the models is fixed, as the dimensionality d increases there are fewer datapoints relative to the number of parameters being estimated, and thus, one can anticipate increased variance (and less reliability) in the higher-dimensional parameter estimates.

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 datapoint 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 3. It is clear that there is a high correlation between the maps obtained using only the first two EOF dimensions and each of the maps obtained using k EOF dimensions, $3 \leq k \leq 12$. This indicates that as the dimensionality of the EOF space grows beyond $d = 2$, the clusters in any of these dimensional spaces are essentially the same as for the two-dimensional sub-space.

4.3.4. 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 6 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 500 mb 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.

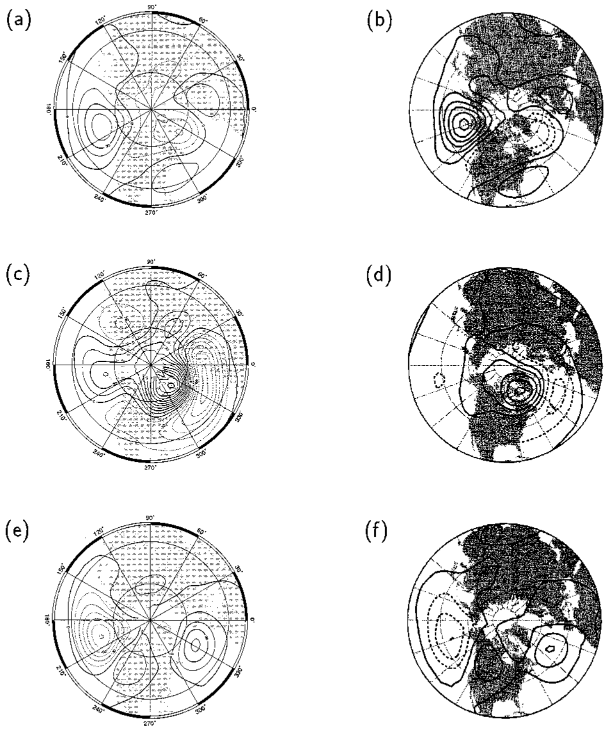

Fig. 6. *Geopotential height maps for the 3 cluster centers of the mixture model (left: panels a, c and e) and of Cheng and Wallace's (1993) hierarchical cluster model (right: panels b, d and f) which are reproduced in Wallace (1996) (panels b, d, and f reproduced by permission)*

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 results. 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, Ide and Ghil (1999).

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 Fig. 6 (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 (1999) extend the framework in this paper to *model averaging* of mixture models for density estimation, 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 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. 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. The word "practical" here is intended in the sense that the cross-validation method is relatively straight-forward to implement and the results have a direct interpretation. The method is particularly useful in cases when data and computational resources are relatively plentiful.

Acknowledgments

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.

References

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

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

Celeux G. and Govaert G. 1995. Gaussian parsimonious clustering models. Pattern Recognition 28: 781–793.

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

Chickering D.M. and Heckerman D. 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Machine Learning 29(2/3): 181–244.

Cover T.A. and Thomas J.M. 1991. Elements of Information Theory, New York, John Wiley.

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

Diebolt J. and Robert C.P. 1994. Bayesian estimation of finite mixture distributions. J. R. Statist. Soc. B 56: 363–375.

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

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

Fraley C. and Raftery A.E. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal 41: 578–588.

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

Hjorth J.S.U. 1994. Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap, Chapman and Hall, UK.

Kass R.E. and Raftery A.E. 1995. Bayes factors. J. Am. Stat. Assoc. 90. 773–795.

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

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): 2625–2643.

Lavine M. and West M. 1992. A Bayesian method for classification and discrimination. Can. J. Statist. 20: 451–461.

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

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

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

McLachlan G.J. and Peel D. 1997. On a resampling approach to choosing the number of components in normal mixture models. In: L. Billard and N.I. Fisher (Eds.). Computing Science and Statistics (Vol. 28), Fairfax Station, Virginia, Interface Foundation of North America, pp. 260–266.

McLachlan G.J. and Peel D. 1998. MIXFIT: An algorithm for the automatic fitting and testing of normal mixture models. In: Proceedings of the 14th International Conference on Pattern Recognition, Vol. I, Los Alamitos, CA, IEEE Computer Society, pp. 553–557.

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. 1988. Cluster analysis of multiple planetary flow regimes. J. Geophys. Res. 93, D9: 10927–10952.

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

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

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

Schwarz G. 1978. Estimating the dimensions of a model. Annals of Statistics 6: 461–462.

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

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

Smyth P. 1996. 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.

Smyth P. 1997. Clustering sequences using hidden Markov models. In: Mozer M.C., Jordan M.I., and Petsche T. (Eds.), Advances in Neural Information Processing 9. Cambridge, MA: MIT Press, 648–654.

Smyth P., Ide K., and Ghil M. 1999. Multiple regimes in Northern hemisphere height fields via mixture model clustering. Journal of Atmospheric Sciences 56(21): 3704–3723.

Smyth P. and Wolpert D. 1999. Linearly combining density estimators via stacking. Machine Learning 36(1): 59–83.

Symons M. 1981. Clustering criteria and multivariate normal mixtures. Biometrics 37: 35–43.

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

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

Wallace J.M. 1996. Observed Climatic Variability: Spatial Structure. In: Anderson D.L.T. and Willebrand J. (Eds.), Decadal Climate Variability, NATO ASI Series, Springer Verlag.

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