Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Entropic component analysis and its application in geological data

Chih-Yuan Tseng[a], Chien-Chih Chen[b,*]

[a] Department of Oncology, University of Alberta, Edmonton, AB, Canada T6Z 1G2
[b] Department of Earth Sciences and Graduate Institute of Geophysics, National Central University, Jhongli 320, Taiwan

ARTICLE INFO

Article history:
Received 30 March 2010
Received in revised form
27 October 2010
Accepted 16 November 2010
Available online 19 January 2011

Keywords:
Model selection
Maximum entropy
Logistic likelihood function
Sample classification
Oil-field brines

ABSTRACT

We present an entropic component analysis for identifying key parameters or variables and the joint effects of various parameters that characterize complex systems. This approach identifies key parameters through solving the variable selection problem. It consists of two steps. First, a Bayesian approach is utilized to convert the variable selection problem into the model selection problem. Second, the model selection is achieved uniquely by evaluating the information difference of models by relative entropies of these models and a reference model. We study a geological sample classification problem, where a brine sample from Texas and Oklahoma oil field is considered, to illustrate and examine the proposed approach. The results are consistent with qualitative analysis of the lithology and quantitative discriminant function analysis. Furthermore, the proposed approach reveals the joint effects of the parameters, while it is unclear from the discriminant function analysis. The proposed approach could be thus promising to various geological data analysis.

1. Introduction

Identifying key parameters or variables, which characterize complex systems from experimental datum, is a necessary and yet difficult data analysis process for further investigations in science and engineering. Many methods such as principal component analysis (PCA) (Jolliffe, 2002), independent component analysis (ICA) (Comon, 1994; Stone, 2004), discriminant function analysis (DFA) (Davis, 2002), etc. have been proposed to provide an objective approach to accomplish this goal. These approaches basically project possible parameters either into principal components using the PCA or the DFA or independent components using the ICA. These components, linear combinations of the parameters with specific weightings, are then used to characterize the complex systems of interest. Yet there are no assessments regarding effects of individual parameter on the complex systems directly from these approaches and does not directly identify key parameters either.

Alternatively, one can treat the identification of key parameters as a variable selection problem. One standard approach to tackle the variable selection problem is to convert the problem into a model selection problem. A model constructed from the data is associated with the possible parameters or variables and experimental responses. Therefore, selecting variables is identical to selecting models.

Many strategies have been proposed for model selection. In standard statistics, the *P*-value method selects a model based on a hypothesis-testing procedure. However, it is limited to two-model selection problems (Raftery, 1995). Some other approaches, such as those based on the Akaike information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (BIC) (Schwarz, 1978; Raftery, 1995; Kieseppa, 2000; Forbes and Peyrard, 2003), the C information criterion (CIC) (Rodriguez, 2005), generalizations of the BIC and the AIC, and the relative entropy, the mutual information, or Kullback–Leibler distance (Bonnlander and Weigend, 1994; Dupuis and Robert, 2003), attempt to provide another appropriate selection criterion for multiple models. Alternatively, one can select a model according to a quantity that we have called the "preference". Tseng (2006) showed that the preference of a model could be uniquely obtained by evaluating the relative entropy of the model and a reference model.

In this work, we propose an entropic component analysis (ECA) based on the concept of variable selection to identify key parameters to optimally characterize the complex systems. The proposed approach consists of two steps. First, we utilize a Bayesian approach to model the systems of interest instead of seeking out common regression approaches. The idea is applying Bayes' theorem to update our prior knowledge of a system according to experimental data of input parameters to obtain a posterior probability distribution model of observing responses. Second, we consider the method of maximum entropy as a tool to

* Corresponding author. Tel.: +886 3 422 7151x65653; fax: +886 3 422 2044.
E-mail addresses: chih-yuan.tseng@ualberta.ca, richard617@gmail.com (C.-Y. Tseng), chencc@earth.ncu.edu.tw (C.-C. Chen).

0098-3004/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2010.11.016

rank the model. Based on the axiomatic approach used in developing the maximum entropy method as a tool for assigning a probability to a system (Jaynes, 1957) and for updating probabilities (Shore and Johnson, 1980, 1981; Skilling, 1988, 1989, 1990; Caticha, 2004), we will show that preferences of the parameters can be uniquely determined by evaluating the relative entropy of the models with respect to a reference.

We examine the ECA by studying a geological sample classification problem, whether some brine samples from oil field waters in Texas and Oklahoma (Davis, 2002) belong to the Grayburg Dolomite. In addition, the results obtained from the ECA are compared with a qualitative analysis of the lithology and a quantitative analysis based on a discriminant function analysis. We also apply the receiver operator characteristic analysis (ROC) (Johnson and Albert, 1999) to examine the performance of the ECA.

The structure of paper is as following. Section 2 presents the ECA. Section 3 discusses the example to examine the ECA. The first two sections present a qualitative discussion and a quantitative analysis based on DFA. The remaining sections then present the ECA analysis. Finally, the conclusions are given.

2. Entropic component analysis

2.1. From variable selection to model selection

To appropriately model the system of interests, we apply a Bayesian approach. First, we only consider a binary-output problem here. The positive response of the ith observation is denoted by $y^i=1$ and the negative response is $y^i=0$. Next, we assume the information regarding properties of the system of interest before any studies on that system are conducted is described by a prior probability distribution of observing positive responses $P(y^i=1)$. Finally, suppose N parameters that characterize the system are measured and denoted by $X^i=\{x_j{}^i\}$, where subscript $j=1,\ldots,N$ labels parameter and superscript $i=1,\ldots,l$ labels the observations. Furthermore, l corresponding responses or dependent variables measured are denoted by $\widehat{Y}=\{y^i\}$.

Given the above three considerations, the posterior probability updated from the prior distribution $P(y^i=1)$ according to measurements $X^i=\{x_j^i\}$ can be obtained from Bayes' theorem, $P(y^i=1|X^i,\hat{\beta})=P(y^i=1)L_{y^j}(X^i,\hat{\beta})$, where $L_{y^i=1}(X^i,\hat{\beta})$ is the likelihood function and $\hat{\beta}=\{\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_N\}$ are maximum likelihood estimates (MLE) (Johnson and Albert, 1999). This probability distribution is then considered to be the preferred probability model to associate the observations with the systems. For N parameters, there are 2^N-2 different combinations (sets) S_i of parameters $X_{Si}\in X$. Similarly, one also can define submodels as $P(y^i=1|X^i_{S_i},\hat{\beta})$. Thus selecting the parameters $X_{Si}\in X$ is identical to selecting the corresponding $P(y^i=1|X^i_{S_i},\hat{\beta})$. The ranking scheme of parameters can be pertinently obtained through determining the ranking scheme of the posterior probability models.

2.2. Ranking scheme

The axiomatic approach suggested a solution to rank models by the preference of models. As shown in Shore and Johnson (1980, 1981), Skilling (1988, 1989, 1990), and Caticha (2004), it uniquely determines the preference S of probability models $p(x)$. In this approach, the axioms used reflect the condition that one must not change mind frivolously when ranking the probability distributions, and that whatever information was originally codified into the probability distribution is important and should be preserved. Three axioms, locality, coordinate invariance and consistency chosen for this purpose (Caticha, 2004) are briefly discussed below.

The locality axiom states that *local information has local effects.* This axiom results with non-overlapping domains of parameter x contribute additively to the quantity $S[p]=\int dxF(p(x),x)$, where F is some unknown function.

The second axiom, coordinate invariance, states that *the ranking should not depend on the system of coordinates*. The coordinates or parameters x used to label the system are arbitrary. In certain situations, we might have explicit reasons to believe that a particular choice of coordinates should be preferred over other possibilities. But unless evidence has in fact been given, we should not assume it. Namely, if there is no evidence to indicate that a set of parameters x is preferred over another set, the ranking of probabilities defined by these two sets of parameters should be independent of the coordinates, the two sets of parameters. A consequence of this axiom is that $S[p]=\int dxm(x)f(p(x)/m(x))$ involves coordinate invariant such as $dxm(x)$ and $p(x)/m(x)$, where the density $m(x)$ represents a Jacobian and f are, at this point, undetermined.

The third axiom states that when a system is composed of subsystems that are believed to be independent, it should not matter whether we treat them separately or jointly. Namely, probability distributions of the system and the independent subsystems must satisfy the product rule of probability theory. This axiom restricts the function f to being a logarithm.

The overall consequence of these three axioms is that the probability distribution $p(x)$ should be ranked relative to $m(x)$ according to the functional form

$$S[p,m]=-\int dxp(x)\log\frac{p(x)}{m(x)}\leq 0, \tag{1}$$

where this quantity is called relative entropy.

To determine $m(x)$, in principle, the real underlying function that associates with input parameters and responses should be chosen as the reference. However, it is difficult to determine such a real underlying function practically. As suggested by Tseng (2006), we also consider a uniform distribution $\mu=1/l$, which represents the complete ignorance of the system of interests and contains no information. Rewriting Eq. (1) gives

$$S[p,\mu]=\log\mu+S[p], \tag{2}$$

where $S[p]=-\int dxp(x)\log p(x)$. Increasing $S[p,\mu]$ indicates that $p(x)$ tends to become information free. Therefore, given a family of all possible models, the preferred one that contains mostly information relevant to the system will minimize $S[p,\mu]$. Namely, parameters considered in such a model will be the preferred key parameters.

2.3. Analysis strategy

The above derivation has singled out $S[p,\mu]$ as the unique entropy to be used for the purpose of ranking probability distributions. Other expressions may be useful for other purposes, but they are not a generalization from the simple cases described in axioms above. Furthermore, it can be applied to all kinds of probability models. However, when the number of input parameters gets large, ranking all possible submodels is unwieldy and slow. We propose a two-step approach, ECA, to efficiently identify key parameters and analyze the joint effects of various combinations of parameters. The first step evaluates the ranking scheme of all single variables. From which, one can determine variables that are highly likely responsible to responses of the system. Given these parameters, the second step enumerates the ranking scheme of various combinations of rest of parameters. The characteristics of the system may then be inferred from these results.

Next, we will illustrate the use of the ECA in detail by studying a geological sample classification problem (Davis, 2002).

3. Example: a geological sample classification problem

3.1. The problem

Saltwater is trapped in sedimentary rocks when they are formed in a marine environment. The chemical composition of the connate water is subsequently modified by several mechanisms including ion exchange, mixing with other brines, dilution, and infiltrating surface water. Brines recovered during drillstem tests of wells may have relict compositional characteristics that provide clues to the origin or depositional environment of their source rocks. Two questions need to be addressed before further investigations. The first question is "How does one classify such samples through compositional analysis?" and the second is "Which compositions play key roles in the depositional process?"

To illustrate the application of the ECA scheme to answer both questions, we consider an example of a set of analyses of oil field water from three groups of carbonate units in Texas and Oklahoma (Davis, 2002). The data from one of the analyses are given in Table 1. The fi Th column in Table 1 denotes whether the brine samples belong to a specific carbonate unit, the Grayburg Dolomite, referred to as unit G here for short. There are 19 measurements regarding the responses and the concentrations of the six chemical compounds. These six chemical compounds are treated as parameters. There are five positive experimental responses, $y^i=1$, denoted by the symbol "*Y*", and 14 negative responses, $y^i=0$, denoted by the symbol "*N*". Before we apply the ECA to answer questions, we present a qualitative analysis of the lithology and a quantitative analysis based on the discriminant function method as a benchmark.

3.2. Qualitative analysis of the lithology

Unit G is composed mainly of dolomite ($CaMg(CO_3)_2$) and anhydrite ($CaSO_4$). In ancient geological times, unit G experienced two important sedimentary processes, of dolomitization, which is associated with the dissolution of calcite by acidic fluids, and evaporation (Ostroff, 1967; Roche, 1997; Davis, 2002). Anhydrite is one of the index products of evaporation. Chalcraft and Ward (1988) claimed further that the principal diagenetic processes here included dolomitization, anhydrite occlusion in primary porosity, and leaching. The dolomitization played a crucial role in the formation of unit G, and was followed by the anhydrite occlusion. One can therefore infer that the process of the dissolution of calcite by acidic fluids was more significant than anhydrite occlusion in the formation of unit G. Therefore, the most important chemical species in the process is HCO_3.

Table 1
Chemical analyses (in ppm) of brines recovered from drillstem tests of three carbonate rock units in Texas and Oklahoma (Davis, 2002).

Unit G	HCO_3	SO_4	Cl	Ca	Mg	Na
N	10.4	30	967.1	95.9	53.7	857.7
N	6.2	29.6	1174.9	111.7	43.9	1054.7
N	2.1	11.4	2387.1	348.3	119.3	1932.4
N	8.5	22.5	2186.1	339.6	73.6	1803.4
N	6.7	32.8	2015.5	287.6	75.1	1691.8
N	3.8	18.9	2175.8	340.4	63.8	1793.9
N	1.5	16.5	2367	412	95.8	1872.5
Y	25.6	0	134.7	12.7	7.1	134.7
Y	12	104.6	3163.8	95.6	90.1	3093.9
Y	9	104	1342.6	104.9	160.2	1190.1
Y	13.7	103.3	2151.6	103.7	70	2054.6
Y	16.6	92.3	905.1	91.5	50.9	871.4
Y	14.1	80.1	554.8	118.9	62.3	472.4
N	1.3	10.4	3399.5	532.3	235.6	2642.5
N	3.6	5.2	974.5	147.5	69	768.1
N	0.8	9.8	1430.2	295.7	118.4	1027.1
N	1.8	25.6	183.2	35.4	13.5	161.5
N	8.8	3.4	289.9	32.8	22.4	225.2
N	6.3	16.7	360.9	41.9	24	318.1

3.3. Quantitative analysis—discriminant function analysis

Davis (2002) proposed the use of discriminant function analysis (DFA) to quantitatively analyze the problem above. The DFA is designed to find a set of linear weights for the parameters that causes a multivariate analog of the F-ratio to be a maximum. It combines a rationale similar to that of analysis of the variance of data with computational procedures based on eigenvector calculations, for example the PCA. A succession of discriminant functions along which the samples are as distinct as possible can thus be calculated. Therefore, each function represents successively the most possible efficient discriminator. One can then use DFA on analyzing multivariate measurements made on the samples alone to find combinations of measurements that allow various categories of samples to be distinguished. For detailed calculations, see Davis (2002).

DFA was applied to determine whether the data given in Table 1 were distinctive. The first discriminant function was found to be $-0.3765\cdot[HCO_3]$ $-0.0468\cdot[SO_4]$ $+0.0112\cdot[Cl]$ $-0.0148\cdot[Ca]$ $-0.0174\cdot[Mg]$ $-0.0110\cdot[Na]$, which distinctively separates the samples of unit G from other units. Note that the notation of square bracket represents the concentration of a compound. One may argue that the coefficients in the discriminant function represent weighting factors that indicate effects of the corresponding parameters. Therefore, since parameters HCO_3 and SO_4 have two largest factors in magnitude among the six parameters, it suggests that both parameters play the most dominant role in the classification. Although this result consists with the qualitative analysis from the lithology, one should be cautious of using DFA to select key parameters. There is no rigorous justification of identifying preference of parameters through magnitudes of coefficients of a discriminant function. Besides, it is not straightforward to determine the joint effects of parameters from DFA.

3.4. Entropic component analysis

3.4.1. Analysis procedure

The first step of ECA is to apply a Bayesian approach to convert the variable selection problem into the model selection problem. The success of the ECA hinges on the appropriate choice of the likelihood functions. Because Cox and Snell (1989) have shown empirically, the likelihood function for a generic binary-response system is properly given by the logistic function, we also defined the logistic likelihood function for the six chemical species, $\{HCO_3, SO_4, Cl, Ca, Mg, Na\}$, which are denoted by parameters $X^i=\{x_j{}^i;\ i=1,\ldots,19 \text{ and } j=1,\ldots,6\}$, respectively, with corresponding binary responses in this geological example as

$$L_{y^i=1}(X^i,\hat{\beta})=\frac{\exp\sum_{j=1}^{6}\beta_j x_j^i}{\exp\sum_{j=1}^{6}\beta_j x_j^i+1} \tag{3}$$

Note that x_j^i represents the *i*th concentration measurement for chemical species *j*. Afterward, the posterior distribution updated from a prior distribution $P(y^i=1)=1/19$, which denotes complete ignorance of the occurrence of the positive response, is given by $P(y^i=1\,|\,X^i,\hat{\beta})=P(y^i=1)L_{y^i}(X^i,\hat{\beta})$ based on data of Table 1. The coefficients β_j were determined through the MLE method. A MATLAB code given in Johnson and Albert (1999) was used. Note that the probability model that includes all six variables is called the full model. One can have, $2^6-2=62$ probability submodels $P(y^i=1|X^i_{S_k},\hat{\beta})$, in which various combinations of parameters $X_{S_k} \in X$ are considered.

After obtaining these 62 submodels, the second step of ECA then evaluates the ranking scheme of 62 submodels $P(y^i=1|X^i_{S_k},\hat{\beta})$ and one full model based on Eq. (2). For each submodel, we calculated probabilities of positive response for nineteen measurements. Therefore, the entropy of each submodel, $S[P]=-\sum_{i=1}^{19} P(y^i=1|X^i_{S_k},\hat{\beta})\log P(y^i=1|X^i_{S_k},\hat{\beta})$, was calculated by averaging the natural logarithm of corresponding probabilities of positive response over all nineteen measurements x^i_j. The detailed analysis procedure for identifying key parameters and joint effects is illustrated in the next section.

3.4.2. Results

The ranking scheme of 62 submodels is shown in Fig. 1, which plots the entropy value against submodels. The inset plots result from submodels 40–62. One can roughly group the 62 submodels into four groups marked in the figure. The label 63 denotes the full model. The group 1 contains models that have the five largest entropy values distributed within 6.8 and 7.0; these are (000001), (000010), (001000), (001010), and (000011), where "0" denotes that the corresponding parameter is not included in the model, and "1" denotes the presence of the parameter. Note that the first (HCO_3), second (SO_4), and fourth (Ca) parameters are not included in this group. The group 2 consists of models 21–30 with entropy values around 2.5, in which the occurrence of the second (SO_4) and the fourth (Ca) parameters are the most (six and seven out of ten, respectively). When the first (HCO_3) and the fifth (Mg) parameters are considered, we obtain models 40–46, which form the group 3; this has the second lowest entropy values distributed within 1 and 1.5. The last 16 models, which consider the first, (HCO_3) and second (SO_4) parameters simultaneously, form the group 4, which has the minimum entropy about 0.001. We list only submodels from the fourth group and two models that include either the first (HCO_3) or the second (SO_4) parameter in Table 2, and the corresponding entropy values. The first column denotes the model label as used in Fig. 1 and the second to seven columns indicate which of the six chemical species are included in the submodel. The entropy values of submodels, $P(y^i=1|X^i_{S_i},\hat{\beta})$ are shown in the last column. The ranking scheme of these 22 submodels is in the order of decreasing entropy value.

We subsequently analyzed this ranking scheme in two steps. First, we analyzed the ranking scheme of the single parameters, as shown in the first six rows in Table 2, which states that HCO_3 $(1.508) < SO_4$ $(3.577) <$ Ca $(4.875) <$ Cl $(6.88) <$ Mg $(6.906) <$ Na (6.909), where numerical values in the bracket are entropy values. Since the minimum number of significant figures in the experimental data in Table 1 is three, the entropy value should also have three significant figures and the fourth digit is just an estimate. The ranking scheme indicates that the first parameter HCO_3 should play a more important role than the second parameter SO_4 in the model.

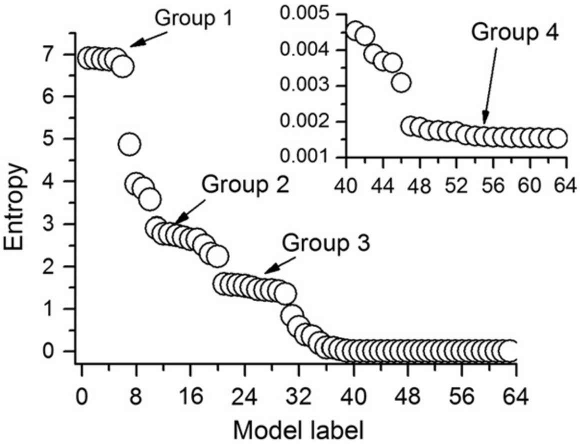

Fig. 1. Plot of the entropy calculated from Eq. (2) given the probability model $P(y^i=1|X^i_{S_i},\hat{\beta})$ versus the model label. The models are numbered in order of entropy value so that model 1 has the largest entropy and is followed by model 2, and so on. Table 2 gives an example of an exact representation of the model corresponding to a specific model label.

Table 2
Ranking scheme of the six chemical species. A number "1" denotes a species considered and "0" denotes a species neglected. The value of the entropy $S[P(y^i=1|X^i_{S_i},\hat{\beta})]$ is given by Eq. (2) with $P(y^i=1|X^i_{S_i},\hat{\beta})$. Each row represents a submodel. Only 22 submodels are listed.

Model	HCO_3	SO_4	Cl	Ca	Mg	Na	$S[P(y^i=1\|X^i_{S_i},\hat{\beta})]$
1	0	0	0	0	0	1	6.909(39)
2	0	0	0	0	1	0	6.906(98)
3	0	0	1	0	0	0	6.880(6)
7	0	0	0	1	0	0	4.875(2)
10	0	1	0	0	0	0	3.577(66)
24	1	0	0	0	0	0	1.508(13)
47	1	1	0	0	0	0	0.001(87)
48	1	1	0	1	0	0	0.001(85)
49	1	1	1	1	0	0	0.001(76)
50	1	1	0	1	0	1	0.001(74)
51	1	1	1	0	0	0	0.001(73)
52	1	1	0	0	0	1	0.001(72)
53	1	0	1	1	1	1	0.001(63)
54	1	1	1	0	0	1	0.001(59)
55	1	1	0	0	1	0	0.001(59)
56	1	1	1	1	0	1	0.001(57)
57	1	1	1	0	1	0	0.001(56)
58	1	1	0	0	1	1	0.001(56)
59	1	1	0	1	1	0	0.001(56)
60	1	1	1	0	1	1	0.001(55)
61	1	1	0	1	1	1	0.001(55)
62	1	1	1	1	1	0	0.001(55)
63	1	1	1	1	1	1	0.001(54)

Second, we examined this result when the joint effects from combinations of the parameters are included. The last 16 submodels in Table 2 all have the minimum entropy value, 0.001. The preferences of these 16 submodels are indistinguishable. The digits in parentheses show numerical results, where the number of significant figures was not considered. This just indicates that if the number of significant figures were higher, the resolution of the entropy would be better. In this case, preferences of these 16 submodels could be identified.

In order to determine the most dominant parameters in these 16 submodels, the frequencies of the six parameters appearing in these 16 submodels were used as weighting factors. The frequencies of observation of the first and second parameters are 16 and 15, respectively, and 8 for the rest of the parameters. This result suggests that our ability to interpret the experimental measurements by use of the logistic model is strongly dominated by the first parameter, HCO_3, and the second parameter, SO_4. Parameters 3–6 seem to play a minor role here and preference of joint effects of these parameters are indistinguishable.

This is exactly the same result obtained by using qualitative and DFA analyses described earlier. However, the ECA provides more information about the significance of different combinations of parameters.

3.4.3. Performance examination

At last, we examined the performance of the full model, covariate free model, and six submodels that associate six single parameters correspondingly by the ROC analysis (Fig. 2) (Johnson and Albert, 1999; Chen et al., 2005, 2006, 2007). The ROC diagram

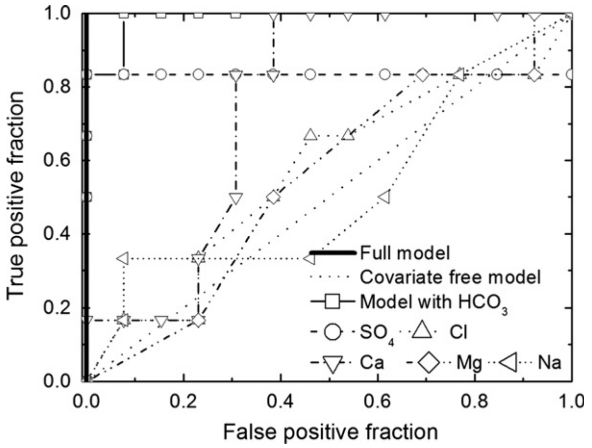

Fig. 2. ROC curves of the full model, covariate model, and six submodels as noted in the legend. A larger area under the ROC curve indicates higher predicting power of the model. For the details, please refer to the text.

is a well-established way to examine the performance of the model predictor. A ROC graph is a plot with the false positive fraction on the horizontal axis and the true positive fraction on the vertical axis. The true positive fraction is the fraction of positive occurrences of unit G (Table 1, "Y" in the first column) that were correctly predicted as unit G, while the false positive fraction is the fraction of not unit G cases (Table 1, "N" in the first column) that were incorrectly predicted as unit G. The point (0, 1), which means the false positive fraction is 0 and the true positive fraction is 1, on an ROC graph is the perfect predictor. It predicts all occurrences and non-occurrences of unit G correctly. When the area below a ROC curve of a model (ROC area) is close to 1, the model highly likely predicts all occurrence and non-occurrence of unit G correctly. Contrary, when the ROC area is close to 0.5, the model becomes worthless, i.e. the predicting power of the model is identical to a random distribution model. Based on the ROC curve (Fig. 2), we found the predicting power of the full model, the submodels 24, 10, and 7 (models with HCO_3, SO_4, and Ca, respectively) are significantly better than rest of submodels and the worthless covariate free model. This test also indicates appropriateness of the usage of the logistic function as the likelihood function in this problem.

3.4.4. *Summary*

Based on the above ECA studies, we conclude that the formation of unit G may strongly involve chemical processes associated with HCO_3. The chemical processes associated with SO_4 may play a minor role in the formation. This is precisely the same result inferred from the qualitative analysis. Furthermore, the joint effect analysis shows the indistinguishability of submodels with and without Cl, Ca, Mg, and Na given HCO_3 and SO_4. This result suggests that the formation may be likely independent of chemical processes with Cl, Ca, Mg, or Na. One may conduct further analysis in a similar way to extract more information, but this will not be pursued here. While the formation of rocks involves complicated processes, classifying the rocks becomes very difficult. Our results indicate the ECA may provide a quantitative guidance for investigating formation of the brine samples, which is not straightforward from the DFA.

4. Conclusions

An entropic component analysis is proposed to identify key parameters and the joint effects of various combinations of parameters of a complex system. This approach includes two steps, a Bayesian approach for converting the variable selection into the model selection problem and an entropic model selection approach. We have shown that, after the experimental responses and the parameters have been associated by means of probability models, the preferences or the ranking scheme of the probability models can be uniquely obtained by evaluating relative entropy of each model with respect to a reference model, Eq. (1). Since a reference model $m(x)$ is usually difficult to acquire in practice, we propose to use a uniform probability distribution as the reference. Thus the preferences of the models are given by the order of decreasing entropy, Eq. (2), and the preferences of the corresponding parameters can be obtained as well.

We have illustrated the ECA by studying a geological sample classification problem, of whether some brine samples from oil field waters in Texas and Oklahoma (Davis, 2002) belong to the Grayburg Dolomite. The results obtained from ECA are consistent with a qualitative analysis of the lithology and a quantitative analysis based on the DFA. In addition to giving consistent results, however, the ECA also provides a complete analysis of the significance of different combinations of parameters.

Finally, we remark that in the present work, the parameters' significance analysis is based on the model selection approach. However, a model-free selection method proposed by Li et al. (2005) offers an alternative route. Since it is model-free, this approach is suitable for data sets that cannot be modeled easily. Can our ECA be extended to become a model-free analysis? This question is beyond the scope of the present work, but will be addressed in the future.

Acknowledgment

CCC is grateful for research support from the National Science Council and the Institute of Geophysics, National Central University, Taiwan (ROC).

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 14, 716–723.

Bonnlander, B.V., Weigend, A.E., 1994. Selecting input variables using mutual information and nonparametric density estimation. In: Proceedings of the 1994 International Symposium on Artificial Neural Networks, 42–50.

Caticha, A., 2004. Relative entropy and inductive inference. In: Erickson G, Zhai Y (Ed.) Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings 707, Melville, New York, 75–96, arXiv:physics/0311093.

Chalcraft, R.G., Ward, R.F., 1988. McElroy field, Permian basin, West Texas: cyclic sequence dolomite reservoir of central basin platform. Am. Assoc. Pet. Geol. Bull. 72, 171.

Chen, C.C., Rundle, J.B., Holliday, J.R., Nanjo, K.Z., Turcotte, D.L., Li, S.C., Tiampo, K.F., 2005. The 1999 Chi-Chi, Taiwan, earthquake as a typical example of seismic activation and quiescence. Geophys. Res. Lett. 32 (22), L22315. doi:10.1029/2005GL023991.

Chen, C.C., Rundle, J.B., Li, H.C., Holliday, J.R., Nanjo, K.Z., Turcotte, D.L., Tiampo, K.F., 2006. From tornadoes to earthquakes: forecast verification for binary events applied to the 1999 Chi-Chi, Taiwan, earthquake. Terr. Atmos. Oceanic Sci. 17, 503–516.

Chen, C.C., Tseng, C.Y., Dong, J.J., 2007. Variable selection based on entropic criterion and its application to the debris-flows triggering. Eng. Geol. 94, 19–26.

Comon, P., 1994. Independent component analysis, a new concept? Signal Process. 36, 287–314.

Cox, D.R., Snell, E.J., 1989. Analysis of Binary Data second ed. Chapman and Hall, New York, 231pp.

Davis, J.C., 2002. Statistics and Data Analysis in Geology third ed. Wiley, New York, 656 pp.

Dupuis, J.A., Robert, C.P., 2003. Variable selection in qualitative models via an entropic explanatory power. J. Stat. Plann. Inference 111, 77–94.

Forbes, F., Peyrard, N., 2003. Hidden Markov random field model selection criteria based on mean field-like approximations. IEEE Trans. Pattern Anal. Mach. Intell. 25, 1089–1101.

Jaynes, E.T., 1957. Information theory and statistical mechanics. Phys. Rev. 106, 620–630.

Johnson, V.E., Albert, J.H., 1999. Ordinal Data Modeling. Springer, New York, 255 pp.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer, New York, 487 pp.

Kieseppa, I.A., 2000. Statistical model selection criteria and Bayesianism. Philos. Sci. 68, S141–152.

Li, L., Cook, R.D., Nachtsheim, C.J., 2005. Model-free variable selection. J. R. Stat. Soc. B 67, 285–299.

Ostroff, A.G., 1967. Comparison of some formation water classification systems. Bull. Am. Assoc. Pet. Geol. 51, 404–416.

Raftery, A.E., 1995. Bayesian model selection in social research. Sociol. Methodol. 25, 111–163.

Roche, S.L., 1997. Time-lapse, multicomponent, three-dimensional seismic characterization of a San Andres shallow shelf carbonate reservior, vacuum field, Lea County, New Mexico. Ph.D. Dissertation, Colorado School of Mines, USA.

Rodriguez, C., 2005. The ABC of model selection: AIC, BIC, and the new CIC. In: Knuth K.H., Abbas A.E., Morris R.D., Castle J.P. (Ed.) Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings 803, Melville, New York, 80–87.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Shore, J.E., Johnson, R.W., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans. Inform Theory IT-26, 26–37.

Shore, J.E., Johnson, R.W., 1981. Properties of cross-entropy minimization. IEEE Trans. Inf. Theory IT-27, 472–482.

Skilling, J., 1988. The axioms of maximum entropy. In: Erickson, G.J., Smith, C.R. (Eds.), Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Kluwer, Dordrecht, pp. 173–187.

Skilling, J., 1989. Classic maximum entropy. In: Skilling, J. (Ed.), Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht, pp. 45–52.

Skilling, J., Fougere, P.F. (Eds.), 1990. Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht.

Stone, J.V., 2004. Independent Component Analysis: A Tutorial Introduction. MIT Press, Cambridge MA, 193 pp.

Tseng, C.Y., 2006. Entropic criterion for model selection. Physica A 370, 530–538.