Machine Learning (2006) 63: 43–67
DOI 10.1007/s10994-005-5931-5

Data-guided model combination by decomposition and aggregation

Mingyang Xu · Michael W. Golay

Received: 23 March 2005 / Accepted: 1 November 2005 / Published online: 9 March 2006

Abstract Model selection and model combination is a general problem in many areas. Especially, when we have several different candidate models and also have gathered a new data set, we want to construct a more accurate and precise model in order to help predict future events. In this paper, we propose a new data-guided model combination method by decomposition and aggregation. With the aid of influence diagrams, we analyze the dependence among candidate models and apply latent factors to characterize such dependence. After analyzing model structures in this framework, we derive an optimal composite model. Two widely used data analysis tools, namely, Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are applied for the purpose of factor extraction from the class of candidate models. Once factors are ready, they are sorted and aggregated in order to produce composite models. During the course of factor aggregation, another important issue, namely factor selection, is also touched on. Finally, a numerical study shows how this method works and an application using physical data is also presented.

Keywords Model selection · Model combination · Model dependence · Model structure · Model decomposition · Principal component analysis · Independent component analysis · BIC · Cross-validation

Editor: Dan Roth

M. Xu (✉) · M. W. Golay
Department of Nuclear Science and Engineering, Massachusetts Institute of Technology,
77 Massachusetts Ave., Cambridge, MA 02139
e-mail: xumy@alum.mit.edu
Tel.: +212-412-1827
Present address:
1 River Court Apt. #2506, Jersey City, NJ 07310

M. W. Golay
e-mail: golay@mit.edu

1. Introduction

A model, which is usually in a mathematical form, is a proposed explanation of a particular phenomenon. This proposed explanation is also used to predict future events. As more evidence is gathered by subsequent observations, that model can be validated against data. If the prediction error, i.e. difference between the prediction and the observation, is not tolerable, the model has to be calibrated or modified in terms of its structure by incorporating newly observed data. This cycle is then repeated iteratively as new observations become available until the model provides a satisfactory explanation of all observed events.

Most often scientists propose models in order to explain a phenomenon from different perspectives, based upon different theories or different data sets. For example, in thermal hydraulics many different models were put forward to describe the behavior of two-phase flows and to predict their pressure drops through a flow passage. Another example can be the probabilistic seismic hazard analysis, where many ground motion attenuation models were developed independently to predict the ground shaking given an earthquake. In such a situation, we are facing the thorny problem of choosing the best model to predict the future events. In general, we may also have some, but often sparse, data at hand. Thus we can test the models against these data in order to select an optimal one. Several model selection methods and procedures have been developed in order to achieve this goal.

However, selecting a single best model is not so desirable as it does not make efficient use of the information at hand, e.g. a class of competing models and a new data set. Therefore, alternatively model combination is proposed in order to improve model performance. The benefits of model combination for augmenting model accuracy and reducing model uncertainty has been noted in the literature (cf. Madigan & Raftery 1994; Clemen & Winkler, 1986). As we know, model uncertainty, categorized into "epistemic uncertainty", stems from incomplete or imprecise knowledge and can be reduced by improvements in data measurements and model formulation. It is not surprising to see that combining different information sources including candidate models and data could result in a better model. To date several model combination methods have been proposed, for example the equally-weighted combination and Bayesian Model Averaging (BMA) (Hoeting et al., 1999) or Bayes factor (Kass & Raftery, 1995) weighting method.

The basic idea behind model combination is to aggregate all available information, which, however, may contain errors or noise, and then to build a new composite model as well as possible. From this perspective, a good model combination method should have the following properties:

(i) It should be able to aggregate information in all competing models and thereby reduce model bias and uncertainty.
(ii) It should be able to detect errors in competing models to some degree, thereby reducing model bias.
(iii) It should model dependencies among competing models and thus reduce information redundancy.
As pointed out by Hogarth (1987), the poor performance of human judges relative to statistical models stems largely from an inability to recognize and process redundant information appropriately. Furthermore, reducing information redundancy helps to reduce model dimensionality, e.g. the number of factors in a factor model, and thus reduce model uncertainty.
(iv) It should be able to combine different kinds of information, including models and data.

(v) It should perform robustly when treating different sets of data.
(vi) It should be objective, involving no subjective judgment.

Ideally, a model selection process should be objective and therefore repeatable.

In order to achieve the above goals, basically improving both accuracy and precision, we propose a new model combination method by utilizing decomposition and aggregation based upon data. This method is mainly suitable to the situations where there is no well-founded theory and only sparse data are available. Otherwise we might be able to derive a more exact theoretical model. For example, in seismic risk analysis there are many seismic attenuation models available for estimating ground motion given an earthquake. Some of them are empirical models based upon historical data and others are created based upon some theories in earth science such as geognosy. Our method can be applied to combine these competing models so as to obtain more accurate ground motion estimations.

This paper is organized as follows. In Section 2, a brief review of related work is presented. In Section 3, a new model combination method is proposed. Meanwhile, dependence among candidate models is analyzed and factor model is proposed in order to model such dependencies. In Section 4, we discuss how to decompose a class of candidate models to factors. After that, Section 5 presents a regression method to aggregate factors based upon data. Finally, examples are presented in section 6 to demonstrate the performance of this model combination method.

2. Related work

Closely related problems include model evaluation, model selection and model combination. To date much effort has been devoted to these problems. In this section, we briefly review some of them.

A model can be evaluated based upon how well the resulting prediction agrees with future observations (Dawid, 1984). In the case where the same group of data is used for both model calibration and validation, model selection method or its variants are widely applied.

By now a variety of model selection methods have been developed, including classical hypothesis testing, penalized maximum likelihood, Bayesian methods, information criteria and cross-validation. All these methods, which overlap with one another, provide an implementation of Occam's razor (Madigan & Raftery 1994) in one way or another, in which parsimony or simplicity is somehow balanced against goodness-of-fit.

Among those model selection methods, information criteria are considered to be novel and promising, and thus draw much attention. The name, i.e., information criterion, arises from its close connection to the information theory. This class of model evaluation and selection methods was pioneered by Akaike's Information Criterion (AIC) (Akaike, 1973). Later many other similar information criteria were derived from different perspectives, for example, Bayesian Information Criterion (BIC) (Schwarz, 1978), Takeuchi's Information Criterion (TIC) (Takeuchi, 1976), Minimum Description Length (MDL) (Rissanen, 1978), Hannan and Quinn criterion (HQ) (Hannan and Quinn, 1979). Basically, all these criteria can be expressed as

$$IC = -2\log(\textit{Maximum likelihood}) + \textit{penalty}\,(k, n) \tag{1}$$

where the maximum likelihood is the likelihood $f(\theta, x)$ evaluated at the maximum likelihood estimate $\hat{\theta}$, and the penalty term is a function of the model dimension k, the number of model

parameters, and the sample size n. From Eq. (1), it is easily seen that this class of information theoretic criteria can be viewed as modified maximum likelihood or penalized maximum likelihood methods.

All of these approaches select the model that minimizes this quantity based upon available data. The only difference between them lies in the second term, that is, different evaluation methods use different penalty terms as corrections.

As we mentioned, in addition to model selection another class of approaches is model combination, which includes, for instance, equally weighted combinations, combinations based upon information criteria evaluation, Bayesian model averaging. The equally weighted combination is the simplest treatment in this class, because each model is assigned the same weight. This approach does not involve new data, and thus is usually applied in cases where there are no data available and all competing models have the same preference. When some data are gathered, this approach is ready to be extended to a weighted combination. For example, each model can be evaluated using Akaike's information criterion (AIC) and assigned different weights based on their AIC value (Burnham and Anderson 2002), for example

$$w_i = \frac{\exp\left(-\frac{1}{2}AIC_i\right)}{\sum_{j=1}^{K}\exp\left(-\frac{1}{2}AIC_j\right)}. \tag{2}$$

A recently developed model combination method is that of Bayesian Model Averaging (BMA) (Hoeting et al., 1999) or Bayes factor (Kass and Raftery 1995) weighting, which became computationally possible since the invention of the Markov Chain Monte Carlo algorithm (Gilks et al. 1998). The basic idea of BMA is very straightforward, that is, to calibrate the probabilities of competing models using Bayesian updating method. After obtaining the posterior model probability, the composite model can be expressed as

$$f(y \mid D) = \sum_{i=1}^{K} f_i(y)\Pr(M_i|D), \tag{3}$$

where D is the observed data, K is the number of competing models that are assumed to be mutually exclusive, $f_i(y)$ is the ith model, and according to the Bayesian formula the posterior probability $\Pr(M_i|D)$ can be calculated as

$$\Pr(M_i \mid D) = \frac{\Pr(D \mid M_i)\Pr(M_i)}{\sum_{i=1}^{K}\Pr(D \mid M_i)\Pr(M_i)}, \tag{4}$$

where $\Pr(M_i)$ is the prior probability of model M_i. The difficulty of implementing BMA partly consists in the computation of the integral

$$\Pr(D \mid M_i) = \int \Pr(D \mid \theta_i, M_i)\Pr(\theta_i \mid M_i)\,\mathrm{d}\theta_i, \tag{5}$$

where $\Pr(\theta_i \mid M_i)$ is the prior density and θ_i is the vector of parameters of model M_i.

Another class of methods of model combination is Bayesian information-aggregation, which is also based upon the Bayesian method (Morris 1977; Clemen and Winkler 1993). Suppose θ is a continuous quantity to be estimated, and we obtain a group of estimates $x_1, \ldots, x_K$ from a class of competing models, say, $M_1, \ldots, M_K$, respectively. According to the

Bayesian formula, the posterior distribution of θ is

$$\Pr(\theta|x_1,\ldots,x_K)=\frac{\Pr(x_1,\ldots,x_K,\theta)}{\Pr(x_1,\ldots,x_K)}=\frac{\Pr(x_1,\ldots,x_K,\theta)}{\int\Pr(x_1,\ldots,x_K,\theta)d\theta}, \tag{6}$$

where according to the Markov's property

$$\Pr(x_1,\ldots,x_K,\theta)=\Pr(x_K|x_{K-1},\ldots,x_1,\theta)\cdots\Pr(x_2|x_1,\theta)\Pr(x_1|\theta)\Pr(\theta). \tag{7}$$

The central idea of these method lies in modeling the dependence among models, which is termed the conditional mean dependence assumption (CMDA) in Clemen and Winkler (1993), that is,

$$E(X_i|X_{i-1},\ldots,X_1,\theta)=\beta_{i,0}+\beta_{i,1}X_1+\cdots+\beta_{i,i-1}X_{i-1}+\alpha_i\theta. \tag{8}$$

In Eq. (8), the knowledge about the information sources is incorporated in the aggregation. Thus, if we know the distribution of X_i in advance, we can obtain its conditional distribution $\Pr(X_i|X_{i-1},\ldots,X_1,\theta)$ with the expected value determined by Eq. (8). Finally, we obtain the posterior distribution of θ.

Unfortunately, none of the above methods can give us a satisfactory solution to the problem mentioned earlier. For example, the model selection methods can only choose a single best model, the BMA method does not model the dependence in model structure among candidate models and is computationally expensive, and Bayesian information-aggregation methods cannot incorporate information in new data. These weaknesses are part of reasons that motivated the work of this paper.

3. Model combination by decomposition and aggregation

Since the existing methods cannot achieve those model combination objectives discussed earlier, in this section we propose a new model combination method by means of decomposition and aggregation.

Before we proceed, it is time to further clarify our problem. Suppose that we have a set of competing models, denoted as $M_1,\ldots,M_K$, which can be expressed in mathematical forms as $f_1(x),\ldots,f_K(x)$, and we gather a new set of data, i.e. $\{(x_i, y_i)\colon i=1,\ldots,n\}$, where x_i and y_i can be vectors of input variables and response variables, respectively. Now our question is "how can we construct a more accurate composite model with lower uncertainty, given a class of competing models and a set of sparse data?"

Note that here we evaluate a model using both accuracy and uncertainty criteria, being consistent with the goals of Section 1.

From a theoretical point of view, it is advantageous to view models as sets of probabilistic, or statistical, hypotheses (Forster, 2000), that is, given inputs a statistical model produces the distribution of outputs $p_{Y|X}(y|x)$. In fact, more generally a model only delivers mean values $f(x)$ rather than a distribution function. For example, in physics in order to interpret models in a statistical context error distributions are associated with models. In fact, any measurement involves measurement errors. In addition, the input X, which can be a vector, are often assumed to be randomly drawn from distribution $p_X(x)$. Thus, in any case a deterministic equation can be regarded as providing mean values of the dependent response variables given a set of input variables. In this sense, a model can be reduced to a probability distribution

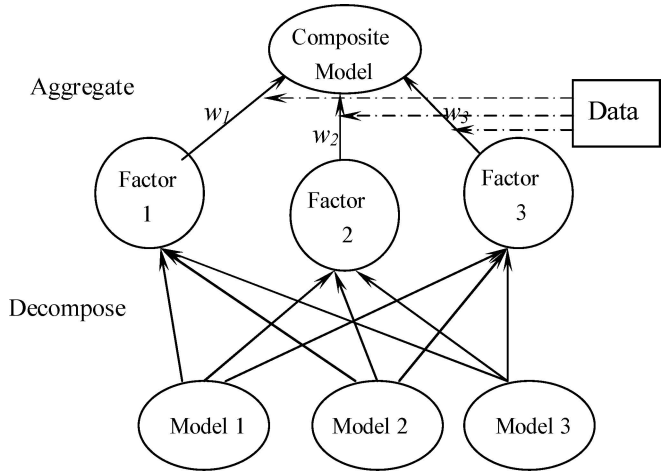

Fig. 1 Model decomposition and aggregation (The dashed lines mean that the pointed arc is under the guide of data)

governed by a group of parameters. Therefore, throughout this paper, we restrict our attention to statistical mathematical models, unless stated otherwise.

Consequently, a data set $\{(x_i, y_i)$: i=1, . . . ,$n\}$ can be viewed as a realization of a random process, and the same for models. Consequently, this model combination problem can be treated in a statistical framework.

The model combination method by decomposition and aggregation that we present in this paper can be summarized as in Fig. 1. Basically, it first decomposes candidate models into common latent factors and aggregate common factors into a composite model.

In the rest of this section, we make two arguments, (i) candidate models are dependent upon each other and (ii) such dependencies can be modeled using common factors, and then we will propose a model structure formula, based upon which an ideal model is then obtained.

3.1. Model dependence

Intuitively, the dependencies among competing models are obvious because they are intended to describe the same phenomena and predict the same future events. When considering how models are built, we also notice that each model is built upon the basis of some theories and data, which may be shared with other models.

In order to explain information sources and the dependence among competing models, it is useful to introduce influence diagram (Howard and Matheson 1984; Shachter 1986, 1988), which offers a convenient graphical tool to model the dependence among different information sources. For example, Fig. 2 shows a typical example, where each circle or oval represents nodes, which can be the truth or the full reality, a theory, a set of data or a model, and each directed arc refers to conditional dependence between a chance node and a decision node, which conveys information from a node to another and implies causality. Note that here we use the term, theory, to denote any set of statements or principles devised to explain a group of facts or phenomena, while model refers to mathematical models in particular. Different theories are proposed to explain the same phenomena, termed as truth in Fig. 2, and different sets of data are generated by the same true model, so in fact there exist dependencies even

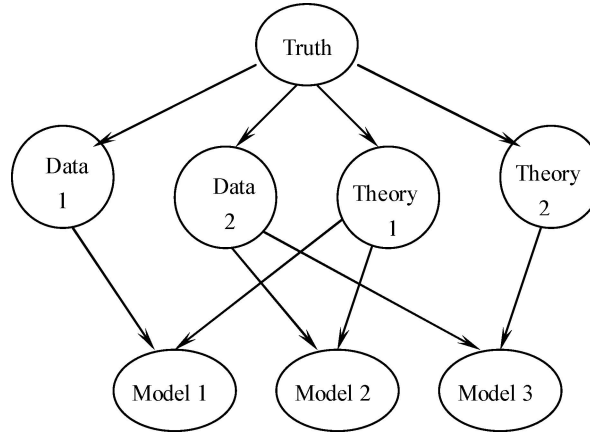

Fig. 2 Influence diagram for models

among different theories and different sets of data. Furthermore, the overlapping information source, including theory and data, serves as a vehicle for representing dependence among the models. Therefore, influence diagrams can give us a clear idea where the information sources of modelers come from.

In reality, such dependence is quite common. For example, the correlation coefficients among economic forecasters are usually around 0.9 (Clemen & Winkler, 1986; Figlewski & Urich, 1983).

3.2. Factor model

In the previous subsection, we have analyzed the dependencies among candidate models, and now we apply a latent factors model to treat such dependencies. Our key argument is that the propagation of information beginning at the truth and ending with models occurs through latent factors, or components, and the dependencies among candidate models is due to their sharing of common factors. This factor model is not a new idea and it has been applied in many areas. For example, factor analysis (Bartholomew and Knott 1999) is widely used in such areas as psychology, chemistry and economics.

It is easy to modify the influence diagram in Fig. 2 slightly into a factor diagram as in Fig. 3. Note that in this factor diagram, information is characterized by factors and correspondingly each directed arc is associated with a set of pairs of factor and weight, i.e. $Q_{ij} = \{(f_{ij}, w_{ij})\}$. In such a scheme, information is propagated from the node of truth to the nodes of models in the form of factors, but it is obvious that the sets of factors received by candidate models are not necessarily to be the same. The loss of information and misspecification can also be described in terms of factors.

3.3. Optimal model structure

Now that we analyzed the model dependence and how to model it, we are ready to discuss how to build an optimal composite model in this scheme.

First, let us define the full truth or the true model. In terms of factors, the true model can be expressed as

$$M_T(x) = \sum\nolimits_{i=1}^{N} w_i(x) f_i(x), \tag{9}$$

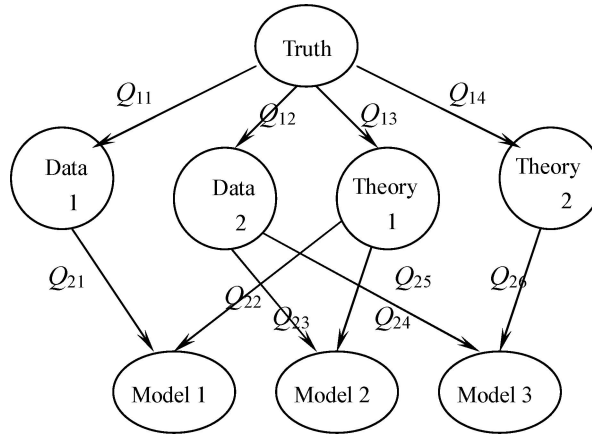

Fig. 3 Factor diagram for candidate models

where $f_i(x) \in F$, the factor set, and $w_i(x)$ is its corresponding weight, intensity, or factor loading as in factor analysis (cf. Bartholomew & Knott 1999), N is the number of total factors, x is an input variable. In a linear case, $w_i(x)$ is constant, independent of x. Actually, in a nonlinear case we can divide the range of input variables and approximate each subrange by a linear model. Therefore, in this paper we will assume that the true model is linear with respect to factors.

We believe that usually "truth" or full reality has essentially infinite dimension, i.e. N tends to infinity, and therefore it cannot be revealed with only finite samples of data and a limited set of candidate models. At best, we can only build a model providing a good approximation to the data available. Thus, a candidate model, an approximate representation of the system, can be expressed in a similar way, for example, the kth candidate model

$$M_k(x) = \sum_{i=1}^{N_k} w_{ki}(x) f_{ki}(x), \tag{10}$$

where $f_{ki}(x) \in F_k$ with F_k the set of factors for kth model and N_k is the number of factors in F_k. Note that in a nonlinear case this mixture of factors model is in spirit similar to the Mixture of Experts model by Jacobs et al. (1991).

As noted, a model is only a simplification or approximation of the reality and hence certainly reflects the full truth to some degree. Whether a model is good depends on the quality of the data and the theoretical foundation underlying the modeling. In the factor model framework, the disagreement between a candidate model and the true model can be caused by the following:

(i) The set of factors contained in a certain candidate model is incomplete, i.e. $F_k \subset F$;
(ii) The factor loadings are biased or imprecise, i.e. $w_{ki} \neq w_i$ for the same factor;
(iii) A candidate model incorporates an erroneous or spurious factor, i.e. $f_{ki} \notin F$.

In addition, distinct competing models might incorporate different subsets of components, which are overlapping as illustrated by the Venn diagram of Fig. 4.

It is convenient to divide a set of factors into two parts, i.e. common factors, which are shared by all the candidate models, and unique factors. Actually, unique factors are so called only in the sense that they are not shared by all the candidate models. This differentiation can be easily understood because candidate models are created based on common knowledge

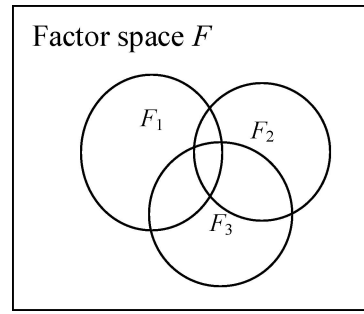

Fig. 4 Model factor space diagram

and individual intelligence. Therefore, every competing model can potentially contribute to the composite model. It is clear that the union of $F_1, \ldots, F_K$ give us a better approximation to F than any single subsets.

Meanwhile, as pointed out a while ago, each candidate model may contain erroneous or spurious factors. For example, unique factors may be attributed to personal bias and be incorrect. Such bias or error is also what we must try to eliminate in forming a combination. At this point, data come to play their crucial roles just as in a general modeling process data are used to calibrate and validate a model. The detection of erroneous factors is inherent in composite model construction.

In order to aggregate information in all competing models, it is important to model dependence among multiple competing models and reduce information redundancy. Our factor model enables us to extract those important common factors, thus reducing model dimensionality and redundancy with the minimum loss of information by discarding those trivial components.

Once a set of factors is ready, we can construct an optimal composite model by linearly aggregating factors, i.e.

$$M(x) = \alpha + \sum\nolimits_{i \in S_c} w_i f_i(x). \tag{11}$$

where S_c denotes the set of selected factors.

In the above Eq. (11), the factor weights and constant α can be determined based upon data. In the course of aggregating factors, whether a factor is valid or not is determined by its agreement with the data.

With these manipulations, the incompleteness, imprecision, error, information redundancy as well as model dimensionality can be reduced, and at last we can obtain an optimal composite model, which is closer to the true model than any other candidate models individually. But, we still have not solved two difficult key issues, namely, (1) How can we extract factors from a class of candidate models? and (2) How should we integrate factors and detect erroneous ones?

In the next sections, we propose some methods to attack these two obstacles.

4. Model decomposition

Our model combination method consists of two stages, decomposition and aggregation. In this section we propose different approaches to commit model decomposition.

4.1. Model factor extraction

In Section 3, we model the dependencies among candidate models by means of common factors. Thus, factors can be extracted from candidate models by taking advantage of this relationship. Before introducing the factor extraction method, let us further clarify model structure and also assume some simplification to make it mathematically tractable.

If we fuse all of the unique factors of the kth candidate model into a single factor, $f_{ku}(x)$, we obtain a simpler model structure as

$$M_k(x) = \sum_{i=1}^{Nc} w_{ki}(x) f_{ci}(x) + f_{ku}(x), \tag{12}$$

where $f_{ci}(x)$'s are common factors and $f_{ku}(x)$ is the single unique factor of kth candidate model, and N_c is the number of common factors extracted, and $k = 1, \ldots, K$. If we rewrite the above Eq. (12) in a matrix notation, we have

$$M = W f_c + f_u, \tag{13}$$

where candidate model vector $M=[M_1, \ldots, M_K]^{\mathrm{T}}$, common factor vector $f_c=[f_{c1}, \ldots, f_{cN}]^{\mathrm{T}}$, the unique factor $f_u=[f_{u1}, \ldots, f_{uK}]^{\mathrm{T}}$, and the factor loading matrix $W=[w_1, \ldots, w_K]^{\mathrm{T}}$ with $w_k=[w_{k1}, \ldots, w_{kN}]^{\mathrm{T}}$. In particular, Eq. (13) reduces to a linear transformation from common factors to candidate models when assuming no unique factors.

To further simplify, we might assume that the unique factors follow the same probability distribution and are independent as in factor analysis, that is,

$$M = W f_c + f_r, \tag{14}$$

where f_r is the random factor which generates different unique factors.

Such decomposition is similar to the Mosleh and Apostolakis (1986) additive error model of experts. As a result, the average of the unique factors becomes an estimate of another common factor.

Actually, we have not precisely defined model dependence yet. In the following discussion we define dependence in two different ways and propose a second-moment as well as a higher-order statistical method for performing factor extraction.

4.2. Principal component analysis (PCA)

Principal Component Analysis (PCA) (Jolliffe, 1986; Christensen, 2001) is the most commonly used subspace-related techniques for dimensionality reduction, filtering and data modeling. The basic idea of PCA is to find the components that can explain the maximum amount of variance of original variables, e.g. $M_1, \ldots, M_K$ in the current case.

PCA can be defined in a recursive way as described below. The direction of the first principal component (PC) is so defined that the variance of the projection on that direction is maximized, i.e.

$$w_1 = \arg\max_{\|w\|=1} Var(w^T M) = \arg\max_{\|w\|=1} \left\{ E\left[\left(w^T M\right)^2\right] - E\left[w^T M\right]^2 \right\}, \tag{15}$$

where w is a vector of same dimension as M. The first principal component is then given by $f_1 = w_1{}^{\mathrm{T}} M$.

In general, after determining the first $k-1$ principal components, the kth component can be determined similarly as the principal component of the residual:

$$w_k = \arg\max_{\|w\|=1} Var\left(w^T\left(M - \sum\nolimits_{i=1}^{k-1} w_i w_i^T M\right)\right). \tag{16}$$

This process continues until the dimensions of the space are used up. Note that all principal components are sorted naturally by their variances.

Alternatively, the computation of the w_i can be accomplished simultaneously by the singular value decomposition (SVD) of the (sample) covariance matrix of M, i.e. $\sum_M = E[(M-E(M))(M-E(M))^{\mathrm{T}}]$, which can be decomposed as :

$$\sum\nolimits_{\mathrm{M}} = W^{\mathrm{T}} \Lambda W \tag{17}$$

where Λ is a diagonal matrix composed of eigenvalues and W is composed of eigenvectors. Those eigenvectors correspond to $w_1, \ldots, w_k$ found by Eq. (15) and (16). Finally, all principal components, or factors, are obtained as $f = W^{\mathrm{T}} M$.

PCA can at least serve two purposes in our case. First, it helps in reducing model dimension and reducing information redundancy, since the first several components, having the largest variance, also contain most information. Second, noise or error may be reduced by removing the principal components ranking in the tail, which are more likely due to error or bias.

4.3. Independent component analysis (ICA)

Under the classical assumption of Gaussianity, lack of correlation is equivalent to statistical independence. However, for non-Gaussian variables, it is not the case any more. Rather, uncorrelated variables are only partially independent. Therefore, in the case of non-Gaussian random variables, much more sophisticated techniques have to be devised to achieve independence, which can incorporate the information of higher-order moments. Independent Component Analysis (ICA) algorithms are developed to satisfy this purpose.

To date many different ICA methods have been developed. In spite of such diversity, the basic idea of ICA remains the same, that is, components or factors $f_c = W^{-1} M$ are so determined that f_{ci} is independent of f_{cj} for $i \neq j$. The diversity of ICA methods is due to different independence measures and various optimization algorithms used to maximize these independence measures. For details, please refer to Hyvärinen (1999). In the work of this paper, the FastICA algorithm by Hyvärinen and Oja (2000) is applied.

In FastICA, the dependence is measure by negentropy or equivalently nongaussianity (Hyvärinen & Oja, 2000), which is approximated by

$$J(Y) \approx \sum\nolimits_{i=1}^{p} k_i \left(E[G_i(Y)] - E[G_i(Y_G)]\right)^2, \tag{18}$$

where k_i are some positive constants, both Y and Y_G are of zero mean and unit variance, and the functions $G_i(\cdot)$ are some nonquadratic functions. Hyvärinen and Oja (2000) suggested

two nonquadratic functions:

$$G_1(u) = \frac{1}{a_1} \log \cosh a_1 u, \quad G_2(u) = -\exp\left(-\frac{u^2}{2}\right), \tag{19}$$

where $1 \leq a_1 \leq 2$ is some suitable constant.

Then, the FastICA algorithm applies a fixed-point iteration scheme to find a direction, i.e. a unit vector w, such that the projection $w^T M$ maximizes nongaussianity measured by Eq. (18). Similar to PCA, FastICA can extract independent components either one by one or all once in a symmetric way. For details on FastICA, please refer to Hyvärinen and Oja (2000). In our simulation, we use the FastICA package downloaded from the authors' website.

ICA is often considered a tool for explanatory data analysis. This is not surprising because causes can be defined in terms of conditional probability or dependence in some circumstance (cf. Ellery, 1991; Forster, 1984). ICA is also efficient for redundancy reduction as each components (features) are independent from each other, that is, they provide no information to predict one variable using another one (Deco & Obradovic 1995). Another wide application of ICA is in noise reduction. Such denoising capability of the ICA was particularly noted in blind source separation (Jutten & Herault 1991). With these desirable properties, ICA can serve as a good tool for us to extract factors from candidate models.

5. Factor selection and aggregation

Applying the methods proposed earlier, we can extract factors from a class of candidate models. With factors being available, the next step is to select a subset of factors and integrate them based upon available data.

The factors are aggregated in a linear form as shown in Eq. (9), in the same manner that the candidate models are decomposed. The multiple linear regression method is used to estimate factor loadings. The basic idea of factor selection is to check whether a factor is supported by the empirical data. In other words, if inclusion of a factor makes the resultant composite model worse, it is likely to be an erroneous one and should be ruled out. Some criteria must be introduced to accomplish factor selection.

As is mentioned earlier, factor selection is not a separate activity that precedes the model calibration; rather, it is a critical and integral part of model building. In the context of multiple regression analysis, it is especially known as variable selection.

5.1. Sorting factors

In factor selection and assembly, the importance rank of factors becomes an important matter, especially when the pool of factors is quite large. For example, suppose we have N factors, then the total number of subsets of factors is equal to 2^N, which means that we will have to compare 2^N possible composite models in order to choose the optimal one. However, if factors are ranked, a stepwise factor selection can be applied, which makes the procedure of factor selection and aggregation substantially easier and accordingly save computing costs dramatically, because we have only N possible composite models. Also, we are able to construct a sequence of subsets of factors, which are nested, and therefore some statistical model selection method can be applied.

The principal components are naturally sorted by their capability of explaining variance of original variables, or equivalently the variance of components. Typically, the variances of

components, or the eigenvalues of the covariance matrix in Eq. (17), decrease very fast. This implies that a principal component contains more information about a system and therefore more important than those ranked below it.

As pointed out by some authors (cf. Hyvärinen, 1999; Cheung & Xu, 2001), one of the drawbacks of ICA is that components resulting from ICA are not sorted, having zero mean value and unit variance. Here we propose two simple methods to order independent components (ICs) based upon their contribution to reconstruction of original data, which is similar to the ordering of principal components.

The first method is based upon the mixing matrix W as in Eq. (13). An assumption behind this method is that the candidate models are close to the true model and thus factors make similar contributions to both the true model and the candidate models. Meanwhile, we can expect that the larger the absolute value of an entry W_{ji} in the mixing matrix W, the greater the contribution in terms of variation that the ith factor makes to the jth candidate model, because all the ICs are normalized to have zero mean and unit variance. Therefore, we define as a component importance measure (*CIM*) the average coefficients of an IC in reconstructing the candidate models,

$$CIM_i = \frac{1}{K} \sum\nolimits_{j=1}^{K} \left| W_{ji} \right|, \tag{20}$$

where $|\cdot|$ stands for the absolute value.

The second method is based upon the sample correlation between a factor and the observed data. As we will see later on, the contribution of an IC to the composite model is determined by how it is supported by the data. Thus, it is reasonable to rank ICs based upon their agreement with data. In this method, the *CIM* is defined as

$$CIM_i = \frac{1}{n} \sum\nolimits_{j=1}^{n} \left(f_i(x_j) - \bar{f}_i \right) \cdot \left(y_j - \bar{y} \right), \tag{21}$$

where $\bar{f}_i = \frac{1}{n} \sum_{j=1}^{n} f_i(x_j)$ and $\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j$, and (x_j, y_j), $j = 1, \ldots, n$ are observed data. This is closely related to the statistics R^2 in regression analysis.

Although these two measures are based upon different information, one on original candidate models and the other on new observations, they produce similar results under the assumption that candidate models are close to the true model.

The ordering of the importance of independent components can be verified using a little more complicated method. The rank checking can be accomplished both forwards and backwards. The forward method starts with an empty queue and ranks components based on their squared error reduction worth (ERW) ΔL_{2-}, that is, based upon how much the squared error defined in Eq. (26) is reduced by adding a certain component to a set of factors. The larger is the ERW, then the more important is a factor, also. In contrast, the backward method begins with a full queue and ranks factors based upon their squared error achievement worth (EAW) ΔL_{2+}, i.e., how much the squared error is increased by deleting a particular factor from a set of factors. Once again, the larger is the EAW, the more important is a factor. In our empirical study, it shows that all the above methods give us consistent results.

However, we have to point out that the ranking of ICs is far from so simple. As we note subsequently in our numerical study, the ordering of the non-dominant ICs according to the methods introduced above is quite subtle and even changes with regard to data, although the ordering of the dominant ICs is in good agreement with that resulting from the above component importance measures. Here, we define dominant ICs as those whose *ICMs* are

significantly larger than those of others. Therefore, ICs can only be partially ranked in advance.

Another important issue is that unlike PCs ICs cannot be determined uniquely, as demonstrated in the FastICA algorithm (Hyvärinen & Oja 2000). Applying a different a nonlinearity function $G(\cdot)$ leads to a different group of ICs, and furthermore even with the same function $G(\cdot)$, different initial guesses of w in iteration also leads to different optima although those dominant ICs will remain very similar. Thus, it is beneficial to repeat ICA many times so as to choose a group of ICs which include as many dominant ICs as possible.

5.2. Model calibration

As in Eq. (3), a composite model of factors can be expressed as

$$M_c(x) = \alpha + \sum_{i=1}^{N} w_i f_i(x) = w^T f(x), \tag{22}$$

where $w = [w_1, \ldots, w_N]^T$ and $f(x)=[f_1(x), \ldots, f_N(x)]^T$.

In the above equation, $f_i(x)$ is a function of input variable x, and so is the composite model. Furthermore, $f_i(x)$, $i=1, \ldots, K$, forms a set of orthogonal base functions. The factor loadings w_i are assumed to be constant over the range of input variable x. Now the task of model calibration is to estimate factor weights, i.e. w_i, given a data set $\{(x_i, y_i): i = 1, \ldots, n\}$. In the face of such a problem, a general solution is first to define a loss function and then design an algorithm to search for parameters such that minimize the loss function. The most widely used loss function is the mean squared error loss function, i.e.

$$L_2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - M_c(x_i))^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - w^T f(x_i)\right)^2. \tag{23}$$

Thus, the estimated factor weights are

$$\hat{w} = \arg\min_{w} L_2 = \arg\min_{w} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - w^T f(x_i)\right)^2. \tag{24}$$

The above equation can be solved analytically, and we obtain

$$\hat{w} = (F^T F)^{-1} F^T y, \tag{25}$$

where $F=[f_1, \ldots, f_N]$ with $f_i=[f_i(x_1), \ldots, f_i(x_n)]^{\mathrm{T}}$ and $y = [y_1, \ldots, y_n]^{\mathrm{T}}$. This is exactly the well-known Ordinary Least Squares (OLS) method.

With factor weights estimated, we are able to calculate the estimated mean squared loss as

$$\hat{L}_2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{w}^T f(x_i)\right)^2. \tag{26}$$

If we denote a subset of factors as Γ, then we designate $\hat{L}_2(\Gamma)$ as the estimated squared loss of the composite model using all factors in Γ, whose factor weights are estimated by Eq. (25). It is easy to see that

$$\hat{L}_2(\Gamma) \geq \hat{L}_2(\Omega), \quad \text{for } \forall \Omega \supset \Gamma. \tag{27}$$

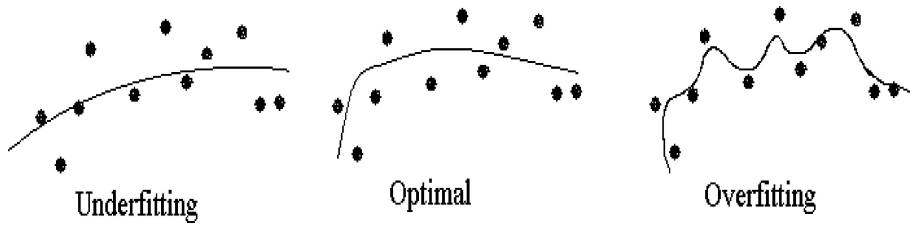

Fig. 5 Example of models fitted to a set of data

This means adding more factors will definitely reduce the estimated loss using the same data set as for calibration, but the capability of predicting the future data is not necessarily improved, and is likely deteriorated. This phenomenon is called overfitting in statistical literature (Burnham & Anderson 2002).

The danger of overfitting is that it tends to identify spurious features unique to a single data set and so calibrated model cannot be generalized. In contrast to overfitting, an underfitted model fails to identify effects or factor that are actually supported by the data set. Generally, a fitted model starts with underfitting and end up with overfitting with the number of variables increasing. Quantitatively, the prediction error decreases at first and goes up at last by adding more predictors. The balance point between underfitting and overfitting is considered optimal. To understand this, it is helpful to take a look at the bias-variance tradeoff.

Let us first define the expected prediction error at x as $E[(y-M_c(x))^2]$, which can be decomposed as follows:

$$\begin{aligned} E\left[(y(x) - M_c(x))^2\right] &= E\left[\left(y(x) - E\left(M_c(x)\right) + E\left(M_c(x)\right) - M_c(x)\right)^2\right] \\ &= \left(y(x) - E\left(M_c(x)\right)\right)^2 + E\left[\left(E\left(M_c(x)\right) - M_c(x)\right)^2\right], \\ &= \left\{Bias\left(M_c(x)\right)\right\}^2 + Variance\left(M_c(x)\right) \end{aligned} \tag{28}$$

where $E[M_c(x)]$ is the expected composite model given a certain subset of factors and given the sample size. The expectation and variance of $M_c(x)$ is formulated with respect to observed data, because the composite changes from data set to data set.

In general, adding more factors can reduce the model bias, the first term, or in other words achieve better fit, but in the meantime model variance is increased because the sample size becomes smaller relative to the number of model parameters to be estimated. In the case of underfitting, the bias in parameter estimation is generally substantial while the variance is underestimated. As for overfitting, the parameter estimation is usually free of bias but has a large variance. In view of this trade-off, we need to identify a balance point in this tradeoff, which is considered optimal, thereby minimizing expected predictive squared error in the future. Figure 5 may provide an intuitive sense of the relationship between underfitting and overfitting.

5.3. Factor selection

In the current situation factor selection is actually the same as variable selection in regression. However, since the factors have been already ranked in terms of their importance, the factor selection process is much simpler. We present a stepwise factor selection procedure to complete this.

Before designing a factor selection procedure, let us first formulate how to evaluate a composite model. To this end, we must apply some statistical model selection method or criterion. The first criterion we use is Schwarz's Bayesian Information Criterion (BIC) (Schwarz, 1978; Kass & Raftery, 1995), which is simple in computation and was proven to be consistent (Woodroofe 1982). Similar to a general information criterion as in Eq. (1), BIC is expressed as

$$BIC = -2\log L(\hat{\theta}|x) + k \cdot \log(n), \tag{29}$$

where $\log L(\hat{\theta}|x)$ is the maximum log-likelihood of a model with k model parameters based on data $x = (x_1, \ldots, x_n)$, that is,

$$\log L(\hat{\theta} \mid x) = \sum\nolimits_{i=1}^{n} \log f(x_i|\hat{\theta}), \tag{30}$$

and where $f(\cdot \mid \hat{\theta})$ is a conditional pdf and $\hat{\theta}$ is the maximum likelihood estimate of that model. In the current case, $\theta = (w_1, \ldots, w_k)$. For a linear regression model, under the assumption of Gaussian error the BIC can be derived as follows:

The log-likelihood can be expressed as

$$L(x, y, \theta) = -\frac{n}{2}\left[\log\left(\frac{2\pi}{n}\right) + \log(RSS) + 1\right] \tag{31}$$

where $RSS = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$ is the residual sum of squared error.

Thus, we obtain the BIC as

$$BIC = -2L(x, y, \theta) + k\log n = -n\big[\log(n/2\pi) - \log RSS - 1\big] + k\log n \tag{32}$$

In our case, the model dimension is equal to the number of factors plus 2 (one constant α and σ^2 are also estimated). According to this criterion, a model having a smaller BIC value is thought of as better than others with larger BIC values.

Another resampling model evaluation method is Cross-Validation (CV), which is usually considered to be a natural treatment. It estimates the generalization error or expected prediction error by mimicking future observations. In a simple version of cross-validation, the data set is divided into two parts: one part for model calibration, training set, and the other part for model validation, test set. That is,

$$\hat{L}_2 = \frac{1}{L}\sum\nolimits_{i=1}^{L} \left(y_l - \hat{w}^T f(x_l)\right)^2, \tag{33}$$

where y_l are data points in the test set and L is the size of test set, and $\hat{w}$ is estimated using the training data set.

As a rule of thumb, one-third of the data should be used for the purpose of validation. Although simple, this version requires that the data set be large. In a more complicated L-fold cross-validation scheme, the data set is randomly broken into L partitions, and then one trains on all the points not in the lth partition with the lth partition serving as test set, and at last one finds the average testing error. In L-fold cross-validation, the procedure of model calibration and cross-validation test should be repeated L times. In this version, the

estimated generalization error is

$$\hat{L}_2 = \frac{1}{L}\sum\nolimits_{l=1}^{L}\frac{1}{J}\sum\nolimits_{j=1}^{J}\left(y_{lj} - \hat{w}_l^T f(x_{lj})\right)^2, \tag{34}$$

where L is the total number of data partitions and J is the partition size, (x_{lj}, y_{lj}) are data points in the lth partition, the model parameters $\hat{w}_l$ are estimated using the $L-1$ partitions of data excluding the lth partition.

In a situation where only sparse data are available, L-fold cross-validation is more data-efficient, and thus a better choice. Therefore, in the work of this paper we set L equal to 5.

With the model evaluation approach ready, now let us start the design of factor selection procedure, supposing that the factors $f_1, \ldots f_K$ are already been ordered. The stepwise factor selection procedure starts with an empty subset of factors and let $k = 0$, and then goes through the following steps:

(1) Add factor f_{k+1} to the subset and estimate a composite model M_{ck+1} by OLS;
(2) Evaluate the newly created model M_{ck+1} by BIC or cross-validation;
    In the case of ICs, each unused IC can be a candidate for f_{k+1}, and therefore we have to try several different M_{ck+1} correspondingly and then choose the best one among them.
(3) If according to the above assessment the result M_{ck+1} is worse than M_{ck}, then we stop; otherwise we go back to step (1).
(4) $f_1, \ldots, f_k$ are selected as good factors and correspondingly M_{ck} is considered to be the optimal composite model.

If we apply cross-validation, the above step (4) is slightly different. That is, after an optimal subset of factors is determined, we use the whole data set, instead of L-1 partitions, to fit a composite model, M_{ck}.

By use of such a procedure, we can avoid confronting exhaustive combination of factors, and thus, can save computation cost. In such a stepwise factor selection, those factors ranked in the tail have a much smaller chance to be included in the optimal subset of factors, which seems reasonable. This is because those factors assigned smaller importance have smaller contributions to explaining the observed data and are more likely to be corrupted by noise or error.

6. Numerical results

By now we have already developed the data-guided model combination method by decomposition and aggregation, and in this section we will demonstrate its performance with both artificial and physical examples.

6.1. An artificial example

In the following we will first use an artificial example to illustrate the model combination method and also show how it works. Based upon Monte Carlo simulation results, some general conclusions are drawn. For the purposes of demonstration, we would like to use an artificial example, where the true model is supposed to be known.

Suppose for some particular system the true model is already known and a set of observations are also obtained somehow. Let us assume that the true model can be expressed as

$$ly(x) = 150 - 150\exp(-2x) + x^2 - 0.1x^3 + 4x + 30\exp(-x/3)\cdot\sin(x) + 15\sin(1.5x) - 20\ln(x+1), \tag{35}$$

where the real number $x \in [0, 10]$.

Correspondingly, its realistic data generating model can be written as

$$y = y(x) + \varepsilon, \tag{36}$$

where ε is supposed to assume a normal distribution, i.e. $N(0, \sigma^2)$, where σ^2 is set as being equal to 64 in the current example. From this generative model, we gathered a set of data with the sample size $n = 20$, i.e. (x_i, y_i), where x_i is evenly distributed within [0, 10].

Meanwhile, suppose that we also collected a class of competing models, all of which can predict the system behavior to a similar degree.

$$\begin{aligned}
&y_1(x) = 150 - 150\exp(-2x) + 4x + 15\sin(1.5x) - 20\log(x+1);\\
&y_2(x) = 150 - 150\exp(-2x) + x^2 - 0.1x^3;\\
&y_3(x) = 150 - 150\exp(-2x) + x^2 - 0.1x^3 + 30\exp(-x/3)\cdot\sin(x);\\
&y_4(x) = 150 - 150\exp(-2x) + 6x + 30\exp(-x/3)\cdot\sin(x) - 20\cdot\log(x+1);\\
&y_5(x) = 150 - 150\exp(-2x) + x^2 - 0.1x^3 + 15\sin(1.5x);\\
&y_6(x) = 150 - 150\exp(-2x) + 15\cos(2x) - 15 + 0.004x^2;
\end{aligned} \tag{37}$$

Note that each candidate model is either incomplete or erroneous, or both. Now we will apply our model combination method to derive a composite model, which will gives us better predictions of future data.

If we plot both the truth and the candidate models in a single figure as in Fig. 6, we see that each candidate model does approximate the true model to some degree. Next we apply both PCA and ICA to extract orthogonal factors from the class of candidate models. The mixing matrices obtained in PCA and ICA, respectively, are as follows:

$$W_P = \begin{bmatrix}
0.3572 & -0.2513 & 0.6192 & -0.2198 & -0.1678 & -0.5911\\
0.4335 & -0.0932 & -0.42 & 0.5593 & 0.3426 & -0.4436\\
0.4119 & -0.2443 & -0.4771 & -0.1913 & -0.701 & 0.123\\
0.3888 & -0.0061 & -0.1883 & -0.6863 & 0.5695 & 0.1339\\
0.4326 & -0.3134 & 0.3968 & 0.3563 & 0.1096 & 0.6468\\
0.4202 & 0.8776 & 0.1403 & 0.0656 & -0.1636 & 0.0497
\end{bmatrix}$$

and

$$W_I = \begin{bmatrix}
-0.1349 & -0.2476 & 0.1194 & -0.0169 & 0.2302 & 0.0390\\
-0.0977 & -0.0689 & -0.0261 & 0.0319 & 0.1082 & 0.0056\\
-0.1817 & -0.1311 & 0.0182 & 0.1508 & 0.1524 & -0.0279\\
-0.0720 & 0.1390 & -0.3287 & 0.2578 & 0.0731 & -0.0681\\
-0.2512 & -0.1761 & 0.0450 & 0.0580 & 0.2294 & 0.0673\\
-0.3456 & -0.1843 & 0.0539 & 0.0397 & 0.3815 & 0.0039
\end{bmatrix}.$$

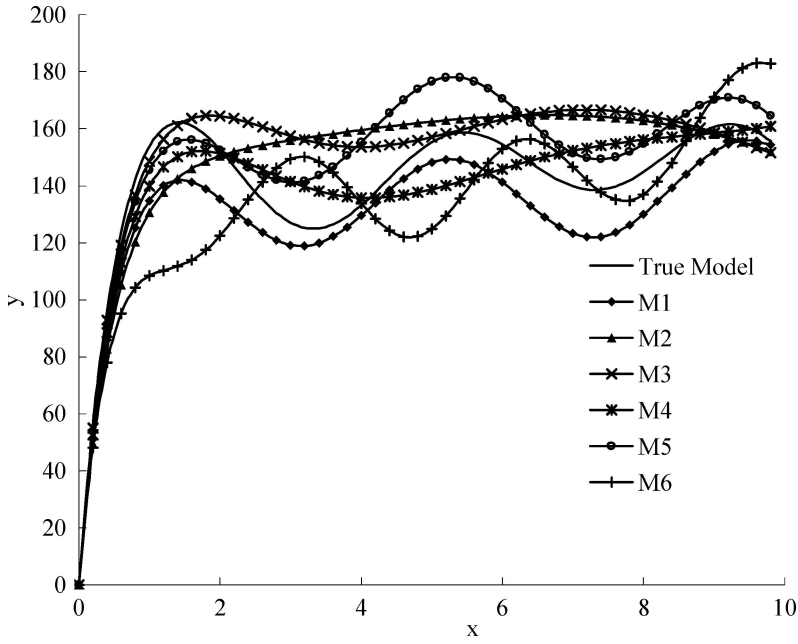

Fig. 6 Artificial candidate models

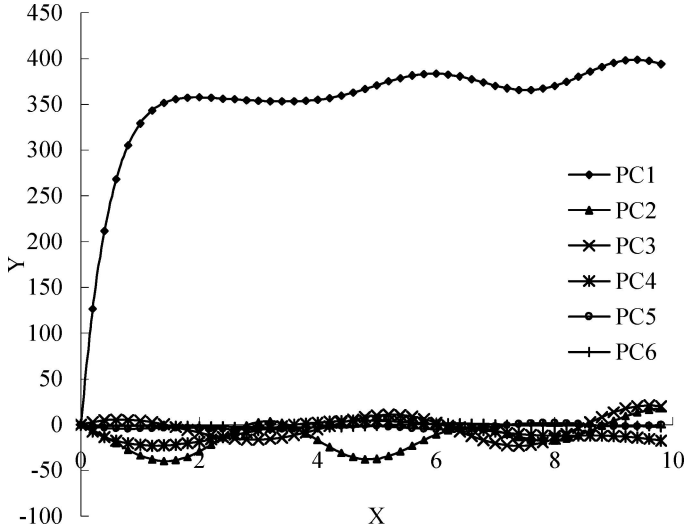

Fig. 7 Principal components

The corresponding separating matrices are the inverse of the mixing matrices and thus factors $f = W^{-1} M$.

The resultant principal components and independent components are shown in Figs. 7 and 8, respectively.

Table 1 Residual sum of squared errors of composite models

Models	−IC2*	−IC1	−IC4	−IC3	−IC5	−IC6
RSS	19838	3665	1278.7	717.2	679	671.3

*Sign minus means that a specific IC is excluded.

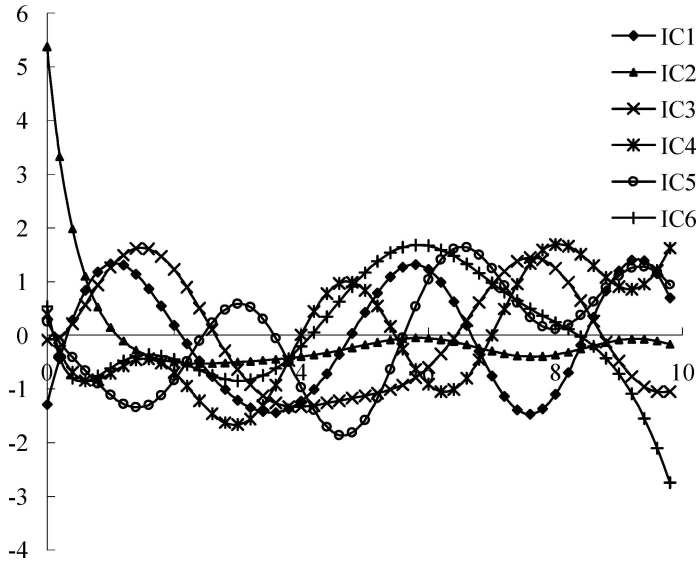

Fig. 8 Independent components

As for independent components, before proceeding we must rank them first. Based upon the mixing matrix W_I, the ICs can be ordered as IC2, IC1, IC4, IC3, IC5 and IC6 according to the labels in Fig. 8. This is consistent with the result from the backward approach shown in Table 1. According to this ordering, IC6 is the least important one, but as noted earlier, the ordering of those non-dominant ICs, such as IC3, IC5, and IC6, is rather subtle.

Alternatively, if we regress on the data using all the six factors, we obtain:

$$y = 143.97 + 12.0926 \cdot \text{IC1} - 24.4966 \cdot \text{IC2} - 0.6979 \cdot \text{IC3} + 5.6838 \cdot \text{IC4} - 1.5551 \cdot \text{IC5} + 0.3299 \cdot \text{IC6}$$

Therefore, the ordering of the estimated factor loading is the same as the result implied by Table 1, but this provides us a simpler way.

After factor ordering, we are ready to construct a composite model by aggregating factors. The result is shown in Tables 2 and 3.

Note that the IC models listed in the above table are the best ones among those having the same number of factors.

Tables 2 and 3 also show the evaluation of composite models using both BIC and Cross-Validation together with global mean squared error (GMSE), which is computed against the

Table 2 Evaluation of composite models using PCA

PC models	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs
RSS	2418.8	1585.4	955.2	671	670.2	669.5
k	3	4	5	6	7	8
BIC	161.7	156	149	145	148	151
CV	108.8	58.9	49.9	28	22.4	22.6
GMSE	107.8	67.9	40.9	17.7	17.1	18.3

Table 3 Evaluation of composite models using ICA

C models	1 IC	2 ICs	3 ICs	4 ICs	5 ICs	6 ICs
RSS	4289.2	1354.8	728.9	681.4	671.3	669.5
k	3	4	5	6	7	8
BIC	171	153	146	145.3	148	151
CV	179.4	21.48	12.2	16.2	18.8	22.5
GMSE	204.5	35	16.1	16.6	19	18.3

true model as

$$GMSE = \frac{1}{m} \sum_{i=1}^{m} (y(x_i) - M_c(x_i))^2, \tag{38}$$

where m is so large as to provide a satisfactory approximation.

Based upon this information, an optimal composite model can be determined for both cases. As for the PC models, four PCs should be selected in terms of BIC and five PCs should be chosen in terms of CV to be optimal; while for the IC models, both BIC and CV suggest that three dominant ICs form the optimal subset. Since the true model is supposed to be known, and thus we can compare our composite models directly with the truth, which tells us that the optimal numbers of factors are five and three for PC model and IC model, respectively. Therefore, both model evaluation methods, BIC and CV, are acceptable, except that BIC chooses the second best option in the case of PC models.

In addition, the optimal IC model has better performance, or smaller GMSE, than the optimal PC model. This may tell us that ICA is potentially more efficient in terms of factor extraction. What is more, the optimal number of ICs is two less than the optimal number of PCs, which means that ICA is more efficient in term of information redundancy reduction.

Finally, let us compare the performance of the two composite models with any single model. Before comparison, let us suppose that a candidate model can also be calibrated based upon data in the following way:

$$M_i'(x) = a_i + b_i M_i(x). \tag{39}$$

After such calibration, we found that the best-calibrated model had GMSE 40.4, which is much greater than those of the two composite models. This means that combination does improve the model performance.

Through the above demonstration we present the performance of this new method by an example. However, usually in statistics a single specific case might not be so meaningful. Thus, in order to obtain a general result the above procedure is repeated many times with different training data sets by the means of Monte Carlo simulation. The average errors of the

Table 4 Monte Carlo simulation results of average errors

Sample size	All models	PCA	ICA
20	24.95	19.82	11.8
50	9.78	9.62	7.30

Table 5 Comparison of models.

	A single best model	Linear combination of all models	New method with PCA	New method with ICA
Test error	0.1935	0.163	0.146	0.140

resultant composite models are listed in Table 4, for two different factor extraction methods and with different sample sizes.

From the above table, we can draw some conclusions. First, the new model combination method outperforms the simple linear combination of all models. Second, ICA leads to better composite models than PCA. Third, the smaller the sample size, the more effective the new method is and also the more advantageous ICA is relative to PCA.

6.2. Physical example

In the above subsection, we demonstrated our method using manufactured data. Now let us apply our method to a real physical example in order to see if it works there also.

The physical example we use here is that of ground motion attenuation models used in seismology. In this example, the purpose is to build a more accurate composite model which is applicable to south California in the United States. A sample data set of size 102 is obtained from the literature (Steidl & Lee 2000). Correspondingly, the candidate attenuation models include the attenuation relations by Boore et al. (1997), Sadigh et al. (1997), Abrahamson and Silva (1997), Campbell and Bozorgnia (1997), Spudich et al. (1997) and Idriss (1995). All of these attenuation relations may be found in Seismological Research Letters, Volume 68, Number 1, January/February, 1997. All these attenuation relationships were developed for shallow crustal earthquakes in active tectonic regions, and thus they should be applicable to southern California.

Both the candidate models and the sample data are plotted together in the same Fig. 9. From Fig. 9, it is easy to note that all of the models are close to be a straight line, which means that unlike the artificial example the dependencies among the candidate models are mostly linear.

Once the candidate models and sample data are ready, we apply the same procedure as in the artificial example to combine candidate models under the guidance of the sample data, namely decomposing the candidate models, selecting factors and aggregating the factors into a composite model by use of the multiple linear regression method. In order to evaluate the resultant composite model, the cross-validation is applied, in which two-thirds of the data set is used for model calibration and the remaining one third is used to test the model. The results are shown in Table 4, where the test error is the mean squared error.

In this example, the same conclusion can be drawn that this new method outperforms both a single best model and simple linear combination of all models. Meanwhile, ICA seems work better than PCA again. However, it is noteworthy that since the non-uniqueness of FastICA, ICA is used several times and the best result is chosen. Compared to the artificial example, the advantage of ICA over PCA is not so significant in the current case. In fact, this

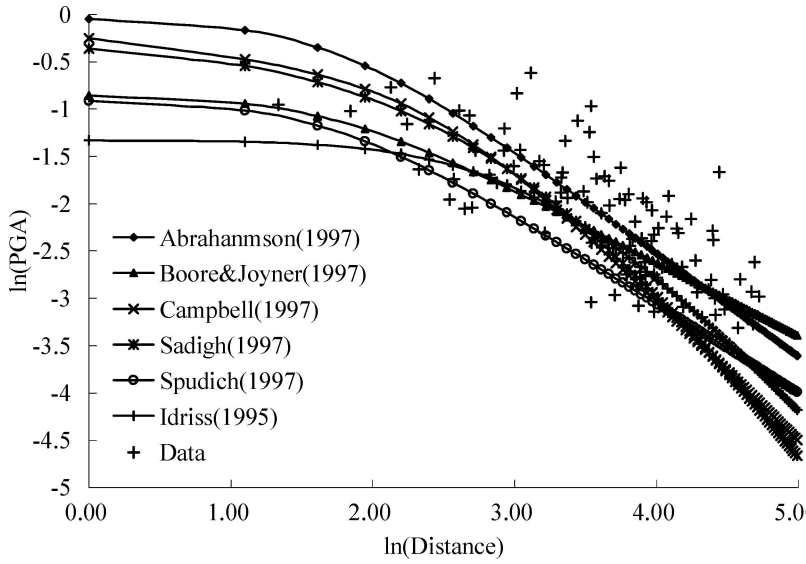

Fig. 9 Candidate ground motion attenuation models and data

observation is in agreement with our expectation. In general, the advantage of ICA compared to PCA is to incorporate higher order nonlinear dependence, but in the current case the models are close to being straight lines and there is only very little if any nonlinear dependence. As a result, ICA simply reduces to PCA. Therefore, in cases where more nonlinear dependence is involved, the strength of ICA will be more significant.

Meanwhile, in the case of ICA, three independent factors are chosen while with PCA four uncorrelated factors are used. This once again verifies our expectation that ICA is more efficient in information compression, which leads to the use of less valid factors.

7. Conclusions and discussion

In this paper we present a model combination method by taking advantage of the factor extraction, noise reduction and information redundancy reduction capability of both PCA and ICA. By some numerical results, we also show that this method works well. But, some problems still remain unsolved, which include

(1) Nonlinear factor loadings, which depend upon the input variables x: For example, factors play different roles over the range of input variables. In our current method, we only suppose that the linear assumption is valid based upon our assumption that the candidate models are similar to the true model to some degree. In considering this nonlinear possibility, a design of mixture of composite local models is helpful (Jacobs et al., 1991; Jordan & Jacobs, 1994).

(2) Unique factor issue: In our current method, we reduce the general model structure to that of a linear transformation by treating equally weighted unique factors as another common factor, but obviously in so doing we may loss some useful information. In order to address this problem, a hierarchical model structure may help, which is similar to hierarchical factor analysis (Schmid & Leiman, 1957; Ghahramani & Hinton, 1997).

(3) Explanation of factors: In our discussion above, although we extracted some factors from a class of candidate models, we have no idea what these factors are physically, or what effects they provide a proxy for. Although explanation of factor entails knowledge about a specific system, it will help us to interpret factors and further refine a composite model, for instance, in factor selection.
(4) Factor selection: Factor selection is always a difficult task just as in general model selection. Although in our numerical study both BIC and CV seem satisfactory, we need a more robust factor selection method, thereby helping to reduce model uncertainty.
(5) Composite model uncertainty. Model uncertainty is central to the performance of a model. How to reduce such model uncertainty is a current active research problem.

Solving these problems can further refine this method or extend it to more general cases. These remaining issues are important future work.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Czáki (Eds.), *2nd International Symposium on Information Theory*. (pp. 267–281) Budapest: Akademiai Kiadó.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold; New York: Oxford University Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Chan, ai-Wan, & Cha Siu-Ming (2001). Selection of independent factor model in finance. In *Proceedings of 3rd International Conference on Independent Component Analysis and blind Signal Separation,* December 9–12. California, USA: San Diego.

Cheung, Y.-M., & Xu, L. (2001). Independent component ordering in ICA time series analysis. *Neurocomputing*, *41*, 145–152.

Christensen, R. (2001). *Linear models for multivariate, time series, and spatial data*. New York: Springer.

Clemen, R. T., & Winkler, R. L. (1993). Aggregating point estimates: A flexible modeling approach. *Management Science*, *39*, 4501–515.

Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. *J. Business and Economic Statistics*, *4*, 39–46.

Clemen, R. T. (1986). Combing overlapped information. *Management Science*, *33:3*, 373–379.

Dawid, A. P. (1984). Present position and potential developments: Some personal views. Statistical theory. The prequential approach (with discussion). *J. R. Statistical Soc. A, 147*, 178–292.

Deco, G., & Obradovic, D. (1995). Linear redundancy reduction learning. *Neural Networks*, *8:5*, 751–755.

Diaconis, P., & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.*, *12*, 793–815.

Ellery, E. (1991). *Probabilistic causality*. Cambridge: Cambridge University Press.

Figlewski, S., & Urich, T. (1983). Optimal aggregation of money supply forecasts: accuracy, profitability and market efficiency. *J. Finance*, *28*, 695–710.

Forster, M. R. (1984). *Probabilistic causality and the foundation of modern science*. Ph.D. Thesis, University of Western Ontario.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*, 205–231.

Ghahramani, Z., & Hinton, G. E. (1997). Hierarchical non-linear factor analysis and topographic maps. *Advances in Neural Information Processing Systems* 10, NIPS∗97, 486–492.

Gilks, W. R. S Richardson., & Spiegelhalter, D. J. (1998). *Markov Chain Monte Carlo in Practice*. Boca Raton, FL. Chapman & Hall.

Hannan, E. J., & Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B 41*, 190–191.

Hoeting, J., Madigan, D., Raftery, A., & Volinskym, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, *14:4*, 382–417.

Hogarth, R. M. (1987). *Judgment and choice: The psychology of decision*, 2nd ed. New York: Chichester [West Susses]. Wiley.

Howard, R. (1989). Knowledge maps. *Management Science*, *35*, 903–922.

Howard, R., & Matheson, J. (1984). Influence diagrams. In R. Howard & J. Matheson, (Ed.), *The principles and applications of decision analysis, SDG systems*, (pp. 719–762). Menlo Park, CA.
Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, *2*, 94–128.
Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13:4–5*, 411–430.
Hyvärinen, A., & Erkki, O. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9:7*, 1483–1492.
Hyvärinen, A. (1998). New approximation of differential entropy for independent component analysis and project pursuit. In *Advance in Neural Information Processing Systems*, *19*, 273–279.
Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput*., *3:1*, 79–87.
Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comput*, *6:2*, 181–214.
Jolliffe, I. T. (1986). *Principal component analysis*. Springer-Verlag.
Jones, M. C., & Sibson, R. (1986). What is projection pursuit? *Journal of the Royal Statistical Society, Ser. A*, *150*, 1–36.
Jutten, C., & Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, *24*, 1–10.
Karhunen, J., Oja, E., Wang, L., Vigário, R., & Joutsensalo, J. (1997). A class of neural networks for independent component analysis. *IEEE Trans. On Neural Networks*, *8:3*, 486–504.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90:430*, 773–795.
Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *J. Amer. Statist. Assoc*., *89*, 1535–1546.
Morris, P. (1977). Combining expert judgments: A bayesian approach. *Management Science*, *23*, 679–69
Papoulis, A. (1991). *Probability, random variables, and stochastic processes*. New York: McGraw-Hill.
Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.
Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
Shachter, R. (1986). Evaluating influence diagrams. *Oper. Res*., *34*, 871–882.
Shachter, R. (1988). Probabilistic inference and influence diagrams. *Oper. Res*., *36*, 589–604.
Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* (*Mathematical Sciences*), *153*, 12–18 (in Japanese).
Woodroofe, M. (1982). On model selection and the arcsine laws. *Ann. Statist*., *10*, 1182–1194.