A Graphical Technique for Model Selection Diagnosis

Hidetoshi Shimodaira

and

Ying Cao

The Institute of Statistical Mathematics, Tokyo, JAPAN

Email: shimo@ism.ac.jp and cao@ism.ac.jp

RM–680 May 1998

Abstract

A graphical method is presented for understanding the relations among the parametric models with respect to data. The models are represented by their predictive densities, and they are drawn in Euclidean space preserving approximately the symmetrized divergence between these densities. This direct visualization of models is very simple and useful for diagnosis of the model selection, especially for nonnested models. Problems such as overparametrization or misspecification of the models are identified from the configuration of the points. Structural patterns in good models are also found as clusters. This is complementary to the methods for assessing the uncertainty of model selection; a brief review is given for the confidence set of models derived from the variance of log-likelihood differences, the bootstrap selection probability, and the simultaneous likelihood ratio test. Illuminating examples from the variable selection in multiple regression as well as practical examples from the phylogeny tree reconstruction are given to illustrate the methodology.

Keywords: Akaike information criterion, graphical exploratory method, Kullback-Leibler divergence, nonnested models, phylogenetic inference, variable selection.

of the points approximately proportional to the symmetrized Kullback-Leibler divergence. This method is quite simple and has a wide range of applicability. In this paper, only the i.i.d. case is treated, but the time series of AR model, for example, can be treated similarly by considering the innovation series.

In Section 2, we give the specific definition of the model map, and show the relationship to the divergence. In Section 3, we review the methods assessing the uncertainty of model selection. However, a large part of the description of the methodology is given in examples of Section 4 step by step from the simplest case to practical cases. We explain how the model map is used by giving illuminating examples from the variable selection problem of multiple regression using datasets taken from textbooks. We also show practical examples from the evolutionary tree reconstruction in phylogeny, where the origin of hominoids and the origin of the tetrapods are discussed. Some remarks are given in Section 5.

2 Model map

Let $q(x)$ be the unknown true density of random variable x, and $x_1, \ldots, x_n$ be i.i.d. observations of sample size n from $q(x)$. Consider the situation that we have a set of candidate probabilistic models $p_\alpha(x|\theta_\alpha)$, $\alpha \in \mathcal{M}$, where α indexes models, $\mathcal{M}$ is the candidate set, and $\theta_\alpha \in \Theta_\alpha$ is the parameter vector. We assume $p_\alpha(x|\theta_\alpha)$ has the same support in x as $q(x)$ does. However, we do not assume the existence of a correctly specified model in the candidate set.

The predictive density of model α is $p_\alpha(x|\hat{\theta}_\alpha)$, where $\hat{\theta}_\alpha$ is an estimated parameter vector. In the examples of Section 3, $\hat{\theta}_\alpha$ is the maximum likelihood estimate (MLE), that is, the maximizer of the log-likelihood function $L_\alpha(\theta_\alpha) = \sum_{t=1}^n \log p_\alpha(x_t|\theta_\alpha)$, but it is not required for the model map.

The definition of the model map for the predictive densities $p_\alpha(x|\hat{\theta}_\alpha)$, $\alpha \in \mathcal{M}$ is quite simple: Let the position of model α in n-dimensional Euclidean space be given by the column vector

$$\xi_\alpha = (\log p_\alpha(x_t|\hat{\theta}_\alpha) : t = 1, \ldots, n). \tag{1}$$

Each model, or its predictive density, is represented as a point in $\mathcal{R}^n$. The maximum log-likelihood is obtained by projecting ξ_α to $1_n = (1, \ldots, 1)'$ direction, $L_\alpha(\hat{\theta}_\alpha) = 1_n' \xi_\alpha$. The model map represents the relative positions of the predictive densities in all the directions, not just in the 1_n direction. To draw the model map, we project the points ξ_α, $\alpha \in \mathcal{M}$ into a lower dimensional space, say $\mathcal{R}^2$ or $\mathcal{R}^3$, using the principal component analysis.

We interpret the model map based on the following property of the Jeffrey symmetric divergence of densities:

$$J(p_1(\cdot);\, p_2(\cdot)) \approx \int q(x)(\log p_1(x) - \log p_2(x))^2\, dx, \tag{2}$$

where the densities $p_1(x)$ and $p_2(x)$ are assumed to be close enough to $q(x)$, and the symmetric divergence $J(p_1(\cdot);\, p_2(\cdot))$ is defined by

$$\int (p_1(x) - p_2(x))(\log p_1(x) - \log p_2(x))\, dx. \tag{3}$$

A detailed discussion is deferred to Appendix, but just notice that $p_1(x) - p_2(x) \approx q(x)(\log p_1(x) - \log p_2(x))$ to derive (2) from (3).

Applying the relation (2) to the predictive densities, and considering that $x_1, \ldots, x_n$ are independent samples from $q(x)$, we have

$$\|\xi_\alpha - \xi_\beta\|^2 \approx nJ(p_\alpha(\cdot|\hat{\theta}_\alpha); p_\beta(\cdot|\hat{\theta}_\beta)). \tag{4}$$

In other words, the squared distance of two points in the model map is approximately proportional to the symmetric divergence of the two predictive densities. However, we should remember that the model map is distorted in the following reasons: (i) We have to project the points into a lower dimensional space to draw them for the visualization. (ii) The relation (4) holds only when the predictive densities are close enough to the unknown true density. (iii) Because the symmetric divergence does not form a metric space, any attempt to represent the whole space of densities in a finite dimensional Euclidean space may fail.

3 Uncertainty of model selection

Although the main subject of this paper is the model map, we briefly review AIC and other related statistics for assessing the uncertainty of model selection. These statistics are tabulated for datasets of Section 4.

The information criterion of Akaike (1974) of model α is

$$\mathrm{AIC}_\alpha := -2L_\alpha(\hat{\theta}_\alpha) + 2\dim\theta_\alpha. \tag{5}$$

AIC is an unbiased estimate (up to the second term) of the expected prediction error, which is the average of $-2\int q(x)\log p(x|\theta_\alpha)\, dx$ taken over the distribution of $\theta_\alpha = \hat{\theta}_\alpha$. Often chosen as the "best model" is the minimum AIC estimate (MAICE), the minimizer of AIC.

However, the difference of AIC values may not be significant, and the other models would be better than MAICE in their average performance. We think

model α is significantly better/worse than model β if AIC_α is significantly smaller/larger than AIC_β. Linhart (1988) and Vuong (1989) considered a normal approximation test using the standardized difference

$$T_{\alpha,\beta} := \frac{\mathrm{AIC}_\alpha - \mathrm{AIC}_\beta}{2\hat{\sigma}_{\alpha,\beta}}, \tag{6}$$

where

$$\hat{\sigma}^2_{\alpha,\beta} := \frac{n}{n-1}\left(\|\xi_\alpha - \xi_\beta\|^2 - \frac{1}{n}(1_n'\xi_\alpha - 1_n'\xi_\beta)^2\right) + \frac{1}{2}v_{\alpha,\beta} \tag{7}$$

is an estimate of $\mathrm{var}(L_\alpha(\hat{\theta}_\alpha) - L_\beta(\hat{\theta}_\beta))$. The higher order term $v_{\alpha,\beta} := \dim\theta_\alpha + \dim\theta_\beta - 2\dim(p_\alpha(\cdot)\bigcap p_\beta(\cdot))$ is introduced in Shimodaira (1997a, 1998).

The confidence set consists of those models which are not significantly worse than MAICE, or the models whose p-values are larger than the significance level. The p-value of each model is $P_\alpha^{(L)} := 1 - \Phi(T_{\alpha,\mathrm{MAICE}})$, where $\Phi(\cdot)$ is the standard normal distribution function. Since Kishino and Hasegawa (1989), this method has been used in phylogeny tree reconstruction, and proved useful in practice.

$P_\alpha^{(M)}$ is another p-value given in Shimodaira (1993, 1998) using a variant of Gupta's subset selection procedure. This is similar to $P_\alpha^{(L)}$, but each model is compared with the minimizer of the expectation of AIC rather than MAICE. The randomness of MAICE is taken into account in $P_\alpha^{(M)}$, while MAICE is regarded as fixed in $P_\alpha^{(L)}$.

$P_\alpha^{(B)}$ is an estimate of the probability for each model to be selected as MAICE; the bootstrap estimate has been used in phylogeny since Felsenstein (1985). This is not a p-value based on hypothesis testing, but it is very useful in practice. Rather the straightforward bootstrap, we employed the normal approximation of Kishino et al. (1990); $P_\alpha^{(M)}$ is also calculated from the same bootstrap samples.

$P_\alpha^{(S)}$ is the p-value of LR test, which is also useful when we have the full model, in which all the candidates are nested. Each model is separately tested against the full model, and the non-rejected models are identified as the adequate models. For simultaneous testing, all those models, in which any non-rejected models are nested, are automatically included in the set of adequate models; this closure method is discussed in Spjϕtvoll (1977).

Note that these four types of p-values are calculated from the log-likelihood vectors (1) of the candidate models.

4 Numerical examples

4.1 Variable selection in multiple regression

The variable selection is a typical example of model selection. We describe the model below, and illustrate the model map with three datasets taken from textbooks.

Let x be partitioned as (y, z), where y is the response variable and $z = (z_1, \ldots, z_m)$ is the vector of the predictors. The normal regression model specifies the conditional density of y_t given z_t as

$$y_t = \eta_0 + \eta_1 z_{1,t} + \cdots + \eta_m z_{m,t} + \epsilon_t, \tag{8}$$

where ϵ_t, $t = 1, \ldots, n$, are distributed as normal with mean zero and variance σ^2. The MLE of $\theta = (\sigma, \eta_0, \eta_1, \ldots, \eta_m)$ is obtained by the least square method, and $\hat{\sigma}^2 = \sum_{t=1}^n \hat{\epsilon}_t^2/n$, where $\hat{\epsilon}_t$ is the residual under the model. The position of the model (8) is given by

$$\log p(y_t|z_t, \hat{\theta}) = -\frac{1}{2}\Big(\log(2\pi\hat{\sigma}^2) + \frac{\hat{\epsilon}_t^2}{\hat{\sigma}^2}\Big). \tag{9}$$

We can ignore $\log p(z_t)$ term, because it just moves the origin.

The full model, which uses all the predictor variables $z_1, \ldots, z_m$, may not lead to the best experimental formula for prediction. A better prediction will be obtained if good predictors, which are strongly related to the response variable y, are picked out for estimation of the coefficients η_i and the others are set to zero. This is an example of model selection; a subset of predictors corresponds to a model.

4.1.1 *Case I: Weight of newborn baby*

This dataset is taken from Sawa (1979, p. 57). The response variable is the weight of a newborn baby. The four predictor variables are the weight of the baby's mother (z_1), the age of the mother (z_2), the number of days of the pregnancy (z_3), and a dummy variable indicating whether the mother is a habitual smoker or not (z_4). These predictors are labeled as `a`, `b`, `c`, and `d`, respectively, and each submodel is denoted by those letters in angle brackets. We consider all possible $2^4 = 16$ combinations of the predictors as the candidate models for selection. The sample size is $n = 15$.

Figure 1 is the model map drawn for the sixteen candidate models. The points ξ_α, $\alpha \in \mathcal{M}$ are given in $\mathcal{R}^n$ by calculating (9) for all the submodels, and then they are projected to the three major principal component vectors u_i, $i = 1, 2, 3$: Let $u_1, \ldots, u_M$ and $\lambda_1^2 \geq \cdots \geq \lambda_M^2 \geq 0$ be the unit eigen vectors and eigen values of the

Table 1: AIC and p-values for Case I. Models are ranked by their AIC values. Only the difference of AIC from MAICE is shown for models other than MAICE. The four types of p-values are explained in Section 3: $P_\alpha^{(L)}$ is p-value of Linhart's model selection test. $P_\alpha^{(M)}$ is p-value of the multiple comparison method. $P_\alpha^{(B)}$ is the bootstrap selection probability. $P_\alpha^{(S)}$ is p-value of LR test. The bootstrap replicate size is $N_b = 10^4$ for $P_\alpha^{(M)}$ and $P_\alpha^{(B)}$.

	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$	$P_\alpha^{(S)}$
1	<abcd>	200	0.500	0.982	0.458	–
2	<abc>	+1.49	0.362	0.699	0.337	0.062
3	<acd>	+4.58	0.099	0.346	0.063	0.010
4	<ac>	+7.18	0.077	0.299	0.020	0.004
5	<abd>	+8.07	0.119	0.365	0.081	0.002
6	<ab>	+8.12	0.078	0.363	0.015	0.002
7	<ad>	+8.74	0.082	0.333	0.022	0.002
8	<a>	+9.86	0.046	0.257	0.005	0.001
9	<>	+23.60	0.000	0.003	0.000	0.000
10	<d>	+24.63	0.000	0.000	0.000	0.000
11	<c>	+25.37	0.000	0.001	0.000	0.000
12	<b>	+25.58	0.000	0.000	0.000	0.000
13	<cd>	+26.38	0.000	0.000	0.000	0.000
14	<bd>	+26.47	0.000	0.000	0.000	0.000
15	<bc>	+27.36	0.000	0.000	0.000	0.000
16	<bcd>	+28.23	0.000	0.000	0.000	0.000

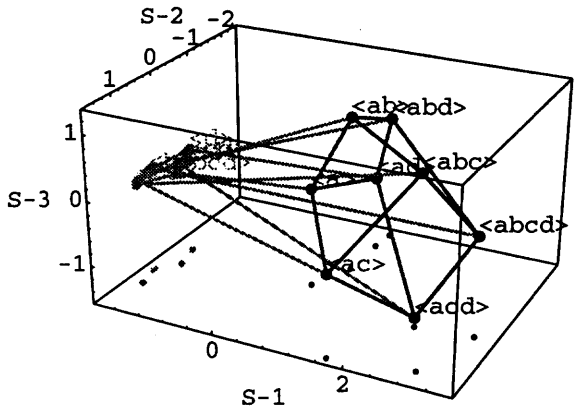

Figure 1: Model map for the candidate models in Case I. The segments indicate the model structure. The major three principal components are used for the drawing; the cumulative contribution ratio (CCR) = 0.94. The two major components are represented by the small points on the bottom. The origin is moved to the center of gravity of the points. The gray-level of points is proportional to the log-likelihood value; light gray means smaller value, and dark gray means larger value.

$n \times n$ matrix $\sum_{\alpha \in \mathcal{M}} (\xi_\alpha - \xi_0)(\xi_\alpha - \xi_0)'$, where $M = |\mathcal{M}|$ and $\xi_0 = \sum_{\alpha \in \mathcal{M}} \xi_\alpha / M$. In the model map, the i-th axis (S-i) denotes the principal component $u_i'(\xi_\alpha - \xi_0)$ for each $\alpha \in \mathcal{M}$. The contribution ratio of the axis is $\lambda_i^2 / \sum_{j=1}^{M} \lambda_j^2$, and the cumulative contribution ratio (CCR) of the three axes is $\sum_{i=1}^{3} \lambda_i^2 / \sum_{j=1}^{M} \lambda_j^2$. In this case, CCR = 0.94, which is large enough to carry the features in most aspects. Note that we calculate u_i's directly from ξ_α's by the singular value decomposition in practice.

The eight models which do not include the predictor z_1 are found as a cluster of light gray points, indicating small values of log-likelihood. On the other side of S-1, the eight models which include z_1 are found as dark gray points, indicating large values of log-likelihood. These eight models are placed approximately on the vertices of a slightly warped cube, representing the simple structure of the boolean lattice formed by inclusion of the other three predictors. The shape implies the three predictors z_3, z_2 and z_4 have about the same size and slightly correlated explanation power on the response y given z_1.

The same pattern of clusters is also found in Table 1, which shows AIC and the four different types of p-values for the candidate models ranked by their AIC values. We see the clear jump of AIC values from <a> to <>, which divides the candidates into two parts corresponding to the two clusters found in the model

map. The cluster of good models is also identified as a confidence set of models using $P_\alpha^{(M)}$ at, say, level = 0.1, while $P_\alpha^{(L)}$ implies a smaller confidence set. $P_\alpha^{(B)}$ implies the same confidence set as $P_\alpha^{(L)}$, if models with smaller values of $P_\alpha^{(B)}$ are eliminated until the sum of them becomes 0.1.

MAICE is <abcd>, the full model. All the other submodels are rejected by $P_\alpha^{(S)}$ at the level = 0.1. This result of LR test may also be seen in the model map; the points of the submodels are far enough from the full model for their rejection. This interpretation is based on

$$\|\xi_\alpha - \xi_\beta\|^2 \approx 2L_\alpha(\hat{\theta}_\alpha) - 2L_\beta(\hat{\theta}_\beta), \tag{10}$$

which holds if model β is nested in model α, denoted as $p_\beta(\cdot) \subset p_\alpha(\cdot)$, and if their predictive densities are close enough to $q(x)$. This is derived from (4), by the Taylor expansion of $L_\alpha(\theta_\alpha)$ around $\hat{\theta}_\alpha$.

4.1.2 *Case II: Heat evolved from cement*

This dataset is taken from Draper and Smith (1981, p. 629). The response variable is heat evolved in calories per gram of cement, and the four predictor variables are the amounts of four major ingredients in percentage. $\mathcal{M}$ consists of $2^4 = 16$ models. The sample size is $n = 13$.

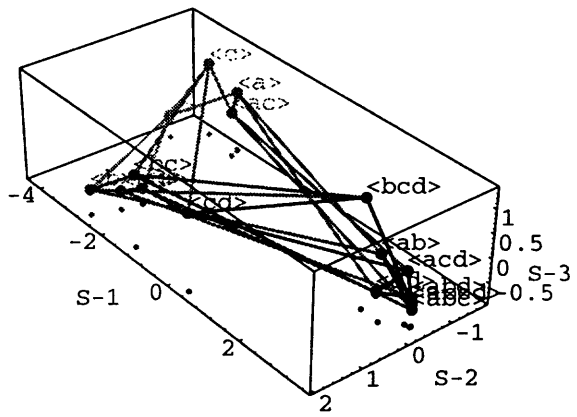

Figure 2: Model map for the candidate models in Case II. CCR = 0.94. A cluster of good models is found in the right side of the map.

The model map is shown in Figure 2. The pattern of points is quite different from that of the previous dataset. A cluster of good seven models is identified

Table 2: AIC and p-values for Case II.

	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$	$P_\alpha^{(S)}$
1	<abd>	64	0.500	0.972	0.180	0.863
2	<abc>	+0.04	0.493	0.938	0.256	0.796
3	<ab>	+0.45	0.438	0.911	0.255	0.290
4	<acd>	+0.75	0.382	0.924	0.158	0.376
5	<abcd>	+1.97	0.088	0.451	0.002	–
6	<ad>	+3.77	0.169	0.540	0.094	0.055
7	<bcd>	+5.60	0.136	0.353	0.052	0.018
8	<cd>	+14.88	0.019	0.074	0.004	0.000
9	<bc>	+26.06	0.001	0.000	0.000	0.000
10	<d>	+33.88	0.000	0.000	0.000	0.000
11	<b>	+34.20	0.000	0.000	0.000	0.000
12	<bd>	+35.66	0.000	0.000	0.000	0.000
13	<a>	+38.55	0.000	0.000	0.000	0.000
14	<ac>	+40.14	0.000	0.000	0.000	0.000
15	<c>	+44.09	0.000	0.000	0.000	0.000
16	<>	+46.47	0.000	0.000	0.000	0.000

clearly, in which all the four submodels with three predictors are included. Especially, <abd> and <abc> are almost indistinguishable from the full model <abcd>. The same cluster is also confirmed in Table 2.

This degenerate shape is a result of the multicollinearity; the sum of the four predictors is approximately a constant in this dataset. Thus any set of three predictors works mostly fine as the full model for the given dataset. This overparametrization is easily identified in the model map.

4.1.3 *Case III: Housing price in Boston*

This dataset is taken from Belsley et al. (1980, p. 244); the computer file is downloaded from StatLib Datasets Archive and some variables are transformed. The response variable is the logarithm of the median value of owner-occupied homes for each area in Boston. There are thirteen predictor variables z_1 to z_{13}, which are labeled as a to m. For example, z_1 (= a) is per capita crime rate by town, z_6 (= f) is average number of rooms per dwelling, z_{11} (= k) is pupil-teacher ratio by town, and z_{13} (= m) is logarithm of the proportion of the population that is lower status. The sample size is $n = 506$. We consider 286 candidates that contain three of the thirteen predictors.

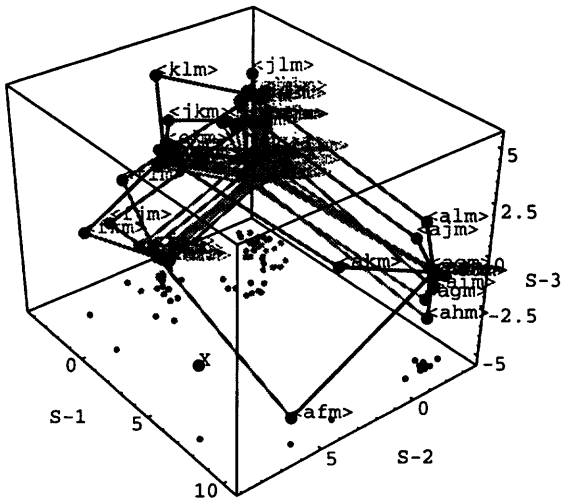

Figure 3: The models which include z_{13} are drawn in $\mathcal{R}^3$ for Case III. The full model is labeled as X. CCR = 0.72.

Table 3: AIC and p-values for the best 20 models ranked by AIC in Case III.

	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$	$P_\alpha^{(S)}$
1	<afm>	−156	0.500	0.994	0.487	0.000
2	<akm>	+1.72	0.461	0.986	0.381	0.000
3	<ahm>	+18.67	0.180	0.774	0.069	0.000
4	<agm>	+24.76	0.098	0.559	0.006	0.000
5	<adm>	+25.32	0.110	0.570	0.016	0.000
6	<alm>	+32.26	0.055	0.256	0.002	0.000
7	<ajm>	+35.31	0.028	0.061	0.000	0.000
8	<abm>	+38.42	0.020	0.122	0.000	0.000
9	<aim>	+39.51	0.016	0.104	0.000	0.000
10	<acm>	+40.72	0.013	0.040	0.000	0.000
11	<aem>	+40.94	0.013	0.048	0.000	0.000
12	<klm>	+46.86	0.079	0.441	0.022	0.000
13	<fjm>	+48.73	0.019	0.280	0.005	0.000
14	<jkm>	+53.99	0.028	0.140	0.000	0.000
15	<fkm>	+54.11	0.023	0.331	0.004	0.000
16	<flm>	+59.78	0.027	0.299	0.006	0.000
17	<hjm>	+60.70	0.023	0.227	0.002	0.000
18	<dkm>	+62.01	0.023	0.177	0.001	0.000
19	<bkm>	+62.69	0.018	0.148	0.000	0.000
20	<ekm>	+64.75	0.013	0.078	0.000	0.000

From a simple plot of $L_\alpha(\hat{\theta}_\alpha)$, $\alpha \in \mathcal{M}$, we find that 66 models which include z_{13} form a cluster of good models. Figure 3 shows the model map for them. Additionally, we drew the other 13 submodels with one or two predictors, as well as the full model labeled as `X`.

`<afm>` and `<akm>` are relatively close to `X`. The other nine models that include z_1 and z_{13} form a cluster, and `<am>` is located at the center of the cluster. These models are the best 11 models in terms of AIC as shown in Table 3. The model map implies that the best two models show distinctively good predictive performances, but the other nine models show a next best performance similar to each other.

The five models ranked from 12 to 16, namely, `<klm>` to `<flm>`, are easily identified as isolated points in the model map. These models have relatively high $P_\alpha^{(L)}$ values, larger than some of the best 11 models. In other words, $T_{\alpha,\mathrm{MAICE}}$ is relatively small for them. This result can be read in the model map; these models are far from `<afm>` and the differences of their AIC values from MAICE are regarded as uncertain. This interpretation is based on

$$\hat{\sigma}^2_{\alpha,\beta} \approx \|\xi_\alpha - \xi_\beta\|^2, \tag{11}$$

which is derived from (7) by ignoring the higher order terms.

Note that $P_\alpha^{(S)} = 0.000$ for all $\alpha \in \mathcal{M}$; all the candidate models are very far from the full model. It is often the case that the set of adequate models is empty if the full model is not a member of $\mathcal{M}$. The methods of Section 3 other than $P_\alpha^{(S)}$ are appropriate in such a case. On the other hand, $P_\alpha^{(S)}$ is often appropriate if the full model is a member of $\mathcal{M}$. The model selection tests ($P_\alpha^{(L)}$ and $P_\alpha^{(M)}$) are very conservative for nested models as discussed in Shimodaira (1997a).

4.2 Evolutionary tree reconstruction in phylogeny

The branching order (i.e. topology) of evolutionary tree is inferred from the genetic information (i.e. molecular sequences) of contemporary species. This is another example of model selection. Given the topology of the tree, the nucleotide or the amino acid substitution is modeled as the time reversible Markov process along the genealogy; it is a probabilistic model with the parameters of the substitution process and the edge lengths of the tree measured by the expected number of substitutions. Until now, many such schemes have been proposed in the literature for the substitution process, and the parameters and the topology have been estimated by the maximum likelihood principle since Felsenstein (1981, 1983). Because the direction of the evolution is not specified in the model, the root of the tree must be inferred by using an outgroup species; such as paleontological evidence clearly indicates that it has branched off earlier than the other taxa under study.

The dataset consists of aligned molecular sequences of n sites for s species. Let the status of species i at site t be $x_{i,t} \in \mathcal{A}$, $i = 1, \ldots, s$, $t = 1, \ldots, n$, where $\mathcal{A} =$ {A, T, G, C} for nucleotide sequences, or $\mathcal{A}$ consists of 20 amino acid code letters for protein sequences. We assume the substitution process is independent and identical for every site, and the dataset is regarded as an outcome of a submodel of the multinomial distribution of n trials with $|\mathcal{A}|^s$ categories (Efron et al., 1996).

Table 4: Ten splits of five groups. Each split corresponds to the internal edge which separates the groups.

$$
\begin{aligned}
\texttt{a} &= \{G_1, G_2, G_3 | G_4, G_5\} \\
\texttt{b} &= \{G_1, G_3, G_4 | G_2, G_5\} \\
\texttt{c} &= \{G_2, G_3, G_4 | G_1, G_5\} \\
\texttt{d} &= \{G_1, G_2, G_4 | G_3, G_5\} \\
\texttt{e} &= \{G_1, G_2 | G_3, G_4, G_5\} \\
\texttt{f} &= \{G_3, G_4 | G_1, G_2, G_5\} \\
\texttt{g} &= \{G_2, G_4 | G_1, G_3, G_5\} \\
\texttt{h} &= \{G_1, G_3 | G_2, G_4, G_5\} \\
\texttt{i} &= \{G_2, G_3 | G_1, G_4, G_5\} \\
\texttt{j} &= \{G_1, G_4 | G_2, G_3, G_5\}
\end{aligned}
$$

In the examples below, we consider $m = 5$ groups of species, $\mathcal{G} = \{G_1, G_2, G_3, G_4, G_5\}$. The number of unrooted tree topologies is $M = (2m-5)!/(2^{m-3}(m-3)!) = 15$. Selection of a topology is essentially the same as the selection of a set of predictors in Section 4.1; each topology is uniquely specified by a combination of *splits*, which are partitions of $\mathcal{G}$ into two parts. (See Table 4.) For example, the topology $((G_1, (G_2, G_3)), G_4)$ given in Figure 4 is denoted by <ai>. The number of splits is $B = 2^{m-1} - (m+1) = 10$, and each bifurcating tree consists of $N = m - 3 = 2$ splits. Not all the combinations of N splits out of the B splits are allowed to construct trees. Denoting $x^c = \mathcal{G} \setminus x$, for two splits $\{x|x^c\}$ and $\{y|y^c\}$ to construct a tree, one of $x \bigcap y$, $x^c \bigcap y$, $x \bigcap y^c$, or $x^c \bigcap y^c$ must be the empty set. This constraint makes the algebraic structure of tree topologies much interesting than the usual variable selection.

4.2.1 *Case IV: Mitochondrial nucleotide sequences of primates*

The code table of DNA is redundant, because $4^3 = 64$ codons, triplets of nucleotides, are translated into 20 amino acids. The third position of codon often

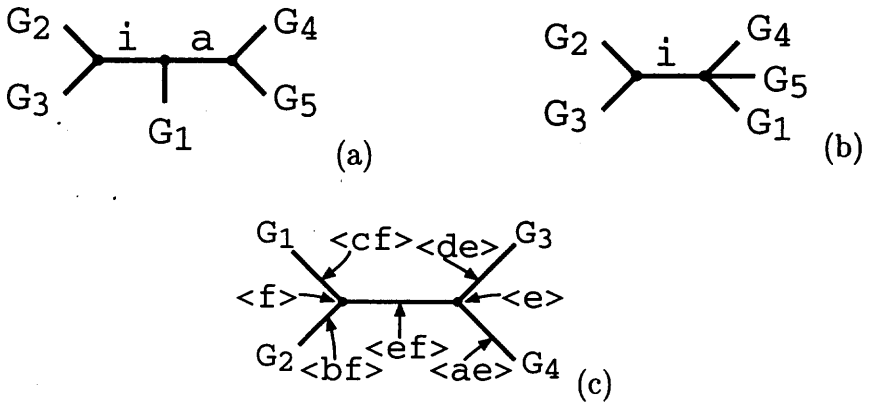

Figure 4: (a) The bifurcating topology <ai>; (b) the multifurcating topology <i>. They are denoted by $((G_1,(G_2,G_3)),G_4)$ and $(G_1,(G_2,G_3),G_4)$ respectively, assuming G_5 is the outgroup. <ai> reduces to <i> if the internal edge length corresponding to a is zero. Similarly, <i> reduces to the star topology <> if the internal edge i is zero. (c) The five bifurcating topologies and the two multifurcating topologies, which are associated with the split $\{G_1,G_2|G_3,G_4\}$. The arrows indicate where G_5 stick to.

does not affect the amino acid, and so carries the information of the genetic diversity; those $n = 1669$ four-fold degenerate sites for $s = 5$ primates are extracted from the dataset of Horai et al. (1995). See Adachi and Hasegawa (1996) for details of the dataset and the Markov models used in the analysis. We are interested in the branching order of G_1 = Human, G_2 = Chimpanzee, G_3 = Bonobo, G_4 = Gorilla, and G_5 = Orangutan; G_5 is the outgroup here.

Table 5 shows that the topology <ai> is significantly better than the others in terms of $P_\alpha^{(L)}$ and $P_\alpha^{(B)}$, while $P_\alpha^{(M)}$ suggests the conventional topologies <ci> and <ij> are on the border of rejection at 5% level. Anyway, these conventional topologies reduce to the multifurcating topology <i>, because their MLE's imply that the internal edge lengths corresponding to the splits c and j are zero. This result agrees with that of Hasegawa and Yano (1984), Sibley and Ahlquist (1984) and Bar-Hen and Kishino (1998). We employed HKY model of Hasegawa et al. (1985) for the substitution process, in which some symmetricity is assumed and one parameter for the process plus three parameters for the base compositions are estimated from data. All the parameters including the edge lengths are estimated by the program package MOLPHY of Adachi and Hasegawa (1995), which generates the the site-wise log-likelihood (1). The result is almost the same even if we use another substitution process, called TN model of Tamura and Nei (1993), in which an additional parameter is estimated from data.

Table 5: AIC and p-values for the bifurcating tree topologies of primates. Some topologies are degenerated to multifurcating topologies, because the internal edge lengths are restricted to be non-negative: <ci> and <ij> to <i>, <ae> and <ah> to <a>. The slight difference of $P_\alpha^{(M)}$ and $P_\alpha^{(B)}$ for them is due to the bootstrap sampling error.

	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$
1	<ai>	11352	0.500	1.000	0.983
2	<ci>	+28.32	0.011	0.049	0.006
3	<ij>	+28.32	0.011	0.052	0.007
4	<ae>	+50.65	0.004	0.019	0.002
5	<ah>	+50.65	0.004	0.020	0.002
6	<bj>	+66.67	0.000	0.002	0.000
7	<dj>	+67.59	0.000	0.001	0.000
8	<cf>	+90.70	0.000	0.001	0.000
9	<cg>	+91.31	0.000	0.001	0.000
10	<bf>	+100.86	0.000	0.002	0.000
11	<ef>	+101.46	0.000	0.002	0.000
12	<bh>	+101.61	0.000	0.002	0.000
13	<dg>	+102.25	0.000	0.002	0.000
14	<gh>	+102.26	0.000	0.002	0.000
15	<de>	+102.37	0.000	0.002	0.000

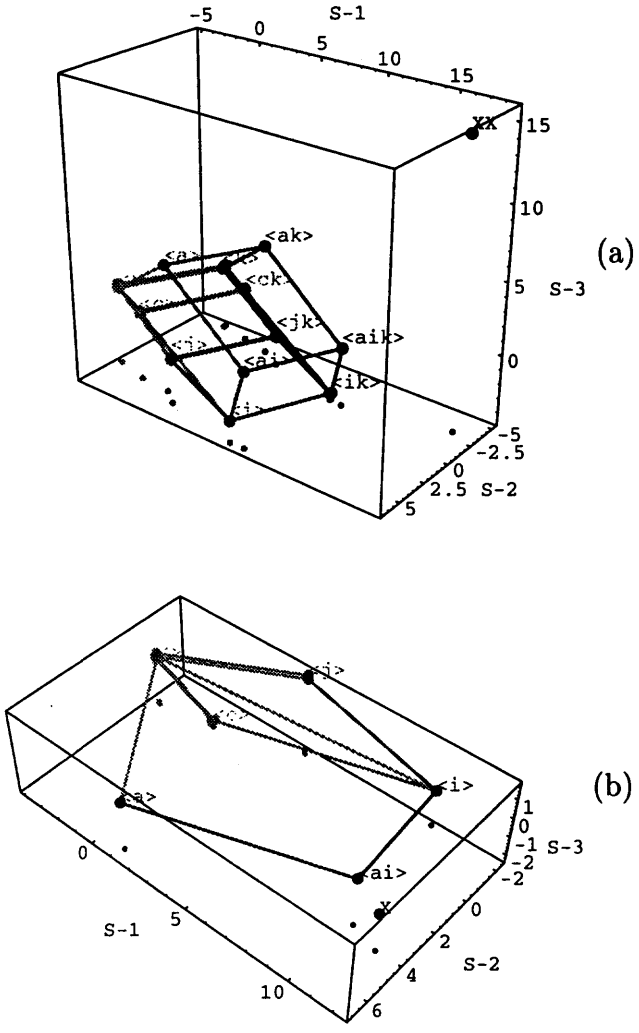

Figure 5: (a) Model map for the 15 bifurcating topologies of primates as well as their submodels (multifurcating topologies). Most of the topologies are almost degenerated to the submodels, and the labels are given only for non-degenerated models. Not only HKY model, but also TN model is considered for the substitution process; the factor `k` denotes the additional parameter for TN model. The points for the topologies estimated under TN model are placed in parallel with those for HKY model. The observed data is indicated by the point `XX`, which is obtained from the non-restricted multinomial distribution. The principal component vectors are calculated from all the points in the map. CCR = 0.97. (b) Model map for the topologies under HKY model. CCR = 0.98. The principal component vectors are calculated from all the points but `XX`. These points span a hyperplane, and `XX` is projected to it as indicated by `X`. Here `X` is very close to `<ai>`.

Figure 5 shows the model maps. In Panel (a), we drew the points for both of HKY model and TN model. The point XX is the predictive density obtained under the non-restricted multinomial distribution. The degenerated topologies are indistinguishable from their submodels in the map. All the points are very far from the data point XX as similar to Case III; actually, Adachi and Hasegawa (1996) showed <ai> and <aik> are rejected by a goodness-of-fit test.

On the other hand, X is very close to <ai> in Panel (b), suggesting the fitting of <ai> is very good in the "full model" as explained below. Here the model map is once drawn using the points of topologies (including the multifurcating ones), then the data point XX is projected to the principal component vectors u_i, $i = 1, 2, 3$, and labeled as X. This configuration of the points appears to be robust, because almost the same configuration is obtained if we use the points of TN model rather than HKY model. X is approximately regarded as the predictive density of the full model; it is the minimal model containing all the non-degenerated topologies, which is uniquely specified by the local linearization around the star topology <> (Shimodaira, 1997b). The non-restricted multinomial distribution also contains all those topologies, but it is unnecessarily large and includes redundant dimensions to compare the topologies each other.

The model map leads to tests of the topologies against the full model. One possibility is to apply the simultaneous confidence region test; topologies are rejected if they are outside of the confidence sphere of radius $\sqrt{\chi^2_{3,1-\mathrm{level}}}$ centered at X. The p-value of <ai> is 0.54, and those for the other topologies are less than 10^{-9}. Another possibility is to apply an approximate version of the LR test. For example, the dimension of <ai> relative to <> is two, and so the test statistic $\sum_{i=1}^{4}(u_i'(\xi_{\texttt{<ai>}} - \xi_{\texttt{X}}))^2$ is approximately distributed as χ^2_{4-2}. We should consider the model map in $\mathcal{R}^4$ (CCR = 0.99) for this test, since the four non-degenerated bases <a>, <i>, <j>, and <c> span the full model around <>. The p-value of <ai> drops to 0.039, and those for the others are less than 10^{-10}, where the segments are long enough to allow us to ignore the non-negative boundary of edge length. In either case, <ai> is the only topology which cannot be rejected clearly. Note that these tests are difficult to obtain without resorting to the model map approach, because the full model does not give a tree topology but a web topology of species, and its construction is not obvious.

4.2.2 *Case V: Mitochondrial protein sequences of vertebrates*

The third position of codon is quickly randomized and the information about the topology is lost for distantly related species. Thus we show another example by using protein sequences dataset of $n = 3274$ sites for $s = 22$ species to solve the

debate of the origin of tetrapods: G_1 = Tetrapods (15 species), G_2 = Lungfish, G_3 = Coelacanth, and G_4 = Ray-finned fish (4 species). The outgroup is G_5 = Lamprey. The parameters are estimated by the program package PAML of Yang (1997). The transition rate matrix of the substitution process is obtained empirically from 20 mammalian species. See Cao (1998) and Cao et al. (1998) for the details of the dataset.

Table 6: AIC and p-values for the bifurcating tree topologies of vertebrates. The substitution rate is assumed to be the same for all the sites.

	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$
1	<bf>	113877	0.500	1.000	0.953
2	<cf>	+80.82	0.023	0.120	0.021
3	<bj>	+85.98	0.022	0.114	0.020
4	<bh>	+124.82	0.000	0.003	0.000
5	<ef>	+139.41	0.000	0.000	0.000
6	<ci>	+141.73	0.010	0.062	0.006
7	<cg>	+195.38	0.000	0.002	0.000
8	<de>	+223.18	0.000	0.000	0.000
9	<ij>	+231.38	0.000	0.001	0.000
10	<dj>	+232.40	0.000	0.001	0.000
11	<ai>	+242.50	0.000	0.000	0.000
12	<ae>	+253.26	0.000	0.000	0.000
13	<dg>	+267.05	0.000	0.000	0.000
14	<ah>	+303.83	0.000	0.000	0.000
15	<gh>	+313.46	0.000	0.000	0.000

Table 6 shows that the topology <bf> is significantly better than the others, while only $P_\alpha^{(M)}$ implies the possibility of <cf> or <bj> being the best topology. These topologies are inconsistent with the conventional idea that G_4 is the outgroup of $\{G_1, G_2, G_3\}$. The conventional topologies <ae> $= (((G_1, G_2), G_3), G_4)$, <ah> $= (((G_1, G_3), G_2), G_4)$, and <ai> $= (((G_2, G_3), G_1), G_4)$ are rejected clearly by all the statistics. Despite this surprising result, a warning is given in the model maps of Figure 6. Panel (a) shows that all the topologies are very far from the full model X, though <bf> is relatively close to X; all the topologies are rejected against the full model. This suggests the substitution process is misspecified significantly even for resolving the topologies. The situation is different from Case IV, because the substitution process was misspecified also there (XX was far from the topologies), but the fitting of one of the topologies was quite nice in the full

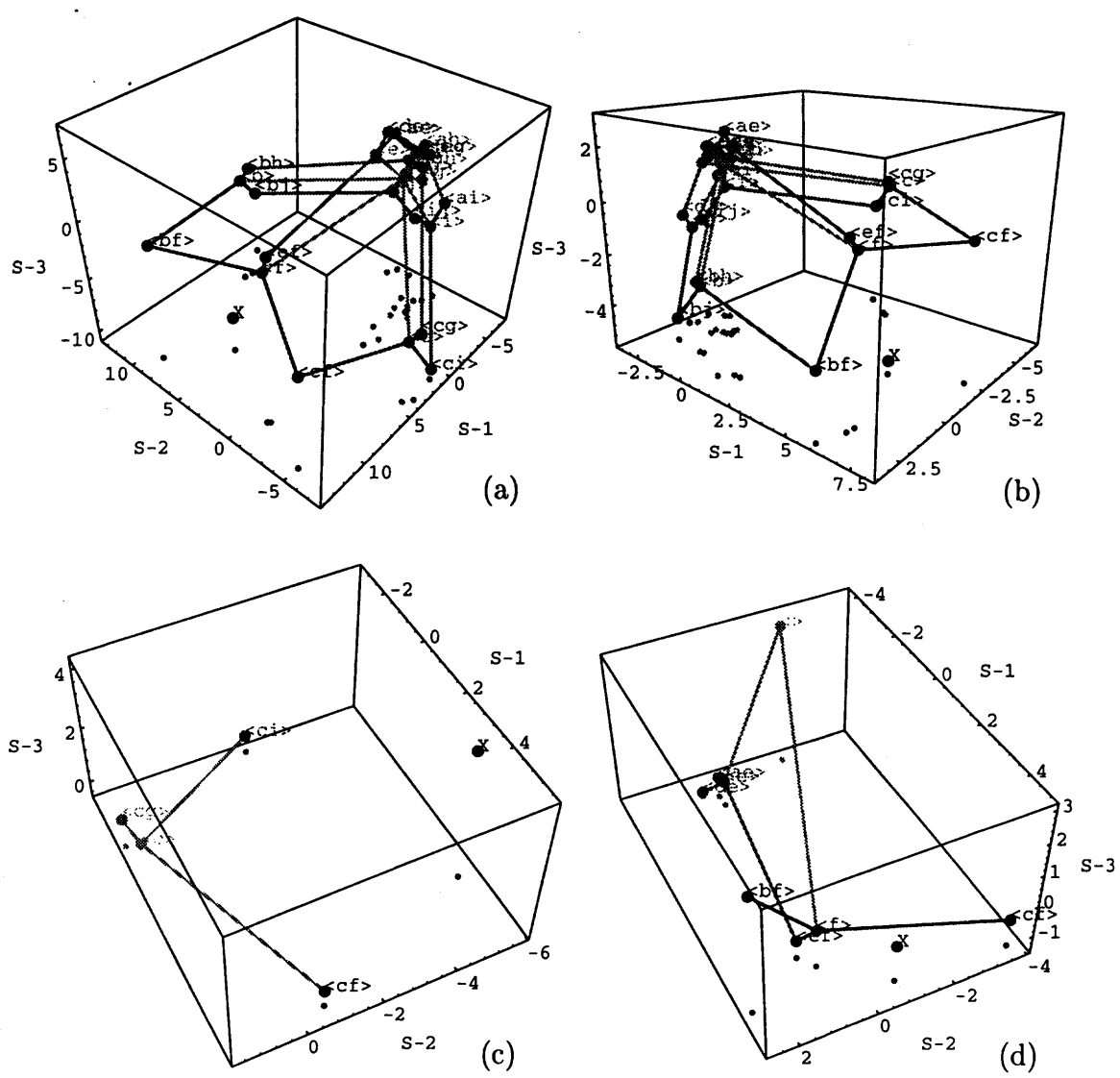

Figure 6: (a) Model map of the vertebrates drawn for the 15 bifurcating topologies and 11 multifurcating topologies. CCR = 0.72. X denotes the full model spanned by the 10 splits around the star topology <>. (b) Among-site rate variation is modeled by the gamma distribution. CCR = 0.77. (c) Model map drawn for the topologies including the split c. The full model is spanned by the three splits f, g, and i around the model <c>. CCR = 1. (d) The seven topologies in Panel (c) of Figure 4. The full model is spanned by the 6 splits a, b, c, d, e, and f around the star topology. CCR = 0.90.

Table 7: AIC and p-values for the bifurcating tree topologies of vertebrates. Among-site rate variation is modeled by the gamma distribution.

	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$
1	<cf>	104790	0.500	0.946	0.516
2	<bf>	+5.46	0.385	0.794	0.354
3	<ci>	+23.83	0.135	0.443	0.108
4	<ef>	+26.31	0.036	0.178	0.002
5	<cg>	+44.41	0.007	0.041	0.000
6	<bj>	+45.73	0.074	0.226	0.015
7	<de>	+55.19	0.015	0.103	0.002
8	<bh>	+60.76	0.018	0.032	0.000
9	<ij>	+62.76	0.022	0.145	0.001
10	<dj>	+64.17	0.020	0.132	0.001
11	<ae>	+64.40	0.004	0.036	0.000
12	<ai>	+68.56	0.012	0.078	0.000
13	<dg>	+78.18	0.003	0.027	0.000
14	<ah>	+88.12	0.001	0.008	0.000
15	<gh>	+90.82	0.000	0.004	0.000

model (X was close to <ai>).

Yang (1996) discussed that the relative substitution rates among sites can be modeled by the gamma distribution. We reanalyzed the vertebrates data with this substitution model, in which an additional gamma-shape parameter is estimated. Table 7 shows that the fitting was improved significantly compared with the constant rate model; the difference of AIC values is 9×10^3. However, it turned out that AIC differences among the topologies decreased and the uncertainty became large. The scales of the axes in Panel (b) are reduced about half of Panel (a), while the relative positions of the topologies are not very different. This implies that the introduction of the gamma distribution did not effect the model in terms of resolving the topologies; but it eliminated the spurious significance by reducing the effective sequence size. The configuration of the points implies that the uncertainty came from the misspecification (as in Case III) rather than the overparametrization (as in Case II). The two model maps give us warning that the substitution process still needs to be improved.

Panel (c) is the model map to resolve the conventional three hypotheses associated with the splits $\{G_1, G_2|G_3, G_4\}$ (= e, f), $\{G_1, G_3|G_2, G_4\}$ (= g, h), and $\{G_2, G_3|G_1, G_4\}$ (= i, j). <cf>, <cg>, and <ci>, respectively, are representatives for them taken from the good models of Table 7. The model map suggests $\{G_1, G_2|G_3, G_4\}$ is better than the others, but it is dubious because of the misspecification.

In the conventional hypotheses, G_4 is assumed to constitute the outgroup with G_5, but it is rejected in the analysis above. Panel (d) shows the possible roots of the four groups assuming the split $\{G_1, G_2|G_3, G_4\}$ is true. <ef> is very close to <f>, which implies the root should be very close to the node separating G_1 and G_2 under <ef>; it follows from (10) that the distance $\|\xi_{<f>} - \xi_{<ef>}\| = 1.11$ approximates the estimated edge length divided by its standard error. <cf>, <bf>, and <ef> are the most likely topologies. But it does not mean the multifurcating topology <f> is supported by them. $\|\xi_{<f>} - \xi_{<cf>}\| = 5.46$ and $\|\xi_{<f>} - \xi_{<bf>}\| = 4.45$ are rather large and their corresponding edge lengths are significantly larger than zero. Obviously <cf> and <bf> are inconsistent, which came from the misspecification shown in the model maps.

5 Concluding remarks

In Case V, the full model X is calculated without using XX. According to (10),

$$\frac{\kappa^2}{2}\|\xi_\alpha - \xi_c\|^2 \approx 1_n'(\xi_\alpha - \xi_c) \tag{12}$$

holds for $\alpha \in \mathcal{M}$, where $\kappa = 1$, and c is the model nested in all the other models (i.e. the star topology). Because of the distortion, however, (12) holds nicely with $\kappa = 0.72$ for Panel (a) of Figure 6, or $\kappa = 0.79$ for Panel (b). The deviation of κ from unity implies that XX is too far from the models, and X may not be obtained properly by the projection of it. Thus, we calculated $\xi_{\mathtt{X}}$ such that

$$\frac{\kappa^2}{2}\|\xi_{\mathtt{X}} - \xi_\alpha\|^2 \approx 1_n'(\xi_{\mathtt{X}} - \xi_\alpha) \tag{13}$$

holds for $\alpha \in \mathcal{M}$ with κ obtained from (12). In the model maps of Case V, we rescaled the axes by the factor κ for the calibration. This technique is useful to calculate approximately the composite model spanned by a given set of models.

The model map approach to the approximate LR test against the comprehensive model is presented in Case IV. The idea is similar to the Lagrange multiplier test based on the *artificial regression* of Davidson and MacKinnon (1987) in which the score vectors are used rather than the log-likelihood vectors of (1). Both of the approaches make use of the local linearization, and give a similar result if the full model is not very far from the null model. Our approach is practically advantageous if existing program packages output only the log-likelihood vectors but the score vectors.

Acknowledgment

The idea of model map was inspired by M. Ishiguro. The authors thank M. Hasegawa, H. Kishino, Z. Yang, and Joe Felsenstein for the helpful discussions in phylogenetic problems.

Appendix

Consider a family of densities parameterized by $\phi = (\phi^1, \phi^2) \in \mathcal{R}^2$ such that the three densities $q(x)$, $p_1(x)$, and $p_2(x)$ are included in it. For example,

$$p(x|\phi) = \exp((1 - \phi^1 - \phi^2)\log q(x) + \phi^1 \log p_1(x) + \phi^2 \log p_2(x) - c(\phi)),$$

where $c(\phi)$ is defined by $\int p(x|\phi)\,dx = 1$. In the following, l denotes $\log p$.

Let ϕ_α and ϕ_β be two distinct parameter values of ϕ. Using Taylor expansion of $l(x|\phi_\beta)$ around ϕ_α and that of $p(x|0,0) = q(x)$, we have

$$\begin{aligned} & p(x|0,0)(l(x|\phi_\alpha) - l(x|\phi_\beta)) \\ = \; & (p(x|\phi_\alpha) + O_p(||\phi_\alpha||))(\sum_i \partial_i l(x|\phi_\alpha)(\phi_\alpha^i - \phi_\beta^i) + O_p(||\phi_\alpha - \phi_\beta||^2)) \end{aligned}$$

$$= \sum_i p(x|\phi_\alpha)\partial_i l(x|\phi_\alpha)(\phi_\alpha^i - \phi_\beta^i) + O_p((||\phi_\alpha|| + ||\phi_\beta||)||\phi_\alpha - \phi_\beta||)$$
$$= p(x|\phi_\alpha) - p(x|\phi_\beta) + O_p((||\phi_\alpha|| + ||\phi_\beta||)||\phi_\alpha - \phi_\beta||),$$

where $\partial_i = \partial/\partial\phi^i$. In the last equation, $p\,\partial_i l = \partial_i p$ was used. Therefore,

$$\begin{aligned} & \int q(x)(l(x|\phi_\alpha) - l(x|\phi_\beta))^2\,dx \\ = & \int (p(x|\phi_\alpha) - p(x|\phi_\beta))(l(x|\phi_\alpha) - l(x|\phi_\beta))\,dx \\ & \qquad + O((||\phi_\alpha|| + ||\phi_\beta||)||\phi_\alpha - \phi_\beta||^2) \\ = & (1+\lambda)J(p(\cdot|\phi_\alpha);\, p(\cdot|\phi_\beta)), \end{aligned}$$

where $\lambda = O\Big(\sqrt{J(q(\cdot);\, p(\cdot|\phi_\alpha))} + \sqrt{J(q(\cdot);\, p(\cdot|\phi_\beta))}\Big)$. Finally put $\phi_\alpha = (1,0)$ and $\phi_\beta = (0,1)$ to obtain (2).

References

ADACHI, J. AND HASEGAWA, M. (1995). *MOLPHY: Programs for Molecular Phylogenetics, ver. 2.3.* Institute of Statistical Mathematics, Tokyo.

ADACHI, J. AND HASEGAWA, M. (1996). Tempo and mode of synonymous substitutions in mitochondrial DNA of primates. *Mol. Biol. Evol.* **13**, 200–208.

AITKIN, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics* **16**, 221–227.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**, 716–723.

BAR-HEN, A. AND KISHINO, H. (1998). Comparing the likelihood functions of phylogenetic trees. *Ann. Inst. Statist. Math.* (in press).

BELSLEY, D. A., KUH, E. AND WELSCH, R. E. (1980). *Regression Diagnostics.* Wiley, New York.

CAO, Y. (1998). *Molecular Phylogeny and Evolution of Vertebrates.* Doctor of science, Dept. of Bioscience, Faculty of Bioscience and Biotechnology, Tokyo Institute of Technology, Tokyo, Japan.

CAO, Y., WADDELL, P. J., OKADA, N. AND HASEGAWA, M. (1998). The complete mitochondrial DNA sequence of the shark mustelus manazo: Evaluating rooting contradictions to living bony vertebrates. *Mol. Biol. Evol.* submitted.

DAVIDSON, R. AND MACKINNON, J. G. (1987). Implicit alternatives and the local power of test statistics. *Econometrica* **55**, 1305–1329.

DRAPER, N. R. AND SMITH, H. (1981). *Applied Regression Analysis.* Wiley, New York, second edition.

EFRON, B., HALLORAN, E. AND HOLMES, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* **93**, 13429–13434.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.

FELSENSTEIN, J. (1983). Statistical inference of phylogenies. *J. Roy. Statist. Soc. Ser. A* **146**, 246–272.

FELSENSTEIN, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

HASEGAWA, M., KISHINO, H. AND YANO, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.

HASEGAWA, M. AND YANO, T. (1984). Phylogeny and classification of Hominoidea as inferred from DNA sequence data. *Proc. Japan Acad.* **B60**, 389–392.

HORAI, S., HAYASAKA, K., KONDO, R., TSUGANE, K. AND TAKAHATA, N. (1995). The recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA* **92**, 532–536.

KISHINO, H. AND HASEGAWA, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170–179.

KISHINO, H., MIYATA, T. AND HASEGAWA, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **30**, 151–160.

LINHART, H. (1988). A test whether two AIC's differ significantly. *South African Statist. J.* **22**, 153–161.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

SAWA, T. (1979). *Kaiki Bunseki (Regression Analysis).* Asakura Shoten, Tokyo. (in Japanese).

SHIMODAIRA, H. (1993). A model search technique based on confidence set and map of models. *Proc. Inst. Statist. Math.* **41**, 131–147. (in Japanese).

SHIMODAIRA, H. (1997a). Assessing the error probability of the model selection test. *Ann. Inst. Statist. Math.* **49**, 395–410.

SHIMODAIRA, H. (1997b). Testing multiple nonnested hypotheses using score statistics and its application to phylogenetic inference. Research Memorandum 648, The Institute of Statistical Mathematics, Tokyo, Japan.

SHIMODAIRA, H. (1998). An application of multiple comparison techniques to model selection. *Ann. Inst. Statist. Math.* **50**, 1–13.

SIBLEY, C. AND AHLQUIST, J. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J. Mol. Evol.* **20**, 2–15.

SPJϕTVOLL, E. (1977). Alternatives to plotting C_p in multiple regression. *Biometrika* **64**, 1–8.

TAMURA, K. AND NEI, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.

VUONG, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–333.

YANG, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecol. & Evol.* **11**, 367–372.

YANG, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**, 555–556.

REGISTRATION
of
Research Memorandum

(month) (day) (year)
Research Memorandum NO. 680 Received on May 29, 1998 by

Section of Research Information, the Institute of Statistical Mathematics

(*Do not write above this line*)

(*Fill in the blanks below and return to Section of Research Information*)

Title

A Graphical Technique for Model Selection Diagnosis

Author(s) and Affiliation(s)

Hidetoshi Shimodaira
and
Ying Cao
The Institute of Statistical Mathematics, Tokyo, JAPAN
Email: shimo@ism.ac.jp and cao@ism.ac.jp

Key Words Akaike information criterion, graphical exploratory method, Kullback-Leibler divergence, nonnested models, phylogenetic inference, variable selection.

Abstract

A graphical method is presented for understanding the relations among the parametric models with respect to data. The models are represented by their predictive densities, and they are drawn in Euclidean space preserving approximately the symmetrized divergence between these densities. This direct visualization of models is very simple and useful for diagnosis of the model selection, especially for nonnested models. Problems such as overparametrization or misspecification of the models are identified from the configuration of the points. Structural patterns in good models are also found as clusters. This is complementary to the methods for assessing the uncertainty of model selection; a brief review is given for the confidence set of models derived from the variance of log-likelihood differences, the bootstrap selection probability, and the simultaneous likelihood ratio test. Illuminating examples from the variable selection in multiple regression as well as practical examples from the phylogeny tree reconstruction are given to illustrate the methodology.