Ann. Inst. Statist. Math.
Vol. 49, No. 3, 395–410 (1997)

ASSESSING THE ERROR PROBABILITY OF THE MODEL SELECTION TEST*

HIDETOSHI SHIMODAIRA**

Department of Mathematical Engineering and Information Physics, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan

(Received September 25, 1995; revised May 13, 1996)

Abstract. The asymptotic error probability of Linhart's model selection test is evaluated, and compared with the nominal significance level. We examine the case where the expected discrepancies of the candidate models from the true model are asymptotically equal. The local alternatives method is employed in the limiting operation of the asymptotic evaluation. Although the error probability under the null hypothesis is actually shown to be equal to or less than the level for most situations, intolerable violations of the error control are observed for nested models: It is often erroneously concluded that the smaller model is significantly better than the larger model. To prevent this violation, a modification of Linhart's test statistic is proposed. The effectiveness of the proposed test is confirmed through theoretical analysis and numerical simulations.

Key words and phrases: Akaike's information criterion, model selection, fixed alternatives, local alternatives, test on expected discrepancies, error probability, canonical correlation coefficient.

1. Introduction

Akaike's information criterion (AIC) is an estimator of the Kullback-Leibler (K-L) discrepancy between the true density and its estimate. The best model among given candidates is defined to be the minimizer of the expected discrepancy. The minimizer of AIC over the candidates is regarded as an estimate of the best model.

Linhart (1988) considered a test of whether two AIC's differ significantly for a pair of models among the candidates. The test statistic is a standardized difference of AIC between the two models. It converges to $N(0,1)$, the standard normal distribution, as the sample size n goes to infinity under the null hypothesis

* This work was supported in part by Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture.

** Now at The Institute of Statistical Mathematics.

that the two expected discrepancies are equal. Linhart's model selection test is based on this asymptotic normality of the test statistic.

In practice, the null hypothesis does not hold, and the difference between the two expected discrepancies is of order $O(1)$. In this case the test statistic goes to infinity, or minus infinity, as $n \to \infty$. To evaluate the error probability, a "moderate" sample size may be assumed: The sample size is large enough to ensure that the usual expansions of maximum-likelihood theory are good approximations, but not so large that the test is performed with negligible probability of error (Cox (1962)).

Rather than the moderate sample size under *fixed alternatives* mentioned above, a mathematically strict way to handle this difficulty is the asymptotic theory under *local alternatives*. We consider a sequence of densities converging to a point in the intersection of the two models. The rate of convergence is of order $O(1/\sqrt{n})$ so that the test statistic will be bounded in probability even if $n \to \infty$. The aim of this article is to discuss the error probability of the model selection test under this setup.

Steiger *et al.* (1985) derived the joint distribution of AIC's for nested models under local alternatives. We give the general results including nonnested cases in Section 2. Under local alternatives, the discrepancy reduces to just the Euclidean squared distance. Then the maximum likelihood estimator (MLE) of the true density will be the projection of a standard normal vector onto a linear space. Using this standard normal vector and the projection operators for the two models, we can asymptotically express both the AIC difference and an estimate of its variance. Consequently the test statistic reduces to its *canonical form*, which is given by the canonical correlation coefficients (Hotelling (1936)) between the two MLE's. The canonical form is the basis for the error probability assessment in Section 3.

Theoretical analysis of the canonical form suggests that the model selection test is rather conservative for most situations. Numerical simulations, however, indicate that the error probability of the test can be larger than the nominal significance level for nested models. Ideally, the error probability under the null hypothesis should be equal to the level; while a conservative test is not ideal, it is much better than violation of the error control. To prevent such violation, a modification of Linhart's test statistic is considered in Shimodaira (1996). The second order term is added to the variance estimator of the difference between the two AIC's, whereas only the first order term is used in Linhart's test statistic. This modification remedies the violation to a considerable extent.

The construction of this article is as follows. In Subsection 2.1, the MLE is discussed under local alternatives; results are summarized in Proposition 2.1. In Subsection 2.2, the asymptotic distribution of the test statistic is discussed, and its canonical form is given in Proposition 2.2. The error probability assessments through theoretical analysis and numerical simulations are given in Subsection 3.1 and Subsection 3.2, respectively. Concluding remarks are made in Section 4. All proofs are deferred to Appendix A.

2. Asymptotic distribution

2.1 *Models and maximum likelihood estimator*

Consider a parametric family of densities of random variable x, $p(\cdot) = \{p(x \mid \xi) \mid \xi \in \Xi\}$, where $\Xi \subset \mathcal{R}^m$ is the parameter space. Let $x_1^{(n)}, \ldots, x_n^{(n)}$ be i.i.d. observations of sample size n from the unknown true density $p(x \mid \xi^{(n)})$, where $\xi^{(n)} \in \Xi$ is the true parameter value. We consider $\xi^{(n)}$ depends on n, and it converges to ξ^*, an interior point of Ξ. The rate of the convergence is of order $O(1/\sqrt{n})$, that is, $\lim_{n\to\infty} \sqrt{n}(\xi^{(n)} - \xi^*) = \xi^\diamond \in \mathcal{R}^m$ exists. Let Ξ^* denote a generic neighborhood of ξ^* in Ξ. In the following, we will see that the space of distributions in Ξ^*, whose scale is magnified by $\sqrt{n}$ times, reduces asymptotically to a linear space as $n \to \infty$.

Assume regularity conditions that $\mathcal{E}_\xi\{|\partial_i \partial_j \log p(x \mid \xi)|\} < \infty$, $\mathcal{E}_\xi\{|\partial_i \log p(x \mid \xi)|^2\} < \infty$, and $\mathcal{E}_\xi\{|\log p(x \mid \xi)|^2\} < \infty$ for $\xi \in \Xi^*$, where $\mathcal{E}_\xi\{\cdot\}$ denotes the expectation with respect to $p(x \mid \xi)$, and $\partial_i = \partial/\partial\xi^i$ denotes the partial differentiation with respect to the i-th element of ξ. Let $G(\xi)$ be Fisher's information matrix, whose elements are $G_{ij}(\xi) = \mathcal{E}_\xi\{\partial_i \log p(x \mid \xi)\partial_j \log p(x \mid \xi)\} = -\mathcal{E}_\xi\{\partial_i\partial_j \log p(x \mid \zeta)\}$. Assume $G(\zeta)$ is of full rank, and all the elements are of C^1, once continuously differentiable for $\xi \in \Xi^*$. Define $L(\xi_1, \xi_2) = \mathcal{E}_{\xi_1}\{\log p(x \mid \xi_2)\}$ for $\xi_1, \xi_2 \in \Xi$. Then the K-L discrepancy is $D(\xi_1, \xi_2) = L(\xi_1, \xi_1) - L(\xi_1, \xi_2)$, and Jeffery's discrepancy is $J(\xi_1, \xi_2) = D(\xi_1, \xi_2) + D(\xi_2, \xi_1)$. Assume $D(\xi_1, \xi_2)$ is of C^2 for $\xi_1, \xi_2 \in \Xi^*$.

LEMMA 2.1. *Let $\xi_\alpha^{(n)} \in \Xi$, $\alpha = 1, 2$, be sequences converging to ξ^* such that $\lim_{n\to\infty} \sqrt{n}(\xi_\alpha^{(n)} - \xi^*) = \xi_\alpha^\diamond$, $\alpha = 1, 2$, exist. Then we have*

$$\lim_{n\to\infty} nD(\xi_1^{(n)}, \xi_2^{(n)}) = (\xi_1^\diamond - \xi_2^\diamond)' G^* (\xi_1^\diamond - \xi_2^\diamond)/2, \tag{2.1}$$

where $G^* = G(\xi^*)$. This immediately implies $\lim_{n\to\infty} nD(\xi_1^{(n)}, \xi_2^{(n)}) = \lim_{n\to\infty} nD(\xi_2^{(n)}, \xi_1^{(n)}) = \lim_{n\to\infty} nJ(\xi_1^{(n)}, \xi_2^{(n)})/2$.

Let α index models, and $\mathcal{M}$ be the set of α's for the candidate models. Consider a parametric family of densities $p_\alpha(\cdot) = \{p_\alpha(x \mid \theta_\alpha) \mid \theta_\alpha \in \Theta_\alpha\}$, $\Theta_\alpha \subset \mathcal{R}^{m_\alpha}$ for each $\alpha \in \mathcal{M}$. We assume that $p_\alpha(\cdot)$ is a subset of $p(\cdot)$, and that $p(x \mid \xi^*)$ is interior to $p_\alpha(\cdot)$. Thus, using a function $\xi_\alpha : \Theta_\alpha \to \Xi$, we can write $p_\alpha(x \mid \theta_\alpha) = p(x \mid \xi_\alpha(\theta_\alpha))$ for $\theta_\alpha \in \Theta_\alpha$, and $\xi^* = \xi_\alpha(\theta_\alpha^*)$ for some $\theta_\alpha^* \in \Theta_\alpha$, where θ_α^* is interior to Θ_α.

In a neighborhood of θ_α^*, we assume $\xi_\alpha(\theta_\alpha)$ is of C^1 and $m \times m_\alpha$ matrix $B_{\alpha j}^i(\theta_\alpha) = \partial\xi_\alpha^i/\partial\theta_\alpha^j$ is of rank m_α. Let $\theta_\alpha^{(n)} = \operatorname{argsup}_{\theta_\alpha \in \Theta_\alpha} L(\xi^{(n)}, \xi_\alpha(\theta_\alpha))$, and assume $\lim_{n\to\infty} \sqrt{n}(\theta_\alpha^{(n)} - \theta_\alpha^*) = \theta_\alpha^\diamond$ exists. In $p_\alpha(\cdot)$, $p_\alpha(x \mid \theta_\alpha^{(n)})$ is regarded as the "closest point" to the true density in the sense that it minimizes the K-L discrepancy. Write $B_\alpha^* = B_\alpha(\theta_\alpha^*)$, $\xi_\alpha^{(n)} = \xi_\alpha(\theta_\alpha^{(n)})$, and $\xi_\alpha^\diamond = \lim_{n\to\infty} \sqrt{n}(\xi_\alpha^{(n)} - \xi^*)$ for brevity.

LEMMA 2.2. *The asymptotic limit $\xi_\alpha^\diamond$ of the point closest to the true density in model-α satisfies*

$$\xi_\alpha^\diamond = B_\alpha^* \theta_\alpha^\diamond, \qquad B_\alpha^{*\prime} G^* (\xi_\alpha^\diamond - \xi^\diamond) = 0, \tag{2.2}$$

and thus we have

$$\theta_\alpha^\diamond = (B_\alpha^{*\prime} G^* B_\alpha^*)^{-1} B_\alpha^{*\prime} G^* \xi^\diamond. \tag{2.3}$$

Note that $\xi_\alpha^\diamond$ is the projection of $\xi^\diamond$ onto $\operatorname{Im} B_\alpha^*$, the linear space spanned by the column vectors of B_α^*, using G^* as the metric.

Let $L^{(n)}(\xi) = n^{-1}\sum_{t=1}^n \log p(x_t^{(n)} \mid \xi)$ denote the log-likelihood function (divided by n), and $\hat{\xi}^{(n)} = \operatorname{argsup}_{\xi\in\Xi} L^{(n)}(\xi)$ denote the MLE of $\xi^{(n)}$ for the model $p(\cdot)$. We assume the MLE is asymptotically bounded in probability, that is, $\operatorname{plim}_{n\to\infty}\sqrt{n}(\hat{\xi}^{(n)} - \xi^*) = \hat{\xi}^\diamond = O_p(1)$, where plim denotes the convergence in probability.

LEMMA 2.3. *The asymptotic distribution of the MLE is normal*;

$$\hat{\xi}^\diamond \sim N(\xi^\diamond, G^{*-1}). \tag{2.4}$$

Let $\hat{\theta}_\alpha^{(n)} = \operatorname{argsup}_{\theta_\alpha\in\Theta_\alpha} L^{(n)}(\xi_\alpha(\theta_\alpha))$ denote the MLE for $p_\alpha(\cdot)$. Assume $\operatorname{plim}_{n\to\infty}\sqrt{n}(\hat{\theta}_\alpha^{(n)} - \theta_\alpha^*) = \hat{\theta}_\alpha^\diamond = O_p(1)$. Write $\hat{\xi}_\alpha^{(n)} = \xi_\alpha(\hat{\theta}_\alpha^{(n)})$ and $\hat{\xi}_\alpha^\diamond = \operatorname{plim}_{n\to\infty}\sqrt{n}(\hat{\xi}_\alpha^{(n)} - \xi_\alpha^*)$ for brevity. We will see $\hat{\xi}_\alpha^\diamond$ is the projection of $\hat{\xi}^\diamond$ onto the model, exactly the same manner as in Lemma 2.2.

LEMMA 2.4. *The asymptotic limit $\hat{\xi}_\alpha^\diamond$ of the MLE for model-α satisfies*

$$\hat{\xi}_\alpha^\diamond = B_\alpha^* \hat{\theta}_\alpha^\diamond, \qquad B_\alpha^{*\prime} G^* (\hat{\xi}_\alpha^\diamond - \hat{\xi}^\diamond) = 0, \tag{2.5}$$

and thus we have

$$\hat{\theta}_\alpha^\diamond = (B_\alpha^{*\prime} G^* B_\alpha^*)^{-1} B_\alpha^{*\prime} G^* \hat{\xi}^\diamond. \tag{2.6}$$

Consider the square root decomposition $G^* = (G^{*1/2})' G^{*1/2}$. Let $\xi^\circ = G^{*1/2}\xi^\diamond$, $\hat{\xi}^\circ = G^{*1/2}\hat{\xi}^\diamond$, $\xi_\alpha^\circ = G^{*1/2}\xi_\alpha^\diamond$, $\hat{\xi}_\alpha^\circ = G^{*1/2}\hat{\xi}_\alpha^\diamond$, and $B_\alpha^\circ = G^{*1/2}B_\alpha^*$. These variable transformations will lead to the orthogonalization of the metric.

PROPOSITION 2.1. *Define $P_\alpha^\circ = B_\alpha^\circ (B_\alpha^{\circ\prime} B_\alpha^\circ)^{-1} B_\alpha^{\circ\prime}$, the projection operator of $\operatorname{Im} B_\alpha^\circ$. Letting $\xi_\alpha^\circ = G^{*1/2}\xi_\alpha^\diamond$ in Lemma 2.1, (2.1) will be*

$$\lim_{n\to\infty} nD(\xi_1^{(n)}, \xi_2^{(n)}) = \|\xi_1^\circ - \xi_2^\circ\|^2/2. \tag{2.7}$$

Similarly, Lemma 2.2 and Lemma 2.4 will be rewritten as

$$\xi_\alpha^\circ = P_\alpha^\circ \xi^\circ, \qquad \hat{\xi}_\alpha^\circ = P_\alpha^\circ \hat{\xi}^\circ. \tag{2.8}$$

Also, Lemma 2.3 will be

$$\hat{\xi}^\circ \sim N(\xi^\circ, I_m), \tag{2.9}$$

where I_m denotes the identity matrix of size m.

2.2 *Statistics for model selection*

Let $\hat{L}_\alpha^{(n)} = L^{(n)}(\hat{\xi}_\alpha^{(n)})$ denote the maximum log-likelihood for $p_\alpha(\cdot)$. Then, AIC (divided by $2n$) for $p_\alpha(\cdot)$ is

$$C_\alpha^{(n)} = -\hat{L}_\alpha^{(n)} + m_\alpha/n. \tag{2.10}$$

Let $r_\alpha^{(n)} = \mathcal{E}\{D(\xi^{(n)}, \hat{\xi}_\alpha^{(n)})\}$ denote the expected discrepancy, or equivalently the *expected prediction error*. The best model with respect to the population is defined to be the minimizer of $r_\alpha^{(n)}$. Since $C_\alpha^{(n)}$ is regarded as an estimate of $r_\alpha^{(n)}$, the minimizer of $C_\alpha^{(n)}$ is regarded as an estimate of the best model. Note that only the differences of $\hat{L}_\alpha^{(n)}$, or $C_\alpha^{(n)}$, between models are needed for model selection; the asymptotic distributions of them are given here.

LEMMA 2.5. *For $\alpha \in \mathcal{M}$,*

$$\operatorname*{plim}_{n\to\infty} n(\hat{L}_\alpha^{(n)} - L^{(n)}(\xi^*)) = \|\hat{\xi}_\alpha^\circ\|^2/2. \tag{2.11}$$

Thus, for $\alpha, \beta \in \mathcal{M}$, we have

$$\operatorname*{plim}_{n\to\infty} n(\hat{L}_\alpha^{(n)} \quad \hat{L}_\beta^{(n)}) = \hat{\xi}^{\circ\prime}(P_\alpha^\circ - P_\beta^\circ)\hat{\xi}^\circ/2, \qquad \textit{and} \tag{2.12}$$

$$\operatorname*{plim}_{n\to\infty} n(C_\alpha^{(n)} - C_\beta^{(n)}) = -\hat{\xi}^{\circ\prime}(P_\alpha^\circ - P_\beta^\circ)\hat{\xi}^\circ/2 + \operatorname{tr}(P_\alpha^\circ - P_\beta^\circ). \tag{2.13}$$

Let a $m \times 1$ random vector X be distributed as $N(b, I_m)$. Then, for any symmetric $m \times m$ matrix A, we have $\mathcal{E}\{X'AX\} = b'Ab + \operatorname{tr} A$ and $\operatorname{var}\{X'AX\} = 4b'A^2b + 2\operatorname{tr} A^2$. Applying these facts to (2.13), we obtain

$$L\mathcal{E}\{nC_\alpha^{(n)} - nC_\beta^{(n)}\} = -\xi^{\circ\prime}(P_\alpha^\circ - P_\beta^\circ)\xi^\circ/2 + \operatorname{tr}(P_\alpha^\circ - P_\beta^\circ)/2, \qquad \text{and} \tag{2.14}$$

$$L\operatorname{var}\{nC_\alpha^{(n)} - nC_\beta^{(n)}\} = \xi^{\circ\prime}(P_\alpha^\circ - P_\beta^\circ)^2\xi^\circ + \operatorname{tr}(P_\alpha^\circ - P_\beta^\circ)^2/2, \tag{2.15}$$

where $L\mathcal{E}\{\cdot\}$ and $L\operatorname{var}\{\cdot\}$ denote the asymptotic expectation and variance, respectively. The following lemma tells us that $C_\alpha^{(n)}$ is asymptotically unbiased for the estimation of $r_\alpha^{(n)}$, ignoring a term common to all the models.

LEMMA 2.6. *For $\alpha \in \mathcal{M}$,*

$$L\mathcal{E}\{nD(\xi^{(n)}, \hat{\xi}_\alpha^{(n)})\} = \|\xi^\circ - \xi_\alpha^\circ\|^2/2 + m_\alpha/2. \tag{2.16}$$

Thus, for $\alpha, \beta \in \mathcal{M}$, it is easy to verify that $L\mathcal{E}\{nD(\xi^{(n)}, \hat{\xi}_\alpha^{(n)}) - nD(\xi^{(n)}, \hat{\xi}_\beta^{(n)})\}$ equals (2.14). Only $C_\alpha^{(n)}$, or its equivalent up to the $O(1/n)$ term in (2.10), has this asymptotic unbiasedness. Note that $\lim_{n\to\infty} nr_\alpha^{(n)}$ is not necessarily $L\mathcal{E}\{nD(\xi^{(n)}, \hat{\xi}_\alpha^{(n)})\}$, but we assume these are equivalent throughout.

As has been seen, $C_\alpha^{(n)}$ is unbiased, but it still has the variance (2.15). Here we consider an estimator of this variance and its asymptotic distribution. We will investigate

$$V_{\alpha\beta}^{(n)}/n + \kappa v_{\alpha\beta}^{(n)}/n^2 \tag{2.17}$$

as an estimator of $\mathrm{var}\{C_\alpha^{(n)} - C_\beta^{(n)}\}$, where $V_{\alpha\beta}^{(n)}$ and $v_{\alpha\beta}^{(n)}$ are described below, and $\kappa \geq 0$ is a constant, set to zero in Linhart's test statistic, or set to unity in our modified test statistic. The first term in (2.17) is

$$V_{\alpha\beta}^{(n)} = n^{-1}\sum_{t=1}^{n}(\log p(x_t^{(n)} \mid \hat{\xi}_\alpha^{(n)}) - \log p(x_t^{(n)} \mid \hat{\xi}_\beta^{(n)}))^2 - (\hat{L}_\alpha^{(n)} - \hat{L}_\beta^{(n)})^2, \tag{2.18}$$

and the second term is

$$v_{\alpha\beta}^{(n)} = (m_\alpha + m_\beta)/2 - \mathrm{tr}(G_{\alpha\beta}^{(n)} G_{\beta\beta}^{(n)-1} G_{\beta\alpha}^{(n)} G_{\alpha\alpha}^{(n)-1}), \tag{2.19}$$

where

$$(G_{\alpha\beta}^{(n)})_{ij} = n^{-1}\sum_{t=1}^{n}\left\{\frac{\partial \log p(x_t^{(n)} \mid \xi_\alpha(\hat{\theta}_\alpha^{(n)}))}{\partial \theta_\alpha^i}\frac{\partial \log p(x_t^{(n)} \mid \xi_\beta(\hat{\theta}_\beta^{(n)}))}{\partial \theta_\beta^j}\right\}. \tag{2.20}$$

LEMMA 2.7. *The two terms in* (2.17) *are asymptotically*

$$\operatorname*{plim}_{n\to\infty} nV_{\alpha\beta}^{(n)} = \hat{\xi}^{\circ\prime}(P_\alpha^\circ - P_\beta^\circ)^2\hat{\xi}^\circ, \quad \textit{and} \tag{2.21}$$

$$\operatorname*{plim}_{n\to\infty} v_{\alpha\beta}^{(n)} = \mathrm{tr}(P_\alpha^\circ \quad P_\beta^\circ)^2/2. \tag{2.22}$$

Note that the sum of (2.21) and (2.22), if $\hat{\xi}^\circ$ is replaced with ξ°, gives (2.15). This implies that (2.17) is a reasonable estimate of (2.15) when $\kappa = 1$.

Using the results obtained above, we will investigate a statistic

$$T_{\alpha\beta}^{(n)} = \frac{C_\alpha^{(n)} - C_\beta^{(n)}}{\sqrt{V_\alpha^{(n)}/n \ + \ \kappa v_{\alpha\beta}^{(n)}/n^2}} \tag{2.23}$$

for testing of $H_{\alpha\beta}^{(n)} : r_\alpha^{(n)} \leq r_\beta^{(n)}$ against $\bar{H}_{\alpha\beta}^{(n)} : r_\alpha^{(n)} > r_\beta^{(n)}$, or $H_{\beta\alpha}^{(n)}$ against $\bar{H}_{\beta\alpha}^{(n)}$. The model selection test is derived under the assumption that (2.23) is normally distributed with unit variance. To see the deviation from the normality, we will introduce a canonical form of the test statistic. Let $T_{\alpha\beta}^\circ = \mathrm{plim}_{n\to\infty} T_{\alpha\beta}^{(n)}$, which can be written explicitly using $\hat{\xi}^\circ$ and $P_\alpha^\circ - P_\beta^\circ$. Applying the following lemma to it, we will obtain the canonical form.

LEMMA 2.8. *Assume $m_\alpha \geq m_\beta$ without losing generality. Then, the eigen values of $P_\alpha^\circ - P_\beta^\circ$ are $\pm\lambda_i$, $0 \leq \lambda_i \leq 1$, $i = 1, \ldots, m_\beta$, and unity of $m_\alpha - m_\beta$ multiplicity.*

Note that $\varphi_i = \sin^{-1}\lambda_i$ represents the angle between $\operatorname{Im} P_\alpha^\circ$ and $\operatorname{Im} P_\beta^\circ$ confined to the corresponding eigen space. Note also that $\gamma_i = \cos\varphi_i$, $i = 1, \ldots, m_\beta$, are the canonical correlation coefficients of Hotelling (1936) between $\hat{\theta}_\alpha^\circ$ and $\hat{\theta}_\beta^\circ$.

Let $r = m_\beta$ and $l = m_\alpha - m_\beta \geq 0$. Let $v_{\pm i}$ be the unit eigen vector corresponding to $\pm\lambda_i$ for $i = 1, \ldots, r$, and v_i, $i = r+1, \ldots, r+l$, be those of unity. Let $\mu_{\pm i} = v_{\pm i}'\xi^\circ$, $i = 1, \ldots, r$, and $\mu_i = v_i'\xi^\circ$, $i = r+1, \ldots, r+l$. Define $\mu' = (\mu_{-r}, \ldots, \mu_{-1}; \mu_{+1}, \ldots, \mu_{+r}; \mu_{r+1}, \ldots, \mu_{r+l})$ and $\Lambda = \operatorname{diag}(-\lambda_r, \ldots, -\lambda_1; \lambda_1, \ldots, \lambda_r; 1, \ldots, 1)$.

PROPOSITION 2.2. *The canonical form of* $T_{\alpha\beta}^\circ$ *is*

$$T_{\alpha\beta}^\circ = \frac{(\mu \mid z)'\Lambda(\mu \mid z)/2 \mid \operatorname{tr}\Lambda}{\sqrt{(\mu+z)'\Lambda^2(\mu+z) + \kappa \operatorname{tr}\Lambda^2/2}}, \tag{2.24}$$

where $z \sim N(0, I_{2r+l})$. *Using* μ *and* Λ, (2.14) *will be rewritten as*

$$\lim n(r_\alpha^{(n)} - r_\beta^{(n)}) = -\mu'\Lambda\mu/2 + \operatorname{tr}\Lambda/2, \tag{2.25}$$

which is the expectation of the numerator of (2.24). *Also,*

$$\lim_{n\to\infty} n(D(\xi^{(n)}, \xi_\alpha^{(n)}) - D(\xi^{(n)}, \xi_\beta^{(n)})) = -\mu'\Lambda\mu/2, \qquad \textit{and} \tag{2.26}$$

$$\lim_{n\to\infty} nJ(\xi_\alpha^{(n)}, \xi_\beta^{(n)}) - \mu'\Lambda^2\mu \tag{2.27}$$

will be easily verified.

If $\lambda_i = 0$ for some i, then this element can be removed from the canonical form. Thus, we can assume $0 < \lambda_i \leq 1$ without losing generality. Let m_c denote the dimension of $\operatorname{Im} B_\alpha^* \cap \operatorname{Im} B_\beta^*$. Then, $\lambda_i = 0$ for m_c elements, and $r = m_\beta - m_c$. Note that $r = 0$ for the nested case $p_\beta(\cdot) \subset p_\alpha(\cdot)$, since $m_c = m_\beta$.

3. Assessing the error probability

3.1 *Theoretical results*

Let $c = \Phi^{-1}(1 - \text{level}) \geq 0$ denote a critical constant for the one-sided normal test, where $\Phi(x)$ is the standard normal distribution function, and level denotes the nominal significance level of the test. Linhart (1988) considered a test such that $H_{\alpha\beta}^{(n)}$ is rejected if $T_{\alpha\beta}^{(n)} > c$, and similarly $H_{\beta\alpha}^{(n)}$ is rejected if $T_{\alpha\beta}^{(n)} < -c$, where $\kappa = 0$ in (2.23). Our modification of these tests is to add the second term in (2.17), or to set $\kappa = 1$. In the following, we will investigate their probabilities of false rejection, which are supposed to be less than the level. Note that $H_{\alpha\beta}^{(n)}$ and $H_{\beta\alpha}^{(n)}$ are tested separately above. When these are tested simultaneously, the two-sided constant $c = \Phi^{-1}(1 - \text{level}/2)$ is used.

Let $F^{(n)}(t) = \Pr\{T_{\alpha\beta}^{(n)} \leq t\}$. The actual error probability of testing of $H_{\alpha\beta}^{(n)}$ is $e_{\alpha\beta}^{(n)} = 1 - F^{(n)}(c)$ under $H_{\alpha\beta}^{(n)}$, and that of $H_{\beta\alpha}^{(n)}$ is $e_{\beta\alpha}^{(n)} = F^{(n)}(-c)$. Taking the

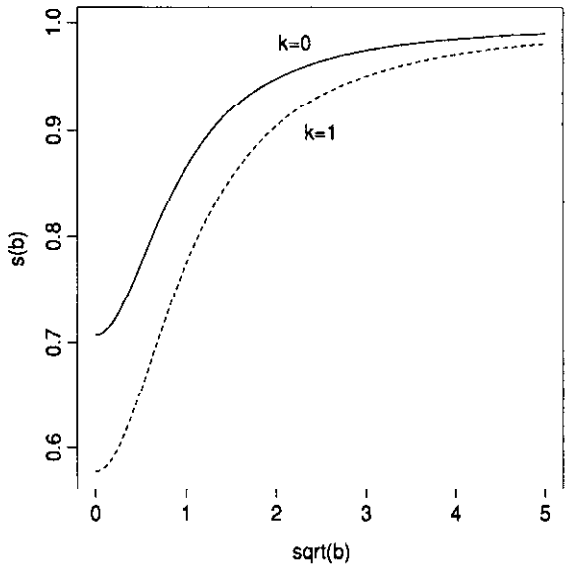


Fig. 1. Curves of $s(b)$ drawn for $\kappa = 0$ and 1.

asymptotic limit of Section 2, we have $\lim_{n\to\infty} F^{(n)}(t) = F^{\circ}(t)$, where $F^{\circ}(t) = \Pr\{T^{\circ}_{\alpha\beta} \le t\}$. Let $H^{\circ}_{\alpha\beta}$ denote the hypothesis $\lim_{n\to\infty} n(r^{(n)}_{\alpha} - r^{(n)}_{\beta}) \le 0$, and define $H^{\circ}_{\beta\alpha}$ similarly. Then, the corresponding asymptotic error probabilities are $e^{\circ}_{\alpha\beta} = 1 - F^{\circ}(c)$ and $e^{\circ}_{\beta\alpha} = F^{\circ}(-c)$, respectively. In the theoretical results below, we will take the limit of $F^{\circ}(t)$, such as $\|\mu\| \to \infty$ or $\operatorname{tr}\Lambda^2 \to \infty$, to obtain rough estimates of the error probabilities.

LEMMA 3.1. *Consider the limit* $\|\mu\| \to \infty$ *such that* $\lim_{\|\mu\|\to\infty} -\mu'\Lambda\mu/\sqrt{4\mu'\Lambda^2\mu} = a$ *exists. Then,* $\lim_{\|\mu\|\to\infty} F^{\circ}(t) = \Phi(t-a)$.

Note that $H^{\circ}_{\alpha\beta}$ implies $a \le 0$ and $H^{\circ}_{\beta\alpha}$ implies $a \ge 0$ when $\|\mu\| \to \infty$. Then we find $e^{\circ}_{\alpha\beta} = 1 - F^{\circ}(c) \le 1 - \Phi(c) =$ level under $H^{\circ}_{\alpha\beta}$ and $e^{\circ}_{\beta\alpha} = F^{\circ}(-c) \le \Phi(-c) =$ level under $H^{\circ}_{\beta\alpha}$, where these inequalities become equalities if $a = 0$. This corresponds to the result of Linhart (1988), in which $F^{(n)}(t) \to \Phi(t)$ is derived under $r^{(n)}_{\alpha} = r^{(n)}_{\beta}$ and $J(\xi^{(n)}_{\alpha}, \xi^{(n)}_{\beta})$ fixed. Note that the fixed $J(\xi^{(n)}_{\alpha}, \xi^{(n)}_{\beta})$ leads to $\|\mu\| \to \infty$, because of (2.27).

LEMMA 3.2. *Consider the limit* $l + 2r \to \infty$ *such that* $\operatorname{tr}\Lambda^2 \to \infty$. *Let* $a^{(\Lambda)} = (-\mu'\Lambda\mu + \operatorname{tr}\Lambda)(4\mu'\Lambda^2\mu + 2\operatorname{tr}\Lambda^2)^{-1/2}$, $b^{(\Lambda)} = \mu'\Lambda^2\mu/\operatorname{tr}\Lambda^2$. *Assume both of* $\lim_{\operatorname{tr}\Lambda^2\to\infty} a^{(\Lambda)} = a$ *and* $\lim_{\operatorname{tr}\Lambda^2\to\infty} b^{(\Lambda)} = b$ *exist. Then,* $\lim_{\operatorname{tr}\Lambda^2\to\infty} F^{\circ}(t) = \Phi(t/s - a)$, *where* $s = s(b) = (1 + (1+\kappa)/(1+2b))^{-1/2}$.

Suppose $\operatorname{tr}\Lambda^2 \to \infty$. Then, the error probabilities are again controlled by the level, since $0 < s(b) \le 1$ for $b \ge 0$ as shown in Fig. 1. For small $b \ge 0$, $s(b)$ is much smaller than unity, which implies the test will be too conservative.

As $b \to \infty$, $s(b) \to 1$, which is consistent with the result of Lemma 3.1. Note that $b^{(\Lambda)} \operatorname{tr} \Lambda^2 = \mu' \Lambda^2 \mu \approx nJ(\xi_\alpha^{(n)}, \xi_\beta^{(n)})$ is approximately regarded as a squared "distance" between the projections of the true density onto the two models; the standard normal approximation is valid if the distance is large, while it will be conservative if the distance is small and $\operatorname{tr} \Lambda^2$ is sufficiently large.

3.2 *Empirical results*

It follows from Lemma 3.1 and Lemma 3.2 that the error probabilities are less than the level regardless of the value of κ, if either $\|\mu\|$ or $\operatorname{tr} \Lambda^2$ is sufficiently large. Here we examine some numerical simulations in which $\|\mu\|$ and $\operatorname{tr} \Lambda^2$ are small enough to observe violations of the error control. The value of μ will be chosen so that the null hypothesis $H^\circ_{\alpha\beta} \bigcap H^\circ_{\beta\alpha}$ holds.

We will confine the discussion to a particular case: All the diagonal elements of Λ are unity or zero. The linear spaces of the two models are mutually orthogonal, since nonzero $\varphi_i = \sin^{-1} \lambda_i$ is $\pi/2$. The canonical form reduces to

$$T^\circ_{\alpha\beta} = \frac{-(W_{(+)} - W_{(-)})/2 + l}{\sqrt{W_{(+)} + W_{(-)} + \kappa(r + l/2)}}, \tag{3.1}$$

where $W_{(+)} \sim \chi^2(l + r, \delta_+)$ and $W_{(-)} \sim \chi^2(r, \delta_-)$, which are independent χ^2 random variables with the noncentrality parameters $\delta^2_{(+)} = \sum_{i=1}^r \mu^2_{+i} + \sum_{i=r+1}^{r+l} \mu_i^2$ and $\delta^2_{(-)} = \sum_{i=1}^r \mu^2_{-i}$, respectively. The null hypothesis implies $\delta^2_{(+)} - \delta^2_{(-)} = l$, since (2.25) reduces to $-(\delta^2_{(+)} - \delta^2_{(-)})/2 + l/2$. Note that $r = 0$ and $\delta^2_{(-)} = 0$ in the nested case.

Table 1. Parameter sets for the simulations.

	Canonical form				Regression**				
Sim. #	l	r	$\delta_{(-)}$	$\delta_{(+)}$	$\mathcal{K}_\alpha$	$\mathcal{K}_\beta$	$\eta_1^\diamond$	$\eta_2^\diamond$	$\eta_3^\diamond$
1	1	0	0	1	$\{0,1\}$	$\{0\}$	1	1	0
2	2	0	0	$\sqrt{2}$	$\{0,1,2\}$	$\{0\}$	1	1	0
3	3	0	0	$\sqrt{3}$	$\{0,1,2,3\}$	$\{0\}$	1	1	1
4	0	1	0–4*	$\delta_{(-)}$	$\{0,1\}$	$\{0,2\}$	$\delta_{(-)}$	$\delta_{(-)}$	0
5	1	1	0–4*	$\sqrt{1+\delta^2_{(-)}}$	$\{0,2,3\}$	$\{0,1\}$	$\delta_{(-)}$	$\sqrt{1+\delta^2_{(-)}}$	0

(*) $\delta_{(-)} = 0, 1, 2, 3, 4$. (**) See Appendix B.

Five simulations are carried out with the parameter sets given in Table 1. Figure 2 shows the results of the first simulation, in which the nested case with one degree of freedom is treated. Curves of $e^\circ_{\alpha\beta} = 1 - F^\circ(c)$ estimated from 10,000 samples of $T^\circ_{\alpha\beta}$ are drawn with solid lines for $\kappa = 0$ and 1. Those of $e^{(n)}_{\alpha\beta} = 1 - F^{(n)}(c)$ are drawn with dashed lines, which are estimated from 10,000

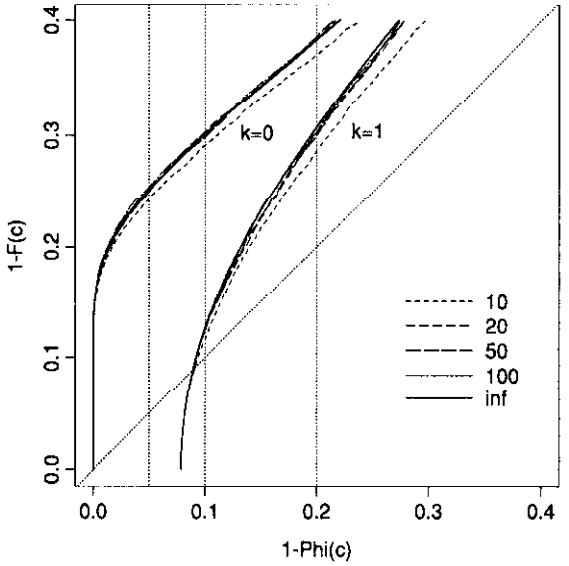


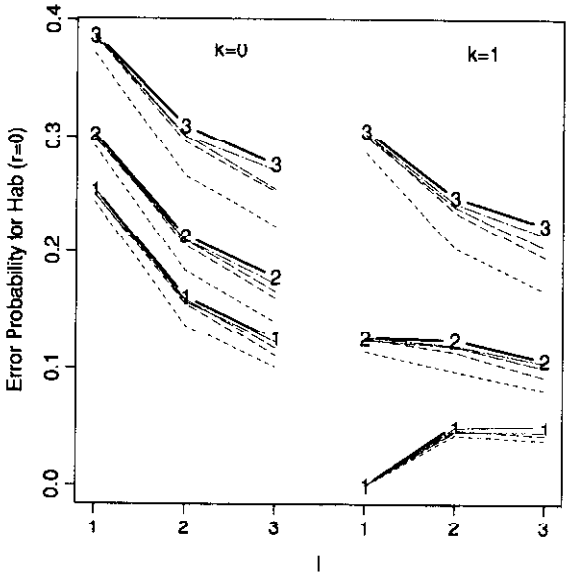
Fig. 2. Error probability curves plotted against level $= 1 - \Phi(c)$. Drawn for $\kappa = 0$ and 1. The solid lines are $e^{\circ}_{\alpha\beta} = 1 - F^{\circ}(c)$ of $(l, r) = (1, 0)$. The dashed lines are $e^{(n)}_{\alpha\beta} = 1 - F^{(n)}(c)$ for $n = 10, 20, 50, 100$, which are almost indistinguishable from $e^{\circ}_{\alpha\beta}$ for $n \geq 20$.

artificial regression data sets; see Appendix B for details. We observe that $e^{\circ}_{\alpha\beta}$ gives a good approximation of $e^{(n)}_{\alpha\beta}$, and so we may confine our attention to $e^{\circ}_{\alpha\beta}$ rather than $e^{(n)}_{\alpha\beta}$ of finite n. The error probability $e^{\circ}_{\alpha\beta}$ for $\kappa = 0$ is larger than the nominal level by approximately 0.2. For example, $e^{\circ}_{\alpha\beta} \approx 0.25$ when the level is 0.05. By letting $\kappa = 1$, the situation is remedied to a considerable extent; $e^{\circ}_{\alpha\beta} \approx 0.1$ at level 0.1, and it will be conservative for smaller levels.

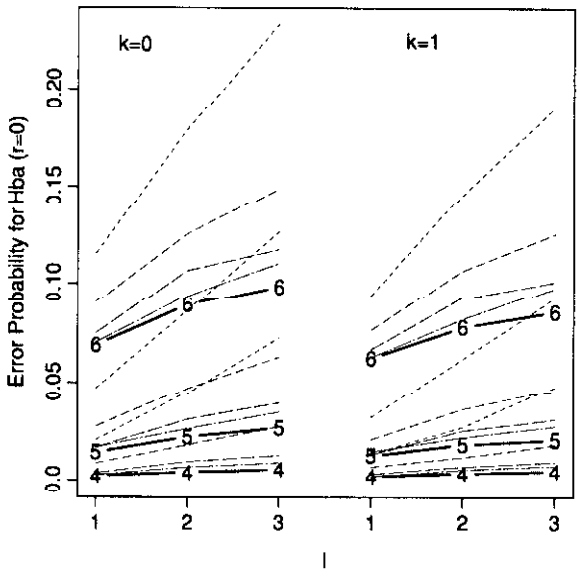
The results of the second and third simulations, together with those of the first, are summarized in Fig. 3. The nested cases with degrees of freedom from one to three, that is, $(l, r) = (1, 0)$, $(2, 0)$, and $(3, 0)$, are treated. For each choice of l, the error probabilities are plotted for levels 0.05, 0.1, and 0.2. Solid lines denote $e^{\circ}_{\alpha\beta}$ or $e^{\circ}_{\beta\alpha}$, and dashed lines $e^{(n)}_{\alpha\beta}$ or $e^{(n)}_{\beta\alpha}$; see the legend of Fig. 2. Figure 3(a) shows that $e^{\circ}_{\alpha\beta}$ for $\kappa = 0$ is larger than the level in the nested cases, but it decreases as l increases. By letting $\kappa = 1$, $e^{\circ}_{\alpha\beta}$ becomes closer to the level. On the other hand, Fig. 3(b) shows that $e^{\circ}_{\beta\alpha}$ is considerably smaller than the level even if $\kappa = 0$. This means the test is very conservative. Letting $\kappa = 1$ makes the situation even worse.

Figure 4 shows the result of the fourth simulation where $(l, r) = (0, 1)$. This case is symmetrical because $e^{\circ}_{\alpha\beta} = e^{\circ}_{\beta\alpha}$ holds. Error probabilities are calculated for $\delta_{(-)} = 0, 1, 2, 3, 4$; they approach the level as $\delta_{(-)}$ increases, which is suggested by Lemma 3.1. For small $\delta_{(-)}$, however, the error probability is very small, which is similar to the behaviour of $e^{\circ}_{\beta\alpha}$ for nested cases. The difference between $\kappa = 0$ and 1 is small.

In the fifth simulation, a situation in between the nested and the symmetric



(a)



(b)

Fig. 3. Error probability in the nested cases: $(l,r) = (1,0),(2,0),(3,0)$, and $\kappa = 0,1$. Drawn for (a) $H_{\alpha\beta}$, and (b) $H_{\beta\alpha}$. The $e^{\circ}_{\alpha\beta}$ are labeled 1, 2, 3 and the $e^{\circ}_{\beta\alpha}$ 4, 5, 6 for levels 0.05, 0.1, and 0.2 respectively.

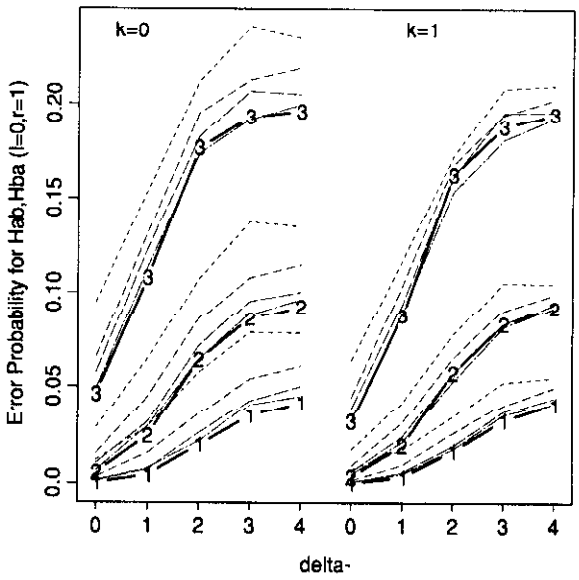


Fig. 4. Error probability in the symmetric case: $(l, r) = (0, 1)$. Plotted for $\delta_{(-)} = 0, 1, 2, 3, 4$.

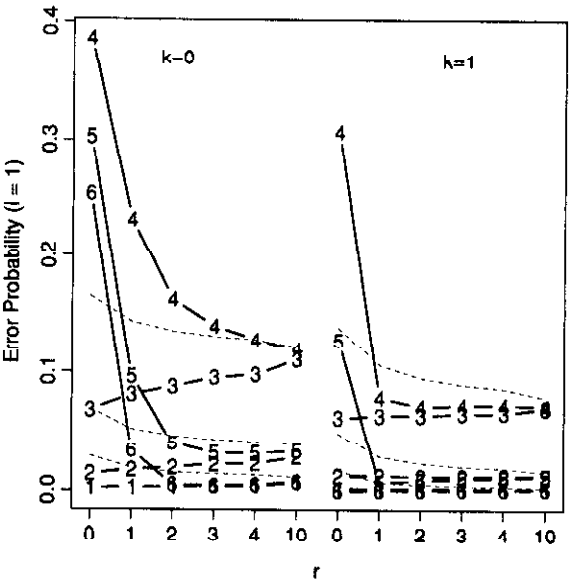


Fig. 5. Error probability at $\delta_{(-)} = 0$. Plotted for $l = 1$, and $r = 0$ to 10. Dashed lines are approximate curves derived from Lemma 3.2.

case is treated. Not surprisingly, the results are also in between those of the nested and symmetric case, and the figures are omitted here. Both $e^{\circ}_{\alpha\beta}$ and $e^{\circ}_{\beta\alpha}$ are very similar to those of the symmetric case on the whole, and they are close to those of the nested case for small $\delta_{(-)}$.

Given (l,r), the largest deviation of the error probability from the nominal level is obtained when $\delta_{(-)} = 0$, especially the case $(l,r) = (1,0)$ which gives the largest $e^{\circ}_{\alpha\beta}$, while $e^{\circ}_{\beta\alpha}$ is much smaller than the level. If either l or r is sufficiently large, $e^{\circ}_{\alpha\beta}$ as well as $e^{\circ}_{\beta\alpha}$ will be smaller than the level, as suggested by Lemma 3.2. This tendency is shown in Fig. 3 and Fig. 5.

4. Concluding remarks

We have examined the error probability of the model selection test under local alternatives. Linhart's test works well if $\|\mu\|$ or $\operatorname{tr}\Lambda^2$ is sufficiently large: The error probability under the null hypothesis agrees well with the nominal level if $\|\mu\|$ is large enough, and it is smaller than the level if $\operatorname{tr}\Lambda^2$ is sufficiently large. However, $e^{\circ}_{\alpha\beta}$ for $\kappa = 0$ can be larger than the level, particularly in the nested case with small $l = m_\alpha - m_\beta > 0$. This violation of the error control can be satisfactorily remedied by letting $\kappa = 1$. On the other hand, $e^{\circ}_{\beta\alpha}$ is considerably smaller than the level for small $\|\mu\|$, even if $\kappa = 0$. When $l = 0$, $e^{\circ}_{\alpha\beta}$ as well as $e^{\circ}_{\beta\alpha}$ is very small.

The large $e^{\circ}_{\alpha\beta}$ may be explained as follows. The test statistic $T^{\circ}_{\alpha\beta}$ for $\kappa = 0$, $l > 0$ can take an extremely large value if $\|\mu + z\|$ is nearly zero, since the denominator of $T^{\circ}_{\alpha\beta}$ is nearly zero while the numerator is positive. This can make the distribution of $T^{\circ}_{\alpha\beta}$ heavy tailed unless $\|\mu\|$ is large enough.

The model selection test shows a general tendency to be conservative; this is partly explained by the bias of the variance estimator. The asymptotic expectation of (2.17) multiplied by n^2 is $\xi^{\circ\prime}(P^{\circ}_{\alpha} - P^{\circ}_{\beta})^2\xi^{\circ} + (2+\kappa)\operatorname{tr}(P^{\circ}_{\alpha} - P^{\circ}_{\beta})^2/2$, which overestimates the true value if $\kappa > -1$. An *ad hoc* modification of the variance estimator, such as $\max(V^{(n)}_{\alpha\beta}/n - v^{(n)}_{\alpha\beta}/n^2, v^{(n)}_{\alpha\beta}/n^2)$, might improve the test, but is not considered in this article.

Although the empirical results are confined to a particular case of Λ, there is an implication for general Λ. In Shimodaira (1995), it is numerically confirmed that the maximum value of $e^{\circ}_{\alpha\beta}$ $(= e^{\circ}_{\beta\alpha})$ for $l = 0$, $\mu = 0$ and a fixed $r > 0$ is attained when all λ_i have the same value. This suggests the error probability for general Λ is still controlled by the level if $l = 0$, $\mu = 0$.

Even if the apparent structure of the pair of models is nonnested, its canonical form may happen to correspond the nested case. Consider the following example. Assume $\mathcal{E}\{X_1X_2\} = \rho$, while $X_i \sim N(0,1)$ and $\mathcal{E}\{X_1X_3\} = \mathcal{E}\{X_2X_3\} = 0$ in the fifth set of regression parameters of Table 1. Then, it is easy to see $\Lambda = \operatorname{diag}(-\sin\varphi, \sin\varphi, 1)$, where $\cos\varphi = \rho$. This example will be the nonnested case with $(l,r) = (1,1)$ if $\rho = 0$. But if $\rho = 1$, it reduces to the nested case with $(l,r) = (1,0)$, since $\sin\varphi \to 0$ as $\rho \to 1$. Note that a degenerate nested case like this example cannot happen if $l = 0$.

As demonstrated, Linhart's model selection test works poorly for the nested case, in which the difference between the two AIC's is asymptotically χ^2 with unknown noncentrality parameter. Although a better test of the noncentrality

can be devised, we considered the modification $\kappa - 1$ to remedy the defect in the model selection test. This is important to the confidence set of models derived in Shimodaira (1996), in which the modified statistics for all pairs of models are simultaneously tested using the normal approximation. The results of the present article suggest that the error probability of the confidence set of models is fairly well controlled by the level. Future work might derive a better model selection test or a confidence set of models as a test on μ, utilizing (estimates of) the canonical correlation coefficients.

Acknowledgements

I appreciate Professor Kaoru Nakano of University Tokyo for the environmental support. I also thank Professor Satoshi Kuriki of the Institute of Statistical Mathematics (ISM), Dr. Avner Bar-Hen visiting researcher of ISM, and Ms. Jinko Graham of University of Washington for valuable comments to improve the manuscript.

Appendix A

Proofs

PROOF OF LEMMA 2.1. Consider $\partial_i L(\xi_1, \xi) \mid_{\xi=\xi_1} = \mathcal{E}_{\xi_1}\{\partial_i \log p(x \mid \xi_1)\} = 0$ for $\xi_1 \in \Xi^*$. Then expand $D(\xi_1, \xi_2)$ with respect to ξ_2 around ξ_1 to find $D(\xi_1, \xi_2) = (\xi_1 - \xi_2)' G(\xi_1)(\xi_1 - \xi_2)/2 + o(\|\xi_1 - \xi_2\|^2)$ for $\xi_1, \xi_2 \in \Xi^*$. Noting $\lim_{n\to\infty} \sqrt{n}(\xi_1^{(n)} - \xi_2^{(n)}) = \xi_1^\diamond - \xi_2^\diamond$ and $\lim_{n\to\infty} G(\xi_1^{(n)}) = G^*$, we obtain (2.1). □

PROOF OF LEMMA 2.2. Consider $\xi_\alpha(\theta_\alpha^{(n)}) = \xi_\alpha(\theta_\alpha^*) + B_\alpha^*(\theta_\alpha^{(n)} - \theta_\alpha^*) + o(\|\theta_\alpha^{(n)} - \theta_\alpha^*\|)$. Then we obtain $\xi_\alpha^\diamond = \lim_{n\to\infty} \sqrt{n} B_\alpha^*(\theta_\alpha^{(n)} - \theta_\alpha^*) = B_\alpha^* \theta_\alpha^\diamond$, which is the first equation in (2.2). By definition, $\theta_\alpha^\diamond = \lim_{n\to\infty} \operatorname{arginf}_{u\in\mathcal{R}^{m_\alpha}} nD(\xi^{(n)}, \xi_\alpha(\theta_\alpha^* + u/\sqrt{n}))$. Because of the smoothness of the K-L discrepancy and $\|\theta_\alpha^\diamond\| < \infty$, the limit and arginf can be exchanged, and then it follows from Lemma 2.1 that $\theta_\alpha^\diamond = \operatorname{arginf}_{u\in\mathcal{R}^{m_\alpha}} (\xi^\diamond - B_\alpha^* u)' G^* (\xi^\diamond - B_\alpha^* u)$. This implies the second equation in (2.2). □

PROOF OF LEMMA 2.3. Note that $\partial_i L^{(n)}(\hat{\xi}^{(n)}) = 0$ for sufficiently large n. Expand it around $\xi^{(n)}$ to obtain $n^{-1/2} \sum_{t=1}^n \partial_i \log p(x_t^{(n)} \mid \xi^{(n)}) + n^{-1} \sum_{t=1}^n \sum_j \partial_i \partial_j \log p(x_t^{(n)} \mid \xi^{(n)}) \sqrt{n}(\hat{\xi}^{(n)} - \xi^{(n)})^j + o_p(1) = 0$. Considering $\operatorname{plim}_{n\to\infty} n^{-1/2} \sum_{t=1}^n \partial_i \log p(x_t^{(n)} \mid \xi^{(n)}) \sim N(0, G^*)$ from the central limit theorem, and $\operatorname{plim}_{n\to\infty} n^{-1} \sum_{t=1}^n -\partial_i \partial_j \log p(x_t^{(n)} \mid \xi^{(n)}) = G^*$ from the law of large numbers, then we have $X + G^*(\hat{\xi}^\diamond - \xi^\diamond) = 0$, where $X \sim N(0, G^*)$. □

PROOF OF LEMMA 2.4. The first equation in (2.5) is easily seen as the same manner as that of (2.2). We will show the second equation in (2.5). Consider $(\partial/\partial\theta_\alpha^i) L^{(n)}(\xi_\alpha(\hat{\theta}_\alpha^{(n)})) = \sum_j B_{\alpha i}^j(\hat{\theta}_\alpha^{(n)}) n^{-1} \sum_{t=1}^n \partial_j \log p(x_t^{(n)} \mid \hat{\xi}_\alpha^{(n)}) = 0$ for sufficiently large n. Expand it with respect to $\hat{\zeta}_\alpha^{(n)}$ around $\hat{\zeta}^{(n)}$ to obtain

$\sum_j B^j_{\alpha i}(\hat{\theta}^{(n)}_\alpha)(\sqrt{n}\partial_j L^{(n)}(\hat{\xi}^{(n)}) + n^{-1}\sum_{t=1}^n \sum_k \partial_j \partial_k \log p(x_t^{(n)} \mid \hat{\xi}^{(n)})\sqrt{n}(\hat{\xi}^{(n)}_\alpha - \hat{\xi}^{(n)})^k + o_p(1)) = 0$. Noting $\partial_j L^{(n)}(\hat{\xi}^{(n)}) = 0$ and taking the limit, then we obtain the desired result. □

PROOF OF LEMMA 2.5. Expand $L^{(n)}(\hat{\xi}^{(n)}_\alpha)$ with respect to $\hat{\xi}^{(n)}_\alpha$ around $\hat{\xi}^{(n)}$ to obtain $L^{(n)}(\hat{\xi}^{(n)}_\alpha) = L^{(n)}(\hat{\xi}^{(n)}) + \sum_i \partial_i L^{(n)}(\hat{\xi}^{(n)})(\hat{\xi}^{(n)}_\alpha - \hat{\xi}^{(n)})^i + (1/2n) \cdot \sum_{t=1}^n \sum_i \sum_j \partial_i \partial_j \log p(x_t^{(n)} \mid \hat{\xi}^{(n)})(\hat{\xi}^{(n)}_\alpha - \hat{\xi}^{(n)})^i(\hat{\xi}^{(n)}_\alpha - \hat{\xi}^{(n)})^j + o_p(n^{-1})$. Noting $\partial_i L^{(n)}(\hat{\xi}^{(n)}) = 0$ and $\operatorname{plim}_{n\to\infty} n^{-1}\sum_{t=1}^n -\partial_i \partial_j \log p(x_t^{(n)} \mid \hat{\xi}^{(n)}) = G^*_{ij}$, we have

$$\operatorname*{plim}_{n\to\infty} n(L^{(n)}(\hat{\xi}^{(n)}) - L^{(n)}(\hat{\xi}^{(n)}_\alpha)) = \|\hat{\xi}^\circ_\alpha - \hat{\xi}^\circ\|^2/2. \tag{A.1}$$

Replacing $\hat{\xi}^{(n)}_\alpha$ above with ξ^*, we also have

$$\operatorname*{plim}_{n\to\infty} n(L^{(n)}(\hat{\xi}^{(n)}) - L^{(n)}(\xi^*)) = \|\hat{\xi}^\circ\|^2/2. \tag{A.2}$$

Noting $\|\hat{\xi}^\circ_\alpha - \hat{\xi}^\circ\|^2 = \|(P^\circ_\alpha \quad I_m)\hat{\xi}^\circ\|^2 = \hat{\xi}^{\circ\prime}(I_m \quad P^\circ_\alpha)\hat{\xi}^\circ$, and taking the difference between (A.1) and (A.2), we obtain (2.11). Noting $P^{\circ 2}_\alpha = P^\circ_\alpha$ and $\operatorname{tr} P^\circ_\alpha = m_\alpha$, we immediately obtain (2.12) and (2.13). □

PROOF OF LEMMA 2.6. Considering (2.7) and $(\xi^\circ - \xi^\circ_\alpha)'(\xi^\circ_\alpha - \hat{\xi}^\circ_\alpha) = (\xi^\circ - \xi^\circ_\alpha)'B^\circ_\alpha(\theta^\circ_\alpha - \hat{\theta}^\circ_\alpha) = 0$, we have $\operatorname{plim}_{n\to\infty} nD(\xi^{(n)}, \hat{\xi}^{(n)}_\alpha) = \|\xi^\circ - \hat{\xi}^\circ_\alpha\|^2/2 = \|\xi^\circ - \xi^\circ_\alpha\|^2/2 + \|\xi^\circ_\alpha - \hat{\xi}^\circ_\alpha\|^2/2$. Noting $\|\hat{\xi}^\circ_\alpha - \xi^\circ_\alpha\|^2 \sim \chi^2(m_\alpha)$, we obtain (2.16). □

PROOF OF LEMMA 2.7. Consider $\sqrt{n}(\log p(x_t^{(n)} \mid \hat{\xi}^{(n)}_\alpha) - \log p(x_t^{(n)} \mid \hat{\xi}^{(n)}_\beta)) = \sum_i \partial_i \log p(x_t^{(n)} \mid \hat{\xi}^{(n)})\sqrt{n}(\hat{\xi}^{(n)}_\alpha - \hat{\xi}^{(n)}_\beta)^i + o_p(1)$ from expansion with respect to $\hat{\xi}^{(n)}_\alpha$ and $\hat{\xi}^{(n)}_\beta$ around $\hat{\xi}^{(n)}$. Noting $n(\hat{L}^{(n)}_\alpha - \hat{L}^{(n)}_\beta)^2 = o_p(1)$ and $\operatorname{plim}_{n\to\infty} n^{-1} \cdot \sum_{t=1}^n \partial_i \log p(x_t^{(n)} \mid \hat{\xi}^{(n)})\partial_j \log p(x_t^{(n)} \mid \hat{\xi}^{(n)}) = G^*_{ij}$, we have $\operatorname{plim}_{n\to\infty} nV^{(n)}_{\alpha\beta} = (\hat{\xi}^\circ_\alpha - \hat{\xi}^\circ_\beta)'G^*(\hat{\xi}^\circ_\alpha - \hat{\xi}^\circ_\beta)$, which implies (2.21). Considering $\operatorname{plim}_{n\to\infty} G^{(n)}_{\alpha\beta} = B^{*\prime}_\alpha G^* B^*_\beta = B^{\circ\prime}_\alpha B^\circ_\beta$, it is easy to verify (2.22). □

PROOF OF LEMMA 2.8. Consider the singular value decomposition $P^\circ_\alpha P^\circ_\beta = U_\alpha \Gamma_{\alpha\beta} U'_\beta$, where $\Gamma_{\alpha\beta} = \operatorname{diag}(\gamma_1, \ldots, \gamma_{m_\beta})$ is $m_\alpha \times m_\beta$ matrix, $U_\alpha = (u_{\alpha\cdot 1}, \ldots, u_{\alpha\cdot m_\alpha})$ is $m \times m_\alpha$ matrix such that $U'_\alpha U_\alpha = I_{m_\alpha}$ and $U_\alpha U'_\alpha = P^\circ_\alpha$. Define $A_i = u_{\alpha\cdot i}u'_{\alpha\cdot i} - u_{\beta\cdot i}u'_{\beta\cdot i}$ for $i = 1, \ldots, m_\beta$ and $A_i = u_{\alpha\cdot i}u'_{\alpha\cdot i}$ for $i = m_\beta + 1, \ldots, m_\alpha$. Then, we have $P^\circ_\alpha - P^\circ_\beta = U_\alpha U'_\alpha - U_\beta U'_\beta = \sum_{i=1}^{m_\alpha} A_i$. Note that $\operatorname{Im} A_i$'s are orthogonal to each other, because $u'_{\alpha\cdot i}u_{\beta\cdot j} = u'_{\alpha\cdot i}P^\circ_\alpha P^\circ_\beta u_{\beta\cdot j} = u'_{\alpha\cdot i}U_\alpha \Gamma_{\alpha\beta}U'_\beta u_{\beta\cdot j} = (\Gamma_{\alpha\beta})_{ij}$. Thus, the eigen values and vectors of $P^\circ_\alpha - P^\circ_\beta$ are those of A_i's put together. It is easy to verify that the eigen values of A_i are $\pm\sin\varphi_i$ for $i = 1, \ldots, m_\beta$, where $\varphi_i = \cos^{-1}\gamma_i$. Note that we can have $0 \le \varphi_i \le \pi/2$, because $u_{\alpha\cdot i}$ can be replaced with $u_{\alpha\cdot i}$ if necessary. □

PROOF OF LEMMA 3.1. Divide both of the numerator and the denominator of (2.24) by $\sqrt{\mu'\Lambda^2\mu}$, and take the limit of the fraction. Then we have

$\operatorname{plim}_{\|\mu\|\to\infty} T^{\circ}_{\alpha\beta} = \operatorname{plim}_{\|\mu\|\to\infty}(a - \mu'\Lambda z/\sqrt{\mu'\Lambda^2\mu})(1 + 2\mu'\Lambda^2 z/\mu'\Lambda^2\mu)^{-1/2}$, because $z'\Lambda z$, $\operatorname{tr}\Lambda$, $z'\Lambda^2 z$, and $\operatorname{tr}\Lambda^2$ are bounded in probability. Since $\mu'\Lambda^4\mu/(\mu'\Lambda^2\mu)^2 \le 1/\mu'\Lambda^2\mu \to 0$, we have $\mu'\Lambda^2 z/\mu'\Lambda^2\mu \to 0$, and thus find $T^{\circ}_{\alpha\beta}$ converges to $N(a,1)$ in distribution. □

PROOF OF LEMMA 3.2. Let $U_k = \operatorname{st}((\mu+z)'\Lambda^k(\mu+z))$, where $\operatorname{st}(X) = (X - \mathcal{E}\{X\})/\sqrt{\operatorname{var}\{X\}}$, and let $c^{(\Lambda)} = (4\mu'\Lambda^4\mu + 2\operatorname{tr}\Lambda^4)^{1/2}/(\mu'\Lambda^2\mu + (1+\kappa/2)\operatorname{tr}\Lambda^2)$. Then, it is easy to verify $T^{\circ}_{\alpha\beta} = s(b^{(\Lambda)})(-U_1 + a^{(\Lambda)})(1 + c^{(\Lambda)}U_2)^{-1/2}$. Noting $\lambda_i^4(1+2\mu_i^2) \le 2\lambda_i^2(1+\mu_i^2)$, we obtain $(c^{(\Lambda)})^2 \le 4/\sum\lambda_i^2(1+\mu_i^2) \le 4/\operatorname{tr}\Lambda^2 \to 0$. Considering λ_i's are bounded, it follows from the central limit theorem that U_k converges to $N(0,1)$ in distribution, which completes the proof. □

Appendix B

Regression data for simulations

Consider the regression that $Y = \sum_{i\in\mathcal{K}}\eta_i X_i + Z$, where Y is the response, X_i's are the predictors, and $Z \sim N(0,\sigma^2)$ is the sampling error. Let $\mathcal{K} = \{0,1,2,3\}$, $X_0 = 1$, and $\xi = (\sigma^2,\eta_0,\eta_1,\eta_2,\eta_3)'$. Each model $\alpha \in \mathcal{M}$ contains a subset of the predictors, and is denoted by $\mathcal{K}_\alpha$, where $0 \in \mathcal{K}_\alpha \subset \mathcal{K}$. In the simulations, X_i's are i.i.d. $N(0,1)$ and $\sigma^2 = 1$, and thus $G^* = \operatorname{diag}(1/2,1,1,1,1)$. Let $\xi^{(n)} = (1,0,\eta_1^\diamond/\sqrt{n},\eta_2^\diamond/\sqrt{n},\eta_3^\diamond/\sqrt{n})'$ and $\xi^* = (1,0,0,0,0)'$. Then, $\xi^\circ = \xi^\diamond = (0,0,\eta_1^\diamond,\eta_2^\diamond,\eta_3^\diamond)'$, and $\xi^\circ_\alpha = (0,0,\delta_{\alpha\cdot1}\eta_1^\diamond,\delta_{\alpha\cdot2}\eta_2^\diamond,\delta_{\alpha\cdot3}\eta_3^\diamond)'$, where $\delta_{\alpha\cdot i} = 1$ for $i \in \mathcal{K}_\alpha$, and zero otherwise. See Appendix of Shimodaira (1996) for details.

REFERENCES

Cox, D. R. (1962). Further results on tests of separate families of hypotheses, *J. Roy. Statist. Soc. Ser. B*, **24**, 406–424.

Hotelling, H. (1936). Relations between two sets of variables, *Biometrika*, **28**, 321–377.

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

Shimodaira, H. (1995). A Study on Statistical Model Selection—Construction of the Confidence Set of Models, Ph.D. dissertation, Department of mathematical engineering and information physics, University of Tokyo (in Japanese).

Shimodaira, H. (1996). An application of multiple comparison techniques to model selection, *Ann. Inst. Statist. Math.* (to appear).

Steiger, J. H., Shapiro, A. and Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics, *Psychometrika*, **50**, 253–264.