Model selection under misspecification using information complexity

Hamparsum Bozdogan and Jan R. Magnus*

August 15, 2005; Revision March 22, 2006

Abstract: This paper extends Akaike's AIC-type model selection in two respects: we use a more encompassing notion of information complexity (ICOMP), and we allow certain types of model misspecification to be detected using the newly proposed criterion. The analytical closed-form expressions of the "sandwich" or "robust" variance matrix and a penalty-bias function within the context of the misspecified multivariate regression models are derived. The theoretical results are then applied to multivariate regression models in subset selection of the best predictors. A Monte Carlo simulation demonstrates the practical utility and the performance of ICOMP, showing the improvements over the AIC-type criterion, both in the case when the fitted models are correctly specified as in the misspecified case. The new approach proposed in this paper thus guards the researcher against some of the harmful effects of misspecification.

AMS Classification: 62H12, 62J05, 62-07.

Keywords: Model selection, Misspecification, Robustness, Multivariate regression, Information complexity.

Authors' addresses:
Hamparsum Bozdogan, Department of Statistics, Operations, and Management Science, The University of Tennessee, Knoxville, TN 37996, USA;
Jan R. Magnus, Department of Econometrics & Operations Research, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands.

*Corresponding author. *E-mail addresses:* bozdogan@utk.edu (Bozdogan) and magnus@uvt.nl (Magnus).

1 Introduction

Statistical models are approximations to reality and so the wrong model is, more often than not, fitted to the observed data. It is therefore important to develop statistical model selection techniques under misspecification. In recent years, since the classic works of Akaike (1973), White (1982), Nishii (1988), and Vuong (1989), there is a growing literature devoted to the study of misspecified models. For example, Nishii (1988) considered the consistency of various penalized likelihoods under the assumption of independently and identically distributed (i.i.d.) observations, and obtained the stochastic orders of the quasi maximum-likelihood estimator and the quasi log-likelihood, while Sin and White (1996) extended Nishii's results to dependent and heterogeneous data, and provided general conditions on the data-generating process under which the penalized likelihood criterion selects a model with lowest Kullback-Leibler divergence.

There are many ways a researcher can misspecify a regression model, and some of these are discussed in Godfrey (1988, p. 100). The most common misspecification errors are: the functional form of the model is not correctly specified; there are near-linear dependencies among the predictor variables (multicollinearity); skewness or kurtosis occurs in the variables, causing nonnormality of the random disturbances; or there is autocorrelation or heteroskedasticity. Specification errors can cause large forecasting errors (White, 1994), so it is of considerable importance to have means of fitting and choosing models in the presence of misspecification. We thus need new model selection techniques which will guard us against the dangers of model misspecification.

In the standard multiple regression models, a number of criteria have been introduced which are related either to misspecification situations or to the use and significance of the inverse-Fisher information matrix, see for example Konishi and Kitagawa (1996), Wei (1992), and Cavanaugh (1999, 2004). Konishi and Kitagawa (1996) provided numerous criteria from the information-theoretic point of view. Cavanaugh (2004) presented a new criterion named KIC. For large samples, KIC is unbiased. For small samples, Cavanaugh provided the exact unbiased estimator and introduced the corrected KIC_c and the modified KIC. Further, Seghouane and Bekara (2005) introduced a criterion for model selection in the presence of incomplete data based on KIC which take certain forms of misspecification into account.

Following the work of Huber (1967) and White (1982), the "sandwich variance matrix" estimation (also known as "robust variance matrix" estimation) has been shown to be the proper variance matrix under misspecification and has been applied widely, because it yields asymptotically consistent variance

matrix estimates without making distributional assumptions, also when the assumed model is misspecified. Sandwich-robust variance estimation can, for example, be used to deal with heteroskedastic errors, since variance estimates are consistent also when the observations are dependent (Kauermann and Carroll, 2001).

So far, we do not have a closed-form expression of the sandwich variance matrix within the context of a misspecified multivariate regression model taking into account possible misspecification in skewness and kurtosis. Our first objective in this paper is therefore to obtain an analytical closed-form expression of the sandwich variance matrix in the misspecified multivariate regression model. Our second objective is to extend Bozdogan and Haughton's (1998) work for the univariate misspecified regression model to the multivariate case. In particular, we shall extend the Akaike (1973) type model selection criteria in two directions: (i) we use a more general notion of information complexity (ICOMP) introduced by Bozdogan (2000, 2004), which is insensitive to possible misspecification in skewness and kurtosis; and (ii) we open the possibility to detect certain types of model misspecification. In this paper, ICOMP penalizes the "lack-of-fit" of a model using the (asymptotic) variance matrix. Thus, misspecification in skewness and kurtosis (and heteroskedasticity) may inflate or possibly deflate the complexity. Although misspecification in (conditional) heteroskedasticity is quite common in the literature (Eicker (1963), Huber (1967), White (1980)), this type of misspecification is not the main objective of this paper.

We do not claim that the ICOMP criterion derived in this paper captures all forms of model misspecification. We only pay attention to the case where the probabilistic distributional form of the fitted model departs from normality within the multivariate regression framework.

Sawa's (1978) BIC also adjusts penalization according to misspecification, but there is no relationship between ICOMP and BIC, except perhaps that the underlying formulation of the two criteria are both based on the Kullback-Leibler (1951) information. Sawa's penalty term is not an entropic function of the complexity of the estimated sandwich variance matrix of the model. On the other hand, the ICOMP criterion can be seen as an approximation to the sum of two Kullback-Leibler distances. Similarly, ICOMP is not necessarily related to Wei's (1992, p. 30) Fisher Information Criterion (FIC) in the standard multiple regression model. In FIC, the incorporation of the determinant of the Fisher information is not based on any theoretical grounds such as the entropic complexity measure of the variance matrix in ICOMP. FIC is more related to the Predictive Least Squares (PLS) criterion, as Wei (1992) demonstrates.

The plan of the paper is as follows. In Section 2, we define ICOMP for

misspecified models and extend it to structural complexity using the inverse Fisher information matrix (IFIM) under the correct specification assumption as well as under misspecification, using the results of White (1982), based on the "Hessian" form and the "outer-product" form of the Fisher information matrix, respectively. The basic idea is that one can use the difference between ICOMP(misspecified model) and ICOMP(correctly specified model) as an indication of possible departures from the distributional form of the model. This brings out the most important weakness of Akaike-type criteria for model selection: these procedures depend crucially on the assumption that the specified family of models includes the true model. We also propose a penalty-bias function under the distributional misspecification. In Section 3 we provide the explicit expression of ICOMP for the misspecified (as well as for the correctly specified) multivariate regression model and we derive the bias of the penalty for the misspecified multivariate regression model under normality. This form is useful in obtaining the amount of bias, based on maximum-likelihood estimation when distributional (or other) assumptions are not satisfied. The resulting penalty-bias function turns out to be a function of skewness and kurtosis coefficients. This section contains the mathematical part of the paper and may be of independent interest. In Section 4 we demonstrate the usefulness of these formulae through a Monte Carlo simulation experiment. Section 5 concludes. A mathematical appendix defines the duplication matrix and presents a new property of this matrix.

2 ICOMP: A new information measure of complexity for model selection

We briefly review the information complexity (ICOMP) criterion of Bozdogan (2000, 2004), defined as

$$\text{ICOMP} = -2\log L(\widehat{\theta}_k) + 2C(\widehat{\text{var}}(\widehat{\theta}_k)), \tag{1}$$

where $L(\widehat{\theta}_k)$ is the maximized likelihood function, $\widehat{\theta}_k$ is the maximum-likelihood estimate of the parameter vector θ_k under the model M_k, C represents a real-valued complexity measure, and $\widehat{\text{var}}(\widehat{\theta}_k)$ is the estimated variance matrix of the parameter vector of the model. Instead of penalizing the number of free parameters directly, ICOMP penalizes the variance complexity of the model. Thus a compromise takes place between the maximized log-likelihood and the complexity of the estimated variance matrix. The "best" model is the one with minimum ICOMP.

The most general form of ICOMP uses the information-based complexity of the inverse-Fisher information matrix (IFIM), and is referred to as

ICOMP(IFIM). As shown in Bozdogan and Haughton (1998) and Bozdogan (2000, 2004), we have, for a multivariate normal (non)linear structural model,

$$\text{ICOMP(IFIM)} = -2\log L(\widehat{\theta}) + 2C_1(\widehat{\mathcal{I}}^{-1}(\widehat{\theta})), \tag{2}$$

where the function $C_1(\mathcal{V})$ denotes the maximal information-theoretic complexity of a matrix $\mathcal{V}$ given by

$$C_1(\mathcal{V}) := \frac{s}{2}\log\left(\frac{\operatorname{tr}\mathcal{V}}{s}\right) - \frac{1}{2}\log|\mathcal{V}|, \quad s := \operatorname{rk}(\mathcal{V}). \tag{3}$$

ICOMP(IFIM) resembles a penalized likelihood criterium similar to AIC and AIC-type criteria, except that the penalty depends on the curvature of the log-likelihood function via the scalar function C_1, which measures the complexity of the estimated inverse-Fisher information matrix. ICOMP(IFIM) thus views complexity not as the number of parameters in the model, but as the degree of interdependence, that is, the correlational structure of the parameter estimates. For more details on ICOMP we refer the readers to Bozdogan (2000, 2004) and Bozdogan and Haughton (1998). Consistency properties of ICOMP have been studied for the multiple regression model, and the probabilities of underfitting and overfitting for ICOMP as the sample size n tends to infinity have been established, From these and other results, we conclude that ICOMP class criteria agree most often with the KL decision, compared to AIC and the Schwarz (1978) Bayesian Criterion (SBC). This goes to the heart of the consistency arguments about information criteria not studied before, since most of the studies are based on the fact that the true model considered is in the model set. The predictive performance of ICOMP has been discussed in Bearse and Bozdogan (1998) in subset selection in vector autoregressive models using the genetic algorithm.

2.1 ICOMP for misspecified models

We now generalize ICOMP to the case of misspecified models. First we define the two forms of the Fisher information matrix which are useful to check misspecification of a model. Based on a set of observation $y_1, \ldots, y_n$, we consider a quasi likelihood function $L(\theta)$ which we maximize with respect to θ to obtain quasi maximum-likelihood estimators. We define the "Hessian" form of the Fisher information matrix as

$$\mathcal{I} = -\operatorname{E}\left(\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right)$$

and the "outer-product" form as

$$\mathcal{R} = \mathrm{E}\left(\frac{\partial \log L(\theta)}{\partial \theta} \cdot \frac{\partial \log L(\theta)}{\partial \theta'}\right),$$

where the expectations are taken with respect to the true but unknown distribution. Under standard regularity conditions (Lehmann, 1983, Chapter 6), and based on an independent and identically distributed sample, we have

$$\sqrt{n}(\widehat{\theta} - \theta^*) \sim \mathrm{N}(0, \mathcal{I}^{-1}\mathcal{R}\mathcal{I}^{-1})$$

as $n \to \infty$, see for example Pawitan (2001, p. 373). The variance matrix $\mathrm{var}(\theta_k^*) = \mathcal{I}^{-1}\mathcal{R}\mathcal{I}^{-1}$ is called "robust", because it is the correct variance matrix regardless whether the assumed model is correct or not.

If the model is correctly specified and certain regularity conditions hold from White (1982, p. 7), we have *second-order regularity*, that is, $\theta^* = \theta_k^*$ and $\mathcal{I} = \mathcal{R}$, so that

$$\mathrm{var}(\theta_k^*) = \mathcal{I}^{-1}\mathcal{R}\mathcal{I}^{-1} = \mathcal{I}^{-1} = \mathcal{R}^{-1}.$$

Thus, when the model is correctly specified, the information matrix can be expressed in either "Hessian" form $\mathcal{I}$ or "outer-product" form $\mathcal{R}$. When the model is misspecified, then $\mathcal{I}$ and $\mathcal{R}$ are not equal. However, the variance matrix of $\widehat{\theta}$ can still be consistently estimated by

$$\widehat{\mathcal{V}} := \widehat{\mathrm{var}}(\widehat{\theta})_{misspec} = \widehat{\mathcal{I}}^{-1}\widehat{\mathcal{R}}\widehat{\mathcal{I}}^{-1}.$$

Therefore, ICOMP can be more generally defined as

$$\mathrm{ICOMP(Model)}_{misspec} = -2\log L(\widehat{\theta}) + 2C_1(\widehat{\mathcal{V}}),$$

where C_1 is defined in (3).

2.2 Bias of the penalty

When we assume that the true distribution does not belong to the specified parametric family of p.d.f.'s, then it is no longer true that the average of the maximized log-likelihood converges to the expected value of the parameterized log-likelihood, that is,

$$\frac{1}{n}\log L(y|\widehat{\theta}) \nrightarrow \mathrm{E}_y\left(\log f(y|\widehat{\theta})\right).$$

The difference between the average of the maximized log-likelihood and the expected maximized log-likelihood (the "bias" b) is given by

$$\begin{aligned} b &= \mathrm{E}_G\left(\frac{1}{n}\log L(y|\widehat{\theta}) - \int \log f(y|\widehat{\theta})\,\mathrm{d}\,G(y)\right) \\ &= \frac{1}{n}\,\mathrm{tr}(\mathcal{I}^{-1}\mathcal{R}) + O(n^{-2}), \end{aligned} \tag{4}$$

where the expectation is taken over the true distribution $G = \prod_{i=1}^{n} \mathrm{d}\,G(y_i)$. We note that $\mathrm{tr}(\mathcal{I}^{-1}\mathcal{R})$ is the well-known Lagrange-multiplier test statistic, see, for example, Takeuchi (1976), Hosking (1980), and Shibata (1989).

When the model is correctly specified, then $\mathcal{I} = \mathcal{R}$ and the bias reduces to

$$b = \frac{1}{n}\,\mathrm{tr}(\mathcal{I}^{-1}\mathcal{R}) + O(n^{-2}) = \frac{k}{n} + O(n^{-2}), \tag{5}$$

which gives AIC as a special case:

$$\mathrm{AIC} = -2\log L(\widehat{\theta}) + 2nb = -2\log L(\widehat{\theta}) + 2k.$$

When the true model is not in the model set considered, AIC will have difficulties to identify the best fitting model, as it does not penalize the presence of skewness and kurtosis. ICOMP works well with both biased or unbiased parameter estimates, and it uses smoothed (or improved) variance estimators. ICOMP-class criteria also legitimize the role of the Fisher information matrix as the natural metric on the parameter manifold of the model.

3 ICOMP and the misspecified multivariate regression model

Consider a set of n vectors $y_1, \ldots, y_n$, each of order $p \times 1$, whose first two moments are given by

$$\mathrm{E}(y_i) = B'x_i, \quad \mathrm{var}(y_i) = \Sigma,$$

where B is a $k \times p$ matrix of unknown coefficients, $X := (x_1, \ldots, x_n)'$ is a nonrandom $n \times k$ matrix of full column rank k, and $\Sigma = (\sigma_{ij})$ is a positive definite unknown $p \times p$ matrix. The full set of $kp + \frac{1}{2}p(p+1)$ coefficients is thus $\theta := ((\mathrm{vec}\,B)', (\mathrm{vech}(\Sigma))')'$. Assume that y_i and y_j are uncorrelated for all $i \neq j$, let $Y := (y_1, \ldots, y_n)'$ of order $n \times p$, and let $n \geq p + k$. These assumptions imply that

$$\mathrm{E}(Y) = XB, \quad \mathrm{var}(\mathrm{vec}\,Y) = \Sigma \otimes I_n.$$

We do *not* assume that the observations are normally distributed. We obtain quasi maximum-likelihood estimators of B and Σ by maximizing the normal log-likelihood function of the sample $y_1, \ldots, y_n$, given by

$$\ell(\theta) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}(Y - XB)\Sigma^{-1}(Y - XB)', \qquad (6)$$

see, for example, Magnus and Neudecker (1988, p. 321). The first differential of the log-likelihood is

$$\begin{aligned} \mathrm{d}\,\ell &= -\frac{n}{2}\operatorname{tr}\Sigma^{-1}\,\mathrm{d}\,\Sigma + \frac{1}{2}\operatorname{tr}(Y - XB)\Sigma^{-1}(\mathrm{d}\,\Sigma)\Sigma^{-1}(Y - XB)' \\ &\quad + \operatorname{tr} X(\mathrm{d}\,B)\Sigma^{-1}(Y - XB)' \\ &= \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(Y - XB)'(Y - XB)\Sigma^{-1} - n\Sigma^{-1}\right)\mathrm{d}\,\Sigma \\ &\quad + \operatorname{tr}\Sigma^{-1}(Y - XB)'X\,\mathrm{d}\,B, \end{aligned} \qquad (7)$$

leading to the first-order conditions

$$\Sigma^{-1}(Y - XB)'(Y - XB)\Sigma^{-1} = n\Sigma^{-1}, \quad X'(Y - XB)\Sigma^{-1} = 0,$$

and hence to the quasi maximum-likelihood estimators

$$\widehat{B} = (X'X)^{-1}X'Y, \quad \widehat{\Sigma} = \frac{(Y - X\widehat{B})'(Y - X\widehat{B})}{n} = \frac{Y'MY}{n}, \qquad (8)$$

where $M := I - X(X'X)^{-1}X'$ is an idempotent matrix.

3.1 The information matrix

Taking the differential of (7), we obtain the second differential of the log-likelihood as

$$\begin{aligned} \mathrm{d}^2\,\ell &= \operatorname{tr}(\mathrm{d}\,\Sigma^{-1})(Y - XB)'(Y - XB)\Sigma^{-1}\,\mathrm{d}\,\Sigma - \frac{n}{2}\operatorname{tr}(\mathrm{d}\,\Sigma^{-1})\,\mathrm{d}\,\Sigma \\ &\quad + 2\operatorname{tr}(\mathrm{d}\,\Sigma^{-1})(Y - XB)'X\,\mathrm{d}\,B - \operatorname{tr}\Sigma^{-1}(\mathrm{d}\,B)'X'X\,\mathrm{d}\,B. \end{aligned}$$

Then, using the fact that $\mathrm{E}(Y - XB) = 0$ and $\mathrm{E}(Y - XB)'(Y - XB) = n\Sigma$, we find

$$\begin{aligned} -\,\mathrm{E}\,\mathrm{d}^2\,\ell &= \frac{n}{2}\operatorname{tr}\Sigma^{-1}(\mathrm{d}\,\Sigma)\Sigma^{-1}\,\mathrm{d}\,\Sigma + \operatorname{tr}\Sigma^{-1}(\mathrm{d}\,B)'X'X\,\mathrm{d}\,B \\ &= \frac{n}{2}(\mathrm{d}\operatorname{vech}(\Sigma))'D_p'(\Sigma^{-1}\otimes\Sigma^{-1})D_p\,\mathrm{d}\operatorname{vech}(\Sigma) \\ &\quad + (\mathrm{d}\operatorname{vec}B)'(\Sigma^{-1}\otimes X'X)\,\mathrm{d}\operatorname{vec}B, \end{aligned}$$

where D_p denotes the $p^2 \times \frac{1}{2}p(p+1)$ duplication matrix defined in the Appendix. Hence, the ("Hessian form" of the) information matrix is

$$\mathcal{I} = \begin{pmatrix} \Sigma^{-1} \otimes X'X & 0 \\ 0 & \frac{n}{2} D_p'(\Sigma^{-1} \otimes \Sigma^{-1}) D_p \end{pmatrix}. \tag{9}$$

We note that the evaluation of $\mathcal{I}$ uses the first two moments of Y only, and is therefore not affected by the misspecification. For the same reason, the expectation of the first differential is zero (first-order regularity). However, it is not the case that $\mathrm{E}(\mathrm{d}\,\ell)^2 = -\,\mathrm{E}\,\mathrm{d}^2\,\ell$ (second-order regularity). This is because the evaluation of $\mathrm{E}(\mathrm{d}\,\ell)^2$ involves third and fourth moments, and these are possibly misspecified.

In order to obtain the "outer-product form" of the information matrix we standardize Y by defining $V := (Y - XB)\Sigma^{-1/2}$, so that

$$\mathrm{E}(V) = 0, \quad \mathrm{var}(\mathrm{vec}\, V) = I_{pn},$$

and introduce matrix generalizations of the usual skewness and kurtosis measures by defining

$$\mathit{\Gamma}_1 := \mathrm{E}(\mathrm{vec}\, V)(\mathrm{vec}(V'V - nI_p))', \quad \mathit{\Gamma}_2 := \mathrm{E}(\mathrm{vec}\, V'V)(\mathrm{vec}\, V'V)'. \tag{10}$$

In the special case of correct specification, these specialize to

$$\mathit{\Gamma}_1 = 0, \quad \mathit{\Gamma}_2 = 2nN_p + n^2(\mathrm{vec}\, I_p)(\mathrm{vec}\, I_p)', \tag{11}$$

where N_p denotes the $p^2 \times p^2$ symmetrizer matrix defined in the Appendix. If $n = p = 1$, the kurtosis further specializes to $\mathit{\Gamma}_2 = 3$, as expected.

We now evaluate $\mathrm{E}(\mathrm{d}\,\ell)^2$. Squaring Equation (7) yields

$$(\mathrm{d}\,\ell)^2 = \left(\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1/2} V'V \Sigma^{-1/2} - n\Sigma^{-1} \right) \mathrm{d}\,\Sigma + \mathrm{tr}\, \Sigma^{-1/2} V'X\, \mathrm{d}\, B \right)^2.$$

Letting $\mathit{\Delta} := D_p'(\Sigma^{-1/2} \otimes \Sigma^{-1/2}) D_p$, we thus obtain

$$\begin{aligned}
\mathrm{E}(\mathrm{d}\,\ell)^2 &= \frac{1}{4} \mathrm{E}\left(\mathrm{tr}\left(\Sigma^{-1/2} V'V \Sigma^{-1/2} - n\Sigma^{-1} \right) \mathrm{d}\,\Sigma \right)^2 + \mathrm{E}\left(\mathrm{tr}\, \Sigma^{-1/2} V'X\, \mathrm{d}\, B \right)^2 \\
&\quad + \mathrm{E}\left(\mathrm{tr}\left(\Sigma^{-1/2} V'V \Sigma^{-1/2} - n\Sigma^{-1} \right) \mathrm{d}\,\Sigma \right) \left(\mathrm{tr}\, \Sigma^{-1/2} V'X\, \mathrm{d}\, B \right) \\
&= \frac{1}{4} (\mathrm{d}\,\mathrm{vec}\,\Sigma)'(\Sigma^{-1/2} \otimes \Sigma^{-1/2})\, \mathrm{var}(\mathrm{vec}\, V'V)(\Sigma^{-1/2} \otimes \Sigma^{-1/2})\, \mathrm{d}\,\mathrm{vec}\,\Sigma \\
&\quad + (\mathrm{d}\,\mathrm{vec}\, B)'(\Sigma^{-1/2} \otimes X')\, \mathrm{var}(\mathrm{vec}\, V)(\Sigma^{-1/2} \otimes X)\, \mathrm{d}\,\mathrm{vec}\, B \\
&\quad + (\mathrm{d}\,\mathrm{vec}\,\Sigma)'(\Sigma^{-1/2} \otimes \Sigma^{-1/2}) \mathit{\Gamma}_1'(\Sigma^{-1/2} \otimes X)\, \mathrm{d}\,\mathrm{vec}\, B \\
&= \frac{1}{4} (\mathrm{d}\,\mathrm{vech}(\Sigma))' \mathit{\Delta} D_p^+ (\mathit{\Gamma}_2 - n^2 (\mathrm{vec}\, I_p)(\mathrm{vec}\, I_p)') D_p^{+\prime} \mathit{\Delta}\, \mathrm{d}\,\mathrm{vech}(\Sigma) \\
&\quad + (\mathrm{d}\,\mathrm{vec}\, B)'(\Sigma^{-1} \otimes X'X)\, \mathrm{d}\,\mathrm{vec}\, B \\
&\quad + (\mathrm{d}\,\mathrm{vech}(\Sigma))' \mathit{\Delta} D_p^+ \mathit{\Gamma}_1' (\Sigma^{-1/2} \otimes X)\, \mathrm{d}\,\mathrm{vec}\, B.
\end{aligned}$$

Hence,

$$-\operatorname{E} \mathrm{d}^2 \ell = (\mathrm{d}\,\theta)'\mathcal{I}\,\mathrm{d}\,\theta, \quad \operatorname{E}(\mathrm{d}\,\ell)^2 = (\mathrm{d}\,\theta)'\mathcal{R}\,\mathrm{d}\,\theta,$$

where $\mathcal{I}$ is the Hessian form of the information matrix defined in (9), $\mathcal{R}$ is the outer-product form,

$$\mathcal{R} = \begin{pmatrix} \Sigma^{-1} \otimes X'X & \frac{1}{2}(\Sigma^{-1/2} \otimes X')\Gamma_1 D_p^{+\prime}\Delta \\ \frac{1}{2}\Delta D_p^+ \Gamma_1'(\Sigma^{-1/2} \otimes X) & \frac{1}{4}\Delta D_p^+ \Gamma_2^* D_p^{+\prime}\Delta \end{pmatrix}, \tag{12}$$

and $\Gamma_2^* := \Gamma_2 - n^2(\operatorname{vec} I_p)(\operatorname{vec} I_p)'$. In the correctly specified case where $\Gamma_1 = 0$ and $\Gamma_2^* = 2nN_p$, one verifies that $\mathcal{R} = \mathcal{I}$.

3.2 Variance of the quasi maximum-likelihood estimator

In the presence of misspecification, the variance of the quasi maximum-likelihood estimator $\widehat{\theta}$ is $\mathcal{V} := \mathcal{I}^{-1}\mathcal{R}\mathcal{I}^{-1}$; see Gouriéroux and Monfort (1995a, p. 237), Gouriéroux and Monfort (1995b, p. 170), Hendry (1995, p. 391), White (1994), and others. The properties of this matrix are therefore of importance. Since $D_p^+ = (D_p'D_p)^{-1}D_p'$, and using Theorem 4.11 of Magnus (1988), we find the inverse of $\mathcal{I}$ as

$$\mathcal{I}^{-1} = \begin{pmatrix} \Sigma \otimes (X'X)^{-1} & 0 \\ 0 & \frac{2}{n} D_p^+ (\Sigma \otimes \Sigma) D_p^{+\prime} \end{pmatrix},$$

and hence

$$\mathcal{V} = \begin{pmatrix} \Sigma \otimes (X'X)^{-1} & \frac{1}{n}(\Sigma^{1/2} \otimes (X'X)^{-1}X')\Gamma_1 D_p \Delta^{-1} \\ \frac{1}{n}\Delta^{-1}D_p'\Gamma_1'(\Sigma^{1/2} \otimes X(X'X)^{-1}) & \frac{1}{n^2}\Delta^{-1}D_p'\Gamma_2^* D_p\Delta^{-1} \end{pmatrix}. \tag{13}$$

Furthermore, a little algebra gives

$$\begin{aligned} \operatorname{tr}\mathcal{V} &= \operatorname{tr}\Sigma \otimes (X'X)^{-1} + \frac{1}{n^2}\operatorname{tr}\Delta^{-1}D_p'\Gamma_2^* D_p\Delta^{-1} \\ &= (\operatorname{tr}\Sigma)(\operatorname{tr}(X'X)^{-1}) \\ &\quad + \frac{1}{n^2}\operatorname{tr} D_p^+(\Sigma^{1/2} \otimes \Sigma^{1/2})\Gamma_2^*(\Sigma^{1/2} \otimes \Sigma^{1/2})D_p^{+\prime}, \end{aligned} \tag{14}$$

and

$$\begin{aligned} |\mathcal{V}| &= |\Sigma \otimes (X'X)^{-1}| \cdot |\frac{1}{n^2}\Delta^{-1}D_p'\left(\Gamma_2^* - \Gamma_1'(I_p \otimes X(X'X)^{-1}X')\Gamma_1\right)D_p\Delta^{-1}| \\ &= 2^{-p(p-1)}n^{-p(p+1)}|\Sigma|^{p+k+1}|X'X|^{-p} \\ &\quad \times |D_p'(\Gamma_2^* - \Gamma_1'(I_p \otimes X(X'X)^{-1}X')\Gamma_1)D_p|. \end{aligned} \tag{15}$$

In the special case of correct specification, one verifies (using Lemma A in the Appendix) that

$$\operatorname{tr}\mathcal{V} = \operatorname{tr}\mathcal{I}^{-1} = (\operatorname{tr}\Sigma)(\operatorname{tr}(X'X)^{-1}) + \frac{1}{2n}\left(\operatorname{tr}\Sigma^2 + (\operatorname{tr}\Sigma)^2 + 2\sum_{j=1}^{p}\sigma_{jj}^2\right)$$

and

$$|\mathcal{V}| = |\mathcal{I}^{-1}| = 2^p n^{-\frac{1}{2}p(p+1)}|\Sigma|^{p+k+1}|X'X|^{-p}.$$

3.3 Derivation of ICOMP

To derive ICOMP for the (misspecified) multivariate regression model, we need the determinant and trace of $\widehat{\mathcal{V}}$, the estimator of $\mathcal{V}$. The matrix $\mathcal{V}$ itself is given in (13), and its trace and determinant in (14) and (15). Thus,

$$\begin{aligned}\operatorname{tr}\widehat{\mathcal{V}} &= (\operatorname{tr}\widehat{\Sigma})(\operatorname{tr}(X'X)^{-1}) \\ &\quad + \frac{1}{n^2}\operatorname{tr} D_p^+(\widehat{\Sigma}^{1/2}\otimes\widehat{\Sigma}^{1/2})\widehat{\Gamma_2}^*(\widehat{\Sigma}^{1/2}\otimes\widehat{\Sigma}^{1/2})D_p^{+\prime},\end{aligned}$$

and

$$\begin{aligned}|\widehat{\mathcal{V}}| &= 2^{-p(p-1)}n^{-p(p+1)}|\widehat{\Sigma}|^{p+k+1}|X'X|^{-p} \\ &\quad \times |D_p'(\widehat{\Gamma_2^*} - \widehat{\Gamma_1}'(I_p\otimes X(X'X)^{-1}X')\widehat{\Gamma_1})D_p|.\end{aligned}$$

As a result we obtain

$$\text{ICOMP(IFIM)}_{misspec} = np\log 2\pi + n\log|\widehat{\Sigma}| + np + 2C_1(\widehat{\mathcal{V}}), \tag{16}$$

where

$$C_1(\widehat{\mathcal{V}}) = \frac{s}{2}\log\left(\frac{\operatorname{tr}\widehat{\mathcal{V}}}{s}\right) - \frac{1}{2}\log|\widehat{\mathcal{V}}| \tag{17}$$

and $s := \operatorname{rk}(\widehat{\mathcal{V}}) = pk + \frac{1}{2}p(p+1)$.

In the special case of correct specification these results simplify to $\operatorname{tr}\widehat{\mathcal{V}} = \operatorname{tr}\widehat{\mathcal{I}}^{-1}$ and $|\widehat{\mathcal{V}}| = |\widehat{\mathcal{I}}^{-1}|$, and $\text{ICOMP(IFIM)}_{misspec}$ reduces to ICOMP(IFIM).

3.4 Derivation of the penalty bias

When the model is correctly specified, the skewness and kurtosis are given by (11), so that $\mathcal{R} = \mathcal{I}$. In general (under misspecification), we obtain

$$\begin{aligned}\operatorname{tr}(\mathcal{I}^{-1}\mathcal{R}) &= \operatorname{tr}(\Sigma\otimes(X'X)^{-1})(\Sigma^{-1}\otimes X'X) \\ &\quad + \frac{1}{2n}\operatorname{tr}\left(D_p^+(\Sigma\otimes\Sigma)D_p^{+\prime}\Delta D_p^+\Gamma_2^*D_p^{+\prime}\Delta\right) \\ &= pk + \frac{1}{2n}\operatorname{tr} N_p\Gamma_2^* = pk + \frac{1}{2n}\operatorname{tr}\Gamma_2^*.\end{aligned} \tag{18}$$

As derived in (4), the bias is then given by

$$b = \frac{1}{n}\operatorname{tr}(\mathcal{I}^{-1}\mathcal{R}) + O(n^{-2}) = \frac{1}{n}\left(pk + \frac{1}{2n}\operatorname{tr}(\Gamma_2^*)\right) + O(n^{-2}),$$

and hence the estimated bias $\widehat{b}$ is

$$\widehat{b} = \frac{1}{n}\operatorname{tr}(\widehat{\mathcal{I}}^{-1}\widehat{\mathcal{R}}) = \frac{1}{n}\left(pk + \frac{1}{2n}\operatorname{tr}(\widehat{\Gamma_2^*})\right), \tag{19}$$

which we compare with $b = k/n$, typically used in subset selection of variables and deletion diagnostics in multivariate regression models.

In the special case when there is no misspecification, we have $\Gamma_2^* = 2nN_p$ and

$$\operatorname{tr}(\Gamma_2^*) = np(p+1).$$

In that case,

$$\operatorname{tr}(\mathcal{I}^{-1}\mathcal{R}) = pk + p(p+1)/2,$$

which is the number of estimated parameters in the multivariate regression model and also the penalty term in AIC. This shows why (corrected) AIC, other AIC-type criteria, Schwarz's (1978) Bayesian criterion (SBC), and cross-validation techniques do not guard the researcher against misspecification of the model.

4 Monte Carlo simulations

4.1 Set-up

We provide some evidence, through Monte Carlo simulations, that the developed theory is useful in practice. The Monte Carlo set-up is based on the regressors

$$\begin{aligned}
x_1 &= 10 + u_1,\\
x_2 &= 10 + 0.3u_1 + \alpha u_2,\\
x_3 &= 10 + 0.3u_1 + 0.5604\alpha u_2 + 0.8282\alpha u_3,\\
x_4 &= -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5u_4,\\
x_5 &= -5 + 0.5x_1 + x_2 + 0.5u_5,
\end{aligned}$$

where $\alpha = \sqrt{1 - 0.3^2} = \sqrt{0.91} = 0.9539$. Each of the u_i represents a series of n independent and identically distributed $\mathrm{N}(0,1)$ variables. In addition, the five errors $u_1, \ldots, u_5$ are independent of each other.

The response variables are generated from

$$Y_{(n\times 2)} = (\imath, X_{or})_{(n\times 4)}\, B_{(4\times 2)} + E_{(n\times 2)},$$

where $\imath$ denotes a vector of ones,

$$X_{or} = (x_1, x_2, x_3) \quad \text{and} \quad B = \begin{pmatrix} -8 & -5 \\ 1 & 0.5 \\ 0.5 & 0 \\ 0.3 & 0.3 \end{pmatrix}.$$

Note that the generation of Y does not contain the variables x_4 and x_5, but x_4 is implicitly related to the variables x_1, x_2, x_3; and x_5 is implicitly related to x_1 and x_2. The value of α controls the amount of multicollinearity.

Denoting the n rows of the $n \times 2$ error matrix E by $\varepsilon_1', \ldots, \varepsilon_n'$, we assume as in Liu and Bozdogan (2005) that the ε_i are independent and identically distributed as a two-dimensional power exponential distribution

$$f(\varepsilon_i) = \text{constant} \times |\Sigma|^{-1/2} \exp -\frac{1}{2}\left(\varepsilon_i' \Sigma^{-1} \varepsilon_i\right)^{\beta} \tag{20}$$

with mean zero and variance

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

The multivariate power exponential (MPE) distribution is a member of the multivariate elliptically contoured family. The parameter $\beta > 0$ is kurtosis-related, indicating the amount of nonnormality of the distribution. When $\beta = 1/2$, Equation (20) specializes to the multivariate generalization of the double exponential distribution. When $\beta = 1$, we obtain the multivariate normal distribution, and when $\beta \to \infty$ the distribution tends to a multivariate generalization of the uniform distribution. An advantage of the MPE-class is that it is adaptive to both peakedness and flatness in the data by varying the values of β; see Gómez-Villegas and Marín (1998).

Some two-dimensional plots of the MPE-distribution are shown in Figure 1.

4.2 Full subset selection

Using our simulation protocol in the presence of both misspecification and multicollinearity, we carry out a subset selection of variables on the full saturated model:

$$Y_{(n\times 2)} = (\imath, x_1, x_2, x_3, x_4, x_5)_{(n\times 6)}\, B_{(6\times 2)} + E_{(n\times 2)}.$$

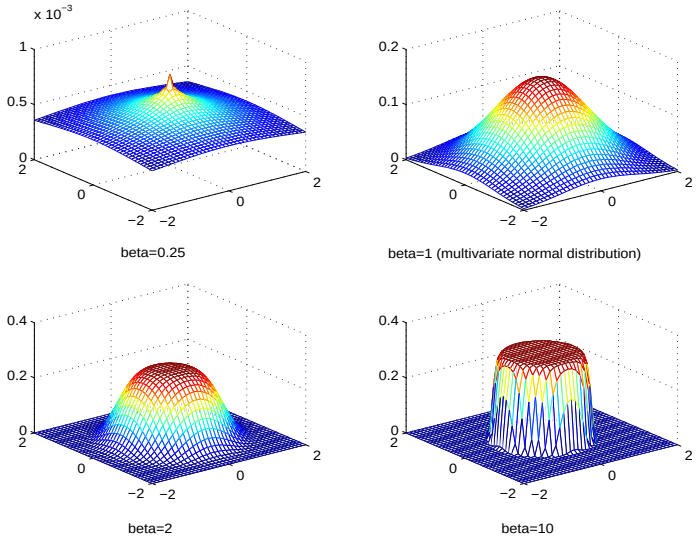

Fig. 1. Plots of the MPE-distribution for different β-values.

This is a difficult environment for any criterium to select the correct variables $X_{or} = (x_1, x_2, x_3)$.

We replicate our Monte Carlo experiment with MPE-distributed random errors 500 times with sample size $n = 1000$, and two values of the "kurtosis" parameter β: 0.75 and 1.50. We record ICOMP$_{misspec}$, AIC under the model with normally distributed errors, ICOMP$_{MPE}$, and AIC$_{MPE}$ across all possible subset models. See Liu and Bozdogan (2005) for some of the derivations. Some representative results are summarized in Table 1. The first panel of Table 1 shows that when $\beta = 0.75$, ICOMP$_{misspec}$ chooses the correct subset model $\{C, 1, 2, 3\}$ 460 times, while the subset model $\{C, 1, 2\}$ is chosen 40 times. On the other hand, under the correct model, ICOMP$_{MPE}$ chooses the correct subset model $\{C, 1, 2, 3\}$ 493 times, as expected, and the subset models $\{C, 1, 2, 3, 5\}$ and $\{C, 1, 3, 5\}$ five and two times, respectively. We note that AIC does not perform well even under the MPE model.

In the second panel of Table 1, we see that when $\beta = 1.50$, ICOMP$_{misspec}$ *always* chooses the correct subset model $\{C, 1, 2, 3\}$, In this case, ICOMP$_{MPE}$ chooses the correct model $\{C, 1, 2, 3\}$ 496 times. Still, AIC does not enjoy a very good performance even under the correct MPE model specification. Hence, the performance of ICOMP$_{misspec}$ is quite remarkable and consistent.

Interestingly, one does not need to derive the complicated form of the inverse Fisher information matrix (as in the MPE case). Instead, one can use the sandwich variance matrix estimator along with ICOMP$_{misspec}$, when

Table 1: Frequency of choosing the best subset multivariate regression model in 500 replications of the Monte Carlo experiment with MPE-distributed random errors, $n = 1000$, $\beta = 0.75$ (top) and $\beta = 1.50$ (bottom).

Subset	$\text{ICOMP}_{misspec}$	AIC	ICOMP_{MPE}	AIC_{MPE}
$C, 1, 2, 3, 4, 5$	0	5	0	2
$C, 1, 3, 4, 5$	0	0	0	0
$C, 1, 2, 3, 5$	0	66	5	70
$C, 1, 2, 3, 4$	0	43	0	51
$\vdots$	$\vdots$	$\vdots$	0	0
$C, 1, 3, 5$	0	0	2	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C, 1, 2, 3$	460	386	493	377
$C, 1, 2, 4$	0	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C, 1, 2$	40	0	0	0
$C, 1, 4$	0	0	0	0
$C, 1$	0	0	0	0
$C, 2$	0	0	0	0
$C, 3$	0	0	0	0
$C, 4$	0	0	0	0
$C, 5$	0	0	0	0

Subset	$\text{ICOMP}_{misspec}$	AIC	ICOMP_{MPE}	AIC_{MPE}
$C, 1, 2, 3, 4, 5$	0	7	0	6
$C, 1, 3, 4, 5$	0	0	0	0
$C, 1, 2, 3, 5$	0	41	4	34
$C, 1, 2, 3, 4$	0	67	0	58
$\vdots$	$\vdots$	$\vdots$	0	0
$C, 1, 3, 5$	0	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C, 1, 2, 3$	500	385	496	402
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C, 1, 2$	0	0	0	0
$C, 1$	0	0	0	0
$C, 2$	0	0	0	0
$C, 3$	0	0	0	0
$C, 4$	0	0	0	0
$C, 5$	0	0	0	0

it is suspected that a model is misspecified.

In contrasting the results in the two panels of Table 1, we see that these results agree with our intuitive notions that a sharply peaked distribution represents less uncertainty than does a broad distribution.

Our results also show that AIC is not able to perform well in the presence of misspecification and multicollinearity. It is possible that in Table 1 AIC's performance is due to the "superconsistency" property of AIC (Shibata, 1976). Superconsistency is a large sample problem and hence the regression coefficients may be biased in finite samples. Therefore, superconsistency may outweigh the impact of bias, especially in large samples.

4.3 Subset selection with genetic algorithm (GA)

In the increasingly important case of high-dimensional data sets, a genetic algorithm (GA) can be used to select the optimal subset of predictors. Using the same simulation protocol as above, we add redundant predictors to the model by generating uniformly distributed noise variables x_4 through x_{10}, and simulate a data set with $n = 1000$ observations. We use the genetic algorithm (GA) with population size 60, number of generations 50, probability of crossover 0.7, and probability of mutation 0.01. We replicate the Monte Carlo simulation experiment 100 times. The true model again contains an intercept and the variables x_1, x_2, and x_3. Most criteria find it difficult to select the right variables. However, the ICOMP(IFIM)$_{misspec}$ criterion always selects the true model (100 times), while AIC only chooses the correct model 45 times.

With our approach, the estimated bias $\widehat{b}$ can be easily computed from Equation (19), and the sampling distribution of the bias from our simulation is depicted in Figure 2. We see from Figure 2 that the sampling distribution of the estimated bias is skewed to the right due to the presence of multicollinearity. We note that the sampling distribution of the theoretical bias in estimation is not uniformly distributed or fixed, that is, the estimated bias $\widehat{b}$ does not equal k/n.

Next, we generate uniformly distributed noise variables x_4 through x_{50} and simulate a data set with (again) $n = 1000$ observations. In this case, there are 2^{50} possible subset models. We use the same genetic algorithm as above and simulate the subset selection process by running the algorithm 100 times for both ICOMP(IFIM)$_{misspec}$ and AIC. The top-five best subsets selected by these two criteria are listed in Tables 2 and 3. In all reported cases the true model, that is the one containing an intercept and the variables x_1, x_2, and x_3, is contained in the selected model. The GA with the ICOMP(IFIM)$_{misspec}$ criterion selects the model with three additional vari-

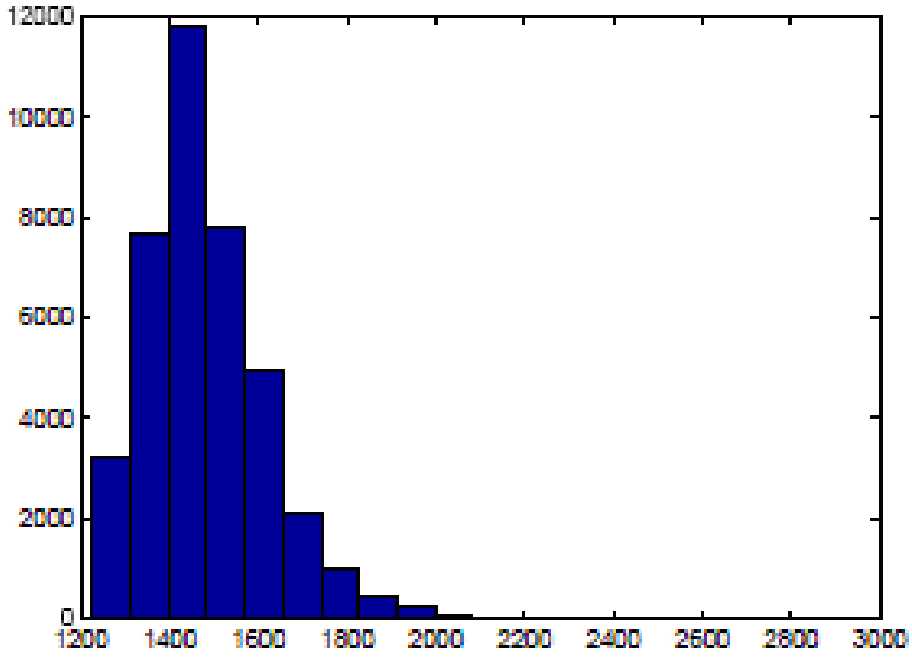

Fig. 2. Sampling distribution of the estimated bias.

ables (x_5, x_{45}, and x_{48}) as the best subset, and the correct model with no additional variables as the second best. The GA with the AIC criterion selects a model with eleven additional variables as the best subset and the correct model with fourteen additional variables as the second best. Clearly, AIC selects significantly more redundant variables than ICOMP(IFIM)$_{misspec}$. One GA run of the subset selection simulation is shown in Figure 3. We see from Figure 3 that GA finds the best-fitting model in generation 10, then finds another best-fitting model after generation 20, and converges between generations 42 and 50.

Table 2: Top-five best subsets selected by ICOMP(IFIM)$_{misspec}$

Rank	Subset	ICOMP$(IFIM)_{misspec}$
1	C,1,2,3,5,45,48	4162.1
2	C,1,2,3	4182.1
3	C,1,2,3,5,6,7,17,18,21	4222.1
4	C,1,2,3,16,34	4227.0
5	C,1,2,3,20,22	4228.6

Table 3: Top-five best subsets selected by AIC

Rank	Subset	AIC
1	C,1,2,3,5,9,14,30,31,35,38,39,45,49,50	4005.4
2	C,1,2,3,4,9,12,13,16,22,26,30,33,35,38,43,46,49	4005.9
3	C,1,2,3,8,11,14,17,26,28,35,38,41,43,44,47,48,49	4018.7
4	C,1,2,3,4,7,12,13,15,16,26,27,28,30,35,36,37,48,50	4025.0
5	C,1,2,3,5,7,8,11,12,17,18,23,25,26,27,33,38,41,46,47	4030.6

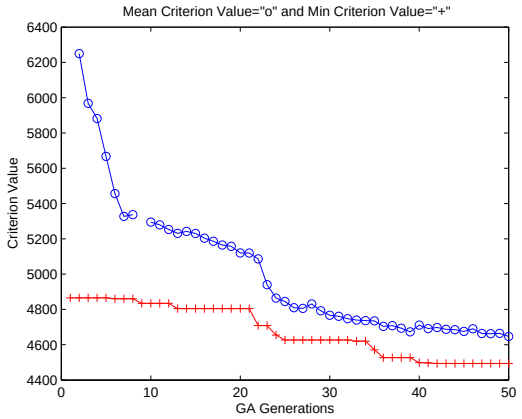

Fig. 3. One run of GA for the simulated data set with 50 variables.

For more on the use and the development of the genetic algorithm for subset selection in vector autoregressive models, we refer the readers to Bearse and Bozdogan (1998), Bozdogan and Bearse (2003), Bozdogan (2004), and Liu and Bozdogan (2005).

5 Conclusions

In this paper we introduced a new technique for subset selection of best predictors in the multivariate regression model, when the true underlying probability model is possibly misspecified. For this we required a closed-form expression of the "sandwich variance matrix" estimator, which is also useful for constructing confidence intervals for parameter estimates in the multivariate regression model and for obtaining standard errors of these estimates. We derived the expression of ICOMP for misspecified (as well as for correctly specified) multivariate regression models, and we gave an explicit expression of the bias of the penalty for misspecified models under normality,

which turns out to be a function of the skewness and kurtosis coefficients. The penalty bias shows the amount of bias when the maximum-likelihood estimator is used and the distributional assumptions are incorrect, and can also be used to assess the extent of the misspecification.

Through a Monte Carlo simulation experiment under nonnormality (specifically under the multivariate power exponential distribution), we demonstrated high stability of $\text{ICOMP}_{misspec}$. When the sample size is large, the performance $\text{ICOMP}_{misspec}$ is superior and consistent. The advantage of model selection under possible misspecification is that it extracts more information from the data and penalizes the presence of skewness and kurtosis when subset selection is performed.

In the increasingly important case of high-dimensional data sets, we introduced a genetic algorithm to select the optimal subset of predictors when the model is misspecified.

Possible extensions of the current paper include modifications of ICOMP to make model selection criteria to cover both robustness and misspecification at the same time by bringing the robust estimators into the lack-of-fit component of the ICOMP and the misspecification to the complexity or the penalty component. Such model selection approaches are important in vector autoregressive models; in principal components, factor analysis, and cluster analysis, including the kernel techniques in machine learning.

Acknowledgements

The first author gratefully acknowledges funding from the Scholarly Research Grant Program (SRGP) Awards of the College of Business Administration at the University of Tennessee in Knoxville, 2001–02. Versions of this paper were presented as an invited paper at the IMPS-2001 Conference, July 15–19, 2001 in Osaka, Japan, and at the workshop on *Reduction of Complexity, Trade-Offs, Methods* at the University of Dortmund, Germany, November 14–15, 2002. We gratefully acknowledge the computational assistance of Paul Gao, Xinli Bao, and Minhui Liu. Finally, we thank the two referees for their constructive comments.

Appendix: The duplication matrix: a new property

Let A be a square matrix of order $p \times p$. The two vectors $\text{vec}\, A$ and $\text{vec}\, A'$ contain the same p^2 components, but in a different order. Hence there exists

a unique permutation matrix that transforms $\operatorname{vec} A$ into $\operatorname{vec} A'$. This $p^2 \times p^2$ matrix is (a special case of) the *commutation matrix* and is denoted K_p; it is implicitly defined by the operation $K_p \operatorname{vec} A = \operatorname{vec} A'$.

Closely related to the commutation matrix is the $p^2 \times p^2$ *symmetrizer matrix* N_p with the property $N_p \operatorname{vec} A = \frac{1}{2} \operatorname{vec}(A + A')$ for every square $p \times p$ matrix A. It is easy to see that $N_p = \frac{1}{2}(I_{p^2} + K_p)$.

We now introduce the half-vec operator $\operatorname{vech}(\cdot)$. For any $p \times p$ matrix A, the vector $\operatorname{vech}(A)$ denotes the $\frac{1}{2}p(p+1) \times 1$ vector that is obtained from $\operatorname{vec} A$ by eliminating all supradiagonal elements of A. For example, for $p = 2$,

$$\operatorname{vec} A = (a_{11}, a_{21}, a_{12}, a_{22})' \quad \text{and} \quad \operatorname{vech}(A) = (a_{11}, a_{21}, a_{22})',$$

where the supradiagonal element a_{12} has been removed. Thus, for symmetric A, $\operatorname{vech}(A)$ only contains the distinct elements of A. Now, if A is symmetric, the elements of $\operatorname{vec} A$ are those of $\operatorname{vech}(A)$ with some repetitions. Hence, there exists a unique $p^2 \times \frac{1}{2}p(p+1)$ matrix D_p, called the *duplication matrix*, that transforms, for symmetric A, $\operatorname{vech}(A)$ into $\operatorname{vec} A$, that is,

$$D_p \operatorname{vech}(A) = \operatorname{vec} A \quad (A = A').$$

The matrices D_p and N_p are connected through $D_p D_p^+ = N_p$. The duplication matrix was introduced by Magnus and Neudecker (1980). A systematic treatment of K_p, N_p, and D_p, among others, is given in Magnus (1988). We now present the following new property.

Lemma A: Let $A = (a_{ij})$ be a square matrix of order $p \times p$. The determinant and trace of the matrix $D_p^+(A \otimes A)D_p^{+\prime}$ are given by

$$|D_p^+(A \otimes A)D_p^{+\prime}| = 2^{-\frac{1}{2}p(p-1)}|A|^{p+1}$$

and

$$\operatorname{tr}\left(D_p^+(A \otimes A)D_p^{+\prime}\right) = \frac{1}{4}\operatorname{tr}(A'A) + \frac{1}{4}(\operatorname{tr} A)^2 + \frac{1}{2}\sum_{j=1}^{p} a_{jj}^2.$$

Proof: Since

$$D_p^+(A \otimes A)D_p^{+\prime} = (D_p'D_p)^{-1}D_p'(A \otimes A)D_p(D_p'D_p)^{-1},$$

we obtain, from Magnus (1988, Theorem 4.11(i)),

$$\begin{aligned} |D_p^+(A \otimes A)D_p^{+\prime}| &= |D_p'D_p|^{-1}|D_p'(A \otimes A)D_p||D_p'D_p|^{-1} \\ &= 2^{-\frac{1}{2}p(p-1)}2^{\frac{1}{2}p(p-1)}|A|^{p+1}2^{-\frac{1}{2}p(p-1)} = 2^{-\frac{1}{2}p(p-1)}|A|^{p+1}. \end{aligned}$$

This proves the first result. To prove the second result, let δ_{st} denote the Kronecker delta, and write $u_{ij} = \text{vech}(e_i e_j')$, where e_i denotes the i-th column of the identity matrix I_p. Then,

$$\begin{aligned}
&\text{tr}\left(D_p^+(A\otimes A)D_p^{+\prime}\right) = \text{tr}\left(D_p^+(A\otimes A)D_p\right)\left(D_p'D_p\right)^{-1} \\
&= \frac{1}{2}\,\text{tr}\left(\sum_{i\geq j}\sum_{s\geq t}(a_{it}a_{js}+a_{is}a_{jt}-\delta_{st}a_{is}a_{js})u_{ij}u_{st}'\right)\left(I_{\frac{1}{2}p(p+1)}+\sum_{k=1}^{p}u_{kk}u_{kk}'\right) \\
&= \frac{1}{2}\sum_{i\geq j}(a_{ij}a_{ji}+a_{ii}a_{jj}-\delta_{ij}a_{ii}a_{jj})+\frac{1}{2}\sum_{j=1}^{p}a_{jj}^2 \\
&= \frac{1}{4}\sum_{ij}a_{ij}a_{ji}+\frac{1}{4}\sum_{ij}a_{ii}a_{jj}+\frac{1}{2}\sum_{j=1}^{p}a_{jj}^2 \\
&= \frac{1}{4}\,\text{tr}(A'A)+\frac{1}{4}(\text{tr}\,A)^2+\frac{1}{2}\sum_{j=1}^{p}a_{jj}^2,
\end{aligned}$$

where the second equality follows from the proof of Theorem 4.9 and Theorem 4.4(ii) in Magnus (1988).

References

H. Akaike, Information theory as an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csaki (Eds.), Second International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973, pp. 267–281.

P.M. Bearse, H. Bozdogan, Subset selection in vector autoregressive models using the genetic algorithm with informational complexity as the fitness function, Systems Analysis Modelling Simulation (SAMS) 31 (1998) 61–91.

H. Bozdogan, Akaike's information criterion and recent developments in informational complexity, J. Math. Psych. 44 (2000) 62–91.

H. Bozdogan, Statistical Data Mining and Knowledge Discovery, Chapman and Hall/CRC, Boca Raton, Florida, 2004.

H. Bozdogan, P.M. Bearse, Information complexity criteria for detecting influential observations in dynamic multivariate linear models using the genetic algorithm, J. Statist. Plann. Inference 114 (2003) 31–44.

H. Bozdogan, D.M.A. Haughton, Informational complexity criteria for regression models, Comput. Statist. Data Anal. 28 (1998) 51–76.

J.E. Cavanaugh, A large-sample model selection criterion based on Kullback's symmetric divergence, Statist. Probab. Lett. 44 (1999) 333–344.

J.E. Cavanaugh, Criteria for linear model selection based on Kullback's symmetric divergence, Aust. N.Z. J. Stat. 46 (2004) 257–274.

F. Eicker, Asymptotic normality and consistency of the least squares estimators for families of linear regression, Ann. Math. Statist. 34 (1963) 447–456.

L.G. Godfrey, Misspecification Tests in Econometrics, Cambridge University Press, Cambridge, 1988.

M.A. Gómez-Villegas, J.M. Marín, A multivariate generalization of the power exponential family of distributions, Comm. Statist. Theory Methods 27 (1998) 589–600.

C. Gouriéroux, A. Monfort, Statistics and Econometric Models, vol. 1, Cambridge University Press, Cambridge, 1995a.

C. Gouriéroux, A. Monfort, Statistics and Econometric Models, vol. 2, Cambridge University Press, Cambridge, 1995b.

D.F. Hendry, Dynamic Econometrics, Oxford University Press, Oxford, 1995.

J.R.M. Hosking, Lagrange-multiplier tests of time-series models, J. Roy. Statist. Soc. Ser. B 42 (1980) 170–181.

P.J. Huber, The behavior of maximum likelihood estimates under nonstandard conditions, in: L.M. LeCam, J. Neyman (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, 1967, pp. 221—233.

G. Kauermann, R.J. Carroll, A note on the efficiency of sandwich covariance matrix estimation, J. Amer. Statist. Assoc. 96 (2001) 1387–1396.

S. Konishi, G. Kitagawa, Generalized information criteria in model selection, Biometrika 83 (1996) 875–890.

S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1951) 79–86.

E.L. Lehmann, Theory of Point Estimation, Wiley, New York, 1983.

M.-H. Liu, H. Bozdogan, Multivariate regression models with power exponential random errors and subset selection using genetic algorithms with information complexity, Submitted for publication, 2005.

J.R. Magnus, Linear Structures, Charles Griffin & Company, London and Oxford University Press, New York, 1988.

J.R. Magnus, H. Neudecker, The elimination matrix: some lemmas and applications, SIAM J. Algebra. Discr. 1 (1980) 422–449.

J.R. Magnus, H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, Wiley, Chichester/New York (1988). Revised edition, 1999.

R. Nishii, Maximum likelihood principle and model selection when the true model is unspecified, J. Multivariate Anal. 27 (1988) 392–403.

Y. Pawitan, In All Likelihood: Statistical Modelling and Inference Using Likelihood, Oxford University Press, Oxford, 2001.

T. Sawa, Information criteria for discriminating among alternative regression models, Econometrica 46 (1978) 1273–1291.

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

A.K. Seghouane, M. Bekara, A criterion for model selection criterion in the presence of incomplete data based on Kullback's symmetric divergence, Signal Processing 85 (2005) 1405–1417.

R. Shibata, Selection of the order of an autoregressive model by Akaike's information criterion, Biometrika 63 (1976) 117–126.

R. Shibata, Statistical aspects of model selection, in: J.C. Willems (Ed.), From Data to Model, Springer, New York, 1989, pp. 215–240 .

C.-Y. Sin, H. White, Information criteria for selecting possibly misspecified parametric models, J. Econometrics 71 (1996) 207–225.

K. Takeuchi, Distribution of information statistics and a criterion of model fitting, Suri-Kagaku (Mathematical Sciences) 153 (1976) 12–18 (in Japanese).

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

C.-Z. Wei, On predictive least squares principles, Ann. Statist. 20 (1992) 1–42.

H. White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica 48 (1980) 817–838.

H. White, Maximum likelihood estimation of misspecified models, Econometrica 50 (1982) 1–26.

H. White, Estimation, Inference and Specification Analysis, Cambridge University Press, Cambridge, 1994.