EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 5, No. 2, 2012, 211-249
ISSN 1307-5543 – www.ejpam.com

Misspecified Multivariate Regression Models Using the Genetic Algorithm and Information Complexity as the Fitness Function

Hamparsum Bozdogan[1,*], J. Andrew Howe[2]

[1] *Department of Statistics, Operations, and Management Science, The University of Tennessee, Knoxville, TN, 37996, USA*
[2] *TransAtlantic Petroleum, Business Analytics, Istanbul Turkey*

Abstract. Model misspecification is a major challenge faced by all statistical modeling techniques. Real world multivariate data in high dimensions frequently exhibit higher kurtosis and heavier tails, asymmetry, or both. In this paper, we extend Akaike's AIC-type model selection criteria in two ways. We use a more encompassing notion of information complexity ($ICOMP$) of Bozdogan for multivariate regression to allow certain types of model misspecification to be detected using the newly proposed criterion so as to protect the researchers against model misspecification. We do this by employing the "sandwich"or "robust"covariance matrix $\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1}$, which is computed with the sample kurtosis and skewness. Thus, even if the data modeled do not meet the standard Gaussian assumptions, an appropriate model can still be found. Theoretical results are then applied to multivariate regression models in subset selection of the best predictors in the presence of model misspecification by using the novel genetic algorithm (GA), with our extended $ICOMP$ as the fitness function.

We demonstrate the power of the confluence of these techniques on both simulated and real-world datasets. Our simulations are very challenging, combining multicolinearity, unnecessary variables, and redundant variables with asymmetrical or leptokurtic behavior. We also demonstrate our model selection prowess on the well-known body fat data. Our findings suggest that when data are overly peaked or skewed - both characteristics often seen in real data, $ICOMP$ based on the sandwich covariance matrix should be used to drive model selection.

2010 Mathematics Subject Classifications: 62H86, 62J05, 62N02, 65G20, 62-07, 62F12, 62E20, 81T80, 78M34, 62B10

Key Words and Phrases: Misspecified multivariate regression models, Information complexity, Robust estimation, Genetic algorithm, Subset selection, Dimension reduction

1. Introduction

> *Model misspecification is a major, if it is not the dominant, source of error in the quantification of most scientific analysis* - Chatfield, 1995.

[*] Corresponding author.

Email addresses: bozdogan@utk.edu (H. Bozdogan), ahowe42@gmail.com (J. Howe)

In this research article, we simultaneously address several issues that affect many statistical modeling procedures. Specifically, this paper deals with the context of multivariate regression (MVR).

Statistical models are typically merely approximations to reality; additionally, for a set of observations, we typically don't know the true data generating process. Therefore, a wrong or *misspecified model* has a high probability of being fit to observed data. In many real-life applications of strategic decision making, large numbers of variables need to be simultaneously considered to build an *operating model* and for real-time data-mining. Application examples would include

- Behavioral and social sciences,
- Biometrics,
- Econometrics,
- Environmental sciences, and
- Financial modeling.

Further, it is often the case that several response variables are studied simultaneously given a set of predictor variables. In such cases it is often desirable to:

- determine which *subsets of the predictors* are most useful for forecasting each response variable in the system,
- interpret simultaneously a large number of regression coefficients, since this can become unwieldy even for a moderately-sized data, and
- achieve parsimony of unknown parameters, allowing both better estimation and clearer interpretation of the parameters.

Our objectives are twofold. The first is to develop a *computationally feasible intelligent data mining and knowledge discovery technique* that addresses the potentially daunting statistical and combinatorial problems of MVR modeling under *model misspecification*. Secondly, we provide new research tools that guide model fit and evaluation for high dimensional data, regardless of whether or not the probability model is misspecified. We employ a three-way hybrid:

- Our approach integrates novel statistical modeling procedures based on a misspecification-robust form of the *information complexity ($ICOMP$) criterion* [5, 4, 6, 7] as the fitness function;
- Multivariate regression models that allow for non-Gaussian random errors, and
- The *genetic algorithm* (GA) as our optimizer to select the best subset of predictors or variables.

To this end, we developed an *easy-to-use* command- and GUI- driven interactive computational toolbox. With this MATLAB® toolbox, we illustrate our new approach on both real and simulated data sets, showing the versatility of these techniques.

The rest of the paper is organized as follows. Section 2 provides the requisite background on multivariate regression (MVR) modeling, which is then followed by Section 3 regarding information complexity and AIC-type criteria for model selection. Here, we define $ICOMP$ under the correct and misspecified models and extend it to structural complexity using the results of [44], based on the Hessian and outer-product forms 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 4, we provide the explicit expression of $ICOMP$ for the correctly specified as well as for misspecified 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. Next, in Section 5 we provide details of the genetic algorithm (GA) and robust covariance estimators (Section 6). Finally, results with both simulated and real datasets are shown in Section 7, followed by concluding remarks. In Appendices 1 through 3, we repeat the analytical matrix calculus derivations of the model covariance matrix for multivariate regression under misspecification, based on the work of [27], for the benefit of the readers. Much of this work was done in collaboration with the first author while he was a Senior Scientist at Tilburg University, in Tilburg, the Netherlands during May of 1999. In Appendix 1, we show the derivation of the outer-product form of the Fisher information matrix. In Appendix 2, we show the derivation of the sandwich model covariance matrix which is a new result in closed-form which does not exist in the literature within the context of a misspecified MVR model. The opened up form of the misspecification resistant $ICOMP$ is obtained and shown, and Appendix 3 derives the penalty bias for multivariate regression.

2. Multivariate Gaussian Regression (MVR) Model

In the usual well known multivariate regression problem, we have a matrix of responses $Y \in \mathbb{R}^{n\times p}$; n observations of p measurements on some physical process. The researcher also has k variables that have some theoretical relationship to Y: $X \in \mathbb{R}^{n\times q}$, of course, we usually include a constant term as an intercept for the hyperplane generated by the relationship, so $q = k + 1$. The predictive relationship between X and Y has both a deterministic and a stochastic component, such that the model is

$$Y = XB + E, \tag{1}$$

in which $B \in \mathbb{R}^{q\times p}$ is a matrix of coefficients relating each column of X to each column of Y, and $E \in \mathbb{R}^{n\times p}$ is a matrix of error terms. The usual assumption in multivariate regression is that the error terms are uncorrelated, homoskedastic Gaussian white noise, or

$$E \sim N_p(\mathbf{0}, \Sigma \otimes I_n), \tag{2}$$

where the entire covariance matrix of the random error matrix E is given by

$$Cov(E) = \Sigma \otimes I_n. \tag{3}$$

This is an $(np \times np)$ matrix, where $\otimes$ denotes the direct or Kronecker product. Stated another way, we require

$$Y \sim N_p(XB, \Sigma \otimes I_n), \tag{4}$$

where

$$\mathrm{E}(Y) = XB, \text{ and } Cov(Y) = \Sigma \otimes I_n. \tag{5}$$

Under the assumption of Gaussianity, the log likelihood of the multivariate regression model is given by

$$\log L(\theta \mid Y) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}[(Y - XB)\,\Sigma^{-1}\,(Y - XB)']. \tag{6}$$

We obtain quasi maximum-likelihood estimators of B and Σ by maximizing the log-likelihood function in (6). From [28, page 321], the first differential of the log-likelihood is

$$\begin{aligned} \mathrm{d}\log L(\theta \mid Y) &= -\frac{n}{2}\operatorname{tr}\Sigma^{-1}\,\mathrm{d}\,\Sigma + \frac{1}{2}\operatorname{tr}[(Y - XB)\Sigma^{-1}(\mathrm{d}\,\Sigma)\Sigma^{-1}(Y - XB)'] \\ &\quad + \operatorname{tr}X(\mathrm{d}B)\Sigma^{-1}(Y - XB)' \\ &= \frac{1}{2}\operatorname{tr}(\Sigma^{-1}(Y - XB)'(Y - XB)\Sigma^{-1} - n\Sigma^{-1})\,\mathrm{d}\,\Sigma \\ &\quad + \operatorname{tr}\Sigma^{-1}(Y - XB)'X\,\mathrm{d}B, \end{aligned} \tag{7}$$

leading to the first-order conditions

$$\Sigma^{-1}(Y - XB)'(Y - XB)\Sigma^{-1} = n\Sigma^{-1}, \text{ and } X'(Y - XB)\Sigma^{-1} = \mathbf{0}, \tag{8}$$

and hence to the quasi maximum-likelihood estimators of B and Σ given by

$$\hat{B} = (X'X)^{-1}X'Y, \tag{9}$$

and

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})'(Y - X\hat{B}) = \frac{1}{n}\hat{E}'\hat{E} = \frac{1}{n}Y'MY, \tag{10}$$

where $M = I - X(X'X)^{-1}X'$ is an idempotent matrix.
To derive the information complexity ($ICOMP$) criteria, we modify the results of [28, page 321], and obtain the estimated *inverse Fisher information matrix* (IFIM) given by

$$\widehat{Cov}(vec(\hat{B}), \operatorname{vech}(\hat{\Sigma})) \equiv \hat{\mathscr{F}}^{-1} = \begin{bmatrix} \hat{\Sigma} \otimes (X'X)^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2}{n}D_p^{+}(\hat{\Sigma} \otimes \hat{\Sigma})D_p^{+\prime} \end{bmatrix}, \tag{11}$$

where $D_p^+ = (D_p'D_p)^{-1}D_p$ is the Moore-Penrose inverse of the duplication matrix, D_p. D_p is a unique $p^2 \times \frac{1}{2}p(p+1)$ matrix that transforms, for symmetric Σ, $\operatorname{vech}(\Sigma)$ into $\overrightarrow{\Sigma}$. For example, for $p = 2$,

$$\overrightarrow{\Sigma} = (\sigma_{11}, \sigma_{21}, \sigma_{12}, \sigma_{22})', \text{ and } \operatorname{vech}(\Sigma) = (\sigma_{11}, \sigma_{21}, \sigma_{22})',$$

where the supradiagonal element σ_{12} has been removed. Thus, for symmetric Σ, vech(Σ) only contains the distinct elements of Σ. That is,

$$D_p \,\mathrm{vech}(\Sigma) = \overrightarrow{\Sigma}, \,(\Sigma = \Sigma').$$

For an excellent treatment of D_p, see [26].

The IFIM provides the asymptotic variance of the ML estimator when the model is correctly specified. Its *trace* and *determinant* provide scalar measures of the asymptotic variance, and they play a key role in the construction of information complexity. It is also very useful, as it provides standard errors for the regression coefficients on the diagonals.

The method of least squares is generally used to estimate the coefficients in regression models. In many applications, the results of a *least-squares fit* are often unacceptable when the model is *misspecified*, or when the model is *wrong*. In most statistical modeling problems, we almost always fit a wrong model to the observed data. This can introduce bias into the model due to model misspecification. There are a number of ways a researcher can misspecify a regression model. Some of these are discussed in [15, page 100]. In the context of regression, one of the most abused assumptions is that of normality. The most common causes of model misspecification include:

- multicollinearity,
- autocorrelation,
- heteroskedasticity,
- incorrect functional form.

Characteristics related to this last bullet point that are easy to visualize include:

- Leptokurtosis - high peak; more variation; higher probability in tails
- Platykurtosis - flatter; less variation; lower probability in tails
- Skewness - asymmetric; higher probability in one tail, lower probability in other.

Examples can be seen in Figures 1, 2, and 3, generated by the multivariate power exponential distribution given in Section 7.1. In the right pane of all three plots, the heavy black dotted contours are from the Gaussian distribution, for comparison. Characteristics in the first and last figures are most common in real multivariate data. Unfortunately, in the literature the

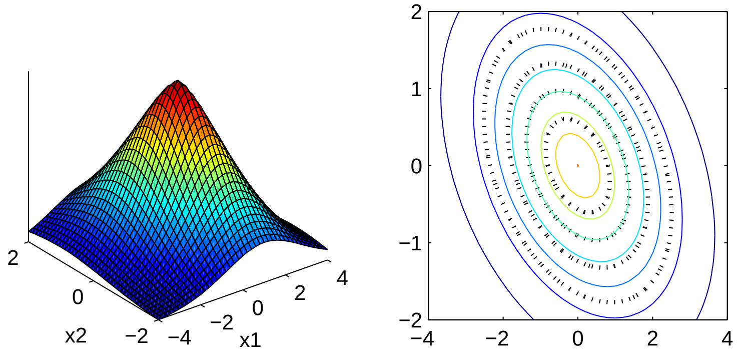

Figure 1: Surface and contour plot for leptokurtic data (black contours = Gaussian).

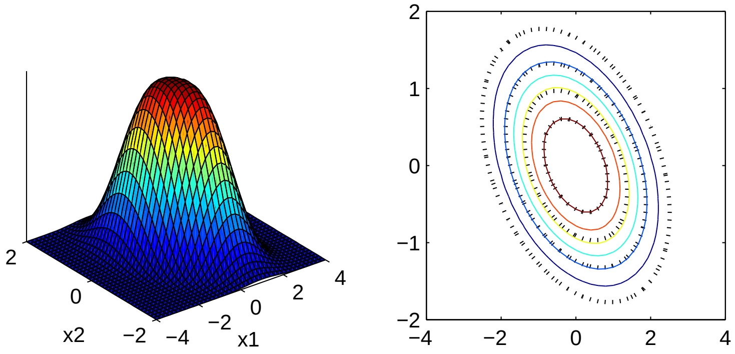

Figure 2: Surface and contour plot for platykurtic data (black contours = Gaussian).

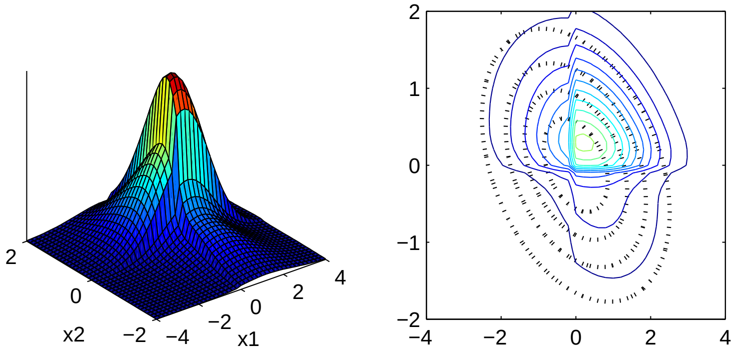

Figure 3: Surface and contour plot for skewed data (black contours = Gaussian).

common answer to nonnormality, has been the utilization of *Box-Cox transformations* of [3], which does not seem to work consistently well, in the context of both univariate and multivariate regression models.

Of course, when performing regression analysis, it is not usually the case that all variables in X have significant predictive power over Y. Choosing an optimal subset model has long been a vexing problem. Typical methods for selecting a subset regression model include:

- Forward stepwise analysis
- Backward stepwise analysis
- Partial sums-of-squares and sequential F-tests
- consideration of reduced rank regression models.

In practice, these methods may not work well in the presence of multicollinearity or model misspecification. Additionally, the hypothesis tests all require somewhat arbitrary selection of significance levels. Finally, it seems doubtful that the stepwise methods can accurately control for Type I Errors across all tests performed. One solution is to use some criterion to perform complete enumeration of all possible subsets of the predictors. As the number of predictors, q, grows, however, this quickly becomes infeasible. For example, consider a simple case of multiple regression with $k = 9$ predictors and the constant term, so that

$q = 10$. There are $2^{10} - 1 = 1023$ possible nontrivial subsets of the predictors. This is clearly too many models to humanly consider, and yet this is a relatively small problem. Consider a dataset we have from a large fractional factorial experiment with $q = 56$; there are $2^{56} - 1 = 72,057,594,037,927,936$ possible subset models. This is far too many for even a computer to automatically evaluate in a timely manner. **Little headway has been made in finding the best subset MVR model from a global perspective**.

In our results, we demonstrate the value of the genetic algorithm (Section 5), driven by information criteria as a fitness function. We substitute the GA as a computationally feasible and efficient approach for complete combinatorial subset analysis. For the subset regression model, we use the notation $X^* \in \mathbb{R}^{n \times q^*}$, where q^* is the number of variables not excluded from the model, such that $q^* \in [1, q]$.

3. $ICOMP$: A New Information Measure of Complexity for Model Selection

Perhaps the most basic information criteria is the *Kullback-Liebler divergence* (KL), first introduced by [24] (also called KL distance, or KL information). This information divergence measures the difference between two probability distributions. If we have some data for which we know the true distribution f, we can use the KL divergence to determine whether f_1 or f_2 more accurately represents the true *data generating process* (dgp). f_1 and f_2 could be different distributions, or the same distribution as f with different parameters. When a true model is known, as in simulation studies, we can compute the KL divergence for all competing models, with the assumption that the distance for the true model will be the closest to 0.

3.1. $ICOMP$ for Correctly Specified Models

Assuming that the model is correctly specified, or the true model is in the model set considered, Akaike [1] introduced his well-known Akaike's Information Criteria (AIC). Acknowledging the fact that any statistical model is merely an approximate representation of the true dgp, information criteria attempt to guide model selection according to Occam's Razor. One restatement of Occam's Razor is:

"*Of all possible solutions to a problem, the simplest solution is probably the best, ceteris paribus*".

This principle of parsimony requires that as model complexity increases, the fit of the model must increase at least as much; otherwise, the additional complexity is not worth the cost. Virtually all information criteria penalize a bad fitting model with negative twice the maximized log likelihood, as an asymptotic estimate of the KL information. The difference, then, is in the penalty for model complexity. The simplest information criteria are AIC and SBC [37], shown below.

$$AIC = -2 \log L(\hat{\theta} \mid y) + 2m \tag{12}$$

$$SBC = -2 \log L(\hat{\theta} \mid y) + \log(n)m \tag{13}$$

In both cases, m is the number of parameters estimated in the model. When using any information criterion to perform model selection, we choose the model corresponding to the lowest score as providing the best balance between good fit and parsimony. It can sometimes be difficult to attach significance (not statistically) to differences in information criteria scores when attempting to select a most appropriate model. To resolve this, we may compute relative weights which can be interpreted as the **probability that a given model is the most appropriate**. The weights are computed as in (14),

$$W_i = \frac{e^{-\frac{IC_i - \min(IC)}{2}}}{\sum_{i=1}^{L} e^{-\frac{IC_i - \min(IC)}{2}}}, \tag{14}$$

where i indexes the L models evaluated.

In contrast to AIC and SBC, $ICOMP$, originally introduced by [4], is a logical extension of AIC and SBC which is based on the structural complexity of an element or set of random vectors via the generalization of the information-based covariance complexity index of [42]. In $ICOMP$, lack-of-fit is still penalized by twice the negative of the maximized log likelihood, while a combination of **lack-of-parsimony** and **profusion-of-complexity** are simultaneously penalized by a scalar complexity measure, C, of the model covariance matrix. In general, $ICOMP$ is defined by

$$ICOMP = -2\log L(\hat{\theta} \mid y) + 2C(\widehat{Cov}(\hat{\theta})), \tag{15}$$

where $\widehat{Cov}(\hat{\theta})$ indicates the estimated model covariance matrix. Each term in (15) approximates one KL distance. There are several forms and justifications of $ICOMP$, two of which are detailed here in what follows.

3.1.1. $ICOMP$ as an Approximation to the Sum of Two KL Distances

For the following, we need the first order maximal entropic complexity of [4] as a generalization of the model covariance complexity of [42], given by

$$C_1(\widehat{Cov}(\hat{\theta})) = \frac{s}{2}\log\frac{\mathrm{tr}(\widehat{Cov}(\hat{\theta}))}{s} - \frac{1}{2}\log|\widehat{Cov}(\hat{\theta})|, s = rank(\widehat{Cov}(\hat{\theta})). \tag{16}$$

The greatest simplicity, that is zero complexity, is achieved when the model covariance matrix is proportional to the identity matrix, implying that the parameters are orthogonal and can be estimated with equal precision.

Proposition 1. *For a multivariate normal linear or nonlinear structural model we define the general form of ICOMP as*

$$ICOMP = -2\log L(\hat{\theta}_M) + 2C_1(\hat{\mathscr{F}}^{-1}), \tag{17}$$

where $\hat{\mathscr{F}}^{-1}$ is the estimated inverse Fisher information matrix of the model.

Proof. Suppose we consider a general statistical model of the form given by

$$y = m(\theta) + \varepsilon, \tag{18}$$

where:

- $y = (y_1, y_2, \ldots, y_n)$ is an $n \times 1$ random vector of response values in $\mathbb{R}^n$;
- θ is a parameter vector in $\mathbb{R}^k$;
- $m(\theta)$ is a systematic component of the model in $\mathbb{R}^n$, which depends on the parameter vector θ, and its deterministic structure depends on the specific model considered; and
- ε is an $n \times 1$ random error vector with

$$\mathrm{E}(\varepsilon) = 0, \text{ and } \mathrm{E}\left(\varepsilon\varepsilon'\right) = \Sigma(\varepsilon). \tag{19}$$

Following [9], we denote θ^* to be the parameter vector of the *true operating model*, and θ to be any other value of the vector of parameters. Let $f(y \mid \theta)$ denote the joint density function of y given θ, and let $f(y \mid \theta^*)$ indicate the true model. Further, let $KL(\theta^* \mid \theta)$ denote the KL distance between the true model. Then, since $y_i, i = 1, 2, \ldots, n$ are independent, we have:

$$\begin{aligned} KL(\theta^*, \theta) &= \int_{\mathbb{R}^n} f(y \mid \theta^*) \log\left[\frac{f(y \mid \theta^*)}{f(y \mid \theta)}\right] dy \\ &= \sum_{i=1}^{n} \int f_i(y_i \mid \theta^*) \log\left[f_i(y_i \mid \theta^*)\right] dy_i \\ &\quad - \sum_{i=1}^{n} \int f_i(y_i \mid \theta^*) \log\left[f_i(y_i \mid \theta)\right] dy_i, \end{aligned} \tag{20}$$

where $f_i, i = 1, 2, \ldots, n$ are the marginal densities of the y_i. Note that the first term in (20) is the usual *negative entropy* $H(\theta^*; \theta^*) = H(\theta^*)$, which is constant for a given $f_i(y_i \mid \theta^*)$. The second term is equal to:

$$-\sum_{i=1}^{n} \mathrm{E}\left(\log f_i(y_i \mid \theta)\right), \tag{21}$$

which can be **unbiasedly estimated** by

$$-\sum_{i=1}^{n} \log f_i(y_i \mid \theta) = -\log L(\theta \mid y). \tag{22}$$

Of course, $\log L(\theta \mid y)$ is the log likelihood of the observations evaluated at θ. In practice, we would estimate the parameter vector for a model M, typically using the MLE $\hat{\theta}_M$, and so use the maximized log likelihood

$$-\sum_{i=1}^{n} \log f_i(y_i \mid \hat{\theta}_M) = -\log L(\hat{\theta}_M \mid y). \tag{23}$$

This gives us an estimate of the first KL distance, which is reminiscent of the derivation of AIC.

On the other hand, a model M gives rise to an *asymptotic covariance matrix*:

$$Cov(\hat{\theta}_M) = \Sigma(\hat{\theta}_M) \tag{24}$$

for the MLE $\hat{\theta}_M$. That is,

$$\hat{\theta}_M \sim N(\theta^*, \Sigma(\hat{\theta}_M) = \mathscr{F}^{-1}). \tag{25}$$

Now invoking the $C_1(\cdot)$ (16) complexity on $\Sigma(\hat{\theta}_M)$ can be seen as the KL distance between the joint density and the product of marginal densities for a normal random vector with covariance matrix $\Sigma(\hat{\theta}_M)$, maximized over all orthonormal transformations of that normal random vector [see 6]. Hence, using the estimated covariance matrix, we define $ICOMP$ as the sum of two Kullback-Liebler distances given by:

$$\begin{aligned} ICOMP(\hat{\mathscr{F}}^{-1}) &= -2\sum_{i=1}^{n} \log f_i(y_i \mid \hat{\theta}_M) + 2C_1(\hat{\Sigma}(\hat{\theta}_M)) \\ &= -2\log L(\hat{\theta}_M \mid y) + 2C_1(\hat{\mathscr{F}}^{-1}). \end{aligned} \tag{26}$$

Some observations:

- The first component of $ICOMP(\hat{\mathscr{F}}^{-1})$ in (26) measures the lack of fit of the model, and the second component measures the complexity of the estimated IFIM, which gives a scalar measure of the celebrated *Cramér-Rao lower bound matrix* (CRLB), taking into account the accuracy of the estimated parameters and implicitly adjusting for the number of free parameters included in the model.
- It is an intrinsic measure of uncertainty, and, furthermore, it is a quality metric of the estimation procedure. For more on this, and for some immediate physical motivation, we refer the readers to the interesting book by [14]. Also, see, [12] and [32, 33, 34].
- $ICOMP(\hat{\mathscr{F}}^{-1})$ contrasts the trace and the determinant of the IFIM, and this amounts to a comparison of the geometric and arithmetic means of the eigenvalues of the IFIM, i.e.:

$$ICOMP(\hat{\mathscr{F}}^{-1}) - 2\log L(\hat{\theta}_M \mid y) + 2\log \frac{\overline{\lambda}_a}{\overline{\overline{\lambda}}_g}. \tag{27}$$

This looks like $CAIC$ of [5], MDL of [35], and SBC of [37], with the exception that $\log n$ is replaced with $\log \frac{\overline{\lambda}_a}{\overline{\overline{\lambda}}_g}$.

3.1.2. $ICOMP$ as an Estimate of Posterior Expected Utility

Following the results from [9], we make the following proposition.

Proposition 2. *$ICOMP$ as a Bayesian criterion close to maximizing a posterior expected utility (PEU) is given by*

$$ICOMP_{PEU} = -2\log L(\hat{\theta}_M \mid y) + m + 2C_1(\hat{\mathscr{F}}^{-1}). \tag{28}$$

Proof. Consider

- Let $L(\theta_M \mid y)$ be the likelihood function of the parameter vector for a given vector y of observations.
- Let $f_{Prior}(\theta \mid M)$ denote the *prior density function* of θ on the model M; $f_{Post}(\theta \mid y)$ is the corresponding *posterior density*.
- Let $\mathscr{F}(\theta_M)$ denote the Fisher information matrix for the n observations corresponding to model M, and let m be the dimension of M.

Following [30], we consider the KL distance between the posterior and the prior densities for model M given by

$$\begin{aligned} &KL(f_{Post}(\theta \mid y), f_{Prior}(\theta \mid M)) \\ &= \int_{\Theta_M} f_{Post}(\theta \mid y) \log f_{Post}(\theta \mid y) d\theta - \int_{\Theta_M} f_{Post}(\theta \mid y) \log f_{Prior}(\theta \mid M) d\theta \\ &= H(f_{Post}(\theta \mid y)) - \int_{\Theta_M} f_{Post}(\theta \mid y) \log f_{Prior}(\theta \mid M) d\theta. \end{aligned} \tag{29}$$

Further following Poskitt's arguments, under regularity conditions which guarantee the asymptotic normality of the posterior distribution, that is, when

$$f_{Post}(\theta \mid y) \cong N(\hat{\theta}, \Sigma(\hat{\theta}) = \hat{\mathscr{F}}^{-1}), \tag{30}$$

it can be shown that

$$KL(f_{Post}(\theta \mid y), f_{Prior}(\theta \mid M)) = -\frac{m}{2}\log(2\pi) - \frac{m}{2} - \frac{1}{2}\log|\hat{\mathscr{F}}^{-1}| - \log f_{Prior}(\theta \mid M). \tag{31}$$

One can argue, as did Poskitt, that a utility U_1 can be defined as $\log U_1$ = the KL distance given by (31). In Bayesian design of experiments, following the suggestion of [25], several authors have considered the Kullback-Liebler divergence as a *utility function*. For more on this, see [10]. In our case, we propose to multiply the utility U_1 by a utility U_2 equal to

$$U_2 = \exp\left[-a \times C_1(\hat{\mathscr{F}}^{-1})\right]. \tag{32}$$

If we have $a = 1$, our utility is $U = U_1 \times U_2$, and the log of that utility equals:

$$\log(U) = -\frac{m}{2}\log(2\pi) - \frac{m}{2} - \frac{1}{2}\log|\hat{\mathscr{F}}^{-1}| - \log f_{Prior}(\theta \mid M) - C_1(\hat{\mathscr{F}}^{-1}), \tag{33}$$

which is the difference of KL distances. Note that our utility U_2 is slightly different from that used by Poskitt. His utility U_2 uses only the trace term in the expression of the complexity, and does not contrast the determinant with the trace. The trace involves only the diagonal elements, analogous to variances, while the determinant involves also the off-diagonal elements, analogous to covariances. Our utility amounts to a comparison of the geometric and arithmetic means of the eigenvalues of *IFIM* already shown in (27).

If we apply Poskitt's Corollary 2.2 (**maximizing posterior expected utility**), or the *Laplace expansion* results of [21], it follows that, under some regularity conditions, if the parameter vector θ lies in model M, the posterior expected utility can be approximated by

$$\log(PEU) \cong \log f(y \mid \hat{\theta}_M) + \frac{m}{2}\log(2\pi) + \frac{1}{2}\log|\hat{\mathscr{F}}^{-1}| + \log(U) + \log f_{Prior}(\hat{\theta}_M \mid M), \quad (34)$$

up to order $O(\frac{1}{n})$ and up to some terms which do not depend on the model M. Replacing $\log(U)$ in this equation by its value in (33), some terms cancel out. We thus obtain a criterion, to be maximized to choose a model:

$$\log f(y \mid \hat{\theta}_M) - \frac{m}{2} - C_1(\hat{\mathscr{F}}^{-1}) + \log f(M), \quad (35)$$

but maximizing this is equivalent to minimizing $ICOMP_{PEU}(\hat{\mathscr{F}}^{-1})$, given by

$$ICOMP_{PEU}(\hat{\mathscr{F}}^{-1}) = -2\log L(\hat{\theta}_M \mid y) + m + 2C_1(\hat{\mathscr{F}}^{-1}) + \log f(M). \quad (36)$$

Finally, assuming that $f(M)$ is constant for all models in (36), we have

$$ICOMP_{PEU}(\hat{\mathscr{F}}^{-1}) = -2\log L(\hat{\theta}_M \mid y) + m + 2C_1(\hat{\mathscr{F}}^{-1}). \quad (37)$$

Note that when we defined the utility

$$U_2 = \exp\left[-a \times C_1(\hat{\mathscr{F}}^{-1})\right], \quad (38)$$

we considered the constant multiplier a to be 1 in obtaining the result shown above. Indeed other choices of a are possible and equally justifiable, giving rise to different penalty functionals. For example, a choice of $a = \log n$ would yield

$$ICOMP_{PEU\_LN}(\hat{\mathscr{F}}^{-1}) = -2\log L(\hat{\theta}_M \mid y) + m + \log(n)C_1(\hat{\mathscr{F}}^{-1}), \quad (39)$$

which clearly enforces a stricter penalty. One can choose yet other forms of the utility U_2 and its exponent to obtain several consistent forms of $ICOMP$, which are justifiable, to penalize overparametrization of the models under consideration. For more on $ICOMP$ we refer the readers to [8].

3.2. $ICOMP$ for Misspecified Models

In order to protect the researcher against this model misspecification, we generalize $ICOMP$ to the case of a misspecified model and introduce $ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$, which can drive effective model selection **even when the Gaussian assumption is invalid for the given dataset**. First we define the two forms of the Fisher information matrix which are useful to check misspecification of a model. We define the Hessian form of the Fisher information matrix as

$$\mathscr{F} = -\mathrm{E}\left(\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right) \tag{40}$$

and the outer-product form as

$$\mathscr{R} = \mathrm{E}\left(\frac{\partial \log L(\theta)}{\partial\theta}\cdot\frac{\partial \log L(\theta)}{\partial\theta'}\right), \tag{41}$$

where the expectations are taken with respect to the true but unknown distribution. Following [13], [17, page 237], [18, page 270], [19, page 391], [45], and others, suppose that the fitted model is incorrectly specified. Let $g(y \mid \theta^*)$ be the true model. Without knowing, suppose we fit $f(y \mid \theta)$ to a random sample $y_1,\ldots,y_n$ of n observations. Under mild conditions, the log likelihood of the fitted model

$$\log L(\theta) = \sum_{i=1}^{n} \log f(y_i \mid \theta) \tag{42}$$

is maximized at the MLE $\hat{\theta}$, and as $n \longrightarrow \infty$ the average maximized log likelihood function

$$\overline{\log L(\hat{\theta} \mid y)} = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid \hat{\theta}) \longrightarrow \int f(y \mid \theta_g^*)\log g(y)dy, \tag{43}$$

where θ_g^* is the value of θ that minimizes the KL information

$$KL = \int \log\left[\frac{g(y \mid \theta^*)}{f(y \mid \theta)}\right] g(y \mid \theta^*)dy, \tag{44}$$

with respect to θ. Thus θ_g^* is the “least bad” value of θ given the misspecified model. Taking the partial derivative of (43) w.r.t θ, we have

$$0 = \int \frac{\partial \log f(y \mid \theta)}{\partial\theta} g(y)dy, \tag{45}$$

with $\hat{\theta}$ obtained from the finite-sample version of the previous equation given by

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(y_i \mid \hat{\theta})}{\partial\theta}. \tag{46}$$

Now expansion of this about θ_g^* yields

$$\hat{\theta} \doteq \theta_g^* + \left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(y_i \mid \theta_g^*)}{\partial\theta\partial\theta'}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(y_i \mid \theta_g^*)}{\partial\theta}\right]. \tag{47}$$

Which provides us with

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(y_i \mid \theta_g^*)}{\partial\theta\partial\theta'} \xrightarrow{p} \left\{ \mathrm{E}\left(\frac{\partial^2 \log f(y \mid \theta_g^*)}{\partial\theta\partial\theta'}\right)\right\} = -\mathscr{F}(\theta_g^*), \tag{48}$$

the inner-product form of the FIM, and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(y_i \mid \theta_g^*)}{\partial\theta} \xrightarrow{p} \left\{ \mathrm{E}\left(\frac{\partial \log f(y \mid \theta_g^*)}{\partial\theta}\right)\cdot\left[\frac{\partial \log f(y \mid \theta_g^*)}{\partial\theta}\right]\right\} = \mathscr{R}(\theta_g^*) \tag{49}$$

the outer-product form of the FIM. These two forms are useful to check for model misspecification. This gives us the following result, with the derivation in the appendix.

Theorem 1. *Based on an iid sample, $y_1, \dots, y_n$, and assuming regularity conditions of the log likelihood function hold, we have*

$$\hat{\theta} \sim N(\theta_g^*, \mathscr{F}^{-1}\mathscr{R}\mathscr{F}^{-1}), \textit{ or } \sqrt{n}(\hat{\theta} - \theta_g^*) \sim N(0, \mathscr{F}^{-1}\mathscr{R}\mathscr{F}^{-1}). \tag{50}$$

Note that this tells us explicitly

$$Cov(\theta_g^*)_{Misspec} = \mathscr{F}^{-1}\mathscr{R}\mathscr{F}^{-1}, \tag{51}$$

which is called the *sandwich* or *robust* covariance matrix, since it is a correct variance matrix whether or not the assumed or fitted model is correct. Of course, in practice the true model and parameters are unknown, so we estimate this with

$$\widehat{Cov}(\hat{\theta}) = \hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1}. \tag{52}$$

If the model is correct, we must have $\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}} = I$, so

$$\widehat{Cov}(\hat{\theta}) = \hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1} = I\hat{\mathscr{F}}^{-1} = \hat{\mathscr{F}}^{-1}.$$

Thus, in the case of a correctly specified model, $\widehat{Cov}(\hat{\theta}) = \hat{\mathscr{F}}^{-1}$. However, when the model is misspecified, this is not the case. Under misspecification, several forms of $ICOMP$ previously defined are given as follows.

$$ICOMP(\widehat{Cov}(\hat{\theta}))_{MISP} = -2\log L(\hat{\theta} \mid y) + 2C_1(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1}). \tag{53}$$

$$ICOMP(\widehat{Cov}(\hat{\theta}))_{MISP\_PEU} = -2\log L(\hat{\theta} \mid y) + \mathrm{tr}(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}) + 2C_1(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1}). \tag{54}$$

$$ICOMP(\widehat{Cov}(\hat{\theta}))_{MISP\_PEU\_LN} = -2\log L(\hat{\theta} \mid y) + \mathrm{tr}(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}) + \log(n)C_1(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1}). \tag{55}$$

In the next section, we will see how m got transformed into $\mathrm{tr}(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}})$ in the latter two.

Result 1. *If*

$$ICOMP \neq ICOMP_{MISP}$$

we say the model is misspecified, or equivalently

$$C_1(\widehat{Cov}(\hat{\theta})) \neq C_1(\hat{\mathscr{F}}^{-1}).$$

3.3. Bias of the Penalty

When we assume that the true distribution does not belong to the specified parametric family of pdfs, that is, if the parameter vector θ of the distribution is unknown and is estimated by maximizing the likelihood, then it is not any longer true that the average of the maximized log likelihood converges to the expected value of the parameterized log likelihood. That is,

$$\frac{1}{n}\log L(\hat{\theta}\mid y)=\frac{1}{n}\sum_{i=1}^{n}\log f(y_i\mid\hat{\theta})\nrightarrow E_y\left(\log f(y\mid\hat{\theta})\right). \tag{56}$$

In this case, the asymptotic bias, b, between these two terms is given by

$$b=E_G\left(\frac{1}{n}\sum_{i=1}^{n}\log f(y_i\mid\hat{\theta})-\int_{\mathbb{R}}\log f(y\mid\hat{\theta})dG(y)\right)=\frac{1}{n}\operatorname{tr}(\mathscr{F}^{-1}\mathscr{R})+O(n^{-2}), \tag{57}$$

where the expectation is taken over the true distribution $G=\prod_{i=1}^{n}G(y_i)$ [22]. We note that $\operatorname{tr}(\mathscr{F}^{-1}\mathscr{R})$ is the well known *Lagrange-multiplier test statistic*. See, for example, [40, 20, 38]. Since we typically have MLE's and not true parameter values, we have to estimate the bias using

$$\widehat{b}=\frac{1}{n}\operatorname{tr}(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}). \tag{58}$$

Thus, we have: Generalized Akaike's information criterion, $GAIC$, defined by

$$\begin{aligned} GAIC &= -2\sum_{i=1}^{n}\log f(y_i\mid\hat{\theta})+2\operatorname{tr}(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}) \\ &= -2\log L(\hat{\theta}\mid y)+2\operatorname{tr}(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}). \end{aligned} \tag{59}$$

In the literature of model selection, $GAIC$ is also known as Takeuchi's [40] information criterion (TIC), or AIC_T.

When the model is correctly specified the asymptotic bias reduces to:

$$\begin{aligned} b &= \frac{1}{n}\operatorname{tr}(\mathscr{F}^{-1}\mathscr{R})+O(n^{-2}) \\ &= \frac{1}{n}\operatorname{tr}(I_m)+O(n^{-2}) \\ &= \frac{1}{n}m+O(n^{-2}), \end{aligned}$$

which gives AIC as a special case of $GAIC$ given by

$$AIC=-2\log L(\hat{\theta}\mid y)+2m,$$

where m is the the number of estimated parameters within the model. We note that the bias in AIC is fixed and has no variability. Other higher order bias correction in model selection criteria is possible, and their asymptotic properties are well explained in [23, p. 176]. **When the true model is not in the model set considered**, which is often the case in practice, AIC will have difficulties identifying the best fitting model, as it does not penalize the presence of *skewness* and *kurtosis*.

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.

In reviewing the literature, we note that Sawa's [see 36] BIC (should not be confused with SBC which is also sometimes referred to as 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 KL information. Sawa's penalty term is not an entropic function of the complexity of the estimated sandwich covariance matrix of the model. As shown above, the $ICOMP$ criterion can be seen as an approximation to the sum of two KL distances. Similarly, $ICOMP$ is not necessarily related to Wei's [see 43, page 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 covariance matrix in $ICOMP$. FIC is more related to the *Predictive Least Squares* (PLS) criterion, as [43] demonstrates.

4. Information Complexity for the Multivariate Regression Models

4.1. $ICOMP$ and Information Criteria for Correctly Specified MVR Models

For multivariate regression, the number of estimated parameters is $m = pq + p(p+1)/2$; we have a coefficient for each of the q covariates for each of the p responses, and allow for a fully general covariance matrix, which has $p(p+1)/2$ unique elements. Thus, the AIC-type criteria are given as

$$AIC = np\log(2\pi) + n\log|\hat{\Sigma}| + np + 2(pq + \frac{p(p+1)}{2}), \text{ and} \tag{60}$$

$$SBC = np\log(2\pi) + n\log|\hat{\Sigma}| + np + \log(n)(pq + \frac{p(p+1)}{2}). \tag{61}$$

The typical $ICOMP(\hat{\mathscr{F}}^{-1})$ is

$$ICOMP(\hat{\mathscr{F}}^{-1}) = np\log(2\pi) + n\log|\hat{\Sigma}| + np + 2C_1(\hat{\mathscr{F}}^{-1}). \tag{62}$$

Rather than compute and store the entire IFIM, we can compute $C_1(\hat{\mathscr{F}}^{-1})$ after "opening it up" as shown in (63):

$$\begin{aligned} ICOMP(\hat{\mathscr{F}}^{-1}) &= -2\log L(\hat{\theta} \mid Y, X) + 2C_1(\hat{\mathscr{F}}^{-1}) \\ &= np\log(2\pi) + n\log|\Sigma| + np \end{aligned}$$

$$
\begin{aligned}
&+ \quad s\log(\frac{1}{s}\left[\operatorname{tr}(\hat{\Sigma})\operatorname{tr}[(X'X)^{-1}] + \frac{1}{2n}(\operatorname{tr}(\hat{\Sigma}^2) + \operatorname{tr}(\hat{\Sigma})^2 + 2\sum_{j=1}^{p}(\sigma_{jj}^2)^2)\right])\\
&- \quad (p+q)\log|\hat{\Sigma}| + p\log|X'X| + \frac{p(p+1)}{2}\log(n) - p\log(2), \qquad (63)
\end{aligned}
$$

where $(\sigma_{jj}^2)^2$ indicates the square of the j$^{\text{th}}$ diagonal element of $\hat{\Sigma}$. To compute $ICOMP(\hat{\mathscr{F}}^{-1})_{PEU}$, one would simply add m to (63). For $ICOMP(\hat{\mathscr{F}}^{-1})_{PEU\_LN}$, multiply the $ICOMP(\hat{\mathscr{F}}^{-1})$ penalty by $\log(n)/2$, then add m.

4.2. $ICOMP$ for Misspecified MVR Models

Model misspecification is an important issue with regression models. To protect the researcher against model misspecification, we need $\widehat{Cov}(\hat{\theta}) = \hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1}$. $\hat{\mathscr{F}}^{-1}$ is repeated in (64)

$$
\hat{\mathscr{F}}^{-1} = \begin{bmatrix} \hat{\Sigma}\otimes(X'X)^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2}{n}D_p^{+}(\hat{\Sigma}\otimes\hat{\Sigma})D_p^{+\prime} \end{bmatrix}, \qquad (64)
$$

$\hat{\mathscr{R}}$ is derived in the appendix, and we show the results here.

$$
\hat{\mathscr{R}} = \begin{bmatrix} \hat{\Sigma}^{-1}\otimes X'X & \frac{1}{2}(\hat{\Sigma}^{-1/2}\otimes X')\hat{\Gamma}_1 D_p^{+\prime}\Delta \\ 1/2\Delta D_p^{+}\hat{\Gamma}_1'(\hat{\Sigma}^{-\frac{1}{2}}\otimes X) & \frac{1}{4}\Delta D_p^{+}\hat{\Gamma}_2^{*}D_p^{+\prime}\Delta \end{bmatrix} \qquad (65)
$$

This matrix takes into consideration the actual sample skewness

$$
\hat{\Gamma}_1 = \operatorname{vec} Z\left[\operatorname{vec}(Z'Z - nI_p)\right]', \qquad (66)
$$

and kurtosis

$$
\hat{\Gamma}_2^{*} = (\operatorname{vec} Z'Z)(\operatorname{vec} Z'Z)' + n^2(\operatorname{vec} I_p)(\operatorname{vec} I_p)'. \qquad (67)
$$

We define

$$
\Delta = D_p'(\hat{\Sigma}^{-1/2}\otimes\hat{\Sigma}^{-1/2})D_p,
$$

and we standardize the response matrix with

$$
Z = (Y - X\hat{\beta})\hat{\Sigma}^{-1/2}.
$$

The vec$(\cdot)$ operator stacks the columns of a matrix on top of each other, such that

$$
Z \in \mathbb{R}^{n\times p} \longrightarrow \operatorname{vec} Z \in \mathbb{R}^{np\times 1},
$$

and $\hat{\Sigma}^{-1/2}$ indicates the inverse of the matrix square root defined by

$$
\hat{\Sigma}^{1/2}\hat{\Sigma}^{1/2} = \hat{\Sigma}.
$$

In cases where the model is correctly specified, we have

$$
E \sim N_{np}(\mathbf{0}, \Sigma\otimes I_n), \qquad (68)
$$

Γ_1 reduces to $\mathbf{0}$, and $\Gamma_2^* \longrightarrow 2nD_pD_p^+$, such that $\hat{\mathscr{R}} = \hat{\mathscr{F}}$, in theory.

Thus, the misspecification-resistant estimator of the model covariance matrix is given in (69)

$$\begin{aligned} \widehat{Cov}(\hat{\theta}) &= \hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1} \\ &= \begin{bmatrix} \hat{\Sigma}\otimes(X'X)^{-1} & \frac{1}{n}(\hat{\Sigma}^{1/2}\otimes(X'X)^{-1}X')\hat{\Gamma}_1D_p\Delta^{-1} \\ \frac{1}{n}\Delta^{-1}D_p'\hat{\Gamma}_1'(\hat{\Sigma}^{1/2}\otimes X(X'X)^{-1}) & \frac{1}{n^2}\Delta^{-1}D_p'\hat{\Gamma}_2^*D_p\Delta^{-1} \end{bmatrix}. \end{aligned} \tag{69}$$

A procedure for "opening up" $\widehat{Cov}(\hat{\theta})$ is shown in Appendix 2. As with the regular $ICOMP$, this procedure simplifies the computation, since the entire covariance matrix does not have to be built and stored.

There is an issue of matrix stability to be addressed with the sandwich covariance matrix, however. In both simulated and real datasets, we have observed that the matrix is consistently rank-deficient. We performed simulation studies similar to those detailed in Section 7.2, in which the model was correctly specified, and observed that $ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$ and $ICOMP(\hat{\mathscr{F}}^{-1})$ did not, in fact, select similar models. It appears that numerical issues with the sandwich covariance matrix prevent it from approximating the IFIM when the model is correctly specified. As an example, consider a dataset in which $p = 2$ and $q^* = 8$. The number of parameters estimated is $m = 19$, and the model covariance matrix is of size 19×19. However, the rank is only 16. Of course, the determinant is 0. We employ the robust covariance estimators (discussed in Section 6) to ensure $\widehat{Cov}(\hat{\theta})$ is of full rank.

4.3. KL Information between True and Fitted Model

Suppose that we denote a true multivariate regression model by

$$M_t : Y = X^*B^* + E^*,\ Cov(vecY^*) = \Sigma^* \otimes I_n \tag{70}$$

and the fitted (or candidate) multivariate regression model by

$$M_f : Y = XB + E,\ Cov(vecY) = \Sigma \otimes I_n. \tag{71}$$

Under the multivariate normal assumption, the log likelihood of the true model M_t is

$$\log L(M_t) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma^*| - \frac{1}{2}\text{tr}[(Y - X^*B^*)\Sigma^{*-1}(Y - X^*B^*)']. \tag{72}$$

The log likelihood of the fitted model M_f is

$$\log L(M_f) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\text{tr}[(Y - XB)\Sigma^{-1}(Y - XB)']. \tag{73}$$

Then the log likelihood of the difference between the true model and the fitted model becomes

$$\Delta \log L(M_t, M_f) = \log L(M_t) - \log L(M_f)$$

$$= \log\left(\frac{|\Sigma|}{|\Sigma^*|}\right) + \frac{1}{n}\operatorname{tr}[(Y - XB)\Sigma^{-1}(Y - XB)'] - \frac{1}{n}\operatorname{tr}[(Y - X^*B^*)\Sigma^{*-1}(Y - X^*B^*)']. \quad (74)$$

Now, taking the expectation with respect to the true model, we obtain the Kullback-Leibler distance as [5]:

$$\begin{aligned} KL &= E\left(\Delta \log L(M_t, M_f)\right) \\ &= \log\left(\frac{|\Sigma|}{|\Sigma^*|}\right) + \operatorname{tr}(\Sigma^{-1}\Sigma^*) + \frac{1}{n}\operatorname{tr}[(X^*B^* - XB)\Sigma^{-1}(X^*B^* - XB)'] - p. \end{aligned} \quad (75)$$

Using the maximum likelihood estimators

$$\hat{B} = (X'X)^{-1}X'Y, \text{ and } \hat{\Sigma} = \frac{(Y - X\hat{B})'(Y - X\hat{B})}{n} = \frac{Y'MY}{n}, \quad (76)$$

we have the estimated KL

$$\widehat{KL} = \log\left(\frac{|\hat{\Sigma}|}{|\Sigma^*|}\right) + \operatorname{tr}(\hat{\Sigma}^{-1}\Sigma^*) + \operatorname{tr}[\hat{\Sigma}^{-1}(X^*B^* - X\hat{B})(X^*B^* - X\hat{B})'] - p. \quad (77)$$

This gives us a yardstick in comparing the performance of model selection criteria between the true and the fitted model in how close they are to the KL distance, especially in simulation protocols, since the true model is known by design. By means of simulation study, we can investigate the finite sample behavior of $ICOMP$ criteria both when the model is correctly specified, and the model is misspecified.

5. Genetic Algorithm (GA)

The GA is a search algorithm that borrows concepts from biological evolution. Biological chromosomes, which determine so much about organisms, are represented as binary words – these determine the composition of possible solutions to an optimization problem. Unlike most search algorithms, the GA simulates a large population of potential solutions. These solutions are allowed to interact over time; random mutations and natural selection allow the population to improve, eventually iterating to an optimal solution. For multivariate regression subsetting, each chromosome is a q-length vector such that each locus represents the presence (1) or absence (0) of a specific predictor. An example chromosome may be [10011001]; in this case, predictors 1,4,5,8 will be used for OLS while 2,3,6,7 will not. The general procedure in the GA is simple and straightforward:

1. Generate initial population of chromosomes
2. Score all members of current population
3. Determine how current population is mated and represented in next generation
4. Perform chromosomal crossover and genetic mutation
5. Pass on offspring to new generation
6. Loop back to step 2 until termination criteria met

As seen in Table 1, there are 8 major parameters used to define the operation of the genetic algorithm. These are all discussed, along with implementation methods, below.

Table 1: Sample Genetic Algorithm parameters.

Parameter	Setting
Number Generations	60
Population Size	30
Generation Seeding	Roulette
Crossover Probability	0.75
Mutation Probability	0.10
Objective Function	Information Criterion

5.1. Number Generations

Each iteration in the GA is called a generation, for obvious reasons. Thus, this parameter is fairly self explanatory. There is an important trade-off to note, when selecting the number of generations through which the GA will run. More generations mean more computation time; however, not allowing the process to go through enough iterations can mean termination with a suboptimal result.

5.2. Population Size

This parameter, P, determines the number of chromosomes are considered in each generation. In general, one would expect convergence times to decrease as population size increases, up to a point. After that point, the computational time (and hence, time to convergence) increases quickly. In other unpublished research, performance of the GA when used for multivariate subsetting was analyzed as operational parameters were purposely varied. In this case, the goal was to maximize the frequency, across all generations, with which the procedure selected the optimal subset. Tests were performed on a dataset for which the optimal subset was known. It was determined that the GA was robust to all parameters tested (not all parameters used here were varied) excepting population size; though in this case, variation in population size only explained half the variation in the response.

5.3. Generation Seeding

From a given population, how do we seed the members of the next generation in preparation for mating? There are two primary methods; in both, the first step is to sort the current population by the objective function values, such that the “most fit” chromosomes are at the beginning of the list. For the simpler “sorted” method, no more preparation is necessary. Chromosomes are mated in sequential pairs (mate(1,2), mate(3,4), etc...). The second method is akin to a biased roulette wheel, in which the individual bins are of varied size as in Figure 4. Bins for all chromosome are computed as $b_i = 2i/(P(P+1)) \mid b_i \in [0,1]$, then a cumulative

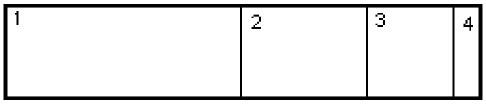

Figure 4: Conceptual roulette "wheel".

sum of these bins is computed. As an example, consider the sorted list of 4 chromosomes - the bin widths are given as $b_i = \begin{bmatrix} 0.40 & 0.30 & 0.20 & 0.10 \end{bmatrix}$, so the bin limits are

bin	1	2	3	4
Lower Limit	0.00	0.40	0.70	0.90
Upper Limit	0.40	0.70	0.90	1.00

.

Clearly, the larger bins are at the beginning, corresponding to the most fit chromosomes. At this point, P random numbers are generated uniformly from $[0,1]$ and placed in the appropriate bin. For each random variate in bin i, chromosome i gets represented in the next generation. In this way, chromosomes with a better objective function value are overrepresented in the mating pool. The last step, then, is to randomly permute the ordering of the chromosomes before mating; after permutation, mating occurs just as in the sorted method.

5.4. Crossover Probability

There are several ways in which crossover can be implemented, including:

- Single-point (fixed or random)
- Multiple-point (fixed or random)
- Uniform (fixed or random)

We've chosen to use the simplest - randomized single-point crossover. For each mating pair, a random uniform variate is selected from integers in the range $[2, q-1]$, this range is used, rather than $[1, q]$, to protect against endpoint crossovers. For a given pair of mating chromosomes, their right-most portions are traded starting from this point. For example, if the crossover point is 2, we have

$$\begin{vmatrix} 1 & 1 & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ 1 & 0 & 0 & 1 & 0 \end{vmatrix} \longrightarrow \begin{vmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & \mathbf{1} & \mathbf{0} & \mathbf{1} \end{vmatrix}.$$

For each mating pair, a random variate from $U(0,1)$ is generated; if it is less than the crossover probability, the mating pair undergo the crossover operation. Otherwise, the offspring are pure genetic replicants of the parents.

5.5. Mutation Probability

The mutation operation is simple. Based on the mutation rate, chromosomes are randomly selected from the current population to undergo mutation. For each chromosome, loci are randomly selected (uniformly), at the same mutation rate, and their bits are switched: $1 \longrightarrow 0$ or $0 \longrightarrow 1$ (a *not* operation).

5.6. Objective Function

More or less, all optimization or search procedures need some objective function to either maximize or minimize. There is a large universe of appropriate functions that could fill this role. For this problem we've chosen to use a class of criteria called information criteria. Specifically, we present the use of Bozdogan's $ICOMP$ which drives effective model selection in the face of model misspecification.

6. Robust Covariance Estimation

In many real-life problems, covariance matrices can become ill-conditioned, non-positive definite, or even singular. This is especially true in cases of regression with highly collinear predictors. It is also seen in situations in which there are not many more observations than there are measurements, or variables - i.e., when it is not the case that $n \gg p$. The usual response to singular or ill-conditioned covariance matrix estimates is ridge regularization,

$$\hat{\Sigma}^{*} = \left[\hat{\Sigma} + \alpha I_p\right], \tag{78}$$

which works to counteract the ill-conditionedness by adjusting the eigenvalues of $\hat{\Sigma}$. Usually, the ridge parameter, α, is chosen to be very small. This, of course, begs the questions

- "*How large should α be?*"
- "*How small can α be?*".

The answer to these questions is to use robust covariance estimators. Many different robust, or smoothed, covariance estimators have been developed as a way to data-adaptively improve ill-conditioned and/or singular covariance matrix estimates. Several of them work by the same mechanism as ridge regularization - perturb the diagonals, and hence, the eigenvalues. Although we use only the MLE/EB covariance estimator in our reported results, several methods implemented in our algorithm include:

- Maximum Likelihood / Empirical Bayes

$$\hat{\Sigma}_{MLE/EB} = \hat{\Sigma} + \frac{p-1}{(n)\,\mathrm{tr}(\hat{\Sigma})} I_p \tag{79}$$

- Stipulated Ridge, [39]

$$\hat{\Sigma}_{SRE} = \hat{\Sigma} + p(p-1)\left[(2n)\,\mathrm{tr}(\hat{\Sigma})\right]^{-1} I_p \tag{80}$$

- Stipulated Diagonal, [39]

$$\hat{\Sigma}_{SDE} = (1-\pi)\hat{\Sigma} + \pi diag(\hat{\Sigma}),\ \pi = p(p-1)\left[2n(\mathrm{tr}(\hat{\Sigma}^{-1}) - p)\right]^{-1} \tag{81}$$

- Convex Sum, [31, 11]

$$\hat{\Sigma}_{CSE} = \frac{n}{n+m}\hat{\Sigma} + (1 - \frac{n}{n+m})\left[\frac{\mathrm{tr}(\hat{\Sigma})}{p}\right] I_p \tag{82}$$

$$0 < m < \frac{2\left[p(1+\beta) - 2\right]}{p-\beta},\ \beta = \frac{\mathrm{tr}(\hat{\Sigma})^2}{\mathrm{tr}(\hat{\Sigma}^2)}$$

- Thomaz Stabilization [41]

$$\hat{\Sigma}_{Thomaz} = V\Lambda^{*}V \tag{83}$$

$$\Lambda^{*} = \begin{bmatrix} \max(\lambda_1,\overline{\lambda}) & & & \mathbf{0} \\ & \max(\lambda_2,\overline{\lambda}) & & \\ & & \ddots & \vdots \\ \mathbf{0} & & & \max(\lambda_p,\overline{\lambda}) \end{bmatrix} \tag{84}$$

$$\lambda_i = \text{i}^{\text{th}} \text{ eigenvalue}, \overline{\lambda} = \text{mean eigenvalue}, V = \text{matrix of eigenvectors}$$

When a small amount of perturbation is all that is required, $\hat{\Sigma}_{MLE/EB}$ has a certain appeal. As is clear in (79), this is the same form of the naive ridge regularization, where α is determined by the data.

For the results reported here, we used the Empirical Bayes estimator, so as to perturb the estimates as little as possible. Even then, we prefer to not change the problem more than necessary, so we perform two tests for matrix condition:

1. Is the reciprocal of the condition number small: $\kappa^{-1}(\hat{\Sigma}) \leq 1e^{-10}$?
2. Is the MLE nonpositive definite?

If the answer to either question is in the affirmative, we apply the robust covariance estimator to give us a well-conditioned estimate, $\hat{\Sigma}^{*}$.

7. Numerical Results

7.1. Simulation with Correlated Redundant Variables

We first demonstrate the performance of misspecification-robust information criteria on a simulated dataset using a complex simulation protocol. We begin by generating five correlated regressors, with x_4 and x_5 redundant.

$$\begin{aligned} x_0 &= 1 \text{ (constant)} \\ x_1 &= 10+\varepsilon_1 \\ x_2 &= 10+\rho\varepsilon_1+\alpha\varepsilon_2 \\ x_3 &= 10+\rho\varepsilon_1+0.5604\alpha\varepsilon_2+0.8282\alpha\varepsilon_3 \\ x_4 &= -8+x_1+0.5x_2+\rho x_3+0.5\varepsilon_4 \\ x_5 &= -5+0.5x_1+x_2+0.5\varepsilon_5 \end{aligned}$$

The ε_i are drawn from a standardized Gaussian distribution, $\rho = 0.3$ controls the correlation, as does $\alpha = \sqrt{1-\rho^2}$. Thus far, we have $X = \left[x_0, x_1, x_2, x_3, x_4, x_5\right]$. We now turn our focus to the response matrix, which we create as:

$$\begin{aligned} Y_{n\times 2} &= X^{*}_{n\times 4}B_{4\times 2} + \epsilon_{n\times 2} \text{ where} \\ X^{*} &= \left[x_0, x_1, x_2, x_3\right] \text{ and} \end{aligned}$$

$$B = \begin{bmatrix} -8 & -5 \\ 1.0 & 0.5 \\ 0.5 & 0.0 \\ 0.3 & 0.3 \end{bmatrix}.$$

To make this a truly misspecified model, we generate the error terms from the multivariate power exponential distribution (MPE). From [16], the density function for the MPE is shown in (85).

$$f(x_i \mid \mu, \Sigma, \beta) = \frac{p\Gamma(\frac{p}{2})|\Sigma|^{-\frac{1}{2}}}{2\pi^{\frac{p}{2}}\Gamma(1+\frac{p}{2\beta})2^{1+\frac{p}{2\beta}}} \exp(-\frac{1}{2}\left[(x_i-\mu)\Sigma^{-1}(x_i-\mu)'\right]^{\beta}) \tag{85}$$

This distribution includes others as special cases:

- When $\beta = 1$, we have the *multivariate Gaussian*
- When $\beta = 1/2$, we have the *multivariate Laplace*
- When $\beta \to \infty$, we have the *multivariate uniform*

We generate the error terms two different ways:

S1: $\mu = [0,0], \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \beta = 0.75$	S2: $\mu = [0,0], \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \lambda = [2,-2]$

In simulation S2, the error terms matrix is created as two independent columns of skewed univariate power exponential distributions, with one skewed right and the other being left-skewed. The skewness is imparted utilizing Azzalini-type skew, with the functional form shown in (86) - see [2]. Figure 5 shows two examples.

$$g(x) = 2f(x)F(\lambda x) \sim Skew(f, \lambda) \tag{86}$$

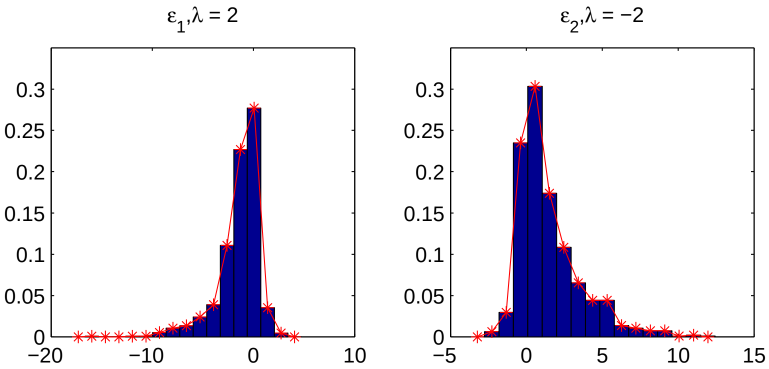

Figure 5: Demonstrating the univariate Azzalini-type skewed PE.

For our first Monte Carlo study, we let $\beta = 0.75$ and ran $M = 500$ trials with both $n = 500$ and $n = 1000$. The results are summarized in Tables 2 and 3. Table 2 summarizes re-

Table 2: Abbreviated % model hits out of $M = 500$ simulations of S1 with $n = 500$.

Criteria	True Model
KL Distance	100.0
$AIC/GAIC$	74.8
SBC	72.8
$ICOMP(\hat{\mathscr{F}}^{-1})$	72.4
$ICOMP_{PEU}(\hat{\mathscr{F}}^{-1})$	73.4
$ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$	85.8
$ICOMP_{MISP\_PEU}(\widehat{Cov}(\hat{\theta}))$	82.6

Table 3: % model hits out of $M = 500$ simulations of S1 with $n = 1000$.

Subset	$AIC/GAIC$	SBC	$ICOMP(\hat{\mathscr{F}}^{-1})/PEU$	$ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))/PEU$
$\{0,1,2,3,4,5\}$	1.6	0.0	0.0/0.0	0.0/0.0
$\{1,2,3,4,5\}$	0.0	0.0	0.4/0.2	0.0/0.0
$\{0,1,3,4,5\}$	0.2	0.0	0.0/0.0	0.0/0.0
$\{0,1,2,3,5\}$	11.0	0.2	1.2/0.4	2.0/0.8
$\{0,1,2,3,4\}$	12.2	0.0	0.0/0.0	1.2/0.8
$\{0,1,3,5\}$	0.0	0.4	0.2/0.2	0.4/0.4
$\mathbf{\{0,1,2,3\}}$	**75.0**	**98.2**	**98.2/99.2**	**96.2/97.6**
$\{0,1,2\}$	0.0	1.2	0.0/0.0	0.2/0.4

sults from the experiment with $n = 500$ observations generated from the simulation protocol. There are, of course, $2^6 - 1 = 63$ possible subsets of the covariates; not counting subsets never selected, there were too many to show in detail here. As such, we're reporting the percent of simulations in which the criteria selected exactly the true model $X^* = \left[x_0, x_1, x_2, x_3\right]$. It is interesting to note that AIC and $GAIC$ performed identically. We also see that AIC seems more appropriate for smaller samples. $ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$ performed well, only selecting a model that did not include the true model in 11% trials. The misspecified $ICOMPs$ hit the true model with the highest frequency of all the criteria.

Next we have the results from the experiment with $n = 1000$ observations in Table 3. As can be seen, the rates at which all forms of $ICOMP$ selected exactly the true model improved dramatically, as good as 99.2% for $ICOMP_{PEU}(\hat{\mathscr{F}}^{-1})$. In this table, only subsets that had any hits at all are included, in the interest of space. It seems that, with a large enough sample, $ICOMP(\hat{\mathscr{F}}^{-1})$ is robust to a degree of misspecification. SBC also demonstrated a dramatic improvement; though it is considered to be a "consistent" criteria, this quality seems to be somewhat lacking when the functional form of the regression model is not correctly specified. It is interesting to note that the hit percentages for AIC and $GAIC$ did not improve much at all, rising slightly to 75.0%. As with the smaller sample size, the KL distance selected the true model 100.0% of the time.

Table 4: Abbreviated % model hits out of $M = 500$ simulations of S2.

Criteria	True Model	
	$n = 500$	$n = 1000$
KL Distance	57.2	67.6
$AIC/GAIC$	64.6	73.0
SBC	29.6	86.8
$ICOMP(\hat{\mathscr{F}}^{-1})$	58.0	93.0
$ICOMP_{PEU}(\hat{\mathscr{F}}^{-1})$	53.2	93.6
$ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$	59.0	91.8
$ICOMP_{MISP\_PEU}(\widehat{Cov}(\hat{\theta}))$	46.0	90.2

Next, we performed to sets of simulation experiments with the S2 model for $n = 500$ and $n = 1000$. In the presence of asymmetry there was a lot more confusion, with 38 different models being selected by different criteria. Table 4 summarizes the results.

In the presence of skewness, AIC and $GAIC$ performed admirably in the tests with the smaller sample size, while SBC did not. The $ICOMP$ criteria performed very well, especially with the larger sample sizes. $ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$ proved better than even the KL distance at both selecting the true model, and a model that included the truth. The KL distance performed much worse than it did when the errors were generated symmetrically. We also note that, like the S1 simulation, $ICOMP(\hat{\mathscr{F}}^{-1})$ is robust against a high degree of skewness, with performance similar to the misspecified form of $ICOMP$.

In summary, we observe that in all experiments, AIC and $GAIC$ performed identically. This suggests that considering just estimated bias does not provide much benefit. Effectively adjusting for model misspecification seems to require the entire sandwich covariance matrix.

7.2. Simulation with Redundant and Unnecessary Variables

This simulation protocol, with the multicollinearity and non-Gaussian errors, is a decent test of the performance of information criteria in the presence of misspecification, but we can do better. Thus, we add 15 unrelated variables to our matrix of regressors, such that

$$X \in \mathbb{R}^{n\times 21} = \left[x_0, x_1, x_2, x_3, x_4, x_5, \{x_i \sim U(0,i) \mid i = 6\ldots 20\}\right].$$

Under this extended simulation protocol, we have $2^{21} - 1 = 2,097,151$ different possible models to evaluate, and the true model is still $X^* = \left[x_0, x_1, x_2, x_3\right]$. This is a very difficult problem for any criteria to perform well under. We use the genetic algorithm to search the subset space with the settings: population size = 30, number generations = 60, crossover rate = 0.75, mutation rate = 0.10. The four Monte Carlo simulation experiments from the previous section, all with $M = 100$, were performed.

For $\beta = 0.75$, as can be seen in Table 5 the KL distance, assuming normality, gave up its

Table 5: Model selection statistics from extended S1 simulations.

Criteria	n	Model Selected	True Model	Avg Length
KL Distance	500	$\{0,1,2,3\}$	86.0	4
	1000	$\{0,1,2,3\}$	92.0	4
AIC	500	$\{0,1,2,3,6,10,12,19\}$	6.0	7
	1000	$\{0,1,2,3,6,12,14,15\}$	4.0	6
SBC	500	$\{0,1,2\}$	54.0	4
	1000	$\{0,1,2,3\}$	94.0	4
GAIC	500	$\{0,1,2,3,5,16,17\}$	7.0	7
	1000	$\{0,1,2,3,15\}$	4.0	6
$ICOMP$	500	$\{1,2,3,4\}$	11.0	7
	1000	$\{1-8,10,14,17,18\}$	63.0	5
$ICOMP_{PEU}$	500	$\{1,4,5\}$	20.0	5
	1000	$\{0,1,2,3\}$	70.0	5
$ICOMP_{MISP}$	500	$\{0,1,2,3\}$	77.0	4
	1000	$\{0,1,2,3\}$	89.0	4
$ICOMP_{MISP\_PEU}$	500	$\{0,1,2,3\}$	75.0	4
	1000	$\{0,1,2,3\}$	93.0	4

perfect track record - at best selecting the true model in 92% of the trials. With more information in the tails, the misspecified $ICOMP$ criteria performed well regardless of the sample size. In fact, minimization of both criteria across all simulations selected $X^* = \left[x_0, x_1, x_2, x_3\right]$ as the best model The PEU version did very well, hitting the true model in 93 of the 100 simulations for the larger sample size. Over all simulations, these two criteria and SBC selected models, on average, with 4 regressors - there was no substantial tendency to overfit. Once again, though, we see the inconsistent behavior of SBC, only doing really well with the large sample. Both AIC and $GAIC$ performed extremely poorly, both tending to pick models with 2 or 3 extra predictor variables. To their credit, they also displayed a strong tendency to pick models that included the true model (eg. $x^* = \left[x_0, x_1, x_2, x_3, x_4, x_5\right]$). The final model selected by $ICOMP(\hat{\mathscr{F}}^{-1})$ for the large sample simulation is truly bizarre - 12 regressors, while the average model length was a respectable 5. This seems like it may have been a vagary of simulation. In Table 6, we see the results from one of the many simulations in which the

Table 6: Summary from one simulation of the extended S1 protocol.

Subset	$ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$	Weights	# Generations
$\mathbf{\{0,1,2,3\}}$	**6789.73**	**0.948**	**30**
$\{0,1,2,3,5\}$	6796.30	0.035	25
$\{0,1,2,3,5,14\}$	6797.80	0.017	1
$\{0,1,2,3,5,9,14\}$	6808.76	0.000	1
$\{0,1,2,3,4,9,14\}$	6819.01	0.000	1
$\{0,1,2,3,5,9,14,18,19\}$	6828.54	0.000	2

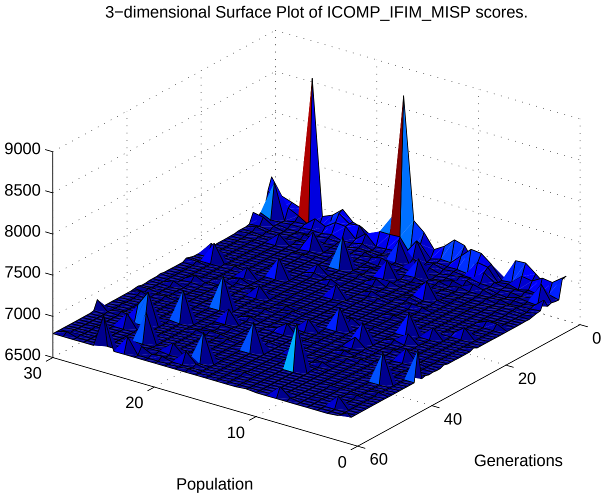

Figure 6: 3-d Surface plot of $ICOMP_{MISP}(\widehat{Cov}(\hat{\theta})$.

misspecified version of $ICOMP$ selected X^* as the best model. Here we see that the criteria was almost 95% certain that its solution was the true model. Finally, we see, in Figure 6, how much the GA smoothed the search surface, with relatively rare spikes in the $ICOMP$ scores. Though this plot doesn't show it, the simulation found it's final solution in the 30th generation.

Secondly, we performed 100 simulations with $n = 500$ and 100 simulations with $n = 1000$ from the extended protocol with the skewed error terms. Results from these simulations can be seen in Table 7. It seems noteworthy that the KL distance, assuming normality, did not perform substantially better with the increase in sample size. With the smaller sample size, $ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$ did the best job of hitting the true model, at 42%; minimizing both SBC and $ICOMP_{MISP\_PEU}(\widehat{Cov}(\hat{\theta}))$ selected the true model. However, when the sample size was increased to $n = 1000$, both misspecified forms of $ICOMP$ had the highest hit rates of 80% and 73%, besting the 58% obtained by the KL distance. The only criteria which selected, via minimization, the true model were three of the four $ICOMPs$.

With these Monte Carlo Simulation studies, we've shown that, in multivariate regression, when the error terms exhibit nonnormal skewness or peakedness, the forms of $ICOMP$ which account for the sample characteristics can consistently pick the correct, or truly best fitting model.

Table 7: Model selection statistics from simulations of extended S2.

Criteria	n	Model Selected	True Model	Avg Length
KL Distance	500	$\{0,1,2,3\}$	48.0	4
	1000	$\{0,1,2,3\}$	58.0	4
AIC	500	$\{0,1,2,3,6,8\}$	6.0	6
	1000	$\{0,1,2,3,14,16\}$	8.0	6
SBC	500	$\{0,1,2,3\}$	17.0	2
	1000	$\{0,1,3,5\}$	63.0	4
GAIC	500	$\{0,1,2,3,4,12,19\}$	5.0	6
	1000	$\{0,1,2,3,5,14,15\}$	3.0	6
$ICOMP$	500	$\{1,3,4,5,6,9\}$	6.0	6
	1000	$\{0,1,2,3,5\}$	47.0	5
$ICOMP_{PEU}$	500	$\{1,3,4,5,9,13,14\}$	9.0	4
	1000	$\{0,1,2,3\}$	49.0	5
$ICOMP_{MISP}$	500	$\{0,1,3\}$	42.0	4
	1000	$\{0,1,2,3\}$	80.0	5
$ICOMP_{MISP\_PEU}$	500	$\{0,1,2,3\}$	38.0	3
	1000	$\{0,1,2,3\}$	73.0	4

7.3. GA Performance under Simulations

Regarding the performance of the GA, recall that, for the extended simulation protocol, there were over two million possible models to evaluate. Each simulation evaluated 30 (not necessarily unique) solutions per generation; and was allowed to progress through 60 generations. Therefore, each simulation evaluated **at most** $30 \times 60 = 1,800$ subset regression models. With the exception of the $ICOMP_{MISP}$ criteria, computation time for each simulation hovered in the $(1.5s, 3.5s)$ range - $(8.5s, 25s)$ for the misspecified criteria. That some of these model statistics could consistently pick out the correct model while only evaluating at most $100 \times (1800/2,097,151) = 0.086\%$ of the subset space, in such a short time, is phenomenal. In fact, to demonstrate this performance and scalability, we performed a series of timing experiments. We set generation count = 60 and used the GA to perform subset regression while varying q, with $P = q + 10$ (P = population size). So as to remove any timing effect of sampling of misspecification, simulations were performed with $\beta = 1.0$ and $n = 500$; for this test, we used SBC to drive model selection. Ten simulations at each level of q were performed, and the average process time was computed; results are summarized in Table 8. While the number of possible models grows exponentially, the number of models evaluated only grows linearly. Note also how well the GA seems to scale in regard to computation time. A 5-fold increase in q only leads to something like a 4-fold increase in time. Keep in mind, of course, that the number of *unique* models evaluated was actually much less than shown. Additionally, the GA was artificially constrained to not allow early termination. Thus, the times are most likely inflated from what they would be in practical use. These results were obtained using our own software written for Mathworks MATLAB$^{®}$, on a non-dedicated **Windows XP PC with a 3.4**

Table 8: Timing test results.

q	P	Possible Models	Models Evaluated per Simulation	Average Time
20	30	$1,048,575$	$\leq 1,800$	$1.89s$
25	35	$33,554,431$	$\leq 2,100$	$1.82s$
35	45	$34,359,738,367$	$\leq 2,700$	$2.34s$
50	60	$1,125,899,906,842,623$	$\leq 3,600$	$2.89s$
75	85	$3.7779e^{22}$	$\leq 5,100$	$4.84s$
100	110	$1.2677e^{30}$	$\leq 6,600$	$7.99s$

GHz processor and 2 GB of RAM while many other processes were running.

7.4. Real Data - Body Fat Measurement

The first real dataset to be analyzed was the familiar body fat dataset This data is composed of body measurement observations from $n = 252$ men. There are $p = 2$ responses and $q = 13$ regressors, listed in Table 9. Where not specified, measurements are in centimeters. Accurately measuring body composition, and specifically the percentage that is fat, is an inconvenient and costly procedure. A method for accurately computing these amounts from simple body measurements without requiring underwater weighing is highly desirable.

Table 9: Body fat dataset variables.

y_1 =Body Density (gm/cm^3)	y_2 =Percent body fat from Siri's equation
x_0 =Constant	x_7 =Hip circumference
x_1 =Age (yrs)	x_8 =Thigh circumference
x_2 =Weight (lbs)	x_9 =Knee circumference
x_3 =Height (in)	x_{10} =Ankle circumference
x_4 =Neck circumference	x_{11} =Extended biceps circumference
x_5 =Chest circumference	x_{12} =Forearm circumference
x_6 =Abdomen 2 circumference	x_{13} =Wrist circumference

We first ran the genetic algorithm with population size = 20, using the $ICOMP(\hat{\mathscr{F}}^{-1})$ as the objective function. In both real datasets analyzed in this paper, we opted not to use SBC, due to the fact that the sample sizes are small, when compared to the $n = 500$ in the simulations. Given that its consistency appears to depend on the type of departure from normality, it seems dangerous to use unless with very large datasets. Minimizing $ICOMP$ led to the best model being: *a constant*, *Weight*, and *Abdomen 2 circumference*, with the estimated coefficients shown in Table 10. The $ICOMP$ score for this model was -644.36. From the univariate plots of the residuals in Figures 7 and 8, it appears that the normality assumption is mostly justified, with only slightly nonnormal tail behavior. However, looking at the plots of each dimension separately disguises the truth that we can discover empirically. Using the tests for multivariate Gaussian skewness and kurtosis of [29], this assumption does not appear to

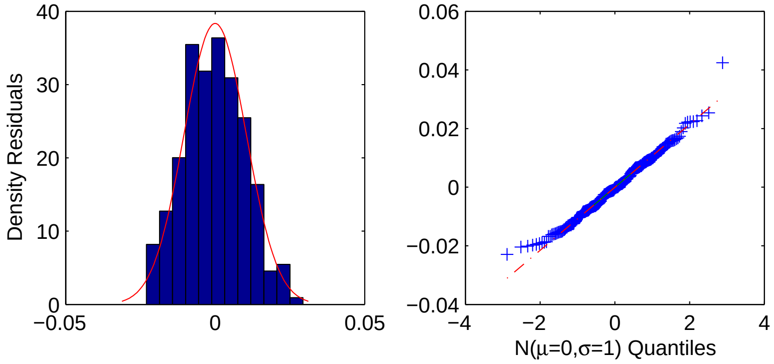

Figure 7: Residuals from the best model fit to the body fat data y_1.

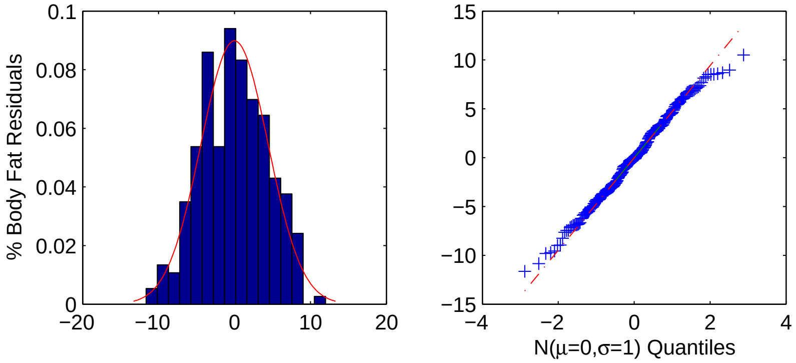

Figure 8: Residuals from the best model fit to the body fat data y_2.

be justified. In the case of multivariate normality, the theoretical population skewness and kurtosis parameters are respectively $\beta_1 = 0$ and $\beta_2 = p(p+2)$, or $\beta_2 = 8$ for $p = 2$. Mardia's sample values can be computed as in (87) and (88).

$$\hat{\beta}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(y_i - \overline{y})\hat{\Sigma}^{-1}(y_j - \overline{y})'\right]^3 \tag{87}$$

$$\hat{\beta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \overline{y})\hat{\Sigma}^{-1}(y_i - \overline{y})'\right]^2 \tag{88}$$

To test the null hypothesis $H_0 : \epsilon \sim N_p(\mu, \Sigma)$ versus the alternative $H_a : \epsilon \nsim N_p(\mu, \Sigma)$, we form the test statistics shown here; the test is one-sided for skewness, while the kurtosis test is two

Table 10: Estimated regression coefficients for body fat data.

	y_1	y_2
b_0	1.202	−45.952
b_2	0.0004	−0.148
b_6	−0.002	0.990

sided. Thus, the latter test can determine if the data exhibit higher or lower peakedness than expected.

$$\chi^{2*} = \frac{n}{6}\hat{\beta}_1 \sim \chi^2(\frac{p(p+1)(p+2)}{6}) \tag{89}$$

$$Z^* = \frac{(\hat{\beta}_2 - \beta_2)}{\sqrt{\frac{8p(p+2)}{n}}} \sim N(0,1) \tag{90}$$

Using the residuals from this model, $\hat{\beta}_2 = 132.13 \gg 8$, indicating substantial peakedness; the test statistic is $Z^* = 246.32$, and the p-value is 0.00000. Additionally, the sample skewness value is $\hat{\beta}_1 = 69.39$; this is significantly different (p-value= 0.00000) from what is expected. Thus, we see that the residuals from this model are not Gaussian white noise. In light of this information, we used the GA to perform the multivariate subset regression with the misspecification-robust form of $ICOMP$. The score for $ICOMP_{MISP}(\widehat{Cov}(\hat{\theta})) = -658.39$ is substantially different than that of $ICOMP(\hat{\mathscr{F}}^{-1})$ (-644.36), also indicative of misspecification. Interestingly, even though $ICOMP(\hat{\mathscr{F}}^{-1}) \neq ICOMP_{MISP}(\widehat{Cov}(\hat{\theta}))$, the same model was selected - $X^* = \left[x_0, x_2, x_6\right]$.

8. Concluding Remarks

Model misspecification is a major challenge faced by all statistical modeling techniques. As compared to the bell curve, real world multivariate data frequently exhibit higher kurtosis and heavier tails, asymmetry, or both. We have extended $ICOMP$ for multivariate regression so as to protect the statistical researcher against model misspecification, using the newly derived model covariance matrix that is appropriate whether or not the specified model is. Once this matrix is regularized to adjust for numerical instabilities, our modified criterion can take into consideration the actual sample kurtosis and skewness. Using this extended $ICOMP$ as the fitness function, we have employed the genetic algorithm to consistently identify the known true subset regression model in the presence of multicolinearity, unnecessary variables, redundant variables, and asymmetrical or leptokurtic behavior. The results from our challenging simulation studies bolster the application of our new criteria to a real dataset. Our findings suggest that when data are overly peaked or skewed, criteria which adjust for sample deviance from normality, such as the new $ICOMP$, should be used to drive model selection.

The world of statistics has too long relied upon the Gaussian distribution for analysis and model selection, and this can lead to suboptimal solutions in medical diagnostics, business analytics and intelligence, bioinformatics, econometric modeling, and engineering applications. With the new methods proposed in this paper, we expect advances in the usefulness and accuracy of statistical models.

ACKNOWLEDGEMENTS Professor Bozdogan would like to thank the hospitality shown by Professor Jan Magnus during May of 1999 when he was a Visiting Senior Scientist at

Tilburg University, in Tilburg, the Netherlands. During his visit the original problem of the misspecification in model selection was raised for the MVR model by Professor Bozdogan. Most of the collaboration between Professors Bozdogan and Magnus took place to derive the misspecification-resistant model covariance matrix which culminated both in this paper and that of [27] in two separate papers. 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. Several 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.

References

[1] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In B.N. Petrox and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Academiai Kiado.

[2] A. Azzalini. A Class of Distributions which Includes the Normal Ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.

[3] G. Box and D. Cox. An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, 26:211–246, 1964.

[4] H. Bozdogan. ICOMP: A New Model-Selection Criteria. In *Classification and Related Methods of Data Analysis*.

[5] H. Bozdogan. Model Selection and Akaike's Information Criteria (AIC): the General Theory and its Analytical Extensions. *Psychometrica*, 52:317–332, 1987.

[6] H. Bozdogan. On the Information-Based Measure of Covariance Complexity and its Application to the Evaluation of Multivariate Linear Models. *Communication in Statistics, Theory and Methods*, 19:221–278, 1990.

[7] H. Bozdogan. Akaike's Information Criterion and Recent Developments in Information Complexity. *Journal of Mathematical Psychology*, 44:62–91, 2000.

[8] H. Bozdogan. A New Class of Information Complexity (ICOMP) Criteria with an Application to Customer Profiling and Segmentation; an Invited Paper. *Istanbul University Journal of the School of Business Administration*, 39(2):370–398, 2010.

[9] H. Bozdogan and D. Haughton. Informational Complexity Criteria for Regression Models. *Computational Statistics and Data Analysis*, 28:51–76, 1998.

[10] K. Chaloner and I. Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273–304, August 1995.

[11] M. Chen. Estimation of Covariance Matrices Under a Quadratic Loss Function. Research Report S-46, Department of Mathematics, SUNY at Albany, 1976.

[12] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New Jersey, 1946.

[13] A. Davison. *Statistical Models*. Cambridge University Press, Cambridge, 2003.

[14] B. Frieden. *Physics from Fisher Information*. Cambridge University Press, Cambridge, UK, 1998.

[15] L. Godfrey. *Misspecification Tests in Econometrics*. Cambridge University Press, Cambridge, 1988.

[16] E. Gomez, M. Gomez-Villegas, and J. Marin. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics - Theory and Methods*, 27(3):589–600, 1998.

[17] C. Gouriéroux and A. Monfort. *Statistics and Econometric Models*, volume 1. Camridge University Press, Cambridge, 1995.

[18] C. Gouriéroux and A. Monfort. *Statistics and Econometric Models*, volume 2. Camridge University Press, Cambridge, 1995.

[19] D. Hendry. *Dynamic Econometrics*. Oxford University Press, Oxford, 1995.

[20] J. Hosking. Lagrange-Multiplier Tests of Time-Series Models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 42:170–181, 1980.

[21] R. Kass, L. Tierney, and J. Kadane. *Bayesian and Likelihood Methods in Statistics and Econometrics*, chapter The Validity of Posterior Expansions Based on Laplace's Method, pages 473–488. 1990.

[22] G. Kitagawa and S. Konishi. Bias and Variance Reduction Techniques for Bootstrap Information Criteria. *Annals of the Institute of Statistical Mathematics*, 62:209–234, 2010.

[23] S. Konishi and G. Kitagawa. *Information Criteria and Statistical Modeling*. Springer, New York, 2008.

[24] A. Kullback and R. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[25] D. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, December 1956.

[26] J. Magnus. *Linear Structures*. Charles Griffin & Company and Oxford University Press, London, 1988.

[27] J. Magnus. The Asymptotic Variance of the Pseudo Maximum Likelihood Estimator. *Econometric Theory*, 23:1022–1032, 2007.

[28] J. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistis and Econometrics*. Wiley, 1988.

[29] K. Mardia. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya*, B36:115–128, 1974.

[30] D. Poskitt. Precision, Complexity and Bayesian Model Determination. *Journal of the Royal Statistical Society, Series B (Methodological)*, 49(2):199–208, 1987.

[31] S. Press. Estimation of a Normal Covariance Matrix. Technical report, University of British Columbia, 1975.

[32] C. Rao. Information and Accuracy Attainable in the Estimation of Statistical Parameters. In *Bulletin of the Calcutta Math Society*, volume 37, page 81, 1945.

[33] C Rao. Minimum Variance and the Estimation of Several Parameters. In *Proceedings of the Cambridge Philosophical Society*, volume 43, page 280, 1947.

[34] C. Rao. Sufficient Statistics and Minimum Variance Estimates. In *Proceedings of the Cambridge Philosophical Society*, volume 45, page 213, 1948.

[35] J. Rissanen. Modeling by Shortest Data Description. *Automatica*, 14:465–471, 1978.

[36] T. Sawa. Information Criteria for Discriminating Among Alternative Regression Models. *Econometrica*, 46:1273–1291, 1978.

[37] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6:461–464, 1978.

[38] R. Shibata. Statistical Aspects of Model Selection. In J.C. Williams, editor, *From Data to Modeling*, pages 216–240. Springer, Berlin, 1989.

[39] A. Shurygin. The Linear Combination of the Simplest Discriminator and Fisher's One. In Nauka, editor, *Applied Statistics*. Moscow, Russia, 1983.

[40] K. Takeuchi. Distribution of Information Statistics and Criterion of Model Fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18, 1976. In Japanese.

[41] C. Thomaz. *Maximum Entropy Covariance Estimate for Statistical Pattern Recognition*. PhD thesis, University of London and for the Diploma of the Imperial College (D.I.C.), 2004.

[42] M. Van Emden. An Analysis of Complexity. In *Mathematical Centre Tracts,* volume 35. Mathematisch Centrum, 1971.

[43] C. Wei. On Predictive Least Squares Principles. *Annals of Statistics*, 20:1–42, 1992.

[44] H. White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50:1–25, 1982.

[45] H. White. *Estimation, Inference, and Specification Analysis*. Cambridge University Press, Cambridge, 1994.

Appendices

In what follows, for the completeness of the paper and the benefit of the readers, we repeat necessary matrix calculus derivations of the outer-product form of the Fisher information matrix and the misspecification resistent sandwich covariance matrix for the MVR model given in [27] to derive the $ICOMP$ criterion and its different forms.

Appendix 1: Outer-Product Form of the Fisher Information Matrix for the MVR Model

When the MVR model is misspecified, in order to obtain the outer-product form of the information matrix we standardize Y by defining $Z = (Y - XB)\Sigma^{-1/2}$, so that

$$\mathrm{E}(Z) = \mathbf{0},\ \mathrm{Var}(\mathrm{vec}\, Z) = I_{pn},$$

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

$$\Gamma_1 = \mathrm{E}\left((\mathrm{vec}\, Z)(\mathrm{vec}\,(Z'Z - nI_p)\right)', \Gamma_2 = \mathrm{E}\left((\mathrm{vec}\, Z'Z)(\mathrm{vec}\, Z'Z)\right)'.$$

In the special case of correct specification, these specialize to

$$\Gamma_1 = \mathbf{0},\ \Gamma_2 = 2nN_p + n^2(\mathrm{vec}\, I_p)(\mathrm{vec}\, I_p)', \tag{91}$$

where N_p denotes the $p^2 \times p^2$ *symmetrizer matrix*. The symmetrizer matrix has the property (for a square matrix A of dimensions p) $N_p\, \mathrm{vec} A = \frac{1}{2}\mathrm{vec}\,(A + A\prime)$. If $n = p = 1$, the kurtosis further specializes to $\Gamma_2 = 3$, as expected.

We now evaluate $\mathrm{E}\left((\mathrm{d} \log L(\theta \mid y))^2\right)$. Squaring (7) yields

$$(\mathrm{d} \log L(\theta \mid y))^2 = (\frac{1}{2}\, \mathrm{tr}(\Sigma^{-1/2} Z'Z \Sigma^{-1/2} - n\Sigma^{-1})\, \mathrm{d}\, \Sigma + \mathrm{tr}\, \Sigma^{-1/2} Z'X\, \mathrm{d} B)^2.$$

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

$$\begin{aligned} \mathrm{E}\left((\mathrm{d} \log L(\theta \mid y))^2\right) &= \frac{1}{4}\mathrm{E}\left(\mathrm{tr}(\Sigma^{-1/2} Z'Z \Sigma^{-1/2} - n\Sigma^{-1})\, \mathrm{d}\, \Sigma\right)^2 + \mathrm{E}\left(\mathrm{tr}\, \Sigma^{-1/2} Z'X\, \mathrm{d} B\right)^2 \\ &\quad + \mathrm{E}\left(\mathrm{tr}(\Sigma^{-1/2} Z'Z \Sigma^{-1/2} - n\Sigma^{-1})\, \mathrm{d}\, \Sigma\right)(\mathrm{tr}\, \Sigma^{-1/2} Z'X\, \mathrm{d} B) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{4}(\mathrm{d}\,\mathrm{vec}\,\Sigma)'(\Sigma^{-1/2}\otimes\Sigma^{-1/2})\,\mathrm{Var}(\mathrm{vec}\,Z'Z)(\Sigma^{-1/2}\otimes\Sigma^{-1/2})\,\mathrm{d}\,\mathrm{vec}\,\Sigma \\
&\quad +(\mathrm{d}\,\mathrm{vec}\,B)'(\Sigma^{-1/2}\otimes X')\,\mathrm{Var}(\mathrm{vec}\,Z)(\Sigma^{-1/2}\otimes X)\,\mathrm{d}\,\mathrm{vec}\,B \\
&\quad +(\mathrm{d}\,\mathrm{vec}\,\Sigma)'(\Sigma^{-1/2}\otimes\Sigma^{-1/2})\Gamma_1'(\Sigma^{-1/2}\otimes X)\,\mathrm{d}\,\mathrm{vec}\,B \\
&= \frac{1}{4}(\mathrm{d}\,\mathrm{vech}(\Sigma))'\Delta D_p^{+}(\Gamma_2-n^2(\mathrm{vec}\,I_p)(\mathrm{vec}\,I_p)')D_p^{+\prime}\Delta\,\mathrm{d}\,\mathrm{vech}(\Sigma) \\
&\quad +(\mathrm{d}\,\mathrm{vec}\,B)'(\Sigma^{-1}\otimes X'X)\,\mathrm{d}\,\mathrm{vec}\,B \\
&\quad +(\mathrm{d}\,\mathrm{vech}(\Sigma))'\Delta D_p^{+}\Gamma_1'(\Sigma^{-1/2}\otimes X)\,\mathrm{d}\,\mathrm{vec}\,B.
\end{aligned}$$

Hence,

$$\mathrm{E}\left(\mathrm{d}\log L(\theta\mid y)\right)^2=(\mathrm{d}\,\theta)'\mathscr{R}\,\mathrm{d}\,\theta,$$

where $\mathscr{R}$ is the outer-product form,

$$\mathscr{R}=\begin{bmatrix}\Sigma^{-1}\otimes X'X & \frac{1}{2}(\Sigma^{-1/2}\otimes X')\Gamma_1 D_p^{+\prime}\Delta \\ \frac{1}{2}\Delta D_p^{+}\Gamma_1'(\Sigma^{-1/2}\otimes X) & \frac{1}{4}\Delta D_p^{+}\Gamma_2^{*}D_p^{+\prime}\Delta\end{bmatrix}, \tag{92}$$

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

Appendix 2: Misspecification-Resistent Sandwich Covariance Matrix for the MVR model and $ICOMP$

In the presence of misspecification, the variance of the quasi maximum-likelihood estimator $\widehat{\theta}$ is

$$\begin{aligned}
Cov(\theta) &= \mathscr{F}^{-1}\mathscr{R}\mathscr{F}^{-1} \\
&= \begin{bmatrix}\Sigma\otimes(X'X)^{-1} & \frac{1}{n}(\Sigma^{\frac{1}{2}}\otimes(X'X)^{-1}X')\Gamma_1 D_p\Delta^{-1} \\ \frac{1}{n}\Delta^{-1}D_p'\Gamma_1'(\Sigma^{1/2}\otimes X(X'X)^{-1}) & \frac{1}{n^2}\Delta^{-1}D_p'\Gamma_2^{*}D_p\Delta^{-1}\end{bmatrix}.
\end{aligned} \tag{93}$$

Furthermore, a little algebra gives

$$\begin{aligned}
\mathrm{tr}\,Cov(\theta) &= \mathrm{tr}\,\Sigma\otimes(X'X)^{-1}+\frac{1}{n^2}\,\mathrm{tr}\,\Delta^{-1}D_p'\Gamma_2^{*}D_p\Delta^{-1} \\
&= (\mathrm{tr}\,\Sigma)(\mathrm{tr}(X'X)^{-1}) \\
&\quad +\frac{1}{n^2}\,\mathrm{tr}\,D_p^{+}(\Sigma^{1/2}\otimes\Sigma^{1/2})\Gamma_2^{*}(\Sigma^{1/2}\otimes\Sigma^{1/2})D_p^{+\prime},
\end{aligned} \tag{94}$$

and

$$\begin{aligned}
|Cov(\theta)| &= |\Sigma\otimes(X'X)^{-1}|\cdot|\frac{1}{n^2}\Delta^{-1}D_p'(\Gamma_2^{*}-\Gamma_1'(I_p\otimes X(X'X)^{-1}X')\Gamma_1)D_p\Delta^{-1}| \\
&= 2^{-p(p-1)}n^{-p(p+1)}|\Sigma|^{p+k+1}|X'^{-p} \\
&\quad \times|D_p'(\Gamma_2^{*}-\Gamma_1'(I_p\otimes X(X'X)^{-1}X')\Gamma_1)D_p|.
\end{aligned} \tag{95}$$

In the special case of correct specification, one verifies that

$$\operatorname{tr} Cov(\theta) = \operatorname{tr} \mathscr{F}^{-1} = (\operatorname{tr} \Sigma)(\operatorname{tr}(X'X)^{-1}) + \frac{1}{2n}(\operatorname{tr} \Sigma^2 + (\operatorname{tr} \Sigma)^2 + 2\sum_{j=1}^{p} \sigma_{jj}^2),$$

and

$$|Cov(\theta)| = |\mathscr{F}^{-1}| = 2^p n^{-\frac{1}{2}p(p+1)} |\Sigma|^{p+k+1} |X'X|^{-p}.$$

To derive *ICOMP* for the misspecified multivariate regression model, we need the determinant and trace of $\widehat{Cov}(\hat{\theta})$, the estimator of $Cov(\theta)$. The matrix $Cov(\theta)$ itself is given in (93), and its trace and determinant in (94) and (95). Thus,

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

and

$$\begin{aligned} |\widehat{Cov}(\hat{\theta})| &= 2^{-p(p-1)} n^{-p(p+1)} |\widehat{\Sigma}|^{p+k+1} |X'X|^{-p} \\ &\quad \times |D_p' (\widehat{\Gamma}_2^* - \widehat{\Gamma_1}' (I_p \otimes X(X'X)^{-1} X') \widehat{\Gamma_1}) D_p|. \end{aligned}$$

As a result we obtain

$$ICOMP(\widehat{Cov}(\hat{\theta}))_{MISP} = np \log 2\pi + n \log |\hat{\Sigma}| + np + 2C_1(\widehat{Cov}(\hat{\theta})), \tag{96}$$

where C_1, repeated from (16), is

$$C_1(\widehat{Cov}(\hat{\theta})) = \frac{s}{2} \log \frac{\operatorname{tr}(\widehat{Cov}(\hat{\theta}))}{s} - \frac{1}{2} \log |\widehat{Cov}(\hat{\theta})|. \tag{97}$$

In the special case of correct specification these results simplify to $\operatorname{tr} \widehat{Cov}(\hat{\theta}) = \operatorname{tr} \hat{\mathscr{F}}^{-1}$ and $|\widehat{Cov}(\hat{\theta})| = |\hat{\mathscr{F}}^{-1}|$, and $ICOMP(\widehat{Cov}(\hat{\theta}))_{MISP}$ reduces to $ICOMP(\hat{\mathscr{F}}^{-1})$.

Appendix 3: Derivation of the Penalty Bias

When the model is correctly specified, the skewness and kurtosis are given by (91), so that $\mathscr{R} = \mathscr{F}$. In general (under misspecification), we obtain

$$\begin{aligned} \operatorname{tr}(\mathscr{F}^{-1}\mathscr{R}) &= \operatorname{tr}(\Sigma \otimes (X'X)^{-1})(\Sigma^{-1} \otimes X'X) \\ &\quad + \frac{1}{2n} \operatorname{tr}(D_p^+ (\Sigma \otimes \Sigma) D_p^{+\prime} \Delta D_p^+ \Gamma_2^* D_p^{+\prime} \Delta) \\ &= pk + \frac{1}{2n} \operatorname{tr} N_p \Gamma_2^* = pk + \frac{1}{2n} \operatorname{tr} \Gamma_2^*. \end{aligned} \tag{98}$$

As derived in (57), the bias is then given by

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

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

$$\hat{b} = \frac{1}{n}\,\mathrm{tr}(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}) = \frac{1}{n}\left(pk + \frac{1}{2n}\,\mathrm{tr}(\hat{\Gamma_2}^{*})\right), \tag{99}$$

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

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

In that case,

$$\mathrm{tr}(\mathscr{F}^{-1}\mathscr{R}) = pk + p(p+1)/2,$$

which is the number of estimated parameters in the multivariate regression model and also the penalty term in AIC. This shows why AIC-type criteria and cross-validation techniques do not guard the researcher against misspecification of the model - the bias is computed under the assumption that the model is correctly specified.