Misspecification Resistant Model Selection Using Information Complexity With Applications

Hamparsum Bozdogan, J. Andrew Howe, Suman Katragadda, and
Caterina Liberati

Abstract In this paper, we address two issues that have long plagued researchers in statistical modeling and data mining. The first is well-known as the "curse of dimensionality". Very large datasets are becoming more and more frequent, as mankind is now measuring everything he can as frequently as he can. Statistical analysis techniques developed even 50 years ago can founder in all this data. The second issue we address is that of model misspecification - specifically that of an incorrect assumed functional form. These issues are addressed in the context of multivariate regression modeling. To drive dimension reduction and model selection, we use the newly developed form of Bozdogan's *ICOMP*, introduced in Bozdogan and Howe (2009b), that penalizes models with a complexity measure of the "sandwich" model covariance matrix. This information criterion is used by the genetic algorithm as the objective function in a two-step hybrid dimension reduction process. First, we use probabilistic principle components analysis to independently reduce the number of response and predictor variables. Then, we use the genetic algorithm with the multivariate Gaussian regression model to identify the best subset regression model. We apply these methods to identify a substantially reduced multivariate regression relationship for an dataset regarding Italian high school students. From 29 response variables, we get 4, and from 46 regressors, we get 1.

Bozdogan, Howe, Katragadda
Department of Statistics, Operations, and Management Science, University of Tennessee, United States of America, e-mail: bozdogan@utk.edu, e-mail: ahowe42@utk.edu, e-mail: katragaddasuman11@gmail.com

Caterina Liberati
Dipartimento di Scienze Statistiche "F. Fortunati", Università di Bologna
e-mail: caterina.liberati@unibo.it

1 Introduction

In this paper, we address two issues that have long plagued researchers in statistical modeling and data mining. The first is well-known as the "curse of dimensionality". Very large datasets are becoming more and more frequent, as mankind is now measuring everything he can as frequently as he can. Statistical analysis techniques developed even 50 years ago can founder in all this data. The second issue we address is that of model misspecification - specifically that of an incorrect assumed functional form. These issues are addressed in the context of multivariate regression modeling, in which we present a novel hybrid dimension reduction technique. We apply these methods to identify a substantially reduced multivariate regression relationship for an dataset regarding Italian high school students. From 29 response variables, we get 4, and from 46 regressors, we get 1.

2 Multivariate Regression Modeling with *ICOMP*

2.1 Multivariate Gaussian Regression

In the usual multivariate regression (MVR) 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:

$$Y \sim N_p(XB, \Sigma \otimes I_n), \text{ where } \mathrm{E}[Y] = XB, \text{ and } Cov(Y) = \Sigma \otimes I_n. \tag{2}$$

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

$$\log L(\theta \mid Y) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}[(Y-XB)\Sigma^{-1}(Y-XB)']. \tag{3}$$

The model covariance matrix, (*inverse Fisher information matrix*) can be derived using the results of Magnus and Neudecker (1988, page 321), and is given by

$$\widehat{Cov}(vec(\hat{B}), \text{vech}(\hat{\Sigma})) \equiv \hat{\mathscr{F}}^{-1} = \begin{bmatrix} \hat{\Sigma} \otimes (X'X)^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2}{n} D_p^+ (\hat{\Sigma} \otimes \hat{\Sigma}) D_p^{+\prime} \end{bmatrix} \tag{4}$$

The IFIM provides the asymptotic variance of the ML estimators 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.

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. The most common causes of model misspecification include: multicollinearity, autocorrelation, heteroskedasticity, and incorrect functional form. This final type is the type of misspecification we address. The common answer in the literature to nonnormality has been the utilization of *Box-Cox transformations* of Box and Cox (1964), which does not seem to work consistently well, especially in the context of multivariate regression. 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, and there are many approaches to this problem. We follow Bozdogan and Howe (2009b) and use the genetic algorithm to select a subset MVR model.

2.2 Robust Misspecification-Resistant Information Complexity Criteria

Acknowledging the fact that any statistical model is merely an approximate representation of the true data generating process, information criteria attempt to guide model selection according to the *principle of parsimony*. 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 poorly 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. In order to protect the researcher against model misspecification, Bozdogan and Howe (2009b) generalized *ICOMP* to the case of a misspecified MVR model and introduce $ICOMP_{MISP}$, which can drive effective model selection **even when the Gaussian assumption is invalid**. Here we show their results without derivations or proofs.

If we note θ_g^* as the value of the parameters vector which minimizes the *Kullback-Liebler* distance (Kullback and Leibler, 1951) for some specified functional model $f\left(\theta_g^*\right)$ to the true functional model $g(\theta)$, and we use $\mathscr{R}$ to indicate the outer-product form of the Fisher information matrix, we have

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

$$\hat{\theta} \sim N(\theta_g^*, \mathscr{F}^{-1}\mathscr{R}\mathscr{F}^{-1}), \text{ or } \sqrt{n}(\hat{\theta} - \theta_g^*) \sim N(0, \mathscr{F}^{-1}\mathscr{R}\mathscr{F}^{-1}). \tag{5}$$

Note that this tells us explicitly

$$Cov(\theta_g^*)_{Misspec} = \mathscr{F}^{-1}\mathscr{R}\mathscr{F}^{-1}, \tag{6}$$

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

$$\widehat{Cov(\theta)} = \hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1}. \tag{7}$$

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

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

Thus, in the case of a correctly specified model, $\widehat{Cov(\theta)} = \hat{\mathscr{F}}^{-1}$.

For multivariate regression, we have already seen the inner-product form of estimated IFIM in (4). The outer-product form $\hat{\mathscr{R}}$ is derived in Magnus (2007), and we show the result in (8).

$$\hat{\mathscr{R}} = \begin{bmatrix} \hat{\Sigma}^{-1} \otimes X'X & \frac{1}{2}(\hat{\Sigma}^{-1/2} \otimes X')\hat{\Gamma}_1 D_p^{+\prime}\Delta \\ \frac{1}{2}\Delta D_p^{+}\hat{\Gamma}_1'(\hat{\Sigma}^{-\frac{1}{2}} \otimes X) & \frac{1}{4}\Delta D_p^{+}\hat{\Gamma}_2^{*} D_p^{+\prime}\Delta \end{bmatrix}. \tag{8}$$

This matrix takes into consideration the actual sample skewness and kurtosis of the data. There is an issue of matrix stability to be addressed with the sandwich covariance matrix, however. Numerical issues with estimating the sandwich covariance matrix prevent it from approximating the FIM when the model is correctly specified. We employ the *Empirical Bayes covariance regularization* procedure

$$\widehat{Cov(\theta)} \leftarrow \widehat{Cov(\theta)} + \frac{p-1}{(n)\,tr(\widehat{Cov(\theta)})} I_p, \tag{9}$$

to ensure $\widehat{Cov(\theta)}$ is of full rank. Thus, the misspecification-resistant form of *ICOMP* for multivariate regression is computed as (10). When the model is correctly specified, we expect $\widehat{Cov(\theta)} = \hat{\mathscr{F}}^{-1}$, we get $ICOMP(\hat{\mathscr{F}}^{-1})$ in (11).

$$ICOMP(\widehat{Cov(\theta)})_{MISP} = np\log 2\pi + n\log|\hat{\Sigma}| + np + 2C_1(\widehat{Cov(\theta)}) \tag{10}$$

$$ICOMP(\hat{\mathscr{F}}^{-1}) = np\log 2\pi + n\log|\hat{\Sigma}| + np + 2C_1(\hat{\mathscr{F}}^{-1}) \tag{11}$$

In both, C_1 is the first order maximal entropic complexity of Bozdogan (1988): a generalization of the model covariance complexity of Van Emden (1971), given by

$$C_1(\widehat{Cov(\theta)}) = \frac{s}{2}\log\frac{\mathrm{tr}(\widehat{Cov(\theta)})}{s} - \frac{1}{2}\log|\widehat{Cov(\theta)}|,\ s = rank(\widehat{Cov(\theta)}). \tag{12}$$

3 Dimension Reduction with the Genetic Algorithm and Probabilistic Principle Components Analysis

3.1 Genetic Algorithm

The genetic algorithm (GA) is a search algorithm that borrows concepts from biological evolution. Unlike most search algorithms, the GA simulates a large population of potential solutions, encoded as binary strings. These solutions are allowed to interact over time; random mutations and natural selection allow the population to improve, eventually iterating to an optimal solution. The GA was popularized by Holland (1975), and it is a widely recognized and popular stochastic search and optimization algorithm. Today, there are many problems in science, economics, and research and development that are solved using the GA. We refer the reader to existing books and articles regarding details of the algorithm. Some excellent books are Goldberg (1989); Haupt and Haupt (2004); Vose (1999). Articles specifically combining the GA with subset regression models would include Bozdogan (2004) in which the GA was implemented for multiple regression subset selection under the normality assumption. Also, Bozdogan and Howe (2009b) extended this work to the case of misspecified multivariate regression.

3.2 Probabilistic Principle Components Analysis

In this paper, we employ Probabilistic Principle Component Analysis (PPCA) as a first step to independently reduce the dimensionality of the independent and dependent matrices. PPCA was developed in the late 1990's and popularized by Tipping and Bishop (1997). Here, we show some results from Tipping and Bishop (1997) and Bozdogan and Howe (2009a) that are relevant to this research. Let $x \in \mathbb{R}^{1\times p}$ be a random vector; assume x can be expressed as a linear combination of *latent variables* and stochastic noise:

$$x = \Lambda f + \mu + \varepsilon, \tag{13}$$

where $f \in \mathbb{R}^{m\times 1}$ holds the latent variables, $\Lambda \in \mathbb{R}^{p\times m}$ is the loading matrix, and $\mu \in \mathbb{R}^{1\times p}$ defines the mean of x. Maximizing the PPCA likelihood function, we get the model covariance matrix in (14)

$$\widehat{Cov}(X) = U_p \hat{L} U_p', \tag{14}$$

where U_p contains all the eigenvectors of $\hat{\Sigma}$. $\hat{L}$ is almost a $(p \times p)$ matrix with eigenvalues of $\hat{\Sigma}$ on the diagonals. Positions corresponding to variables not included in the given subset are replaced with the mean of the left-out eigenvalues. Using this, the inverse Fisher information matrix is given in (15).

$$\hat{\mathscr{F}}^{-1} = \begin{bmatrix} \widehat{Cov}(X) & \mathbf{0} \\ \mathbf{0}' & \frac{2}{n} D_p^+ \widehat{Cov}(X) \otimes \widehat{Cov}(X) D_p^{+\prime} \end{bmatrix}. \tag{15}$$

The heavy-penalty form of *ICOMP* we use here is

$$ICOMP_{PEU}(\hat{\mathscr{F}}^{-1}) = -2\log L(\hat{\Lambda}, \mu, \hat{\sigma}^2 \mid x) + 2\left(\frac{nm}{n-m-2}\right) + \log(n) C_1(\hat{\mathscr{F}}^{-1}), \tag{16}$$

where m is the number of variables included from the original dataset. As with the MVR model, we can use the GA to reduce the dimensionality of a data set, with *ICOMP* as the objective function.

4 Numerical Results

Our dataset is a random sample of $1,400$ students from the ALMALAUREA database. ALMALAUREA was started as a service for addressing the faculty choice of high school students based on interests, skills, and job expectations. All variables have been normalized to vary between -1 and 1. As response variables, we have $Mat_1, Mat_2, \ldots, Mat_{29}$: students judgements about different subjects (math, physics, chemistry, engineering, statistics…). Our regressor matrix is divided into two "sets". Answers regarding what the students think are important for ideal future work - collaboration, time flexibility, …- are measured in variables $Nz_1, Nz_2, \ldots, Nz_{14}$. Variables $Np_1, Np_2, \ldots, Np_{32}$ measure students personal abilities (concentration, time management, curiosity, …). The predictor variables are numbered from 1 to 14 for Nz, and 15 through 46 for Np.

Table 1 *ICOMP* Scores & Subsets of Predictors.

Criteria	Score	Best set of predictors
No Preliminary Dimension Reduction		
$ICOMP(\hat{\mathscr{F}}^{-1})$	64004	$\{1,2,4,5,6,9,12,13,16,17,19-21,25-27,\ldots$ $29-31,34,35,38,41,42\}$
$ICOMP(\widehat{Cov(\theta)})_{MISP}$	59701	$\{1-46\}$
Preliminary Dimension Reduction of Only Dependent Variables Matrix		
$ICOMP(\hat{\mathscr{F}}^{-1})$	9693	$\{1-46\}$
$ICOMP(\widehat{Cov(\theta)})_{MISP}$	9483	$\{1-46\}$
Preliminary Dimension Reduction of Both Responses and Regressors		
$ICOMP(\hat{\mathscr{F}}^{-1})$	10825	45
$ICOMP(\widehat{Cov(\theta)})_{MISP}$	10963	45

For modeling this data, we first used the GA to identify optimal subset MVR models, driven by both $ICOMP(\hat{\mathscr{F}}^{-1})$ and $ICOMP(\widehat{Cov(\theta)})_{MISP}$. If the Gaussian regression model was correctly specified, we would expect the criteria to select very similar models with similar scores. Results shown in the first third of Table

1 do not bear this out. While the substantially lower $ICOMP(\widehat{Cov(\theta)})_{MISP}$ score indicates it has selected a better model, we have not been able to reduce the dimensionality at all. Mardia's tests for multivariate normal skewness and kurtosis (Mardia, 1974), reject the null hypothesis of normality, with results shown in Table 2, confirming the misspecification identified by *ICOMP*. Secondly, we used

Table 2 Normality Test Results for First Identified Model.

Skewness		Kurtosis	
β_1	0	β_2	899
$\hat{\beta}_1$	32.55	$\hat{\beta}_2$	986.84
χ^{2*}	7594.92	Z^*	38.75
95% Region	$[0, 4652.09]$	95% Region	$[-1.96, 1.96]$
p-value	0.00000	p-value	0.00000
Conclusion	$\varepsilon \sim N(\mu, \Sigma)$	Conclusion	$\varepsilon \sim N(\mu, \Sigma)$

PPCA as a preliminary step to reduce the dimensionality of the matrix of responses. Using $ICOMP_{PEU}(\hat{\mathscr{F}}^{-1})$, the GA selected a model with only 4 dependent variables: $Mat_{26} - Mat_{29}$. We then attempted to identify a subset MVR model using just these responses. The *ICOMP* scores indicate that the Gaussian regression model is misspecified, with $ICOMP(\widehat{Cov(\theta)})_{MISP} < ICOMP(\hat{\mathscr{F}}^{-1})$, though both criteria selected the fully saturated model. These results are shown in the middle third of Table 1. Mardia's expected and sample kurtosis values of 24 and 22.6 were very close; the test statistic for skewness, however, was 214 - much higher than the critical value of 31. Once again, we verify the misspecification identified by *ICOMP*.

Finally, we also used PPCA to select a subset of only 4 of the 46 independent variables. Those selected were $Np_{29} - Np_{32}$. We then ran two sets of the GA a third time, using both *ICOMP* versions, with results displayed in the bottom third of Table 1. Note how close the *ICOMP* scores are (relative to the other pairs), and that both criteria selected the same substantially reduced subset MVR model, using only a single predictor for the four responses. Thus, we have gone from an overly complex misspecified multivariate regression model, to a model that is both (very nearly) correctly-specified and parsimonious.

While our end result would suggest the misspecification-resistant *ICOMP* was not needed, recall the first MVR subset model identified. If we had only used $ICOMP(\hat{\mathscr{F}}^{-1})$, we would have had less motivation to use PPCA to reduce the dimensionality of the model. We would have settled upon an MVR model with 32 responses and 24 regressors.

5 Concluding Remarks

In this research, we have applied a novel hybrid dimension reduction technique for multivariate regression. While independently reducing the number of dimensions in

both the matrix of responses and regressors using PPCA and the GA, we used a new misspecification-resistent form of *ICOMP*. These methods allowed us to identify a nearly correctly-specified simple regression relationship with 4 of 29 dependent and 1 of 46 independent variables, rather than a misspecified overly complex relationship.

References

Box, G., Cox, D., 1964. An analysis of transformations. Journal of the Royal Statistical Society, Series B (Methodological) 26, 211–246.

Bozdogan, H., 1988. Icomp: A new model-selection criteria. In: Bock, H. (Ed.), Classification and Related Methods of Data Analysis. North-Holland.

Bozdogan, H., 2004. Intelligent Statistical Data Mining with Information Complexity and Genetic Algorithms. In: Bozdogan, H. (Ed.), Statistical Data Mining and Knowledge Discovery. Chapman & Hall/CRC, Boca Raton, pp. 15–56.

Bozdogan, H., Howe, J., 2009a. The curse of dimensionality in large-scale experiments using a novel hybridized dimension reduction approach. submitted to Technometrics tbd.

Bozdogan, H., Howe, J., 2009b. Misspecification resistant multivariate regression models using the genetic algorithm and information complexity as the fitness function. submitted to Statistical Computing tbd.

Goldberg, D., 1989. Genetic algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Col, Inc.,, Boston, Massachusetts.

Haupt, R., Haupt, S., 2004. Practical genetic algorithms. John Wiley, Hoboken, New Jersey.

Holland, J., 1975. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. The University of Michigan Press, Ann Arbor, Michigan.

Kullback, A., Leibler, R., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Magnus, J., 2007. The asymptotic variance of the pseudo maximum likelihood estimator. Econometric Theory 23, 1022–1032.

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

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

Tipping, M., Bishop, C., 1997. Probabilistic principal component analysis. Tech. Rep. NCRG/97/010, Neural Computing Research Group, Aston University.

Van Emden, M., 1971. An analysis of complexity. In: Mathematical Centre Tracts. Vol. 35. Mathematisch Centrum.

Vose, M., 1999. The simple genetic algorithm: Foundations and Theory. MIT Press, Cambridge, Massachusetts.