To the Graduate Council:
I am submitting herewith a dissertation written by Derek McCrae Norton entitled "Adaptive Data Mining." I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Statistics.

_____

Hamparsum Bozdogan, Major Professor

We have read this dissertation
and recommend its acceptance:

_____

Hamparsum Bozdogan

_____

Committee Member 2

Accepted for the Council:

_____

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

To the Graduate Council:
I am submitting herewith a dissertation written by Derek McCrae Norton entitled "Adaptive Data Mining." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Statistics.

Hamparsum Bozdogan, Major Professor

We have read this dissertation
and recommend its acceptance:

Hamparsum Bozdogan

Committee Member 2

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Adaptive Data Mining

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Derek McCrae Norton

August 2013

dedication...

This dissertation is dedicated to my very supportive wife, Kimberly Dawn Norton, and my beautiful daughters, Audrey Elizabeth Norton and Madelyn McCrae Norton. All three of you are a gift from God, and I thank him everyday for you.

Acknowledgements

I want to thank everyone who helped me along the path to this dissertation. I want to first thank Dr. Hamparsum Bozdogan for all the help and guidance he has provided. I want to also thank my committee members, ... for their guidance and direction. I want to thank Dr. Seaver for pushing me to see the big picture and ask those important questions.

I want to thank Revolution Analytics for dealing with me as I finished this dissertation.

I want to thank my friends and family for their unending support and prayer. I want to thank my lovely wife for pushing me when I said I didn't want to be pushed.

Finally, I want to thank God because without him none of this would be possible.

Some quotation...

Abstract

Often an interest in predictions of future values occurs. Sometimes these future values are based solely on a single predictor, whether that predictor is another variable or simply the lagged variable of interest. However, the case is more often a complex interrelationship of multiple variables where future values for all variables are of interest. Models that adequately describe these complex interrelationships are the focus here. Specifically Vector Autoregressive (VAR) models are to be considered.

There are many potential pitfalls associated with VAR modeling, from subsetting, to non-normality, to misspecification This work aims to provide a unified method to model multiple time series. This begins with extensions to deal with non-normality such as the Power Exponential VAR or Kernel VAR, continues on with how to model under misspecification, and then how to properly subset the saturated model. It is concluded with diagnostics of and prediction with these models. All of these tasks are completed using Information Criteria and modern selection algorithms such as the Genetic Algorithm.

List of Tables

List of Figures

Nomenclature

$\bar{}$	Denotes a dimensional variable
$\mathcal{F}$	Fisher Information Matrix
ψ	Streamfunction
θ	Tangential coordinate
p	Pressure
r	Radial coordinate
u_r	Radial velocity
u_z	Axial velocity
u_θ	Tangential velocity
z	Axial coordinate

Contents

Part I

Adaptive Multivariate Time Series Modeling

Chapter 1

Introduction

Often an interest in predictions of future values occurs. Sometimes these future values are based solely on a single predictor, whether that predictor is another variable or simply the lagged variable of interest. However, the case is more often a complex interrelationship of multiple variables where future values for all variables are of interest. Models that adequately describe these complex interrelationships are the focus here. Specifically Vector Autoregressive (VAR) models are to be considered.

1.1 Motivations for VAR Models

Most of the existing literature treats VAR models as the sole dominion of economic or financial prediction. This is certainly not unreasonable considering the seminal paper by Sims [1980]. There are, of course, some instances in the literature outside of that domain, such as Crescenzi and Enterline [1999] and Freeman et al. [1989]. While many of the possible applications for VAR models are economic in nature, there are certainly abundant non-economic applications. To this end, forecasting will be the main focus of this work, with certain topics such as cointegration noticeably absent. Cointegration, for example, certainly seems justifiable in certain economic applications, but is not necessarily as justifiably in other areas of application, and certainly not desirable if interpretation or explainability of the model and parameters is an issue.

1.2 VAR Model Specification

In the univariate case, a simple autoregressive model of order p, AR(p), can be written as

$$y_t = \gamma d + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t,$$

where d is a matrix of deterministic terms to represent the intercept, linear trend t, quadratic trend t^2, or other terms such as seasonal dummies. The multivariate

extension, VAR(p), is simply

$$\mathbf{y}_t = \Gamma D + \sum_{i=1}^{p} \Phi_i \mathbf{y}_{t-i} + \varepsilon_t. \tag{1.2.1}$$

where $\mathbf{y}_t = (y_{1t}, \ldots, y_{kt})'$ is a $(k \times 1)$ random vector, the Φ_i are fixed $(k \times k)$ coefficient matrices, and $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{kt})'$ is the $(k \times 1)$ error vector such that $\varepsilon_t \overset{i.i.d.}{\sim} (0, \Sigma)$. Here no actual distributional assumptions are made, but that will be addressed in subsequent chapters. Again, as in the univariate case the first term in the model, ΓD, is any deterministic terms included in the model. The addition of the deterministic term is not often seen, but is included here for reasons that will become clear in later chapters.

The main goal of the VAR model is a better understanding of underlying relationships. Whether this understanding is simply to describe the relationships between variables or to predict future outcomes, it is necessary to achieve a model that is both parsimonious and has minimum variance. It becomes clear that a fully ranked VAR model may not satisfy either condition.

Example 1.1.

Consider $\mathbf{y}_t$, a matrix of $k = 3$ time series variables that are hypothesized to depend on $p = 3$ lags of previous values.

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \Phi_3 \mathbf{y}_{t-3} + \varepsilon_t \tag{1.2.2}$$

This model despite its seeming simplicity is not simple, and can be seen if we expand the notation in (1.2.2) to get

$$\begin{aligned} \begin{bmatrix} \mathbf{y}_{1,t} \\ \mathbf{y}_{2,t} \\ \mathbf{y}_{3,t} \end{bmatrix} = & \begin{bmatrix} \phi_{1,1,1} & \phi_{1,2,1} & \phi_{1,3,1} \\ \phi_{2,1,1} & \phi_{2,2,1} & \phi_{2,3,1} \\ \phi_{3,1,1} & \phi_{3,2,1} & \phi_{3,3,1} \end{bmatrix} \mathbf{y}_{t-1} + \begin{bmatrix} \phi_{1,1,2} & \phi_{1,2,2} & \phi_{1,3,2} \\ \phi_{2,1,2} & \phi_{2,2,2} & \phi_{2,3,2} \\ \phi_{3,1,2} & \phi_{3,2,2} & \phi_{3,3,2} \end{bmatrix} \mathbf{y}_{t-2} \\ & + \begin{bmatrix} \phi_{1,1,3} & \phi_{1,2,3} & \phi_{1,3,3} \\ \phi_{2,1,3} & \phi_{2,2,3} & \phi_{2,3,3} \\ \phi_{3,1,3} & \phi_{3,2,3} & \phi_{3,3,3} \end{bmatrix} \mathbf{y}_{t-3} + \varepsilon_t. \end{aligned} \tag{1.2.3}$$

Clearly (1.2.3) has quite a few parameters to be estimated. Specifically it has $k^2 \times p = 3^2 \times 3 = 27$ plus the variance leading to 28 parameters to estimate. Also this does not even include deterministic terms.

1.3 Motivations for Information Criteria

To achieve a more parsimonious model and also a lower variance competing models must be compared. This can be done through a variety of methods, but here the focus is the use of information criteria. An information criteria allows for the comparison of not only different lag lengths, but also subset models of reduced rank. This is

highly advantageous since it allows for a model with a great deal of flexibility. The problem now becomes which criteria to use. There are many options to solve this problem, such as Akaike's Information Criterion (AIC), Schwarz Bayesian Criterion (SBC), Rissanen's Minimum Description Length (MDL), and Bozdogan's Information Complexity (ICOMP) for some examples.

1.4 Motivations for GA

In the previous section, the example of a seemingly simple VAR(3) for $k = 3$ variables was introduced. In light of parsimony and minimal variance, various competing models could be compared using some form of information criteria. However, like the previous example, this is not as easy as on might initially consider. To look at all possible subsets for the example would require looking at $2^{k^2p} - 1 = 2^{27} - 1 =$ over 134 million models. If it took .5 seconds to test each model, this "simple" task would take over 2 years to complete. All possible subsets is the only way to guarantee the best model among competing models, but there are other possibilities for subsetting. Unfortunately, traditional procedures such as stepwise selection do not tend to perform well. This leads to the motivation for the use of an improved search algorithm which comes in the form of the genetic algorithm (GA). The GA provides a means to search the possible subsets and obtain a new optimal model in terms of the information criteria.

1.5 Summary

The goal of the remainder of this work is to offer valid alternatives to the basic VAR model to deal with the so-called "real world" data that often invalidates many of the normal modeling assumptions. The hope is to complete that goal while incorporating the novel combination of model selection and the genetic algorithm to lead to appropriate models that achieve parsimony and minimal variance.

Chapter 2

Model Selection Techniques

Model selection should be viewed as the unification of two distinct parts. The first part is the act of subsetting the data. Specifically, this is the means through which various candidate models are fit. The second part involves scoring the candidate models, which gives an overall rank to the fitness of each model with respect to the other models.

Although there are only two seemingly trivial parts, subsetting and scoring, that is not to say that the process is trivial. On the contrary, there are a plethora of choices for how to complete both the subsetting and scoring parts, and each part is not as trivial as it may first seem. Also, the entire process can be made considerably more complex if ones interest lies not only in what variables to include in the model, but also in other factors such as distributional assumptions, methods of estimation, or possibly different model types all together.

2.1 Information Criteria

Information criteria are the means through which possible candidate models can be scored. The discussion here will be limited to criteria that consist of two parts, which are the lack of fit between the chosen model and the data at hand, and a function which penalizes the first part. In general the first part, or lack of fit, is measured by the likelihood function. This gives a measure of how likely a model is given the data at hand. The second part, or the penalty term, is chosen differently for different criteria, and will be described in detail in the following sections. As a lead in to all of the following criteria, consider

$$y = y_1, y_2, \ldots, y_n \tag{2.1.1}$$

to be a random sample from some process. The underlying data generating process, which is unknown, is denoted f_*. A given candidate model will be written as f_m, where $m = 1, \ldots, M$, and M is the number of models entertained. For a more thorough discussion of these techniques see Bozdogan [1987, 2000] for example.

2.1.1 AIC

Probably the best known and widely used of information criteria is Akaike's Information Criterion (AIC). Originally set forth in Akaike [1974, 1973], AIC has been almost universally accepted as a model selection criteria. The central argument put forth by Akaike is that the model that fits the data the best is the model that minimizes the Kullback-Leibler (K-L) information from Kullback and Leibler [1951]. The K-L information measures how well the candidate model, f_m, approximates the true model, f_*, and is given by

$$\begin{aligned} I\left(f^*, f_m\right) &= E_*\left[\log\left(f_*\right) - \log\left(f_m\right)\right] \\ &= E_*\left[\log\left(f_*\right)\right] - E_*\left[\log\left(f_m\right)\right] \\ &= H\left(f_*; f_*\right) - H\left(f_*; f_m\right) \end{aligned} \tag{2.1.2}$$

where E_* is the expectation with respect to the true model f_*. $H\left(f_*; f_*\right) \equiv H\left(f_*\right)$ is a constant in (2.1.2), so only the cross-entropy

$$H\left(f_*; f_m\right) = E_*\left[\log\left(f_m\right)\right] \tag{2.1.3}$$

needs estimation. One possible estimator for (2.1.3) is $\log L\left(\hat{\theta}_m\right)$, the maximized log likelihood function using the maximum likelihood estimate (MLE), $\hat{\theta}_m$, for the model parameters θ. This is often an overly optimistic estimate of (2.1.3), due to the fact that the data from (2.1.1) is used to estimate both the model parameters and the cross-entropy. It can be shown that this estimation bias is equal to nb, Konishi and Kitagawa [1996]. Akaike made the assumptions that the model would be estimated using maximum likelihood, and that the underlying data generating process is included in the candidate models. With these assumptions, Akaike shows that the bias, nb, is approximately equal to the number of parameters in the model. This leads to an unbiased estimator of (2.1.3), namely

$$AIC = -2\log L\left(\hat{\theta}_m\right) + 2r, \tag{2.1.4}$$

where r is the number of parameters estimated in the candidate model. Thus, AIC is calculated for all candidate model and the minimum over all models is chosen as the best fitting.

2.1.2 SBC / BIC / MDL

Interestingly, Schwarz [1978], Akaike [1978], Rissanen [1978] and all followed very different derivations, yet still came up with the same criterion. Schwarz [1978] and Akaike [1978] both used a Bayesian derivation, while Rissanen [1978] based his derivation on the information theoretic minimum code length for the data along with the model parameters. Due to the similarities within Scwharz's Bayesian Criterion

(SBC), Akaike's Bayesian Information Criterion (BIC), and Rissanen's Minimum Description Length (MDL), the term SBC will be used consistently throughout this work, and the form of this criterion is given by

$$SBC = -2\log L\left(\hat{\theta}_m\right) + r\log\left(n\right). \tag{2.1.5}$$

This particular criterion, unlike AIC, is consistent. This means that if the underlying data generating process is one of the models considered, that the criterion will select it with probability approaching one as sample size approaches infinity.

2.1.3 HQ

Also in an attempt to create a consistent estimator, Hannan and Quinn [1979] derived a criterion denoted (HQ) based on the law of iterated logarithms. The form of this criterion is given by

$$HQ = -2\log L\left(\hat{\theta}_m\right) + r\log\log\left(n\right). \tag{2.1.6}$$

This criterion was derived with the hopes to underestimate the order of autoregressive models to a lesser degree than SBC, and attempts to achieve this through the fact that the second term increases at a slower rate than $k\log\left(n\right)$.

2.1.4 Finite Sample Corrected AIC

Initially developed by Sugiura [1978] and then used by Hurvich and Tsai [1989], AIC_C estimates the exact value of the bias, b, from estimating (2.1.3) in the multiple regression case leading to

$$AIC_{C_A} = -2\log L\left(\hat{\theta}_m\right) + 2\left(\frac{nr}{n-r-1}\right). \tag{2.1.7}$$

Though useful in small sample situations, the bias should be estimated for the model at hand. Hurvich and Tsai [1993] estimate the bias specifically for VAR models leading to

$$AIC_{C_B} = -2\log L\left(\hat{\theta}_m\right) + 2\left(\frac{nr}{n-r-\frac{k+1}{2}}\right) \tag{2.1.8}$$

2.1.5 Consistent AIC

In an attempt to yield a consistent estimator, Bozdogan [1987] follows the derivation of Akaike [1973] and adds the thought that when testing a null hypothesis versus an alternative hypothesis using a parameter, the degrees of freedom is an increasing function of the sample size if the test statistic follows a noncentral chi-squared

distribution. With that thought in mind, Bozdogan [1987] derived an adjustment to AIC which reflects a dependence on the sample size. Namely

$$CAIC = -2\log L\left(\hat{\theta}_m\right) + r\left[\log n + 1\right]. \tag{2.1.9}$$

2.1.6 Consistent AIC with Fisher Information

Exploiting the large sample asymptotic distributional properties of the MLE, Bozdogan [1987] proposes another consistent information criterion which penalizes overparameterization strongly, especially for large sample sizes. This criterion is defined as

$$CAICF = -2\log L\left(\hat{\theta}_m\right) + r\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right|. \tag{2.1.10}$$

This criteria was further extended in Bozdogan and Ueno [2000] in the bayesian framework to

$$\begin{aligned} CAICF_E &= -2\log L\left(\hat{\theta}_m\right) + r\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right), \end{aligned} \tag{2.1.11}$$

which includes AIC, $GAIC$, SBC, and $CAICF$ as special cases. The $CAICF_E$ penalizes overparameterization more harshly than the $CAICF$. One further step taken in Bozdogan and Ueno [2000], is to approximate the last term of $CAICF_E$ to correct for the bias of small sample size. This approximation leads to

$$\begin{aligned} CAICF_C &\cong -2\log L\left(\hat{\theta}_m\right) + r\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2\left(\frac{nr}{n-r-2}\right). \end{aligned} \tag{2.1.12}$$

2.1.7 Generalized AIC

A generalization of AIC attempts to determine the penalty term in (2.1.4) differently for different models and data sets. This is accomplished by replacing k, the number of parameters, with an adaptive value leading to

$$GAIC = -2\log L\left(\hat{\theta}_m\right) + 2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right), \tag{2.1.13}$$

where $\hat{\mathcal{F}}^{-1}$ is the estimated Inverse Fisher Information Matrix (IFIM) in inner product form, and $\hat{R}$ is the estimated Fisher information matrix in outer product form. The penalty term, $tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right)$, reduces to r, the number of parameters, if the estimated model is the actual model, and if certain regularity conditions hold. For more information, see Takeuchi [1976] for example.

2.1.8 Information Complexity

Information Complexity (ICOMP) is a novel approach to the development of an information criterion formulated by Bozdogan [1988, 1990]. ICOMP aims to select models with a form similar to other information criteria, i.e. with a penalized likelihood; however this is where the similarity ends. Instead of penalty terms based on sample size or number of parameters, ICOMP penalizes the likelihood based on a measure of covariance complexity. For an in depth look at this measure of complexity, the reader is referred to Van Emden [1971] or Bozdogan [1990]. For the purposes here, let $C(\cdot)$ be a real-valued measure of complexity of a system. There are various measures of complexity, such as the $C_0(\cdot)$ measure of complexity of Van Emden [1971], however the focus here is the $C_1(\cdot)$ measure of complexity, where

$$C_1\left(\mathbf{cov}\left(\hat{\theta}_m\right)\right) = \frac{k}{2}\log\left(\frac{tr\left(\mathbf{cov}\left(\hat{\theta}_m\right)\right)}{k}\right) - \frac{1}{2}\log\left|\mathbf{cov}\left(\hat{\theta}_m\right)\right|. \tag{2.1.14}$$

Rewritten as

$$C_1\left(\mathbf{cov}\left(\hat{\theta}_m\right)\right) = \frac{1}{2}\log\frac{\left(\frac{tr\left(\mathbf{cov}\left(\hat{\theta}_m\right)\right)}{k}\right)^k}{\left|\mathbf{cov}\left(\hat{\theta}_m\right)\right|},$$

(2.1.14) can be interpreted as a trade-off between the geometric mean of the average total variation of the parameters and the generalized variance of the parameters. From this measure of complexity, a general form of ICOMP is defined as

$$ICOMP(m) = -2\log L\left(\hat{\theta}_m\right) + 2C_1\left(\mathbf{cov}\left(\hat{\theta}_m\right)\right). \tag{2.1.15}$$

The issue now becomes, how to estimate $\mathbf{cov}\left(\hat{\theta}_m\right)$ in practice. Bearse and Bozdogan [2000] argue that when the parametric family of probability models, f_m, is correctly specified,

$$\mathbf{cov}\left(\theta_m^{ML}\right) \approx \mathcal{F}^{-1}\left(\theta_m^{ML}\right),$$

where $\mathcal{F}^{-1}\left(\theta_m^{ML}\right)$ is the inverse Fisher information matrix (IFIM). This leads to the ICOMP(IFIM) criterion

$$ICOMP(IFIM) = -2\log L\left(\hat{\theta}_m\right) + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_m\right)\right) \tag{2.1.16}$$

where $\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_m\right)$ is the estimated Fisher information matrix, as seen in (2.1.13), based on the maximum likelihood estimates.

2.1.9 Bayesian Model Selection Criteria

Bozdogan and Ueno [2000] derive a models selection criteria in the bayesian framework as follows:

$$\begin{aligned} BMS &= -2\log L\left(\hat{\theta}_m\right) - 2\log \pi\left(\hat{\theta}_m\right) + r\log n + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right). \end{aligned} \tag{2.1.17}$$

If a constant prior is assumed, (2.1.17) reduces to:

$$BMS = -2\log L\left(\hat{\theta}_m\right) + r\log n + \log\left|\hat{\mathcal{F}}^{-1}\right| + 2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right).$$

2.2 Genetic Algorithm

Initially developed by Holland [1975], the Genetic Algorithm (GA) is an extension of research into evolutionary algorithms. The GA allows for a search over the entire sample space, i.e. not local, without the need for testing all subsets. The GA is particularly useful when the function to be optimized is not smooth, has multiple optima, or has a large number of parameters, which is certainly true in the case subset selection. Following the methodology in Bearse and Bozdogan [1998, 2000] and Bozdogan [2003], the genetic algorithm implemented here to aid in subset selection is detailed below.

2.2.1 GA Coding Scheme

A binary string is used to indicate which combination of predictor variables are included in a model. A 1 indicates that a particular variable is included, and a 0 indicates that the variable is excluded. Each string is the same length, namely length $kp + d$, where the number of lags, p, and the number of deterministic terms, d, to be examined are determined *a priori*. Using the example in (1.2.2) and assuming that we allow for an intercept term for each variable, the binary string representing a model will be $kp + d = 3 \times 3 + 3 = 12$ digits long. Specifically, (1.2.2) is represented as

$$\underbrace{0 \quad 1 \quad 1 \quad 1}_{\mathbf{y}_1} \quad \underbrace{0 \quad 1 \quad 1 \quad 1}_{\mathbf{y}_2} \quad \underbrace{0 \quad 1 \quad 1 \quad 1}_{\mathbf{y}_3},$$

where the 0's represent the absence of intercept terms. With this method, it is easy to see that the model represented by 111111111111 is the saturated model, or that 100010001000 is the model with only an intercept for each variable.

2.2.2 Initial Population of Models

To create the first pool of potential breeding pairs, first the population size, N_{GA}, must be determined. With that determination, N_{GA} strings are randomly generated

to represent the the models in the initial population. Of course the size of the initial population, as well as a few other parameters related to how the GA explores the model space must be chosen by the user. De Jong [1975] and Grefenstette [1986] both give some idea for these parameters, $N_{GA} = 50 - 100$ and $N = 30$ respectively; however these parameters are highly dependent on each other and the optimization problem at hand.

2.2.2.1 Fitness Function

The fitness function what represents the "survival of the fittest" in the GA. Information criteria attempt to quantify how "fit" a particular model is from a statistical standpoint, and as such are the logical choice for the fitness function. All of the criteria presented in the previous section, are candidates for the fitness function, and will be examined for performance in subsequent chapters. Generally, the higher the fitness function the better, however with information criteria lower is better. This convention will be used throughout.

2.2.3 Selection for Mating Pool

The fitness function gives an idea of how fit a potential mate is, but there must be a procedure for the selection of mating a mating pool and selection of the mating pairs. Two possibilities will be considered here, and as with the fitness function will be examined for performance in subsequent chapters.

2.2.3.1 Rank Based Selection

The first selection procedure is based on the ranking procedure of Baker [1985]. First, the fitness function is computed for all N_{GA} models in the parent population. The models are then sorted by the fitness function from largest to smallest, so that the worst model (i.e. highest criterion value) is ranked 1 and the best model is ranked N_{GA}. Then, following the methodology of Bearse and Bozdogan [1998, 2000], a "weighted roulette wheel" with N_{GA} bins is created where the bin width for the i^{th} ranked model is

$$\frac{i}{N_{GA}\left(N_{GA}+1\right)/2}. \tag{2.2.1}$$

N_{GA} draws are then taken from a $Uniform\left(0,1\right)$ distribution, and a model is included in the mating pool when one of the random numbers falls in that model's bin.

2.2.3.2 Proportional Selection

The second selection procedure is a proportional selection detailed in Bozdogan [2003]. Specifically, after the fitness function (ff) is calculated for each model, the following difference is calculated for each model

$$\Delta ff_i = ff_{\max} - ff_i \tag{2.2.2}$$

for $i = 1, \ldots, N_{GA}$. The average of the differences are then computed.

$$\begin{aligned} \overline{\Delta ff} &= \frac{1}{N_{GA}} \sum_{i=1}^{N_{GA}} \Delta ff_i \\ &= \frac{1}{N_{GA}} \sum_{i=1}^{N_{GA}} ff_{\max} - ff_i \\ &= ff_{\max} - \overline{ff} \end{aligned} \tag{2.2.3}$$

Then the following ratio

$$\frac{\Delta ff_i}{\overline{\Delta ff}} \tag{2.2.4}$$

is computed for each model. (2.2.4) is then used to to pick the models to be included in the mating pool. With this method, the chance of a model being chosen for mating is proportional to (2.2.4). One possible drawback to this procedure is that the least fit model, i.e. the model with the highest ff, has a Δff_i of 0, which in turn leads to that model having no chance of being selected for mating. That shouldn't matter much in later generations, but might make quite a difference in early generations.

2.2.3.3 Boltzmann Selection

In an attempt to prevent the GA from premature convergence, Boltzmann Selection utilizes a continuously varying "temperature" to adjust how quickly the GA converges. In early generations, the temperature start high which in turn increases the randomness of mating pool selection. As generations proceed, the temperature gradually decreases thereby giving more fit models correspondingly larger chance of entering the mating pool. For some examples, refer to Goldberg [1990] or de la Maza and Tidor [1991]. Specifically, de la Maza and Tidor [1991] shows that this method performs better than fitness proportionate selection on a small group of problems. Below is one possible choice for the temperature function, where t is the generation, c_1 and c_2 are arbitrary constants that control speed of convergence and convergence value respectively, and $\sigma^2_{ff_t}$ is the variance of the fitness functions at generation t.

$$T_t = \frac{c_1 \sigma^2_{ff_t}}{t^2} + c_2 \tag{2.2.5}$$

Using the temperature, the following ratio

$$E(i,t) = \frac{\exp(-ff_i/T_t)}{\sum_{i=1}^{n} \exp(-ff_i/T_t)} \tag{2.2.6}$$

is calculated for each model. With this method, the chance of a model being chosen for mating is proportional to (2.2.6).

2.2.4 Mating

2.2.5 Mutation

2.2.6 Elitism

First introduced by De Jong [1975], elitism forces the GA to retain a predetermined number of the "best" models at each generation. This combined with Boltzmann selection will allow for a more thorough search of the model space, but will guarantee that very fit models early on are not lost.

Chapter 3

VAR Models Under Gaussian Errors

3.1 VAR Basics

Classically this is the error distribution associated with VAR as well as most other regression type models, at least initially. For that reason, the discussion here also commences in such a manner. Consider again the initial formulation of the VAR

$$\mathbf{y}_t = \Gamma D + \sum_{i=1}^{p} \Phi_i \mathbf{y}_{t-i} + \varepsilon_t. \tag{3.1.1}$$

Now the following definitions are made, as given by Lütkepohl [1993],

$$\begin{array}{ll} \mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)' & (n \times k) \\ \mathbf{x}_t = \left(D, \mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_{t-p}\right)' & (q \times 1) \\ \mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)' & (n \times q) \\ \mathbf{B} = (\Gamma, \Phi_1, \ldots, \Phi_p)' & (q \times k) \\ \mathbf{E} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)' & (n \times k)\,, \end{array} \tag{3.1.2}$$

where $q = kp + d$, and $d \geq 0$ depends on what deterministic or intercept terms are included. The definitions in (3.1.2) allow for a simplified representation of the VAR model given by

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \tag{3.1.3}$$

(3.1.3) corresponds to the saturated VAR model, however the interest here is in subset VAR models. Following the definitions in Lütkepohl [1993], (3.1.3) can be rewritten in vectorized form as

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon, \tag{3.1.4}$$

where $\mathbf{y} = vec\,(\mathbf{Y})$ is $(nk \times 1)$, $\mathbf{x} = \mathbf{I}_k \otimes \mathbf{X}$ is $(nk \times kq)$, $\beta = vec\,(\mathbf{B})$ is $(kq \times 1)$, and $\varepsilon = vec\,(\mathbf{E})$ is $(nk \times 1)$. A modification can now be applied to (3.1.4) in order to impose linear, specifically zero, constraints that will lead to a subset VAR model, see Lütkepohl [1993, Chap. 5.2]. Specifically β is constrained such that

$$\beta = \mathbf{R}\gamma, \tag{3.1.5}$$

where $\mathbf{R}$ is the $(kq \times l)$ matrix through which the zero constraints are imposed, and l is the number of unconstrained elements of β. This leads to the final form of the subset VAR

$$\mathbf{y} = \mathbf{x}^*\gamma + \varepsilon, \tag{3.1.6}$$

where $\mathbf{x}^* = \mathbf{xR}$.

3.2 Maximum Likelihood (ML) Estimation

Here it is assumed that the errors are normally distributed, that is

$$\varepsilon \sim N_{nk}\left(\mathbf{0}, \Omega\right),$$

where $\Omega = \Sigma \otimes \mathbf{I}_n$. Therefore, the probability density function (pdf) of ε is

$$\begin{aligned} f\left(\varepsilon\right) &= \frac{1}{\left(2\pi\right)^{nk/2}\left|\Omega\right|^{1/2}} \exp\left[-\frac{1}{2}\varepsilon'\Omega^{-1}\varepsilon\right] \\ &= \frac{1}{\left(2\pi\right)^{nk/2}\left|\Omega\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{y}-\mathbf{x}^*\gamma\right)'\Omega^{-1}\left(\mathbf{y}-\mathbf{x}^*\gamma\right)\right]. \end{aligned} \tag{3.2.1}$$

Lütkepohl [1993, Chap. 5.2] shows that under the assumption of normality, the ML estimates are equivalent to the generalized least squares (GLS) estimates. These in turn are estimated by the two-stage feasible generalized least squares (FGLS), namely

$$\widetilde{\gamma} = \left(\mathbf{x}^{*\prime}\widehat{\Omega}^{-1}\mathbf{x}^*\right)^{-1}\mathbf{x}^{*\prime}\widehat{\Omega}^{-1}\mathbf{y} \tag{3.2.2}$$

where $\widehat{\Omega} = \widehat{\Sigma} \otimes \mathbf{I}_n$, and the estimator $\widehat{\Sigma}$ is obtained from the least squares (LS) residuals from the saturated or the subset model, Hamilton [1994]. The variance can then be estimated as

$$\widetilde{\Sigma} = \frac{\widetilde{\mathbf{E}}'\widetilde{\mathbf{E}}}{n}, \tag{3.2.3}$$

where $\widetilde{\mathbf{E}}$ is the FGLS error matrix formed by reshaping

$$\widetilde{\mathbf{e}} = \mathbf{y} - \mathbf{x}^*\widetilde{\gamma}.$$

So the steps to complete the two-stage FGLS Gaussian ML estimation are:

1. Calculate the LS fit.
   (a) Calculate either $\widehat{\beta} = \left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\mathbf{y}$ for the saturated model, or $\widehat{\gamma} = \left(\mathbf{x}^{*\prime}\mathbf{x}^*\right)^{-1}\mathbf{x}^{*\prime}\mathbf{y}$ for the subset model.
   (b) Reshape $\widehat{\beta}$ or $\widehat{\gamma}$ to give $\widehat{\mathbf{B}}$.
   (c) Calculate $\widehat{\Omega} = \widehat{\Sigma} \otimes \mathbf{I}_n$, where $\widehat{\Sigma} = \frac{\left(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}\right)'\left(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}\right)}{n}$.

2. Calculate the GLS fit.

   (a) Calculate $\widetilde{\gamma} = \left(\mathbf{x}^{*\prime}\widehat{\Omega}^{-1}\mathbf{x}^{*}\right)^{-1}\mathbf{x}^{*\prime}\widehat{\Omega}^{-1}\mathbf{y}$.

   (b) Calculate $\widetilde{\Sigma} = \frac{\left(\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{B}}\right)'\left(\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{B}}\right)}{n}$.

By defining the $(l + k(k+1)/2)$-vector $\widetilde{\theta} \equiv \left(\widetilde{\gamma}', vech\left(\widetilde{\Sigma}\right)'\right)'$, the inverse Fisher information matrix is given by

$$\mathcal{F}^{-1}\left(\widetilde{\theta}\right) = \begin{bmatrix} \mathbf{cov}\left(\widetilde{\gamma}\right) & \mathbf{0} \\ \mathbf{0} & \frac{2}{n}\mathbf{D}_p^{+}\left(\widetilde{\Sigma}\otimes\widetilde{\Sigma}\right)\mathbf{D}_p^{+\prime} \end{bmatrix}, \tag{3.2.4}$$

where $\mathbf{D}_p^{+}$ is the Moore-Penrose inverse of the duplication matrix $\mathbf{D}_p$, and $\mathbf{cov}\left(\widetilde{\gamma}\right) = \left(\mathbf{x}^{*\prime}\Omega^{-1}\mathbf{x}^{*}\right)^{-1}$ is the asymptotic covariance matrix of the subset VAR coefficients.

When a portfolio of subset VAR models are considered, the log-likelihood is calculated at the FGLS estimates for each of the M subset models which leads to the following

$$\begin{aligned} \log L\left(\widetilde{\theta}_m\right) &= -\frac{nk}{2}\log(2\pi) - \frac{n}{2}\log\left|\widetilde{\Sigma}_m\right| \\ &\quad -\frac{1}{2}\left(\mathbf{y}-\mathbf{x}_m^{*}\widetilde{\gamma}_m\right)'\left(\widetilde{\Sigma}_m^{-1}\otimes\mathbf{I}_n\right)\left(\mathbf{y}-\mathbf{x}_m^{*}\widetilde{\gamma}_m\right) \\ &= -\frac{nk}{2}\log(2\pi) - \frac{n}{2}\log\left|\widetilde{\Sigma}_m\right| - \frac{nk}{2}. \end{aligned} \tag{3.2.5}$$

Also $\widetilde{\theta}_m \equiv \left(\widetilde{\gamma}_m', vech\left(\widetilde{\Sigma}_m\right)'\right)'$ is the $r_m \equiv (l_m + k(k+1)/2)$-vector of FGLS estimates of the parameters in the m^{th} subset VAR model with the $\mathbf{x}_m^{*}$ predictor matrix.

3.3 Model Selection

Using the results from (3.2.5), the Information Criteria can be calculated.

3.3.0.1 AIC

$$\begin{aligned} AIC(m) &= -2\log L\left(\widetilde{\theta}_m\right) + 2r_m \\ &= nk\log(2\pi) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad +2\left(l_m + k(k+1)/2\right). \end{aligned}$$

3.3.0.2 SBC

$$\begin{aligned} SBC\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m \log\left(n\right) \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\log\left(n\right). \end{aligned}$$

3.3.0.3 HQ

$$\begin{aligned} HQ\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m \log\log\left(n\right) \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\log\log\left(n\right). \end{aligned}$$

3.3.0.4 $\mathbf{AIC_C}$

$$\begin{aligned} AIC_{C_A}\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + 2\frac{r_m}{n - r_m - 1} \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + 2\frac{\left(l_m + k\left(k+1\right)/2\right)}{n - \left(l_m + k\left(k+1\right)/2\right) - 1} \end{aligned}$$

$$\begin{aligned} AIC_{C_B}\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + 2\frac{r_m}{n - r_m - (k+1)/2} \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + 2\frac{\left(l_m + k\left(k+1\right)/2\right)}{n - \left(l_m + k\left(k+1\right)/2\right) - (k+1)/2} \end{aligned}$$

3.3.0.5 CAIC

$$\begin{aligned} CAIC\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\left(\log\left(n\right) + 1\right) \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\left(\log\left(n\right) + 1\right) \end{aligned}$$

3.3.0.6 CAICF

$$\begin{aligned} CAICF\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\left(\log\left(n\right)+2\right) + \log\left|\hat{\mathcal{F}}^{-1}\right| \end{aligned}$$

3.3.0.7 $\mathbf{CAICF_E}$

$$\begin{aligned} CAICF_E\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\left(\log\left(n\right)+2\right) + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \end{aligned}$$

3.3.0.8 $\mathbf{CAICF_C}$

$$\begin{aligned} CAICF_C\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2\left(\frac{nr_m}{n-r_m-2}\right) \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\left(\log\left(n\right)+2\right) + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2\left(\frac{nr_m}{n-r_m-2}\right) \end{aligned}$$

3.3.0.9 GAIC

$$\begin{aligned} GAIC\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + 2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad +2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \end{aligned}$$

3.3.0.10 ICOMP

$$ICOMP\left(m\right)=-2\log L\left(\widetilde{\theta}_{m}\right)+2C_{1}\left(\hat{\mathcal{F}}^{-1}\left(\widetilde{\theta}_{m}\right)\right),$$

where

$$\hat{\mathcal{F}}^{-1}\left(\widetilde{\theta}_{m}\right)=\left[\begin{array}{cc}\widehat{\mathbf{cov}}\left(\widetilde{\gamma}_{m}\right) & \mathbf{0}\\ \mathbf{0} & \frac{2}{n}\mathbf{D}_{p}^{+}\left(\widetilde{\Sigma}_{m}\otimes\widetilde{\Sigma}_{m}\right)\mathbf{D}_{p}^{+\prime}\end{array}\right],$$

and $\widehat{\mathbf{cov}}\left(\widetilde{\gamma}_{m}\right)$ is a consistent estimator of the asymptotic covariance matrix of the subset VAR coefficients. Bearse and Bozdogan [1998] show that ICOMP can be expressed as

$$\begin{aligned}ICOMP\left(m\right) &= nk\left(\log\left(2\pi\right)+1\right)+n\log\left|\widetilde{\Sigma}_{m}\right|\\ &\quad+r_{m}\log\left(\frac{tr\left(\widehat{\mathbf{cov}}\left(\widetilde{\gamma}_{m}\right)\right)+\frac{1}{2n}G_{m}}{r_{m}}\right)\\ &\quad-\log\left|\widehat{\mathbf{cov}}\left(\widetilde{\gamma}_{m}\right)\right|-k\log\left(2\right)+\frac{k\left(k+1\right)}{2}\log\left(n\right)\\ &\quad-\left(k+1\right)\log\left|\widetilde{\Sigma}_{m}\right|\end{aligned}$$

where

$$G_{m}\equiv tr\left(\widetilde{\Sigma}_{m}^{2}\right)+\left(tr\left(\widetilde{\Sigma}_{m}\right)\right)^{2}+2\sum_{i=1}^{k}\widetilde{\sigma}_{i,m}^{2}$$

and $\widetilde{\sigma}_{i,m}^{2}$ is the i^{th} diagonal element of $\widetilde{\Sigma}_{m}$.

3.3.0.11 BMS

$$\begin{aligned}BMS\left(m\right) &= -2\log L\left(\widetilde{\theta}_{m}\right)+r_{m}\log n+\log\left|\hat{\mathcal{F}}^{-1}\right|+2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right)\\ &= nk\log\left(2\pi\right)+n\log\left|\widetilde{\Sigma}_{m}\right|+nk\\ &\quad+\left(l_{m}+k\left(k+1\right)/2\right)\log n\\ &\quad+\log\left|\hat{\mathcal{F}}^{-1}\right|+2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right)\end{aligned}$$

Chapter 4

VAR Models Under Non-Normal Errors

In practice, the assumption of normality in errors is not often satisfied. Simply put, the normal distribution doesn't adequately describe much "real world" data. This is sometimes remedied through arbitrary transformations to achieve near-normality, however, that should not be the goal. It doesn't make sense to manipulate data to fit a model any more than it makes sense to put a square peg into a round hole. The goal of modeling should be to fit the best model possible to the data, not the other way around. To this end, models based on more general distributions are required. There are many more robust and adaptive choices to model the distribution of the errors such as the Multivariate Pearson Family of distributions, the Multivariate Power Exponential (PE) Distribution Gómez et al. [1998], and the Multivariate Generalized t Distribution (MGT) Arslan [2004] to name a few.

4.1 Power Exponential Errors

One possible choice for a more adaptive and robust error term is the Multivariate Power Exponential (PE) Distribution Gómez et al. [1998]. For the multivariate regression model, this has been initially explored Liu and Bozdogan [2004]. Here it is adapted to the VAR framework, which can be written in the multivariate regression format, as seen in (3.1.3), (3.1.4), and (3.1.6).

The univariate PE distribution was initially introduced, as an extension to the normal distribution, by Subbotin [1923], and has been used by Box [1953] and Box and Tiao [1973]. The univariate PE distribution is defined as

$$f\left(x;\mu,\sigma,\beta\right)=\frac{1}{\sigma\Gamma\left(1+\frac{1}{2\beta}\right)2^{1+\frac{1}{2\beta}}}\exp\left(-\frac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^{2\beta}\right), \tag{4.1.1}$$

where $\mu\in\mathbb{R}$ is the location, $\sigma>0$ is the scale, and $\beta>0$ is the shape parameter which is related to kurtosis. The PE distribution is robust in that (4.1.1) can represent many

symmetric unimodal distributions. When $\beta = 1$ (4.1.1) is the Normal distribution, for $\beta = .5$ it is the Laplace, or Double Exponential distribution, and as $\beta \to \infty$ (4.1.1) becomes the Uniform distribution.

Gómez et al. [1998] develop a multivariate generalization of the univariate PE, denoted $PE_k(\mu, \Sigma, \beta)$, defined as

$$f(\mathbf{x}; \mu, \Sigma, \beta) = C |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[(\mathbf{x}-\mu)' \Sigma^{-1} (\mathbf{x}-\mu)\right]^{\beta}\right), \tag{4.1.2}$$

where

$$C = \frac{k\Gamma\left(\frac{k}{2}\right)}{\pi^{\frac{k}{2}}\Gamma\left(1+\frac{k}{2\beta}\right)2^{1+\frac{k}{2\beta}}},$$

and $\mu \in \mathbb{R}^k$ is the location, Σ is the $(k \times k)$ positive definite scale matrix, and $\beta > 0$ is the shape parameter which is related to kurtosis. This formulation is particularly attractive given that when $k = 1$, (4.1.2) reduces to (4.1.1). Figure ?? shows the shape of a bivariate power exponential for various choices of β. To be useful for time series, a further modification is necessary. This is where the matrix variate power exponential distribution, denoted $MPE_{n\times k}(\mu, \Phi, \Sigma, \beta)$ and developed by Sánchez-Manzano et al. [2002], becomes useful.

Here we make the assumption that the errors are distributed as multivariate power exponential, that is

$$\varepsilon \sim PE_{nk}(\mathbf{0}, \Omega, \beta),$$

where $\Omega = \Sigma \otimes \mathbf{I}_n$. Therefore, the probability density function (pdf) of ε is

$$\begin{aligned} f(\varepsilon) &= \frac{nk\Gamma\left(\frac{nk}{2}\right)}{\pi^{\frac{nk}{2}}\Gamma\left(1+\frac{nk}{2\beta}\right)2^{1+\frac{nk}{2\beta}}} |\Omega|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[\varepsilon'\Omega^{-1}\varepsilon\right]^{\beta}\right) \\ &= \frac{nk\Gamma\left(\frac{nk}{2}\right)}{\pi^{\frac{nk}{2}}\Gamma\left(1+\frac{nk}{2\beta}\right)2^{1+\frac{nk}{2\beta}}} |\Sigma \otimes \mathbf{I}_n|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\left[(\mathbf{y}-\mathbf{x}^*\gamma)'\left(\Sigma^{-1}\otimes\mathbf{I}_n\right)(\mathbf{y}-\mathbf{x}^*\gamma)\right]^{\beta}\right) \end{aligned}$$

$$\begin{aligned} l(\theta) \equiv \log L(\theta) &= \log\left(nk\Gamma\left(\frac{nk}{2}\right)\right) - \frac{nk}{2}\log(\pi) \\ &\quad - \log\Gamma\left(1+\frac{nk}{2\beta}\right) - \left(1+\frac{nk}{2\beta}\right)\log(2) \\ &\quad - \frac{n}{2}\log\left|\Sigma^{-1}\otimes\mathbf{I}_n\right| - \frac{1}{2}\left[(\mathbf{y}-\mathbf{x}\beta)'\left(\Sigma^{-1}\otimes\mathbf{I}_n\right)(\mathbf{y}-\mathbf{x}\beta)\right]^{\beta} \end{aligned}$$

$$\frac{\partial l(\theta)}{\partial \beta} = \beta\mathbf{x}'\left(\Sigma^{-1}\otimes\mathbf{I}_n\right)(\mathbf{y}-\mathbf{x}\beta)\left[(\mathbf{y}-\mathbf{x}\beta)'\left(\Sigma^{-1}\otimes\mathbf{I}_n\right)(\mathbf{y}-\mathbf{x}\beta)\right]^{\beta-1}$$

$$\frac{\partial l(\theta)}{\partial \beta}$$

Figure 4.1.1: $PE_2(\mu, \Sigma, \beta)$ density plots with $\mu = [0,0]'$ and $\Sigma = I_2$ for various choices of β.

$$f(\varepsilon) = \frac{nk\Gamma\left(\frac{nk}{2}\right)}{\pi^{\frac{nk}{2}}\Gamma\left(1+\frac{nk}{2\beta}\right)2^{1+\frac{nk}{2\beta}}}\left|\Sigma\right|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}tr\left[\Sigma^{-1}\left(\mathbf{Y}-\mathbf{XB}\right)'\left(\mathbf{Y}-\mathbf{XB}\right)\right]^{\beta}\right)$$

4.1.1 Model Selection

Using the results from (3.2.5), the Information Criteria can be calculated.

4.1.1.1 AIC

$$\begin{aligned} AIC(m) &= -2\log L\left(\widetilde{\theta}_m\right) + 2r_m \\ &= nk\log(2\pi) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad +2\left(l_m + k(k+1)/2\right). \end{aligned}$$

4.1.1.2 SBC

$$\begin{aligned} SBC(m) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\log(n) \\ &= nk\log(2\pi) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad +\left(l_m + k(k+1)/2\right)\log(n). \end{aligned}$$

4.1.1.3 HQ

$$\begin{aligned} HQ\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m \log\log\left(n\right) \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\log\log\left(n\right). \end{aligned}$$

4.1.1.4 $\mathbf{AIC_C}$

$$\begin{aligned} AIC_{C_A}\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + 2\frac{r_m}{n - r_m - 1} \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + 2\frac{\left(l_m + k\left(k+1\right)/2\right)}{n - \left(l_m + k\left(k+1\right)/2\right) - 1} \end{aligned}$$

$$\begin{aligned} AIC_{C_B}\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + 2\frac{r_m}{n - r_m - (k+1)/2} \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + 2\frac{\left(l_m + k\left(k+1\right)/2\right)}{n - \left(l_m + k\left(k+1\right)/2\right) - \left(k+1\right)/2} \end{aligned}$$

4.1.1.5 CAIC

$$\begin{aligned} CAIC\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\left(\log\left(n\right) + 1\right) \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\left(\log\left(n\right) + 1\right) \end{aligned}$$

4.1.1.6 CAICF

$$\begin{aligned} CAICF\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &= nk\log\left(2\pi\right) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad + \left(l_m + k\left(k+1\right)/2\right)\left(\log\left(n\right) + 2\right) + \log\left|\hat{\mathcal{F}}^{-1}\right| \end{aligned}$$

4.1.1.7 $\mathbf{CAICF_E}$

$$\begin{aligned} CAICF_E(m) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \\ &= nk\log(2\pi) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad +\left(l_m + k(k+1)/2\right)(\log(n)+2) + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \end{aligned}$$

4.1.1.8 $\mathbf{CAICF_C}$

$$\begin{aligned} CAICF_C(m) &= -2\log L\left(\widetilde{\theta}_m\right) + r_m\left[\log n + 2\right] + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2\left(\frac{nr_m}{n - r_m - 2}\right) \\ &= nk\log(2\pi) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad +\left(l_m + k(k+1)/2\right)(\log(n)+2) + \log\left|\hat{\mathcal{F}}^{-1}\right| \\ &\quad +2\left(\frac{nr_m}{n - r_m - 2}\right) \end{aligned}$$

4.1.1.9 GAIC

$$\begin{aligned} GAIC(m) &= -2\log L\left(\widetilde{\theta}_m\right) + 2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \\ &= nk\log(2\pi) + n\log\left|\widetilde{\Sigma}_m\right| + nk \\ &\quad +2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \end{aligned}$$

4.1.1.10 ICOMP

$$ICOMP(m) = -2\log L\left(\widetilde{\theta}_m\right) + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\widetilde{\theta}_m\right)\right),$$

where

$$\hat{\mathcal{F}}^{-1}\left(\widetilde{\theta}_m\right) = \begin{bmatrix} \widehat{\mathbf{cov}}\left(\widetilde{\gamma}_m\right) & \mathbf{0} \\ \mathbf{0} & \frac{2}{n}\mathbf{D}_p^{+}\left(\widetilde{\Sigma}_m \otimes \widetilde{\Sigma}_m\right)\mathbf{D}_p^{+\prime} \end{bmatrix},$$

and $\widehat{\mathbf{cov}}\left(\widetilde{\gamma}_m\right)$ is a consistent estimator of the asymptotic covariance matrix of the subset VAR coefficients. Bearse and Bozdogan [1998] show that ICOMP can be

expressed as

$$\begin{aligned} ICOMP\left(m\right) &= nk\left(\log\left(2\pi\right)+1\right)+n\log\left|\widetilde{\Sigma}_m\right| \\ &\quad +r_m\log\left(\frac{tr\left(\widehat{\mathbf{cov}}\left(\widetilde{\gamma}_m\right)\right)+\frac{1}{2n}G_m}{r_m}\right) \\ &\quad -\log\left|\widehat{\mathbf{cov}}\left(\widetilde{\gamma}_m\right)\right|-k\log\left(2\right)+\frac{k\left(k+1\right)}{2}\log\left(n\right) \\ &\quad -\left(k+1\right)\log\left|\widetilde{\Sigma}_m\right| \end{aligned}$$

where

$$G_m \equiv tr\left(\widetilde{\Sigma}_m^2\right)+\left(tr\left(\widetilde{\Sigma}_m\right)\right)^2+2\sum_{i=1}^{k}\widetilde{\sigma}_{i,m}^2$$

and $\widetilde{\sigma}_{i,m}^2$ is the i^{th} diagonal element of $\widetilde{\Sigma}_m$.

4.1.1.11 BMS

$$\begin{aligned} BMS\left(m\right) &= -2\log L\left(\widetilde{\theta}_m\right)+r_m\log n+\log\left|\hat{\mathcal{F}}^{-1}\right|+2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \\ &= nk\log\left(2\pi\right)+n\log\left|\widetilde{\Sigma}_m\right|+nk \\ &\quad +\left(l_m+k\left(k+1\right)/2\right)\log n \\ &\quad +\log\left|\hat{\mathcal{F}}^{-1}\right|+2tr\left(\hat{\mathcal{F}}^{-1}\hat{R}\right) \end{aligned}$$

4.2 Multivariate Generalized t Errors

Another choice for a more adaptive and robust error term, along the same lines of the Multivariate PE distribution, is the Multivariate Generalized t Distribution (MGT) Arslan [2004]. Here, the VAR model with MGT errors is developed.

Chapter 5

Conclusions

Part II

Adaptive Text Mining

Chapter 6

Introduction to Text Mining

6.1 Mixture Modeling

Chapter 7

Document Clustering

7.1 Binary Membership Matrix

7.1.1 Mixtures of Multivariate Bernoulli Distributions

7.2 Document Term Frequency Matrix

7.2.1 Product "Power Law Distribution" – For density estimation

7.2.2 Product Kernel of "Power Law Distribution" – For density estimation

7.3 GA for Parameter Estimation

Chapter 8

Conclusions

Bibliography

Hirotsugu Akaike. Information theory and and extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Academiai Kiado. 9, 10

Hirotsugu Akaike. A new look at statistical model evaluation. *IEEE Transactions on Automatic Control*, AC-19:716–723, 1974. 9

Hirotsugu Akaike. A Bayesian analysis of the minimum AIC procedure. *Annal of the Institute of Statistical Mathematics*, A 30:9–14, 1978. 9

Oclay Arslan. Family of multivariate generalized t distributions. *Journal of Multivariate Analysis*, 89:329–337, 2004. 23, 28

J.E. Baker. Adaptive selection methods for genetic algorithms. In J.J. Grefenstette, editor, *Proceedings of the First International Conference on Genetic Algorithms and Their Applications*, pages 100–111. Lawrence Erlbaum Associates, 1985. 14

Peter Bearse and Hamparsum Bozdogan. Subset selection in vector autoregressive models using the genetic algorithm with informational complexity as the fitness function. *SAMS*, 31:61–91, 1998. 13, 14, 22, 27

Peter Bearse and Hamparsum Bozdogan. A new approach to vector autoregressive (VAR) forecasting using the genetic algorithm with information complexity as the fitness function. Unpublished Manuscript, 2000. 12, 13, 14

G. Box. A note on regions for the tests of kurtosis. *Biometrika*, 40:465–468, 1953. 23

G. Box and G. Tiao. *Bayesian inference in statistical analysis.* Addison-Wesley Publishing Co., Reading, 1973. 23

Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): General theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987. 8, 10, 11

Hamparsum Bozdogan. ICOMP: a new model selection criterion. In Hans Herman Bock, editor, *Classification and Related Methods of Data Analysis*, pages 599–608, Amsterdam, 1988. Elsevier Science Publishers. 12

Hamparsum Bozdogan. On the information-based measure of covariance complexity an its application to the evaluation of multivariate linear models. *Communications in Statistics: Theory and Methods*, 19:221–278, 1990. 12

Hamparsum Bozdogan. Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44(1):62–91, 2000. 8

Hamparsum Bozdogan. Intelligent statistical data mining with information complexity and genetic algorithms. In Hamparsum Bozdogan, editor, *Statistical Data Mining and Knowledge Discovery*, pages 1–42, Boca Raton, 2003. CRC Press. 13, 14

Hamparsum Bozdogan and M. Ueno. A unified approach to information-theoretic and bayesian model selection criteria. Working paper, 2000. 11, 13

Mark J.C. Crescenzi and Andrew J. Enterline. Ripples from the waves? a systemic, time-series analysis of democracy, democratization, and interstate war. *Journal of Peace Research*, 36(1):75–94, 1999. 5

K.A. De Jong. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*. PhD thesis, University of Michigan, Ann Arbor, 1975. 14, 16

M. de la Maza and B. Tidor. Boltzmann weighted selection improves performance of genetic algorithms. Technical Report 1345, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, 1991. 15

John R. Freeman, John T. Williams, and Tse Lin. Vector autoregression and the study of politics. *American Journal of Political Science*, 33(4):842–877, 1989. 5

D.E. Goldberg. A note on boltzmann tournament selection for genetic algorithms and population-oriented simulated annealing. *Complex Systems*, 4:445–460, 1990. 15

E. Gómez, M. A. Gómez-Villegas, and J. M. Marín. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]*, 27(3):589–600, 1998. 23, 24

J.J. Grefenstette. Optimization of control parameters for genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics 16*, 1:122–128, 1986. 14

James D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, 1994. 18

E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B, Methodological*, 41:190–195, 1979. 10

John H. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Princeton, 1975. (Second edition: MIT Press, 1992). 13

Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989. 10

Clifford M. Hurvich and Chih-Ling Tsai. A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis*, 14: 271–279, 1993. 10

Sadanori Konishi and Genshiro Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83:875–890, 1996. 9

S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951. 9

Min-Hui Liu and Hamparsum Bozdogan. Multivariate regression models with power exponential random errors and subset selection using genetic algorithms with information complexity. Submitted for Review, 2004. 23

Helmut Lütkepohl. *Introduction to multiple time series analysis.* Springer-Verlag Inc, New York, 1993. 17, 18

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978. 9

E. G. Sánchez-Manzano, M. A. Gómez-Villegas, and J. M. Marín-Diazaraque. A matrix variate generalization of the power exponential family of distributions. *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]*, 31(12):2167–2182, 2002. 24

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464, 1978. 9

Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980. 5

M. Subbotin. On the law of frequency of errors. *Mathematicheskii Sbornik*, 31:296–300, 1923. 23

Nariaki Sugiura. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]*, 7:13–26, 1978. 10

K. Takeuchi. Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18, 1976. [In Japanese]. 11

M.H. Van Emden. An analysis of complexity. Mathematical Centre Tracts, Amsterdam, Vol. 35, 1971. 12

Vita

Vita goes here...