BOUNDED INFLUENCE APPROACHES TO CONSTRAINED MIXED VECTOR AUTOREGRESSIVE MODELS

by

Mark Amper Gamalo

BS, Ateneo de Manila University, 1998

MS, University of The Philippines, 2001

MA, University of Pittsburgh, 2003

Submitted to the Graduate Faculty of

Faculty of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

Mark Amper Gamalo

It was defended on

June 08, 2006

and approved by

David S. Stoffer

Wesley Thompson

Sati Mazumdar

J. Richard Jennings

Dissertation Director: David S. Stoffer

BOUNDED INFLUENCE APPROACHES TO CONSTRAINED MIXED VECTOR AUTOREGRESSIVE MODELS

Mark Amper Gamalo, PhD

University of Pittsburgh, 2006

The proliferation of many clinical studies obtaining multiple biophysical signals from several individuals repeatedly in time is increasingly recognized, a recognition generating growth in statistical models that analyze cross-sectional time series data. In general, these statistical models try to answer two questions: (i) intra-individual dynamics of the response and its relation to some covariates; and, (ii) how this dynamics can be aggregated consistently in a group. In response to the first question, we propose a covariate-adjusted constrained Vector Autoregressive model, a technique similar to the the STARMAX model (Stoffer, *JASA* **81**, 762-772), to describe serial dependence of observations. In this way, the number of parameters to be estimated is kept minimal while offering flexibility for the model to explore higher order dependence. In response to (ii), we use mixed effects analysis that accommodates modelling of heterogeneity among cross-sections arising from covariate effects that vary from one cross-section to another.

Although estimation of the model can proceed using standard maximum likelihood techniques, we believed it is advantageous to use bounded influence procedures in the modelling (such as choosing constraints) and parameter estimation so that the effects of outliers can be controlled. In particular, we use M-estimation with a redescending bounding function because its influence function is always bounded. Furthermore, assuming consistency, this influence function is useful to obtain the limiting distribution of the estimates. However, this distribution may not necessarily yield accurate inference in the presence of contamination as the actual asymptotic distribution might have wider tails. This led us to investigate boot-

strap approximation techniques. A sampling scheme based on IID innovations is modified to accommodate the cross-sectional structure of the data. Then the M-estimation is applied to each bootstrap sample naively to obtain the asymptotic distribution of the estimates.

We apply these strategies to the extracted BOLD activation from several regions of the brain from a group of individuals to describe joint dynamic behavior between these locations. We used simulated data with both innovation and additive outliers to test whether the estimation procedure is accurate despite contamination.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Writing this research has been a long and arduous endeavor. At first, it seemed like I was just hovering around; searching what topic would be a good one to write about. I kept reading many journals with keen interest on some topics in functional magnetic resonance imaging and its relation with statistical methods. There were times when there seems to be no end and all I can see is that my skies are bruised black and all my hopes fading. Many times I gazed in envy at other people who, like children playing on the sea shore, have their lives all set with their molded castles while I am still taking a long walk with but a washed trail of sand that trickled my hand. But now, when all things are done, I realized that it is not what I have in the end but what I have given...

I am grateful to many people who have influenced me to take this long and arduous path, particularly, Fr. Bobby Buenconsejo and Dr. Augusto Hermosilla who believed in my abilities and encouraged me to pursue my dreams. I am also glad that I came to meet Dr. J. Richard Jennings, Dr. Hernando Ombao, and Dr. Pete Gianaros whose rigor and discipline is paragon to aspiring scientists. Lastly, I am deeply indebted to Dr. David Stoffer whose support had been ubiquitous, guiding me through this work wherever he maybe, while showing me the discipline of an independent thinker.

I also want to thank all those for and with whom I spent the sands of time: my friends Ruth, Lovella and Theo, Lot, Belle and Paul and baby Luke, who have been my acoustic emotional sounding board throughout this research; my aunt Marilou and uncle Edwin, Nanay Erna and Tatay Hernando who were always there when I needed help; my Orpheus who showed me both happiness and grief in a passionate "rollercoaster" ride; my family (Mama and Papa, Alma, Genes, Winston, Ernie and Genelyn) who has protected me and loved me the best they could that sometimes I think I could never equal their kindness; and

lastly, to the One Almighty for giving me strength and hope.

1.0 INTRODUCTION

Economic, financial, environmental, and recently, biophysical data are usually characterized by multiple responses collected at roughly equally spaced time intervals. Such time series observations can come from a single entity or a finite number of individuals that are nested within naturally occurring groups. Analyzing and modelling the series jointly is important not only to improve accuracy of forecasts but also to understand dynamic relationship among them. For example, in studying interaction among distributed neural systems in the brain one may look at measurements of blood oxygenation level dependent (BOLD) signal from several brain regions to see if they are contemporaneously related, if one location leads the others or if there is any feedback relationship among them.

In the past 35 years, since the work of Box and Jenkins [17], the class of Vector Autoregressive (VAR) models of the form

$$\mathbf{z}_t = \boldsymbol{\alpha} + \sum_{h=1}^{n} \Pi_h \mathbf{z}_{t-h} + \boldsymbol{\varepsilon}_t, \tag{1.1}$$

where $\mathbf{z}_t = (z_{1t}, \ldots, z_{pt})$ is the p-dimensional vector of observations at time t, have been investigated extensively and were found useful in representing dynamic serially dependent relationship of time series data. Since then numerous methods and considerations relating them to actual data have been discussed in literature, e.g. accounting for influence of exogenous variables known or suspected to be related to the series of interest using vector ARMAX model [42], removing replication bias in repeated measurements of autoregressions [3], incorporating random effects on coefficients of cross-sectional time series [77], jointly modelling spatio-temporal structure of spatial time series [75, 93].

In this chapter, the problem of modelling interaction of distributed neural systems in the brain and its relation to vector autoregressive models will be introduced. This will

serve as the background from which a general modelling strategy will be developed. We will also present some characteristics of the data set that influence our modelling strategy and estimation.

1.1 MODELLING INTERACTION OF DISTRIBUTED NEURAL SYSTEMS IN THE BRAIN

Historically, functional magnetic resonance imaging (fMRI) was just concerned with localization of neurological function to the neuroanatomy, i.e., mapping where in the brain neural computations mediate a cognitive process of interest. This is accomplished through linear time-invariant models relating the time course of an experimentally controlled stimulus with reports of neural activity detected as blood oxygen level dependent (BOLD) signal [74, 7, 8, 57] for each voxel or three dimensional pixel in the brain. These models may be able to determine focal activations but are incapable of determining causal mechanisms translating local neural dynamics into BOLD signals, i.e., how functionally specialized neuronal systems influence another local neuronal response.

Integration as a principle of organization for distributed neural systems in the brain is best understood in terms of effective connectivity. Effective connectivity is a dynamic and context-dependent causal model that replicates observed timing relationships between the recorded neurons [1]. This definition implies that such analysis is based on hypothesis driven statistical models that restricts inference to networks comprising a number of preselected regions. Early examples of these statistical procedures are linear model variants such as structural equation modelling [71, 19] that makes *a priori* causal semantics between regions that would minimize the discrepancy between the observed and implied correlations. A major criticism in regression models, however, is that they only quantify instantaneous correlations and ignore any temporal information in the measurement of directed influence.

VAR models fit into this scheme of incorporating temporal effects in modelling interregional dependencies in the data in the sense that the nature of interaction is characterized in terms of the historical influence one variable has on another. It must be noted, however,

that these models are really not concerned with causality of the regions of interest *per se*; rather, they address the temporal aspects of causality by explaining regional BOLD signals at time $\mathbf{z}_t$ as a linear combination of n previous vector values, whose contributions are weighted by the parameter matrices $\mathbf{\Pi}_h$, plus an error term $\boldsymbol{\varepsilon}_t$ as given by (1.1). Information regarding directed influence one region has on another is inferred only through their mutual predictability from the past data points.

Although the VAR models has been an established statistical technique, its use in fMRI has only been suggested quite recently with the work of Harrison *et al.* [43]. Their approach augmented the usual p-dimensional vector series with a series obtained by coupling certain components of the vector to represent modulatory effects on connections and used Bayesian schemes to estimate the parameters. Shortly after, deviations from using explicit models of interaction were investigated with the application of classical VAR model on the whole brain. This methodology was presented by Valdes-Sosa [100] with the objective of obtaining whole-brain connectivity maps that also accounts for the underlying continuous spatial manifold of the brain. The same goal but using complementary VAR approach in the context of Granger causality [37, 38] was pursued, just recently, by Roebroeck *et al.* [80]. Their technique calls for the evaluation of linear dependence [34] among all voxel pairs throughout the whole brain image to determine the existence and direction of influence.

The vector autoregressive model may serve well the purpose of quantifying directed influence but it can still be further improved to obtain better results and render accurate conclusions. One such improvement is the evaluation of the context-dependent network of influences in the brain over a sample of individuals to test the hypothesis with greater sensitivity and to determine if the underlying connectivity generalizes to a certain population. Most of the autoregressive techniques mentioned above do this by analyzing data from each subject separately. This approach, however, can lead to inconsistent maps [35] as each individual may adopt varying cognitive strategies or the brain adopts degenerative solutions to perform the same task and may be reflected as changes in the network. On the other hand, pooling observations together as if they come from one subject ignores variation between individuals and, although it may increase sensitivity, a pronounced variation from a single subject can have a dramatic effect on the network. This raises an interesting question of how

to determine consistent network influences while still allowing certain individual variation [72].

Another area of improvement, as suggested by Valdez-Sosa [100], relies on the fact that neuroimaging data is a spatio-temporal data, i.e., it is a vector valued time series sampled over space. Hence, it must be modelled jointly in space and time. However, the additional dimension complicates the modelling of an already highly parameterized model. The challenge then is how to model a network that accommodates the spatial structure while maintaining model parsimony. Moreover, it is desired that this procedure adhere to the autoregressive principle of the mutual predictability of the current value upon its past.

1.2 CHARACTERIZING THE DATASET: ITS INFLUENCE IN MODELING STRATEGY

In this research, we will use data from an fMRI study investigating cortical and subcortical brain regions that are thought to initiate and represent blood pressure reactions to behavioral stressors (see [33] for details). Three of these cortical regions include the anterior and perigenual cingulate and the insula which will serve as a testbed for the technique that will be developed. In each of these regions, representative BOLD activity are extracted from the fMRI scans of the 19 subjects who participated in the experiment. Details of the data set are described in Chapter 4.

Inspecting the three time series plots from each of the subjects reveal some characteristics which are worth noting. While figure 1 exhibits the BOLD activity time course extracted from the three regions of interest for subjects H07, H09, and H11, most of the time series courses often have spike-like transients at any particular time point or range of time points. These spikes are actually effects of magnetic gradient changes during scan acquisition that, from a statistical point of view, are simply outliers. If not accounted for in a statistical model, these outliers can have deleterious effects such as model misspecification, biased parameter estimates and poor forecasts. However, the detection of outliers and the removal of its effects in a multivariate process is a difficult task [96]. For example, an outlier in one component

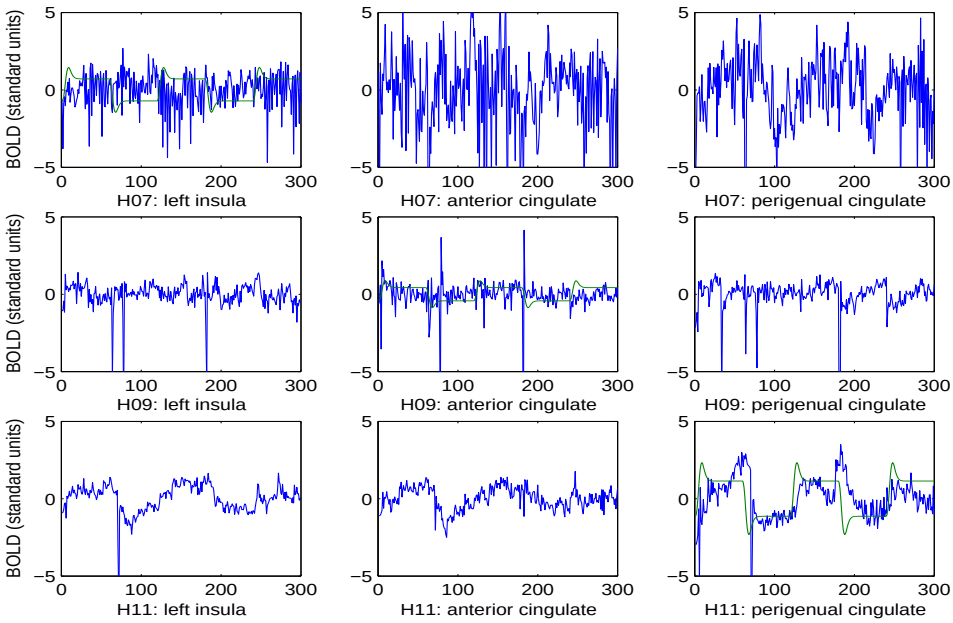

Figure 1: BOLD activity extracted from three regions of interest for subjects H07, H09 and H11.

maybe caused by an outlier in other components so that a moderate size outlier affecting all the others maybe unnoticed if univariate techniques are used.

Another characteristic that should be manifested in each of the time series courses is the intervening effect of the stimulus on the BOLD activity in each region being studied. In fact, it has been reported that greater fMRI BOLD response amplitudes correlate with the stress or stimulus induced increase in mean arterial pressure [33]. Although this effect might be noticeable in some subjects, for example H11 in figure 1 where the curve superimposed along the diagonal plots is the amplitude modulated hypothesized hemodynamic response, others apparently do not exhibit it at all. One reason might be due to the varying strategies the subject employs to complete the task so that activity might be subdued in these regions because it is being mediated by other regions elsewhere. Another reason is the diverging levels of noise to the MR signal that contributes to the variance of the data. However, if one looks at the spectral envelope [94], which is frequency based principal components technique that determines common cyclic component present among a set of time series, the stimulus

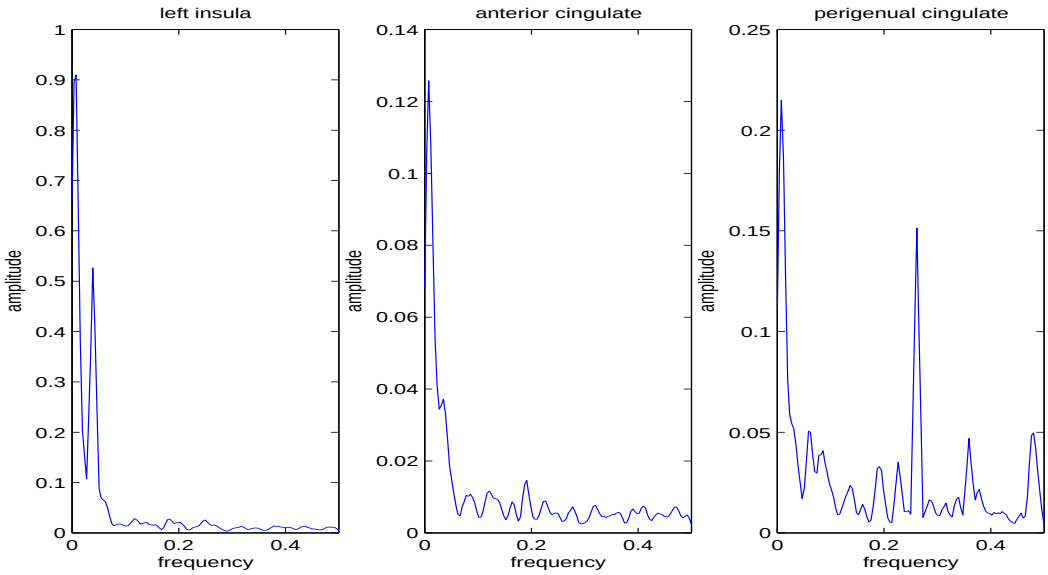

Figure 2: Spectral envelope of the BOLD activity time course at each region of interest.

is actually reflected as the predominant cyclic component in all locations across individuals (see figure 2).

The above observations influence the modelling and model estimation procedure that will be adopted. In particular, a random effect may be appropriate when including the stimulus as an explanatory variable to the variation in the time series BOLD activity. Moreover, the estimation procedure employed must take into account how to impair effects of aberrant observations to obtain accurate parameter estimates.

1.3 TOWARD A GENERAL STATISTICAL MODEL AND ESTIMATION TECHNIQUE

Ultimately, this research aims to develop a general statistical model which can be applied not only for modelling joint dynamic behavior of brain subsystems but also to problems in various disciplines that have a similar challenge. In addition, it is desired that this model allows one

to answer the question of whether significant changes in temporal dynamics can be attributed to differences in group membership, i.e., the model permits comparison of dynamic behavior between groups. These goals can be pursued along the lines of the VAR models which will be modified appropriately to accommodate the following modelling considerations:

1. maintain a manageable number of parameters for flexibility in model specification;
2. incorporate approximate spatial description or hypothesis driven weights among time series components to quantify known or unknown marginal spatial or experimental interaction among them;
3. include random effect for covariates that may have unit-specific effect to accommodate heterogeneity among units arising from covariate effects that vary from one unit to another;
4. estimate the parameters using procedure that is resistant to the effects of outlying observations and departures from prescribed distributions.

One way of reducing the number of parameters is to create a matrix constraint to the autoregressive parameters that carry some information, a method suggested by Stoffer for space-time ARMAX models [93]. In fact, this matrix can be chosen to accommodate spatial description of the component series, constructed through spatial ordering [11, 75] or approximated by geostatistical methods [70, 50, 29]. As a result, this strategy gives both a spatial and temporal structure to the model while reducing the parameters to be estimated.

In general, the model being considered is an extension of (1.1) and is of the form

$$\mathbf{z}_{it} = \boldsymbol{\alpha} + \sum_{h=1}^{n} \mathbf{D}_h \Pi_h^{\text{diag}} \mathbf{z}_{i,t-h} + \Gamma_i \mathbf{u}_{it} + \boldsymbol{\varepsilon}_t, \tag{1.2}$$

where Γ_i, which is randomly selected from a certain distribution, weights the covariate $\mathbf{u}_{it}$ and is associated specifically to unit i, and $\mathbf{D}$ is a matrix constraint. This model is essentially a replicated version of this constrained-VAR, adjusted by exogenous covariates whose sampled effect is specific to a particular replicate. Details of this model will be elaborated in the next chapter.

Replications in autoregressive models have been widely used in many panel and longitudinal data analysis when cross-sections or repeated measurements of a similar process are

observed and a variety of techniques have also been suggested to analyze them. For example, estimation using least squares technique, in simple autoregressions, can be seen in the work of Anderson [3] who also provided asymptotic frameworks for inference. Goodrich and Caines [36] used maximum likelihood (ML) estimation to obtain the parameters of an equivalent model, parameterized in state-space, which includes some input process. Although the least squares and ML estimates are consistent, asymptotically normal, and efficient under Gaussianity, they can be seriously biased under contamination. It is advantageous to use robust methods where bias can be bounded.

Robust estimation, particularly in vector time series, have not received much attention. Most of the treatises that can be found in literature concentrate on the univariate case such as M and GM-estimates (see for e.g. [31, 67, 69]) and residual autocovariance (RA) estimates [23]. To date, the only strategies aimed at robustly estimating VAR or VARMA (MA for moving average) is by bounding residual autocovariance (RA) components pairwise [61] or by weighting multivariate residuals by a function of their distance [9] while all other exogenous and location parameters that may be included in the model are estimated separately. This strategy, however, may not be desirable with the inclusion of a random effect in the exogenous variable. Some useful techniques are actually just based on robustifying the likelihood function itself by bounding the scaled residual by a redescending function in a manner similar to robustifying linear mixed models [47, 48, 78, 101]. In this way, the procedure jointly estimates the autoregressive, location, and exogenous parameters. However, usually the bounding function is applied componentwise to the stacked observations to bound the effect of an observation. In multivariate data, this may not necessarily produce nice, affine equivariant, estimates.

The proposed robust estimation technique, discussed in Chapter 2, is a multivariate extension of redescending M-estimates for univariate AR process (see [31] for a discussion) and is similar to the redescending M-estimates for location and scatter [51] but applied to dependent data. In this method, the redescending function is directly applied to the squared Mahalanobis distance of each row of the design matrix instead of scaling the residuals by the Cholesky factor of the compound covariance matrix. Simplistically, the average of these bounded terms is then minimized to obtain the fixed parameters of the model. Prediction

of random effects through empirical Bayes' method are also modified since these parameters are still susceptible to effects of outliers even if the fixed parameters have been robustly estimated.

The rest of the chapters are organized as follows: In Chapter 3, the asymptotic behavior of the fixed effects are investigated and the bias caused by the contamination will be assessed through the influence function. Assuming consistency of the estimator, we obtain an expression for this influence function and determine whether it is bounded pointwise. Then we can use this expression to obtain the limiting distribution of the estimates. Since the presence of contamination invalidates accurate inference through the asymptotic variance, Chapter 4 discusses the naive bootstrap approximation as an alternative tool for assessing significance of model parameters. Although this is computationally intensive, it has been shown that the true sampling distribution of the estimator, even for a simple AR(1) model, could be heavily skewed for moderate sample sizes so that a naive bootstrap approximation may be a better alternative. Identification of the final model order will also be obtained through a modified version of the Akaike Information Criterion (AIC) for robust autoregressions. Moreover, the asymptotic validity of the bootstrap estimates will be established. Chapter 5 shows some simulation results to test the accuracy of the estimation procedure in the presence of two common types of outliers: innovation and additive outliers. A positive result gives us assurance of the reliability of the procedure when applied to the neuroimaging data. Finally, Chapter 6 summarizes the results obtained from the previous chapters. Some extensions and possible research directions are also pointed out.

2.0 THE CONSTRAINED MIXED-VARX MODEL

In this chapter, we focus on the specification and estimation of the constrained version of the VARX model with mixed-effects. A special interest is geared towards modelling and estimating the parameters of in a manner that is resistant to the effects of atypical observations. In particular, we develop robust techniques for creating constraint matrix, estimating parameters, and predicting individual random effects by imposing weights on the observations that controls the effect of the observation on the value achieved by the estimate.

2.1 MODEL SPECIFICATION

Let the data on each individual or cross-section be denoted by a $p \times T$ matrix $\mathbf{y}_i = \{y_{ijk}\}$, where y_{ijk} is the observation on the kth response variable at the jth time or occasion for the ith cross-section; where $i = 1, \ldots, N$, $j = 1, \ldots, T$, and $k = 1, \ldots, p$. Influencing the response for each of the N cross-sections is a sequence of inputs x_j that varies across time and an $r \times 1$ vector $\mathbf{u}_i = \{u_{il}\}$, $l = 1, \ldots, r$, of non-time-varying exogenous covariates. The inputs x_j may or may not vary for each cross-section and the observation u_{il} is the measurement of lth non-time varying covariate for the ith cross-section. The association of these covariates with each cross-section may give insight about the strength of the relation between them, thereby providing useful information in predicting future values of the response given their values. For instance, in the MR study mentioned previously, a subject's performance rate in a task may have implications in neurologic strategies adopted to perform the task. In the same way, other clinical and demographic factors (eg. age, affect score etc.) may also have an effect in the response.

Assuming that the set of multivariate sequences of observations come from the same covariate adjusted autoregressive vector-valued process, we consider that the model of the current state on column j of $\mathbf{y}_i$, i.e., $\mathbf{y}_{ij} = [y_{ij1}, \ldots, y_{ijp}]^\top$, be described in terms of its previous states $\mathbf{y}_{i,j-1}, \ldots, \mathbf{y}_{i,j-n}$, the time-varying inputs $x_j, \ldots, x_{j-m}$ and the non-time-varying covariates $\mathbf{u}_i = [u_{i1}, \ldots, u_{ir}]^\top$. Furthermore, we also consider constraining the autoregressive parameters by some known weight matrix as well as introducing subject specific input effect so that the desired model has the following parametrization:

$$\mathbf{y}_{ij} = \boldsymbol{\alpha} + \sum_{h=1}^{n} \mathbf{D}_h \Pi_h^{\text{diag}} \mathbf{y}_{i,j-h} + \sum_{k=0}^{m} \Gamma_{ik} x_{j-k} + \Upsilon \mathbf{u}_i + \boldsymbol{\varepsilon}_{ij} \tag{2.1}$$

for $i = 1, \ldots, N$ and $j = n+1, \ldots, T$ and where $\boldsymbol{\alpha}$ is a $p \times 1$ vector which is related to the mean of $\mathbf{y}_{ij}$, Π_h^{diag} $(h = 1, \ldots, n)$ is the $p \times p$ diagonally constrained transition matrix expressing the dependence of the current response and the response at lag h, $\mathbf{D}_h$ is a $p \times p$ matrix of known constraint matrix that expresses the relationship between different dimensions at lag h, Γ_{ik} $(k = 0, \ldots, m)$ is a $p \times q$ matrix of random individual effects such that, marginally, $\text{vec}(\Gamma_{ik}) \sim \text{N}(\tilde{\Gamma}_{ik}, \Sigma_{\Gamma_{ik}})$, Υ is a $p \times r$ matrix of regression coefficients, and $\{\boldsymbol{\varepsilon}_{ij}\}$ is a sequence of random vectors such that $\boldsymbol{\varepsilon}_{ij}$ is independent of $\mathbf{y}_{i,j-1}, \ldots$ with $\text{E}(\boldsymbol{\varepsilon}_{ij}) = \mathbf{0}$ and $\text{E}(\boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{ij}^\top) = \Sigma_\varepsilon$. In addition, we assume that $\boldsymbol{\varepsilon}_{ij}$ is independent both of the input x_j and its effect on an individual $\tilde{\Gamma}_i$. For the moment we take the general case that the input covariate, which is either stochastic or non-stochastic, may have both momentary and building-up effect of activity in each location or component. If it is stochastic, it is assumed that the observed $\mathbf{x}_{ij}$, for all $i = 1, \ldots, N$, is generated by some general linear process, $\mathbf{x}_{ij} = \sum_{h_1=0}^{\infty} \mathbf{A}_{ih} \epsilon_{i,j-h}$, where $\sum_{h=0}^{\infty} ||\mathbf{A}_{ih}|| < \infty$, and that ϵ_{ij} and $\boldsymbol{\varepsilon}_{ij}$ are mutually independent for each i and every $j = 1, \ldots, T$. This assumption ensures that the input process is non-explosive. Furthermore, we assume that $\mathbf{y}_{ij}$ is stationary for each i so that the characteristic roots of

$$| - \lambda^n I + \sum_{j=1}^{n} \lambda^{n-j} \Pi_j | = 0 \tag{2.2}$$

are less than 1 in absolute value.

The model in (2.1) can be extended to permit inquiry whether the set of time series, despite belonging to different groups, is homogeneous. In other words, whether they have

the same vector autoregressive coefficients. In this case, groups specific parameters need to be included to accommodate group specific effect. Then (2.1) can be extended in the following manner:

$$\begin{aligned}\mathbf{y}_{ij} &= \boldsymbol{\alpha} + I\{i \in g_1\}\boldsymbol{\alpha}_1 + \sum_{h=1}^{n} \mathbf{D}_h[\Pi_h^{\text{diag}} + I\{i \in g_1\}\Pi_h^{\text{diag}}]\mathbf{y}_{i,j-h} \\ &\quad + \sum_{k=0}^{m} \Gamma_k x_{j-k} + \Upsilon \mathbf{u}_i + \boldsymbol{\varepsilon}_{ij} \end{aligned} \tag{2.3}$$

where $I\{i \in g_1\}$ is an indicator function which has a value of 1 if the argument is true and 0 if $i \in g_2$. This specification translates inference of group difference to testing whether the group specific term significantly deviates from zero.

The specification of the matrix of constraints can be done in ways that fit the experimenters belief. In fact, this can be exploited to include known variation or physical characteristics coupling the components of the vector time series. In spatial time series, for example, this constraint usually incorporates spatial information that describes the underlying phenomenon relating the different time series locations. The matrix $\mathbf{D}$ can be viewed as the expected weighting measure of inverse "distance" between neighboring locations or components in the target population. This specification assumes that, in every subject, neighboring locations that are close to each other exert the most influence. For instance, the distance can be the order or contiguity that describes locations which is common in regularly spaced systems [11, 75], or it can also be a function of the Euclidean distance between locations [25] in irregularly spaced systems. In particular, this "distance" measure can be the variation of the observed values between locations such as the sample variogram [70, 50, 29] given by

$$2V_l(\delta_{ij}) = \text{var}[y_{i,t+h} - y_{jt}] \tag{2.4}$$

and its sample estimator

$$2\hat{V}(\delta) = \frac{1}{N_l(\delta)} \sum_{k=1}^{N_l(\delta)} (y_{i,k+h} - y_{i+\delta,k})^2 \tag{2.5}$$

where $(y_{i,t+l}, y_{i+\delta,t})$ is a pair of observations that are l time units apart and δ distance apart, and $N(\delta)$ is the number of such pairs. Notice, however, that since distance can never be

negative, "distance" based constraint matrix always describe concordance among spatial locations. From experience, this results in predicted values having larger magnitude than the observed values themselves.

A natural alternative to the constraint matrix regardless of any available spatial description is the expected population cross-correlation matrix, $\rho(h) = \{\rho_{ij}\}$, whose diagonal elements are fixed to 1. If this is not known, it can be estimated using the sample estimator $\hat{\rho}_{ij}(h) = \hat{\gamma}_{ij}(h)/\{\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0)\}^{1/2}$ where $\hat{\gamma}_{ij}(h) = T^{-1}\sum_{k=1}^{T-h}(y_{ik} - \bar{y}_i)(y_{j,k+h} - \bar{y}_j)$ and $\bar{y}_\nu$ is the sample mean of the νth component series. In the stationary case, $\rho(h)$ is actually related to the variogram by the relation $V_l(\delta_{ij}) = \gamma_{ij}(0) - \gamma_{ij}(h)$, but the former can allow inverse co-variation between component series. Note that the problem with using the sample autocorrelation, inherent to all estimators based on unbounded functions of data, is susceptibility to the effects of atypical observations. From the expression of $\hat{\gamma}_{ij}(h)$ it is apparent that any outlier can inflate or deflate the average artificially. A work around would be to modify the sample autocorrelation so that the effect of large observations can be bounded. To this end, let ψ be an odd and bounded continuous function and define the weight function

$$w(x) = \frac{\psi(x)}{x} \tag{2.6}$$

so that w is a non-negative decreasing weight function defined on $[0, \infty)$ with $w(0) = 1$, $w(\infty) = 0$. Some commonly used ψ functions are Huber's monotone ψ-function $\psi_{H,k}(x) = \max\{\min(t, k) - k\}$ and Tukey's redescending bisquare function $\psi_{B,k}(x) = x(1-(x/k)^2)^2, |x| \leq k, \psi_{B,k}(x) = 0, |x| > k$. Then, replace the sample mean adjusted observations $\boldsymbol{\zeta}_{ij} = \mathbf{y}_{ij} - \bar{\mathbf{y}}_i$ by $\tilde{\boldsymbol{\zeta}}_{ij}$ defined as

$$\tilde{\boldsymbol{\zeta}}_{ij} = \boldsymbol{\zeta}_{ij} w\left([\boldsymbol{\zeta}_{ij}^\top \boldsymbol{\zeta}_{ij}]^{\frac{1}{2}}\right) \tag{2.7}$$

The idea behind (2.6) is to reduce the influence of high leverage observations in the sense that observations which are very far from the mean are down-weighted.

To see what is the effect of the constraint on the autoregressive model, we consider an expansion of a specific model by working out the matrix operations at a particular component. For a second order ($n = 2$) Mixed-VARX model that depends on a deterministic stimulus

covariate and a non-time varying covariate, an expansion of (2.3) at the lth component, $1 \leq l \leq p$, yields

$$\begin{aligned} y_{ijl} &= \alpha_l + \alpha_{l(1)} + \sum_{k_1=1}^{p} \{d_{1,lk_1}[\pi_{1k_1} + \pi_{1k_1(1)}]y_{i,j-1,k_1} + d_{2,lk_1}[\pi_{2k_1} + \pi_{2k_1(g)}]y_{i,j-2,k_1}\} \\ &\quad + \gamma_{il}x_j + \sum_{k_2=1}^{r} \tau_{lk_2}u_{ik_2} + \varepsilon_{ijl}. \end{aligned} \tag{2.8}$$

It is clear that the model is essentially a constrained regression of the present state at component l, y_{ijl}, on the past values $y_{i,j-1,1}, \ldots, y_{i,j-1,p}, y_{i,j-2,1}, \ldots, y_{i,j-2,p}$ weighted by d_{h,lk_1} from component l to k at lag h, $1 \leq h \leq n$, $1 \leq l, k \leq p$ and on the covariates x_j, u_{ik_2}, $1 \leq k_2 \leq r$. Consequently, it gives a similar interpretation as the classical covariate adjusted VAR model that the current value at a particular component can be explained jointly by the past values of itself, by the modulated values of the other components, and by the exogenous covariates.

Henceforth, we call model (2.1) and its extension (2.3) as the *constrained Mixed VARX models* where the "X" is for the eXogenous covariates present. The constraining approach offers a way of generalizing both autoregressive time series models and the simultaneous specified physical properties the observations purported to have. It also significantly reduces the number of parameters to be estimated from $O(p^2)$ to $O(p)$.

For ease of notation, we can rewrite (2.1) more compactly using the linear mixed model notation given by

$$\mathbf{y}_{ij} = \Pi^{\mathrm{D}}\mathbf{z}_{ij} + \Gamma_i\mathbf{w}_{ij} + \boldsymbol{\varepsilon}_{ij} \tag{2.9}$$

for $i = 1, \ldots, N$ and $j = n+1, \ldots, T$ and where $\mathbf{z}_{ij} = [1, \mathbf{y}_{i,j-1}^\top, \ldots, \mathbf{y}_{i,j-n}^\top, \mathbf{u}_i^\top]^\top$, $\mathbf{w}_j = [x_j, \ldots, x_{j-m}]^\top$, Π^{D} is the $p \times (pn + r)$ matrix of parameters $\boldsymbol{\alpha}, \mathbf{D}_1\Pi_1^{\mathrm{diag}}, \ldots, \mathbf{D}_n\Pi_n^{\mathrm{diag}}, \Upsilon$, and Γ_i is the $p \times m$ matrix of random effects $\Gamma_{i1}, \ldots, \Gamma_{im}$. Since $\Gamma_i = \tilde{\Gamma} + \Gamma_i^* = [\tilde{\Gamma}_1 + \Gamma_i^*, \ldots, \tilde{\Gamma}_n + \Gamma_{in}^*]$, augment Π^{D} into $\Pi^{\mathrm{D}*} = [\boldsymbol{\alpha}, \mathbf{D}_1\Pi_1^{\mathrm{diag}}, \ldots, \mathbf{D}_n\Pi_n^{\mathrm{diag}}, \tilde{\Gamma}, \Upsilon]$ and $\mathbf{z}_{ij}^* = [1, \mathbf{y}_{i,j-1}^\top, \ldots, \mathbf{y}_{i,j-n}^\top, x_j, \ldots, x_{j-m}^\top, \mathbf{u}_i^\top]^\top$ to accommodate the mean $\tilde{\Gamma}$ so that we have

$$\mathbf{y}_{ij} = \Pi^{\mathrm{D}*}\mathbf{z}_{ij}^* + \Gamma_i^*\mathbf{w}_j + \boldsymbol{\varepsilon}_{ij} \tag{2.10}$$

From hereafter, we will just concentrate our attention on (2.1) since (2.3) can always be written in the form similar to (2.1) by augmenting $\mathbf{z}_{ij}$ appropriately when $\mathbf{y}_{ij} \in g_1$. Then

using the fact that $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\text{vec}(\mathbf{B})$ for any conformable matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, (2.10) can be written into the "workable" linear form

$$\mathbf{y}_{ij} = (\mathbf{z}_{ij}^{*\top} \otimes I_p)\text{vec}(\Pi^{\text{D}*}) + (\mathbf{w}_j^\top \otimes I_p)\text{vec}(\Gamma_i^*) + \boldsymbol{\varepsilon}_{ij} \tag{2.11}$$

where the vec notation concatenates the columns of $\Pi^{\text{D}*}$ and $\otimes$ is the Kronecker product. From this, we translate some of the assumptions made earlier in this chapter into the multivariate regression setting, particularly

1. $\text{vec}(\Gamma_i^*)$ are independent $pm \times 1$ vector with zero mean and $\text{E}[\text{vec}(\Gamma_i^*)\text{vec}^\top(\Gamma_i^*)] = \Sigma_\Gamma$;
2. $\boldsymbol{\varepsilon}_{ij}$ are $p \times 1$ random variables which is independent both across time and cross-section with mean zero and $\text{E}[\boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{ij}^\top] = \Sigma_\varepsilon$;
3. $\boldsymbol{\varepsilon}_{ij}$ and $\mathbf{w}_j$ are independent;
4. $\boldsymbol{\varepsilon}_{ij}$ and $\text{vec}(\Gamma_i^*)$ are uncorrelated, i.e., $\text{E}[\boldsymbol{\varepsilon}_{ij}\text{vec}^\top(\Gamma_i^*)] = 0$ for all $i = 1, \ldots, N$ and $j = n+1, \ldots, T$.

Then, our goal is to estimate Π, Σ_ε, and Σ_Γ, predict the individual random effects Γ_i^* for each i and make inferences about the significance of these parameters.

2.2 REDESCENDING M-ESTIMATION

In the sequel, let $\mathbf{y}_{ij}^k = (\mathbf{y}_{ij}, \ldots, \mathbf{y}_{ik})$, $k \leq j$ and $i = 1, \ldots, N$ denote the finite sets of contiguous $\mathbf{y}_{ij}$, where $\mathbf{y}_{ij}$ given in (2.1) or (2.3). Since the observations are made for $j = 1, \ldots, T$, the model maybe specified by the marginal distribution of $\mathbf{y}_{in}^1$ and the innovations $\mathbf{r}_{i,n+1}, \ldots, \mathbf{r}_{iT}$ where $\mathbf{r}_{ij} = \mathbf{y}_{ij} - (\mathbf{z}_{ij}^{*\top} \otimes I_p)\text{vec}(\Pi^{\text{D}*})$. For simplicity, we assume that the elements in the set $\mathbf{y}_{in}^1$ be independent and identically distributed $N(\mathbf{0}, \mathbf{G})$. From (2.11) it is apparent that $\text{E}[\mathbf{y}_{ij}|\mathbf{z}_{ij}^*] = (\mathbf{z}_{ij}^{*\top} \otimes I_p)\text{vec}(\Pi^{\text{D}*})$ so that $\text{E}[\mathbf{r}_{ij}|\mathbf{z}_{ij}^*] = \mathbf{0}$ and the covariance matrix Ω_j of $\mathbf{r}_{ij}$ is given by

$$\Omega_j = (\mathbf{w}_j^\top \otimes I_p)\Sigma_\Gamma(\mathbf{w}_j \otimes I_p) + \Sigma_\varepsilon. \tag{2.12}$$

Let Θ be the set of all nonzero elements of $\text{vec}(\Pi^*, \Sigma_\varepsilon, \Sigma_\Gamma)$. Then denote by $f^{NT}(\mathbf{y}_{1T}^1, \ldots, \mathbf{y}_{NT}^1; \Theta, \mathbf{G})$ the joint likelihood of $\mathbf{y}_{1T}^1, \ldots, \mathbf{y}_{NT}^1$ and let the innovations be normally

distributed. By the Markovian property of an autoregressive process, we obtain for all $j > n$ that

$$\begin{aligned} f^{NT}(\mathbf{y}_{1T}^1, \ldots, \mathbf{y}_{NT}^1; \Theta, \mathbf{G}) &= f^{Nn}(\mathbf{y}_{1n}^1, \ldots, \mathbf{y}_{Nn}^1; \Theta, \mathbf{G}) \\ &\quad \times \prod_{i=1}^{N} \prod_{j=n+1}^{T} c \exp\{-\frac{1}{2}(\mathbf{r}_{ij}^\top \Omega_j^{-1} \mathbf{r}_{ij})\} |\Omega_j|^{-\frac{1}{2}}. \end{aligned} \tag{2.13}$$

From this, we have the following:

$$\lambda^{NT}(\mathbf{y}_{1T}^1, \ldots, \mathbf{y}_{NT}^1; \Theta, \mathbf{G}) = \frac{\partial}{\partial \Theta} \log f^{NT}(\mathbf{y}_{1T}^1, \ldots, \mathbf{y}_{NT}^1; \Theta, \mathbf{G}) \tag{2.14}$$

$$\kappa(\mathbf{y}_{1,j+n}^j, \ldots, \mathbf{y}_{N,j+n}^j; \Theta) = \frac{\partial}{\partial \Theta} c \exp\{-\frac{1}{2}(\mathbf{r}_{ij}^\top \Omega_j^{-1} \mathbf{r}_{ij})\} |\Omega_j|^{-\frac{1}{2}} \tag{2.15}$$

so that from (2.14) we have

$$\lambda^{NT}(\mathbf{y}_{1T}^1, \ldots, \mathbf{y}_{NT}^1; \Theta, \mathbf{G}) = \lambda^{Nn}(\mathbf{y}_{1n}^1, \ldots, \mathbf{y}_{Nn}^1; \Theta, \mathbf{G}) + \sum_{i=1}^{N} \sum_{j=1}^{T-n} \kappa(\mathbf{y}_{1,j+n}^j, \ldots, \mathbf{y}_{N,j+n}^j; \Theta). \tag{2.16}$$

Condional on $\mathbf{y}_{1n}^1, \ldots, \mathbf{y}_{Nn}^1$, the maximum likelihood estimator (MLE) for Θ based on $\mathbf{y}_{1,T}^{n+1}, \ldots, \mathbf{y}_{N,T}^{n+1}$ is the solution to the following equations:

$$\sum_{i=1}^{N} \sum_{j=1}^{T-n} \kappa(\mathbf{y}_{1,j+n}^j, \ldots, \mathbf{y}_{N,j+n}^j; \hat{\Theta}) = 0. \tag{2.17}$$

It is quite well known that the maximum likelihood estimates are sensitive to the effects of outliers and other atypical observations. This can be directly inferred by taking the logarithm of (2.13) and by noting that it has a component which is the sum of squared Mahalanobis distances. If there are some outliers, some of the terms in the summation will be large and may have considerable influence on the likelihood. Moreover, in practice, we can never be sure that the $\boldsymbol{\varepsilon}_{ij}$ are multivariate normal. So we are usually in the position of using a method which assumes normality when normality does not necessarily hold. In this circumstance, the estimand is no longer the true value of the parameters but the true value plus an unknown bias that for classical or non-robust methods maybe infinite. In fact, non-robust methods can be extremely inefficient when the model distribution does not hold [92]. It is preferable to use a method for which the potential bias is always bounded, that

is both reasonably efficient when normality holds and more efficient than methods derived under normality when normality does not hold [101].

To overcome this problem we bound the squared distances by a function that grows more slowly in the sense that it has a bounded derivative. In this way, the influence of outlying observations can be muted; which is the basic idea behind robust estimation. In the following, we define the redescending M-estimators for the constrained Mixed-VARX model:

Definition 1. Let $\mathbf{Y}_{NT} = \{(\mathbf{y}^*_{11}, \mathbf{u}_1, x_1), \ldots, (\mathbf{y}^*_{NT}, \mathbf{u}_{NT}, x_T)\}$ be a data set in $\mathbb{R}^{p+r+1}$ coming from N subjects at T time points, and let $\mathcal{S}_{NT}$ denote the set of symmetric positive definite pairs of matrices which can be written jointly as $\Sigma = \mathrm{diag}\{\Sigma_\varepsilon, \Sigma_\Gamma\}$. The *redescending M-estimators* for the constrained Mixed-VARX model is the pair $(\hat{\Pi}(\mathbf{Y}_{NT}), \hat{\Sigma}(\mathbf{Y}_{NT}))$ which minimizes

$$\sum_{i=1}^{N}\sum_{j=1}^{T-n} \xi \log|\Omega_j| + \sum_{i=1}^{N}\sum_{j=1}^{T-n} \rho(\mathbf{r}_{ij}^\top \Omega_j^{-1} \mathbf{r}_{ij}) \tag{2.18}$$

among all $(\Pi(\mathbf{Y}), \Sigma(\mathbf{Y})) \in \mathbf{\Theta} = \mathbb{R}^s \times \mathcal{S}_{NT}$.

The constant ξ is an adjustment so that the $\hat{\Omega}_j$ will be a consistent estimator for Ω_j. Its value is usually chosen to be $\mathrm{E}(d\psi(d))$, where $\psi(d) = \rho'(d)$ and the expectation is carried out from the distribution of d. Along with this, we also have the following assumptions for the ρ-function:

(**A1**) ρ is symmetric, twice continuously differentiable and $\rho(0) = 0$.

(**A2**) ρ is strictly increasing on $[0, k]$ and constant on $[k, \infty)$ for some $k < \infty$.

Note that from the definition, the parameters Σ_ε and Σ_Γ are essentially constrained to be symmetric positive definite. We can optimize the objective function (2.18) instead as a function of the nonzero entries of L_ε and L_Γ, which are the Cholesky factors of Σ_ε and Σ_Γ, resp. (i.e., $\Sigma_\varepsilon = L_\varepsilon L_\varepsilon^\top$ and $\Sigma_\Gamma = L_\Gamma L_\Gamma^\top$). This transforms the constrained problem to an unconstrained one and ensures positive definiteness of Σ_ε and Σ_Γ.

For ease of notation, let $\mathbf{S} = [s_1, \ldots, s_r]$ be the set of all non-zero entries of L_ε, and L_Γ. Since ρ is differentiable by **A1**, the derivatives with respect to $\mathrm{vec}(\Pi^{\mathrm{D}*})$ and $\mathbf{S}$ give the

following estimating equations:

$$\sum_{i=1}^{N}\sum_{j=n+1}^{T} u(d_{ij})\mathbf{Q}^{*}\Omega_j^{-1}\mathbf{r}_{ij} = \mathbf{0} \tag{2.19}$$

where $u(d_{ij}) = 2\psi(d_{ij})$, $d_{ij} = \mathbf{r}_{ij}^{\top}\Omega_j^{-1}\mathbf{r}_{ij}$, and $\mathbf{Q}^{*}$ is obtained by deleting some entries in

$$\mathbf{Q} = \begin{bmatrix} 1 & \otimes & I_p \\ \mathbf{y}_{i,j-1} & \otimes & \mathbf{D}_1 \\ & \vdots & \\ \mathbf{y}_{i,j-n} & \otimes & \mathbf{D}_n \\ x_{j-1} & \otimes & I_p \\ & \vdots & \\ x_{j-n} & \otimes & I_p \\ \mathbf{u}_i & \otimes & I_p \end{bmatrix}$$

corresponding to the zero entries in Π^{*} and

$$\sum_{i=1}^{N}\sum_{j=n+1}^{T} u(d_{ij})\mathbf{r}_{ij}^{\top}\Omega_j^{-1}\dot{\Omega}_{jk}\Omega_j^{-1}\mathbf{r}_{ij} - \sum_{i=1}^{N}\sum_{j=n+1}^{T} \xi\mathrm{tr}(\Omega_j^{-1}\dot{\Omega}_{jk}) = \mathbf{0} \tag{2.20}$$

for all s_k and where

$$\dot{\Omega}_{jk} = \frac{\partial\Omega_j}{\partial s_k} = \begin{cases} L_\varepsilon J_k^{\top} + J_k L_\varepsilon^{\top} & s_k \in \{L_\varepsilon\} \\ (\mathbf{w}_j^{\top}\otimes I_p)(L_\Gamma J_k^{\top} + J_k L_\Gamma^{\top})(\mathbf{w}_j \otimes I_p) & s_k \in \{L_\Gamma\} \end{cases} \tag{2.21}$$

J_k denotes a single non-zero entry matrix with 1 at the position of the kth non-zero component in either L_ε or L_Γ.

Using the fact that $\partial\Sigma/\partial\mathrm{vech}(L) = (\partial\Sigma/\partial L)(\partial L/\mathrm{vech}(L))$, where vech concatenates the lower triangular half of a symmetric matrix, (2.20) can be written in the following closed form

$$\sum_{j=n+1}^{T} \mathrm{vec}^{\top}\left(\sum_{i=1}^{N} u(d_{ij})\mathbf{r}_{ij}\mathbf{r}_{ij}^{\top} - \xi N\Omega_j\right)(\Omega_j^{-1}\otimes\Omega_j^{-1})\mathbf{L}_j\mathbf{U} = \mathbf{0} \tag{2.22}$$

where $\mathbf{L}_j = \mathrm{diag}\{\mathbf{L}_\Gamma^{\top}, \mathbf{L}_\varepsilon^{\top}\} = \mathbf{0}$ with $\mathbf{L}_\Gamma = [(\mathbf{w}_j\otimes I_p)\otimes(\mathbf{w}_j\otimes I_p)](I_{p^2} + T_{p,p})(I_p\otimes L_\Gamma^{\top})$ and $\mathbf{F}_\varepsilon = (I_{p^2} + T_{p,p})(I_p\otimes L_\varepsilon^{\top})$, where $T_{p,p}$ is a $p^2\times p^2$ permutation matrix and $\mathbf{U} = \mathrm{diag}\{S_{pm}^{\top}, S_p^{\top}\}$

is a $(p(p+1)+pm(pm+1))/2 \times p^2$ of 0s and 1s so that the relation $\text{vech}A = S_n \text{vec}A$ is true. For unstructured covariance matrices, the above equation is generalized by

$$\sum_{j=n+1}^{T} \text{vec}^\top \left(\sum_{i=1}^{N} u(d_{ij}) \mathbf{r}_{ij} \mathbf{r}_{ij}^\top - \xi N \Omega_j \right) (\Omega_j^{-1} \otimes \Omega_j^{-1}) \mathbf{F}_j = \mathbf{0} \tag{2.23}$$

where $\mathbf{F}_j = \text{diag}\{\mathbf{F}_\Gamma^\top, \mathbf{F}_\varepsilon^\top\}$ with $\mathbf{F}_\Gamma = (\mathbf{w}_j \otimes I_p) \otimes (\mathbf{w}_j \otimes I_p)$ and $\mathbf{F}_\varepsilon = I_{p^2}$.

As mentioned previously, the robustness property actually depend on the derivatives of the ρ-functions denoted by $\psi(x)$. In this research, we are interested in redescending M-estimates, i.e., estimates for which $d^{1/2}u(d)$ is increasing for d near zero and decreasing for d near ∞. This eliminates the need for imposing another set of weights on (2.18), like in GM-estimates [67], since the effect of large $\|z\|$ and, consequently large d, can always be muted by a redescending ψ. This condition, however, does not imply boundedness of (2.20) which only happens when $du(d)$ is also bounded. The latter condition, of course, implies that $d^{1/2}u(d)$ must redescend. The drawback with the use of such kinds of functions is the possibility that the estimating equations may admit multiple solutions for Π^* especially when Σ_ε and Σ_Γ are fixed. Uniqueness is only guaranteed if one imposes a strict condition that $du(d)$ is monotonic [51] which could result in a zero breakdown point [66], i.e., the fraction of outliers which could result in an infinite bias in the estimator. We do not, however, make any assertions here both on positive breakdown points and uniqueness of the estimators. It is a conservative belief that the observed cross-sectional structure in the data, such as the one we will consider, makes it easier for breakdown points to occur so that the procedure on consideration probably has very low breakdown point. In the case of uniqueness, the experimenter can always determine a method on the desirability of a particular estimator from among a set of solutions.

2.3 PREDICTING RANDOM EFFECTS

Using the assumption that the realized random effects which determine the data are just random selections from a conceptual population, it can be shown that the best predictor of Γ_i, in the sense of mean squared error, is the conditional mean $\text{E}[\Gamma_i^* | \mathbf{y}_{iT}^1]$, i.e., the expected

value of the random effect in the light of the observed data. Due to its form, the estimation can be derived by straightforward application of Bayes' theorem [84, 59]. In our model, we have the following distributions:

$$\mathbf{y}_{iT}^{1}|\Pi^{\mathrm{D}*}, \Sigma_{\varepsilon}, \Sigma_{\Gamma}, \Gamma_i^* \sim \mathrm{N}(T^{-1} \sum_{j=n+1}^{T} [(\mathbf{z}_{ij}^{\top} \otimes I_p)\mathrm{vec}(\Pi^{\mathrm{D}*}) + (\mathbf{w}_j^{\top} \otimes I_p)\mathrm{vec}(\Gamma_i^*)], T^{-1}\Sigma_{\varepsilon}) \quad (2.24)$$

and

$$\mathrm{vec}(\Gamma_i^*)|\Sigma_{\Gamma} \quad \sim \quad \mathrm{N}(\mathbf{0}, \Sigma_{\Gamma}). \quad (2.25)$$

Then we obtain

$$\mathrm{vec}(\Gamma_i^*)|\mathbf{y}_{iT}^{1}, \Pi^{\mathrm{D}*}, \Sigma_{\varepsilon}, \Sigma_{\Gamma} \quad \sim \quad \mathrm{N}(\mathrm{vec}(\tilde{\Gamma}_i), \mathbf{U}_i^{-1}) \quad (2.26)$$

where

$$\mathrm{vec}(\tilde{\Gamma}_i) = T^{-1}\mathbf{U}_i \sum_{j=n+1}^{T} (\mathbf{w}_j^{\top} \otimes I_p)\Sigma_{\varepsilon}^{-1} \sum_{j=n+1}^{T} \mathbf{r}_{ij} \quad (2.27)$$

$$= T^{-1}\mathbf{U}_i \sum_{j=n+1}^{T} \left(\sum_{j=n+1}^{T} \mathbf{w}_j \otimes \Sigma_{\varepsilon}^{-1} \right) \mathbf{r}_{ij} \quad (2.28)$$

$$\mathbf{U}_i = \left[\Sigma_{\Gamma}^{-1} + T^{-1} \sum_{j=n+1}^{T} (\mathbf{w}_j^{\top} \otimes I_p)\Sigma_{\varepsilon}^{-1} \sum_{j=n+1}^{T} (\mathbf{w}_j \otimes I_p) \right]^{-1}. \quad (2.29)$$

From the expression in (2.27) it is clear that the predictor for the random effect is susceptible to outlying observations observed through the large residuals. Therefore, it is not sufficient to just use robust estimates Σ_{Γ}, Σ_{ε}, and $\Pi^{\mathrm{D}*}$. To get robust empirical Bayes predictors, we place weights to $\mathbf{r}_{ij}$ given by

$$w(\mathbf{r}_{ij}) = \frac{\psi\left(\sqrt{\mathbf{r}_{ij}^{\top}\hat{\Omega}_{ij}\mathbf{r}_{ij}^{\top}}\right)}{\sqrt{\mathbf{r}_{ij}^{\top}\hat{\Omega}_{j}\mathbf{r}_{ij}^{\top}}} \quad (2.30)$$

so that we have the robustified estimate for the random effect:

$$\mathrm{vec}(\tilde{\Gamma}_i) = T^{-2}\mathbf{U}_i \sum_{j=n+1}^{T} (\mathbf{w}_j^{\top} \otimes I_p)\Sigma_{\varepsilon}^{-1} \sum_{j=n+1}^{T} w(\mathbf{r}_{ij})\mathbf{r}_{ij}. \quad (2.31)$$

Then the estimate is obtained by substituting the values of $\hat{\Pi}^{\mathrm{D}*}, \hat{\Sigma}_{\varepsilon}$ and $\hat{\Sigma}_{\Gamma}$ into $\Pi^{\mathrm{D}*}, \Sigma_{\varepsilon}$ and Σ_{Γ}, respectively.

3.0 ASYMPTOTIC BEHAVIOR OF THE M-ESTIMATES

In order to establish asymptotic behavior of the redescending M-estimates, we focus our attention to a model, general enough, to encompass the model specified in Chapter 2. In particular, we partition the covariates into stochastic and non-stochastic or deterministic component with finite achievable levels and whose effect is random for each cross-section. This makes it easier to specify a measure associated with the process for each level of the deterministic input. The stochastic covariates may vary within subjects in incremental manner across time or constant within but varies across cross-section.

Using the joint empirical marginal distribution of the data, we will find an expression for the influence function of an observation to the estimator of the parameters of the model. This influence function gives a heuristic way of assessing robustness of the estimates against departures from the core distributions and is also useful in obtaining the limiting distribution of the estimates.

In the sequel, keep in mind that $\mathbf{y}_{ij}^{j-n+1} = (\mathbf{y}_{ij}, \ldots, \mathbf{y}_{i,j-n+1})$ as previously used in the last section. The continuation observation $(\mathbf{y}_{ij}^\top, \ldots, \mathbf{y}_{i,j-n}^\top, \mathbf{u}_{ij}^\top) = (\mathbf{y}_{ij}^{j-n+1}, \mathbf{u}_{ij})$ is usually denoted by $\mathbf{z}$ which is in $\mathbb{R}^q$ but if the former is augmented by x then $\mathbf{z}$ resides in $\mathbb{R}^{q+1}$. Denote the unknown fixed and variance parameters by $\mathbf{t}$ which can take on values (i) $\mathbf{T}$ when the underlying distribution is μ^n, or equivalently $\mathbf{T}(\mu^n)$ or $\mathbf{T}(0, \mu^n)$ (ii) $\mathbf{T}_{NT}$ when the underlying distribution is the empirical distribution μ_{NT}^n, or equivalently $\mathbf{T}(\mu_{NT}^n)$ and (iii) $\mathbf{T}(\mu^{n,\epsilon}) = \mathbf{T}(\epsilon, \mu^n)$ when the distribution is $\mu^{n,\epsilon}$. The unknown parameter $\mathbf{t}$ is usually split into $(\boldsymbol{\pi}, \mathbf{S})$ and is considered on the space $\boldsymbol{\Theta} = \mathbb{R}^s \times \mathcal{S}_{NT}$, where $\mathcal{S}$ denotes the set of symmetric positive definite matrices, and $\mathbb{R}^s \times \mathcal{S}_{NT}$ is an open subset of $\mathbb{R}^{s+\frac{1}{2}p(p+1)}$.

The vector function $\boldsymbol{\Psi}_{ij}$ has the same form as $\boldsymbol{\Psi}$ but the subscript in the former is a device to designate the dependence of the function on a particular data. Asymptotically,

$\mathbf{\Psi}_{ij}$ becomes $\mathbf{\Psi}$ and we also drop the indices in the data $(\mathbf{y}_{ij}^{j-n+1}, \mathbf{u}_{ij}, x_j)$ and use $(\mathbf{y}, \mathbf{u}, x)$ instead.

The covariance Ω_j generally depends on x_j so that when $x_j = 1$ it assumes a compound form $(\mathbf{w}_j^\top \otimes I_p)\Sigma_\Gamma(\mathbf{w}_j \otimes I_p) + \Sigma_\varepsilon$. Otherwise it is given by Σ_ε. For brevity, the subscript j in Ω_j is sometimes dropped but this does not mean that it loses its dependence on x_j.

Oftentimes, the symbol $\|\cdot\|$ denotes the Euclidean norm unless, otherwise, indicated, while λ_1 is the smallest eigenvalue of Σ.

3.1 M-FUNCTIONAL OF THE CONSTRAINED MIXED-VARX MODEL

Assume that the sequence of observations $(\mathbf{y}_{ij}, \mathbf{u}_{ij}, x_j)$, $i = 1, \ldots, N$ and $j = 1, \ldots, T$ follow model (2.10) for some arbitrary n. The exogenous input $\mathbf{u}_{ij}$ is a random variable either within and or between cross-sections and that the inputs x_j follow a deterministic box car pattern, i.e.

$$x_j = \begin{cases} 1 & \text{for} \quad j = 1, \ldots, q, 2q+1, \ldots, 3q \\ 0 & \text{for} \quad j = q+1, \ldots, 2q, 3q+1, \ldots, 4q \end{cases} \tag{3.1}$$

From the previous chapter, we assumed that for the duration of any particular input level x, the sequence $(\mathbf{y}_{ij}, \mathbf{u}_{ij})$ is a stationary and ergodic process which is defined on $\mathbb{R}^p(-\infty, \infty) \times \mathbb{R}^r(-\infty, \infty)$, with associated probability space $(\mathbb{R}^{p+r}(-\infty, \infty), \mathscr{B}, \mu)$, $\mathscr{B}$ being the increasing family of Borel sets in $\mathbb{R}^{p+r}(-\infty, \infty)$ and μ in the set $\mathscr{M}$ of all stationary and ergodic measures on $(\mathbb{R}^{p+r}(-\infty, \infty), \mathscr{B})$. The estimators for the parameters of the model are, more generally, obtained through the joint empirical q-dimensional $(q = np + r)$ marginal distribution function μ_{NT}^n defined by

$$\mu_{NT}^n = (NT)^{-1} \sum_{i=1}^{N} \sum_{j=n+1}^{T} \delta(\mathbf{y}_{ij}^\top, \ldots, \mathbf{y}_{i,j-n}^\top, \mathbf{u}_{ij}^\top) \tag{3.2}$$

where $\mathbf{y}_{ij} = \mathbf{y}_{i,j-T}$ for $j > T$ and $\delta_{\mathbf{z}}$ is a pointmass at $\mathbf{z} \in \mathbb{R}^q$. μ_{NT}^n can be viewed as a random measure in $\mathscr{M}^n = \{q\text{-dimensional marginals of a } p\text{-variate stationary process}\}$ and we can then consider our estimator as a functional $\mathbf{T}(\mu^n)$ which can also be defined in a natural way for other measures in $\mathscr{M}^n$. In particular, when $\mathbf{T}$ is defined over μ_{NT}^n we obtain

the model parameters estimates $\mathbf{T}_{NT}$ that are solutions to the estimating equation defined in (2.19-2.20) which can generally be written in the form similar to (2.17), i.e.,

$$\sum_{i=1}^{N}\sum_{j=n+1}^{T}\mathbf{\Psi}_{ij}(\mathbf{y}_{ij}^{j-n+1},\mathbf{u}_{ij},x_j;\mathbf{T}_{NT})=\mathbf{0}. \tag{3.3}$$

where $\mathbf{T}_{NT}=(\hat{\Psi}(\mathbf{Y}_{NT}),\hat{\Sigma}(\mathbf{Y}_{NT}))$ and $\mathbf{\Psi}=(\mathbf{\Psi}_{ij}^{(1)},\mathbf{\Psi}_{ij}^{(2)})$

$$\mathbf{\Psi}_{ij}^{(1)} := u(d_{ij})\mathbf{Q}^{*}\Omega_j^{-1}\mathbf{r}_{ij} \tag{3.4}$$

$$\mathbf{\Psi}_{ij}^{(2)} := u(d_{ij})\mathbf{r}_{ij}^{\top}\Omega_j^{-1}\dot{\Omega}_{jk}\Omega_j^{-1}\mathbf{r}_{ij}-\xi\mathrm{tr}(\Omega_j^{-1}\dot{\Omega}_{jk}) \tag{3.5}$$

The subscript ij in $\mathbf{\Psi}_{ij}$ is a devise used by Martin [68] to denote "edge effects" from the cross-section and continuation of observations within cross-section, which vanishes after a finite number of observations so that a fixed value of $\mathbf{\Psi}$ can be used, eventually.

When $(\mathbf{y}_{ij},\mathbf{u}_{ij})$ comes from a stationary and ergodic process, μ_{NT}^{n} converges by the ergodic theorem weakly to the q-dimensional marginal μ^n almost surely as T goes to infinity. So if $\mathbf{T}$ is continuous in the weak topology then one expects to have, under some regularity conditions, that

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=n+1}^{T}\mathbf{\Psi}_{ij}(\mathbf{y}_{ij}^{j-n+1},\mathbf{u}_{ij},x_j;\mathbf{t}) &= \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\mathbf{\Psi}_i(\mathbf{y}_{i,n+1}^{2n},\mathbf{u}_{i1},x;\mathbf{t})\\ &= \mathrm{E}\mathbf{\Psi}(\mathbf{y},\mathbf{u},x;\mathbf{t}).\end{aligned} \tag{3.6}$$

Therefore we assume that the asymptotic value $\mathbf{T}(\mu^n)$ is given by

$$\int\mathbf{\Psi}(\mathbf{y},\mathbf{u},x,\mathbf{T})d\mu^n=\mathbf{0}. \tag{3.7}$$

We assume that (3.7) has a unique root $\mathbf{t}_0=\mathbf{T}(\mu^n)$, or that a well-defined solution is available in the case of multiple roots. $\mathbf{T}$ is then defined on $\mathscr{M}_0^n$ consisting of all measures μ^n in $\mathscr{M}^n$ for which the integral in (3.7) exists and is finite. We also specify that the structure of $\mathbf{\Psi}$ has the representation $\mathbf{\Psi}(\mathbf{y},\mathbf{u},x;\mathbf{T})=\mathbf{\Psi}_1(\mathbf{y},\mathbf{u},1;\mathbf{T})+\mathbf{\Psi}_0(\mathbf{y},\mathbf{u},0;\mathbf{T})$ and that μ is a product of two measures μ_1 and μ_0 so that $\mathbf{\Psi}_1$ is associated with μ_1 and $\mathbf{\Psi}_0$ is associated with μ_0.

3.2 CONSISTENCY

Important to the problem of determining whether the estimator is consistent is the problem of whether a solution exists for any sample and whether the obtained solution at that sample is unique. Since the redescending M-estimates are solutions to a set of implicit equations we may encounter situations when the solution does not exist for some data, and even if it does exist the solution may not be the only one. It is therefore of interest to know the conditions warranting the existence and uniqueness of solutions.

Given that the parametric equation $\mathbf{G}_{\mu^n}(\mathbf{t}) := \int \mathbf{\Psi}(\mathbf{y}\ \mathbf{u}, x, \mathbf{t}) d\mu^n = \mathbf{0}$ and the empirical equation $\mathbf{G}_{NT}(\mathbf{t}) := (NT)^{-1} \sum_{i=1}^{N} \sum_{j=n+1}^{T} \mathbf{\Psi}_{ij}(\mathbf{y}_{i,j}^{j+n}, \mathbf{u}, x; \mathbf{t})$. By regression equivariance, it is possible to translate the fixed parameters to $\mathbf{0}$ so that the problem boils down to the usual location and scatter problem. Therefore the conditions given by Tyler and Kent [51] and Lopuhaa [63] can always be translated to the linear model case to shed light on what conditions are needed for the existence of solutions. In particular, they agreed that data should not be contained in some arbitrarily thin strips in the space $\mathbb{R}^q$ which is equivalent to the condition that $\lambda_1(\Sigma) \to \infty$, i.e., the distribution does not become degenerate. However, even if this condition is satisfied, it is not guaranteed of a unique solution. One way to overcome this problem is to look at the solution to the estimating equation as that associated to the minimization problem in (2.18). The situation can be seen as the minimization of the negative log likelihood of some elliptical distribution $\exp\{-\rho(\mathbf{r}^\top \Omega \mathbf{r})\}$, with $\mathbf{r} := \mathbf{y} - (\mathbf{z}^{*\top} \otimes I_p)\mathrm{vec}(\Pi^{\mathrm{D}*})$, so that when $dw(d)$, $d = \mathbf{r}^\top \Omega \mathbf{r}$, is increasing for $d > 0$ then the distribution is strongly unimodal and that the minimization is done over a strictly convex parameter space and both the parametric and empirical equation is assured a single solution. For a redescending function ψ, convexity of the parameter space no longer holds and thus do not guarantee uniqueness of solution but generally the most desirable solution can be obtained if we think of the solution not as a zero of the parametric equation but as the minimum of the objective function.

If the solution is not unique, there is a difficulty of identifying a consistent solution sequence $\mathbf{T}_{NT}$ for $\mathbf{T}$, and in practice, one needs to go further than consistency results by Huber [45] and Boos [18] to establish consistency for a particular solution sequence obtained

by a specified algorithm. Boos and Serfling [14] suggested that if ψ is bounded, then the functional $\mathbf{T}(\mu^n)$ may be defined so that it is continuous at μ^n with respect to $\|\cdot\|_\infty$ and thus satisfy $\mathbf{T}_{NT} \to \mathbf{T}$ w.p. 1. This however is contingent on the condition that $\mathbf{G}_{NT}(\mathbf{t}) = \mathbf{0}$ has an isolated root $\mathbf{t}_0 = \mathbf{T}_{NT}$ at a certain neighborhood. Portnoy [76], while assuming μ^n to be symmetric and absolutely continuous with density satisfying certain regularity properties, and that ψ is bounded and has a derivative which is bounded and uniformly continuous almost surely, established consistency of a solution $\mathbf{T}_{NT}$ of $\mathbf{G}_{NT}(\mathbf{t}) = \mathbf{0}$ nearest to a given consistent estimator $\tilde{\mathbf{T}}_{NT}$ for $\mathbf{t}_0 = \mathbf{T}$.

Using the fact that the S-estimators [85] also turns out to be a solution to (3.6), we can use the result of Lopuhaa [63] to establish consistency of the M-estimators.

Theorem 1 (Lopuhaa, 1989). Let $\mathscr{C}$ be the class of all measureable convex sets in $\mathbb{R}^q$ and suppose that every $E \in \mathscr{C}$ is a P_{μ^n}-continuity set, i.e., $P_{\mu^n}(\partial E) = 0$. Suppose that μ^n satisfies (C_ε) for some $0 < \varepsilon < 1-r$, and assume that the solution $\mathbf{T}$ of (3.5) is unique. Then for N, T sufficiently large, $\mathbf{G}_{NT}(\mathbf{t}) = \mathbf{0}$ has at least one solution $\mathbf{T}_{NT}$ and for any sequence of solutions $\mathbf{T}_{NT}$, $\lim_{N,T\to\infty} \mathbf{T}_{NT} = \mathbf{T}$.

Proof. See Lopuhaa [63] □

The above theorem is primarily for the location and scatter problem but can be adopted for the multivariate regression problem. Theorem 1 assures that $\mathbf{G}_{NT}(\mathbf{t}) = \mathbf{0}$ has at least one solution $\mathbf{T}(\mu^n_{NT})$ when $\lambda_1(\hat{\Sigma}) \to \infty$ and that one expects that $\mathbf{T}(\mu^n_{NT}) \to \mathbf{T}(\mu^n)$ as $N, T \to 0$. This implies that there is a neighborhood $\mathscr{O} \subset \mathbf{\Theta}$ around $\mathbf{T}(\mu^n)$ which contains all $\mathbf{T}(\mu^n_{NT})$ for N, T sufficiently large. The proof goes on with showing that all solutions eventually stay within a fixed compact set.

3.3 INFLUENCE FUNCTION

From the functional estimator $\mathbf{T}$ we wish to investigate its behavior when an atypical additional observation is thrown in resulting in a contamination measure $\mu^{n,\epsilon} \in \mathscr{M}^n$. With some

care, this can be approximated by the segment

$$\mu^{n,\epsilon} = (1-\epsilon)\mu^n + \epsilon\nu^n, \quad \nu^n \in \mathscr{M}^n \tag{3.8}$$

where $\epsilon \in [0,1]$. In some cases, $\mu^{n,\epsilon} \notin \mathscr{M}^n$ and we may need to extend the definition of $\mathbf{\Psi}$ so that it also holds for non-stationary and non-ergodic measure. A heuristic way of evaluating the effect of such a contamination is through Hampel's influence function. It has been shown, and by the close relationship between autoregression and ordinary regression, that this definition is still valid [55, 68] and is given in the following:

Definition 2. Let $\mathbf{T}(\cdot)$ be a vector-valued mapping from $\mathscr{M}^n$ into $\mathbf{\Theta}$ and let μ^n lie in the domain of $\mathbf{T}(\cdot)$. If $\nu^n = \delta_{\mathbf{z}}$ denotes the atomic probability distribution concentrated at $\mathbf{z} \in \mathbb{R}^{q+1}$, then the influence function of $\mathbf{T}(\cdot)$ at μ^n is defined pointwise by

$$\mathrm{IF}(\mathbf{y}, \mathbf{u}, x; \mathbf{T}, \mu^n) = \lim_{\epsilon \downarrow 0} \frac{\mathbf{T}(\mu^{n,\epsilon}) - \mathbf{T}(\mu^n)}{\epsilon} \tag{3.9}$$

providing the limit exists.

Note that if μ^n in (3.8) is replaced by the empirical distribution μ_{NT} and ϵ by $1/NT$, we realize that the IF measures a standardized change of the value of the estimator when one additional observation is added to a large sample of size $NT-1$. The importance of the influence function lies in its heuristic interpretation that it describes the effect of an infinitesimal contamination at point $\mathbf{z}$ on the estimate, i.e., it gives an approximation to the effect of the inclusion or deletion of a single observation. It is therefore desired that this effect is bounded so that no observation has a dominant influence on the value achieved by the estimator. As such, bounded influence function is thus considered to be a good robustness property.

Theorem 2. Let $\rho : \mathbb{R} \to [0, \infty)$ satisfy **(A1)** and **(A2)**. Assume that ρ has a second derivative ψ' and suppose that

(**A3**) $\psi'(x)$ and $u(x) = 2\psi(x)$ are bounded and continuous.

Suppose that the conditions similar to that in Theorem 1 hold. Let $\mathbf{\Psi}$ be defined as in (3.4-3.5) and let $\mathbf{G}_{\mu^n}(\mathbf{t}) = \mathrm{E}_{\mu^n}[\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x, \mathbf{t})]$. Suppose that $\mathbf{G}_{\mu^n}(\cdot)$ has a nonsingular derivative

$\mathbf{M}$ at $\mathbf{T}(\mu^n) = (\Pi(\mu^n), \Sigma(\mu^n))$. Then the influence function $\mathrm{IF}(\mathbf{y}, \mathbf{u}, x; \mathbf{T}, \mu^n)$ exists and satisfies

$$\mathrm{IF}(\mathbf{y}, \mathbf{u}, x; \mathbf{T}, \mu^n) = -\mathbf{M}^{-1}\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{T}). \tag{3.10}$$

Theorem 1 assures that $\mathbf{G}_{\mu^{n,\epsilon}}(\mathbf{t}) = \mathbf{0}$ has at least one solution $\mathbf{T}(\mu^{n,\epsilon})$. So we have to specify that we are only interested in the solution of $\mathbf{G}_{\mu^n}(\mathbf{t}) = \mathbf{0}$ given by the functional $\mathbf{T}(\mu^n)$. This also implies that to derive the influence function at a distribution μ^n it is required that, for sufficiently small ϵ, $\mathbf{T}(\cdot)$ is uniquely defined at $(1-\epsilon)\mu^n + \epsilon\delta_{\mathbf{z}}$ for all $\mathbf{z} \in \mathbb{R}^{q+1}$, and that the limit in (3.9) exists. The version of Implicit Function Theorem (IFT) by Lopuhaa [63] applied to $\mathbf{G}_{\mu^n}(\mathbf{t}) = \mathbf{0}$ will assure the uniqueness of $\mathbf{T}(\mu^{n,\epsilon})$ at $\mathbf{G}_{\mu^{n,\epsilon}}(\mathbf{t}) = \mathbf{0}$ for ϵ sufficiently small. The rest of the proof proceeds with the application of the same theorem [63].

Proof. Define the function $\mathbf{W}(\cdot; \mathbf{y}, \mathbf{u}, x) : [0,1] \times \mathscr{O}$, $\mathscr{O} \subset \mathbf{\Theta}$ by

$$\begin{aligned} \mathbf{W}(\epsilon, \mathbf{t}; \mathbf{y}, \mathbf{u}, x) &= \int \mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t}) d\mu^{n,\epsilon} && (3.11) \\ &= (1-\epsilon) \int \mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t}) d\mu^n + \epsilon \mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t}). && (3.12) \end{aligned}$$

Let $\mathbf{W}(\epsilon, \mathbf{t}; \mathbf{y}, \mathbf{u}, x)$ satisfy conditions 1-3 of the Implicit Function Theorem of Lopuhaa [63] (see Appendix) and let $\mathbf{t}_0 = (\boldsymbol{\pi}(\mu^n), \mathbf{S}(\mu^n))$ be the unique solution of 3.7, i.e., $\mathbf{W}(0, \mathbf{t}_0; \mathbf{y}, \mathbf{u}, x) = \mathbf{0}$. Condition 4 of IFT [63] guarantees that, for sufficiently small ϵ, if $G_{\mu^{n,\epsilon}}(\mathbf{t}) = \mathbf{0}$ has two solutions, i.e., they are both a zeros of $\mathbf{W}(\epsilon, \mathbf{t}\mathbf{y}, \mathbf{u}, x)$, then they are contained inside a small neighborhood around $\mathbf{T}(\mu^n)$. Furthermore, it can be argued that they are actually equal for sufficiently small ϵ so that the functional $\mathbf{T}(\mu^{n,\epsilon})$ is thus uniquely defined.

When the limit exists

$$\frac{\partial \mathbf{W}}{\partial \epsilon}(0, \mathbf{t}_0; \mathbf{y}, \mathbf{u}, x) = \lim_{\epsilon \downarrow 0} \frac{\mathbf{W}(\epsilon, \mathbf{t}_0; \mathbf{y}, \mathbf{u}, x) - \mathbf{W}(0, \mathbf{t}_0; \mathbf{y}, \mathbf{u}, x)}{\epsilon} \tag{3.13}$$

and by the applying implicit differentiation the left hand side of (3.13) is given by

$$\frac{\partial \mathbf{W}}{\partial \epsilon}(0, \mathbf{t}_0; \mathbf{y}, \mathbf{u}, x) = \frac{\partial \mathbf{W}}{\partial \mathbf{t}}(0, \mathbf{t}_0; \mathbf{y}, \mathbf{u}, x) \frac{\partial \mathbf{W}}{\partial \epsilon}(0, \mathbf{t}_0; \mathbf{y}, \mathbf{u}, x) \tag{3.14}$$

so that the function $\mathbf{T}(\cdot;\mu^n)$ also has a right derivative at $\epsilon = 0$ with

$$\frac{\partial \mathbf{T}(0,\mu^n)}{\partial \epsilon} = -\left[\frac{\partial \mathbf{W}}{\partial \mathbf{t}}(0,\mathbf{t}_0;\mathbf{y},\mathbf{u},x)\right]^{-1}\frac{\partial \mathbf{W}}{\partial \epsilon}(0,\mathbf{t}_0;\mathbf{y},\mathbf{u},x). \tag{3.15}$$

To show that the above limit exists we have to establish that (1) $\mathbf{\Psi}(\mathbf{y},\mathbf{u},x;\mathbf{t})$ is bounded and continuous on $\mathbb{R}^s \times \mathcal{S}_{NT}$ and that (2) $\partial\mathbf{\Psi}/\partial\mathbf{t}$ is also bounded and continuous on $\mathbb{R}^s \times \mathcal{S}_{NT}$. Requirement (1) is obvious through the function ψ. Hence $\mathbf{W}(\epsilon,\mathbf{t};\mathbf{y},\mathbf{u},x)$ is continuous in $[0,1]\times\mathscr{O}$ for every $(\mathbf{y},\mathbf{u},x)\in\mathbb{R}^{q+1}$. For (2) we compute $\partial\mathbf{\Psi}/\partial\mathbf{t}$:

$$\frac{\partial \mathbf{\Psi}^{(1)}}{\partial \boldsymbol{\pi}} = \mathbf{Q}^*\Omega^{-1}[2u'(d)\mathbf{r}\mathbf{r}^\top - \xi u(d)\Omega]\Omega^{-1}\mathbf{Q}^{*\top} \tag{3.16}$$

$$\frac{\partial \mathbf{\Psi}^{(2)}}{\partial \boldsymbol{\pi}} = \mathbf{Q}^*\Omega^{-1}[u'(d)\mathbf{r}\mathbf{r}^\top\Omega^{-1}\dot{\Omega}_k + \xi u(d)\dot{\Omega}_k]\Omega^{-1}\mathbf{r} = \left\{\frac{\partial \mathbf{\Psi}^{(1)}}{\partial s_l}\right\}^\top \tag{3.17}$$

$$\begin{aligned}\frac{\partial \mathbf{\Psi}^{(2)}}{\partial s_l} &= u'(d)\mathbf{r}^\top\Omega^{-1}\dot{\Omega}_l\Omega^{-1}\mathbf{r}\mathbf{r}^\top\Omega^{-1}\dot{\Omega}_k\Omega^{-1}\mathbf{r} - u(d)\mathbf{r}^\top\Omega^{-1}[\dot{\Omega}_l\Omega^{-1}\dot{\Omega}_k - \dot{\Omega}_l + \dot{\Omega}_k\Omega^{-1}\dot{\Omega}_l] \\ &\quad - \xi\mathrm{tr}[\Omega^{-1}\dot{\Omega}_l\Omega^{-1}\Omega_k + \Omega^{-1}\ddot{\Omega}_{kl}]\end{aligned} \tag{3.18}$$

Notice that $\|\mathbf{Q}^*\| \le \|\mathbf{z}^*\| \le \|\mathbf{r}\| + \|\mathbf{y}\|$ which is bounded in probability since $\mathbf{y}$, and consequently $\mathbf{r}$, is stationary and because $\|\mathbf{r}\mathbf{r}^\top\|/d^2 \le \|\Omega\| \le \|\Sigma_\Gamma\| + \|\Sigma_\varepsilon\|$. Then $\partial\mathbf{\Psi}/\partial\mathbf{t}$ is bounded by a constant that depends only on $\|\Sigma_\varepsilon\|$ and $\|\Sigma_\Gamma\|$. This and by (**A3**) yields requirement 2. Then the dominated convergence, in view of the above result, allows

$$\frac{\partial \mathbf{W}}{\partial \mathbf{t}}(\epsilon,\mathbf{t};\mathbf{y},\mathbf{u},x) = (1-\epsilon)\int \frac{\partial \mathbf{\Psi}}{\partial \mathbf{t}}(\mathbf{y},\mathbf{u},x;\mathbf{t})d\mu^n + \epsilon\frac{\partial \mathbf{\Psi}}{\partial \mathbf{t}}(\mathbf{y},\mathbf{u},x;\mathbf{t}) \tag{3.19}$$

which is also a continuous function on $[0,1]\times\mathscr{O}$ and that

$$\frac{\partial \mathbf{W}}{\partial \mathbf{t}}(0,\mathbf{t}_0;\mathbf{y},\mathbf{u},x) = \int \frac{\partial \mathbf{\Psi}}{\partial \mathbf{t}}(\mathbf{y},\mathbf{u},x;\mathbf{t})d\mu^n|_{\mathbf{t}=\mathbf{t}_0} = \mathbf{M} \tag{3.20}$$

which is nonsingular. Then since $\partial\mathbf{W}/\partial\epsilon = \mathbf{\Psi}(\mathbf{y},\mathbf{u},x;\mathbf{t}_0)$ so the theorem follows. □

A sample estimate of the influence of one observation to the value achieved by the estimator can be obtained by replacing the $\mathbf{M}$ by $\mathbf{M}_{NT}$ given by

$$\mathbf{M}_{NT} = \begin{bmatrix} \mathbf{M}_{\Pi^{\mathrm{D}*}\Pi^{\mathrm{D}*},NT} & \mathbf{M}^{\top}_{\Pi^{\mathrm{D}*}\mathbf{S},NT} \\ \mathbf{M}_{\Pi^{\mathrm{D}*}\mathbf{S},NT} & \mathbf{M}_{\mathbf{SS},NT} \end{bmatrix}$$

where

$$\mathbf{M}_{\Pi^{\mathrm{D}*}\Pi^{\mathrm{D}*},NT} = \sum_{i=1}^{N}\sum_{j=n+1}^{T} \mathbf{Q}^{*}\Omega_j^{-1}[u'(d_{ij})\mathbf{r}_{ij}\mathbf{r}_{ij}^{\top} - \xi u(d_{ij})\Omega_j]\Omega_j^{-1}\mathbf{Q}^{*\top} \tag{3.21}$$

$$\{\mathbf{M}_{\Pi^{\mathrm{D}*}\mathbf{S},NT}\}_k = -\sum_{i=1}^{N}\sum_{j=n+1}^{T} \mathbf{Q}^{*}\Omega_j^{-1}[u'(d_{ij})\mathbf{r}_{ij}\mathbf{r}_{ij}^{\top}\Omega_j^{-1}\dot{\Omega}_{jk} + \xi u(d_{ij})\dot{\Omega}_{jk}]\Omega_j^{-1}\mathbf{r}_{ij} \tag{3.22}$$

$$\begin{aligned} \{\mathbf{M}_{\mathbf{SS},NT}\}_{kl} = & \sum_{i=1}^{N}\sum_{j=n+1}^{T} u'(d_{ij})\mathbf{r}_{ij}^{\top}\Omega_j^{-1}\dot{\Omega}_{jl}\Omega_j^{-1}\mathbf{r}_{ij}\mathbf{r}_{ij}^{\top}\Omega_j^{-1}\dot{\Omega}_{jk}\Omega_j^{-1}\mathbf{r}_{ij} \\ & -u(d_{ij})\mathbf{r}_{ij}^{\top}\Omega_j^{-1}[\dot{\Omega}_{jl}\Omega_j^{-1}\dot{\Omega}_{jk} - \dot{\Omega}_{jl} + \dot{\Omega}_{jk}\Omega_j^{-1}\dot{\Omega}_{jl}] \\ & - \xi \mathrm{tr}[\Omega_j^{-1}\dot{\Omega}_{jl}\Omega_j^{-1}\Omega_{jk} + \Omega_j^{-1}\ddot{\Omega}_{jkl}] \end{aligned} \tag{3.23}$$

Note that the influence function in (3.10) and its sample counterpart is just proportional to $\mathbf{\Psi}$. Thus the principle of M-estimation possesses the nice feature that the desired properties for an influence curve maybe achieved simply by choosing a ψ with the given properties. In particular, if ψ is given by the Tukey's bisquare function then it can be seen that the influence if a single observation on the model parameters is always bounded.

3.4 ASYMPTOTIC NORMALITY

When (3.6) determines $\mathbf{t}_0 = \mathbf{T}$ uniquely or that a well defined solution exists through its corresponding minimization problem, then following Huber [46], Serfling [89], Collins [26], Portnoy [76] and Lopuhaa [63] there exists a sequence of M-estimators $\mathbf{T}_{NT}$ such that $\mathbf{T}_{NT} \to \mathbf{T}$ as $N \to \infty$ and $T \to \infty$ or equivalently there exists M-estimators $\mathbf{T}(\mu^{n,\epsilon})$ such that $\mathbf{T}(\mu^{n,\epsilon}) \to \mathbf{T}(\mu^{n})$ as $\epsilon \to 0$. Since $\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t})$ is differentiable with respect to $\mathbf{t}$ then

so is $\mathbf{G}_{NT}(\mathbf{t})$. Using appropriate regularity conditions, the Mean Value Theorem (MVT) applied to $\mathbf{G}_{NT}(\mathbf{t})$ gives

$$\mathbf{G}_{NT}(\mathbf{T}_{NT}) = \mathbf{G}_{NT}(\mathbf{T}) + \dot{\mathbf{G}}_{NT}(\tilde{\mathbf{T}})(\mathbf{T}_{NT} - \mathbf{T}) \tag{3.24}$$

where $\dot{\mathbf{G}}_{NT}(\tilde{\mathbf{T}}) = [\partial \mathbf{G}_{NT}(\mathbf{t})/\partial \mathbf{t}]_{\mathbf{t}=\tilde{\mathbf{T}}}$ and $\|\tilde{\mathbf{T}} - \mathbf{T}\| \leq \|\mathbf{T}_{NT} - \mathbf{T}\|$ or that $\tilde{\mathbf{T}}$ lies in the segment connecting $\mathbf{T}_{NT}$ and $\mathbf{T}$. Because $\mathbf{G}_{NT}(\mathbf{T}_{NT}) = \mathbf{0}$, we expect upon arrangement that (3.24) yields

$$\sqrt{NT}(\mathbf{T}_{NT} - \mathbf{T}) = [-\dot{\mathbf{G}}_{NT}(\tilde{\mathbf{T}})]^{-1}\sqrt{NT}\mathbf{G}_{NT}(\mathbf{T}). \tag{3.25}$$

The asymptotic normality of $\mathbf{T}_{NT}$ follows if the matrix $\dot{\mathbf{G}}_{NT}(\tilde{\mathbf{T}})$ behaves properly, i.e., if it converges appropriately and if $\sqrt{NT}\mathbf{G}_{NT}(\mathbf{T})$ has the CLT.

Prior to establishing asymptotic normality, we state the following theorem which outlines the conditions for it to happen.

Theorem 3. Let $\rho : \mathbb{R} \to [0, \infty)$ satisfy (**A1**)-(**A3**) and suppose that the conditions similar to Theorem 1 hold. Let $\mathbf{\Psi}$ be defined similar to (3.4-3.5) and let $\mathbf{G}_{\mu^n}(\mathbf{t}) = \mathrm{E}_{\mu^n}[\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x, \mathbf{t})]$ and $\mathbf{G}_{NT}(\mathbf{t}) := (NT)^{-1}\sum_{i=1}^{N}\sum_{j=n+1}^{T} \mathbf{\Psi}_{ij}(\mathbf{y}_{i,j}^{j+n}, \mathbf{u}, x; \mathbf{t})$. Suppose that the solution $\mathbf{T}(\mu^n)$ of $\mathbf{G}_{\mu^n}(\mathbf{t}) = \mathbf{0}$ is unique and let $\mathbf{T}_{NT}$ be the solution of $\mathbf{G}_{NT}(\mathbf{t}) = \mathbf{0}$. Then $\sqrt{NT}(\mathbf{T}_{NT} - \mathbf{T}) \sim \mathrm{N}(\mathbf{0}, \mathbf{M}^{-1}\Lambda\mathbf{M})$, where Λ is the covariance matrix of $\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{T})$.

Proof. Let

$$U(\mathbf{y}, \mathbf{u}, x; \mathbf{t}, d) = \sup_{\|\boldsymbol{\tau} - \mathbf{t}_0\| \leq d} \|\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \boldsymbol{\tau}) - \mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t})\|. \tag{3.26}$$

Theorem 3 of Huber [46] and its corollary suggests that it is sufficient to prove the following conditions:

1. There exists a $\mathbf{t}_0 = \mathbf{T} \in \mathbf{\Theta}$ such that $\mathbf{G}_{\mu^n}(\mathbf{y}, \mathbf{u}, x; \mathbf{T}) = \mathbf{0}$.
2. There exists positive numbers b, c, d, d_0 such that (i) $\mathrm{E}_{\mu^n} U(\mathbf{y}, \mathbf{u}, x; \mathbf{t}, d) \leq bd$ for $\|\mathbf{t} - \mathbf{t}_0\| + d \leq d_0$; and (ii)$\mathrm{E}_{\mu^n}[U(\mathbf{y}, \mathbf{u}, x; \mathbf{t}, d)^2] \leq cd$ for $\|\mathbf{t} - \mathbf{t}_0\| + d \leq d_0$.
3. The expectation $\mathrm{E}_{\mu^n}(\|\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t})\|^2)$ is nonzero and finite.

Condition 1 is clear since a solution $\mathbf{T}(\mu^n)$ if it exists must satisfy $\mathbf{G}_{\mu^n}(\mathbf{t})$, i.e. $\mathbf{t}_0 = (\Pi(\mu^n), \Sigma(\mu^n))$ is a zero of $\mathbf{G}_{\mu^n}(\mathbf{t})$ and that both $\|\Sigma_\varepsilon\|$ and $\|\Sigma_\Gamma\|$ are bounded away from 0 and ∞. On the other hand, part of the proof of theorem 2 requires boundedness of $\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t})$ by a constant and hence proves condition 3.

Let $\mathbf{t}_0 \in K$ where $K \subset \mathbf{\Theta}$, K compact. Since $B_d(\mathbf{t}) \subset K$, the application of MVT yields

$$\|\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t}) - \mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t}_0)\| \leq \|\frac{\partial \mathbf{\Psi}}{\partial \mathbf{t}}(\mathbf{y}, \mathbf{u}, x; \mathbf{t})|_{t=\tilde{\mathbf{T}}}\| \|\mathbf{t} - \mathbf{t}_0\|. \tag{3.27}$$

Again since $\partial \mathbf{\Psi}/\partial \mathbf{t}$ is bounded, there exists $b > 0$ such that $\|\frac{\partial \mathbf{\Psi}}{\partial \mathbf{t}}(\mathbf{y}, \mathbf{u}, x; \mathbf{t})|_{t=\tilde{\mathbf{T}}}\| \leq b$ so that

$$\|\mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t}) - \mathbf{\Psi}(\mathbf{y}, \mathbf{u}, x; \mathbf{t}_0)\| \leq b\|\mathbf{t} - \mathbf{t}_0\| \leq bd \tag{3.28}$$

Hence $\mathrm{E}_{\mu^n} U(\mathbf{y}, \mathbf{u}, x; \mathbf{t}, d) \leq bd$ and consequently $\mathrm{E}_{\mu^n}[U(\mathbf{y}, \mathbf{u}, x; \mathbf{t}, d)^2] \leq cd$ for $\|\mathbf{t} - \mathbf{t}_0\| + d \leq d_0$. □

The rest of the proof follows using Theorem 3 of Huber [46] which establishes that

$$\|\dot{\mathbf{G}}_{NT}(\mathbf{T}_{NT}) - \dot{\mathbf{G}}_{\mu^n}(\mathbf{T})\| \to 0 \tag{3.29}$$

in probability. Because $\mathbf{T}_{NT} \to \mathbf{T}$ and $\tilde{\mathbf{T}}$ is a suitable mean value

$$\|\dot{\mathbf{G}}_{NT}(\tilde{\mathbf{T}}) - \dot{\mathbf{G}}_{\mu^n}(\mathbf{T})\| \to 0, \tag{3.30}$$

i.e., $\dot{\mathbf{G}}_{NT}(\tilde{\mathbf{T}}) \to \mathbf{M}$. The Central Limit Theorem ensures that $\sqrt{NT}\mathbf{G}_{NT}(\mathbf{T}) \to \mathrm{N}(\mathbf{0}, \Lambda)$. A sample estimate of Λ can be obtained using $\Lambda_{NT} = (NT)^{-1} \sum_{i=1}^{N} \sum_{j=n+1}^{T} \mathbf{\Psi}_{ij} \mathbf{\Psi}_{ij}^\top$. The ensuing result has the feature that the asymptotic covariance matrix is actually given by the influence function. Therefore, one way of looking at bounded-influence estimation is to impose a bound on the influence and find an estimator that has the smallest variance subject to the chosen bound.

4.0 BOOTSTRAP APPROXIMATION

In the previous chapter, we note that the M-estimates have a normal limiting distribution. Inference can be based on the asymptotic variance through the sample estimate but this is not expected to give accurate results when outliers are present. Moreover, the expression for the variance using the sample estimate is quite intractable. An alternative approach is through nonparametric bootstrap which does not rely on any distributional assumptions. This can be done either by taking the linear fixed point representation of the estimator or by naive bootstrap. The former is computationally efficient but has the drawback that higher order asymptotic behavior equivalence may not exist. Thus, although the naive bootstrap has more computational burden, its results are generally more accurate [2].

In this chapter, we will describe a scheme to generate bootstrap samples that adopts to the cross-sectional structure of the data and a scheme to approximate the limiting distribution by naive bootstrap estimation. Then it will be shown that this scheme adheres to the bootstrap principle by establishing some asymptotic results for the validity of bootstrap approximation. Moreover, we will discuss model selection, in view of bounded-influence estimation, under a general elliptical response distribution. Inference through bootstrap in any statistical analysis is usually based on the final model which is treated as the "true model". The accuracy of the final model as an approximation to the "true model" thus have an effect on any statistical results. Therefore, it is to our best interest to determine appropriate model order prior to obtaining bootstrap approximation.

Here and elsewhere, $o_P(1)$ stands for convergence in probability while $O_P(1)$ means boundedness in probability. In the course of the discussion, we will use the device that unobservable errors and some resulting constructs associated to it are denoted by $\breve{\boldsymbol{\varepsilon}}$, the empirical unobservables are denoted in plain symbols while bootstrap related observations

and parameters are denoted with an $*$.

4.1 ROBUST MODEL SELECTION

Determining the correct order for the VARX component in model (2.1) and (2.2) is a crucial issue when obtaining reliable bootstrap estimates. Observations derived from simulations in simple univariate ARMA and AR models alone suggest that bootstrap approximation can only be expected to perform well when the parametric model provides a good approximation to the true model [24]. This is also true when underestimating the true model order of a general AR(n) model so that it is imperative that the probability of this happening should be kept asymptotically at zero [53]. The lag order selection criterion for the same model, however, need not be consistent for the lag order for the bootstrap algorithm to be asymptotically valid [53]. This suggests that, providing the range of lag orders considered includes the true lag order, a wide range of information based lag order selection criteria including Akaike Information Criterion (AIC) are potentially valid.

Developing a robustified Akaike-type model selection criterion to determine lag order in univariate autoregression has been investigated by a number of authors (see for example [67], [40], [10], [82]. An analog criterion for linear mixed effects multivariate autoregression models can be also be developed by adopting their methods as well as techniques from multivariate regression. This is because, when inquiry is regarding population or marginal focus the criterion coincides with the usual penalized marginal likelihood [99].

To proceed, we consider fitting the linearized form of (2.1) and (2.2) and make a simplification that the input term is a deterministic function with finite levels as discussed in the previous section. This model can be written in the form of (2.18), i.e. we consider the approximating model to be

$$F_A : \mathbf{y}_{ij} = \Pi^{\mathrm{D}*}\mathbf{z}^*_{ij} + \Gamma^*_i x_j + \boldsymbol{\varepsilon}_{ij} \tag{4.1}$$

where $\Pi^{\mathrm{D}*} = [\boldsymbol{\alpha}, \mathbf{D}_1\Pi_1^{\mathrm{diag}}, \ldots, \mathbf{D}_n\Pi_n^{\mathrm{diag}}, \tilde{\Gamma}, \Upsilon]$ depends implicitly on the lag order n, $\mathbf{z}^*_{ij} = [1, \mathbf{y}^\top_{i,j-1}, \ldots, \mathbf{y}^\top_{i,j-n}, x_j, \mathbf{u}_i^\top]^\top$ and the innovation $\mathbf{r}_{ij}$ has an underlying marginal density cor-

responding to the least favorable distribution proportional to $\exp\{-\tau(\mathbf{y}_{ij}^{j-n+1}, \mathbf{u}_{ij}, x_j; \mathbf{T})\}$, $\mathbf{T} = (\Pi^{\mathrm{D}*}, \Sigma = (\Sigma_\varepsilon, \Sigma_\Gamma))$, which accommodates possible contamination of the core density. Let this approximating model be denoted by $h(\cdot|\mathbf{t})$, $\mathbf{t} \in \mathbf{\Theta}$. Then a useful measure of discrepancy between the operating model F operating and approximating model $h(\cdot|\mathbf{t})$ is the Kullback-Leibler information

$$
\begin{aligned}
I(F, h_{\mathbf{t}}) &= \mathrm{E}_F \left\{-2 \log h(\cdot; \mathbf{t})\right\} && (4.2)\\
&= \mathrm{E}_F \log F(\mathbf{y}, \mathbf{u}, x) - \mathrm{E}_F \log h(\mathbf{y}, \mathbf{u}, x; \mathbf{t}) && (4.3)
\end{aligned}
$$

where the expectation is carried under the true model F. In this setting, smaller values of $I(F, h_{\mathbf{t}})$ correspond to better approximation of F by $h_{\mathbf{t}}$ while the minimum value is obtained for some $\mathbf{t}_0 \in \mathbf{\Theta}$. On the other hand, when F belongs to the fitted class of models $\mathscr{H} = \{h_{\mathbf{t}}, \mathbf{t} \in \mathbf{\Theta}\}$, then $F = h_{\mathbf{t}_0}$ and $I(F, h_{\mathbf{t}_0}) = 0$. However, in general, F many not be in $\mathscr{H}$ and so $I(F, h_{\mathbf{t}}) > 0$.

In practice, $\mathbf{t}$ is estimated from the data. In our case, this estimate is given by the redescending M-estimator defined as a solution to the first order condition in (3.3) where $\mathbf{\Psi}(\mathbf{y}_{ij}^{j-n+1}\mathbf{u}_{ij}, x_j; \mathbf{t}) = \frac{\partial}{\partial \mathbf{T}} \tau(\mathbf{y}_{ij}^{j-n+1}, \mathbf{u}_{ij}, x_j; \mathbf{t})$, and can be thought of as a maximum likelihood estimator with respect to the underlying density. Then, $I(F, h_{\mathbf{t}})$ is approximated by the loss function $I(F, h_{\mathbf{T}_{NT}})$ given by

$$
\begin{aligned}
I(F, h_{\mathbf{T}_{NT}}) &= -2\mathrm{E}_F \log\{(-2\pi)^{-\frac{1}{2}pN(T-n)} \prod_{j=n+1}^{T} |\Omega_j|^{-\frac{Np}{2}} \exp[-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=n+1}^{T} \rho(\mathbf{r}_{ij}^\top \Omega_j^{-1} \mathbf{r}_{ij})]\}\\
&\propto \mathrm{E}_F\{Np \sum_{j=n+1}^{T} |\Omega_j| + \sum_{i=1}^{N} \sum_{j=n+1}^{T} \rho(\mathbf{r}_{ij}^\top \Omega_j^{-1} \mathbf{r}_{ij})\}. && (4.4)
\end{aligned}
$$

Hence, a reasonable criterion for judging the quality of the approximating model is $I(F, h_{\mathbf{T}_{NT}})$ where $\mathbf{T}_{NT}$ is the M-estimator in (3.3). Given a collection of competing approximating models, the one that minimizes $I(F, h_{\mathbf{T}_{NT}})$ is preferred [6].

The robustified Akaike function given by

$$
AIC_R(n, \alpha, \rho_k) = 2 \sum_{i=1}^{N} \sum_{j=n+1}^{T} \tau(\mathbf{y}_{ij}^{j-n+1}, \mathbf{u}_{ij}, x_j; \mathbf{T}_{NT}) + \alpha n, \qquad (4.5)
$$

where αn is the generalized penalty criterion for a fixed α [12], is supposed to give an unbiased estimate and at least approximately mininimize $I(F, h_{\mathbf{T}_{NT}})$. Assuming consistency of the M-estimate, the penalty constant α, following the proposition of Ronchetti, [81], and Behrens [10], can be defined as $\mathrm{tr}(\mathbf{M}^{-1}\Lambda)$ with $\mathbf{M} = -\mathrm{E}[\partial\mathbf{\Psi}/\partial\mathbf{t}]$ and $\Lambda = \mathrm{E}[\mathbf{\Psi}\mathbf{\Psi}^\top]$. This choice of α follows from the asymptotic equivalence of the AIC with cross-validation given in Stone [95]. For the redescending M-estimate with the Tukey's bisquare function $\rho_{B,k}$ as the bounding function $\mathbf{M}$ and Λ can be approximated by $\mathbf{M}_{NT}$ and Λ_{NT}, respectively, which are given in the previous chapter. The danger here is that in the multivariate setting, the finite difference approximation to the Hessian matrix corresponding to the variance components do not result in a positive definite matrix. As an alternative, one can follow the heuristic argument of Martin [67] that the penalty term can be the number of parameters in the model. This follows from the fact that an M-estimator is a maximum likelihood estimate under the least favorable distribution. In this case, a robust maximum-likelihood type order selection criterion would be obtained by choosing n to minimize

$$\begin{aligned} AIC_R(n, 2, \rho_k) &= 2\sum_{i=1}^{N}\sum_{j=n+1}^{T} \tau(\mathbf{y}_{ij}^{j-n+1}, \mathbf{u}_{ij}, x_j; \mathbf{T}_{NT}) + 2\phi(d, n) \\ &= \sum_{i=1}^{T-n} \rho(\hat{\mathbf{r}}_{ij}^\top \hat{\Omega}_j^{-1} \hat{\mathbf{r}}_{ij}) + N\sum_{j=n+1}^{T} \log|\hat{\Omega}_j| + 2(n+1)p + p(p+1) \qquad (4.6) \end{aligned}$$

where $\phi(p, n)$ is the penalty function which in this case is the number of parameters estimated in the model.

4.2 NAIVE BOOTSTRAP APPROXIMATION

It is possible to adapt basic ideas behind bootstrapping autoregressive and linear mixed effects models to obtain an appropriate resampling technique to generate an appropriate bootstrap sample for Mixed-VARX model. In autoregressive bootstrap, the most common resampling technique is to reconstruct the autoregressive process based on conditionally independent innovations. Details on this technique are laid out in earlier papers such as Bose [15, 16], Basawa et al. [5], Kreiss and Franke [54], and Allen and Datta [2]. To

accommodate the mixed model case, this method has to be augmented by incorporating a prior resampling scheme for the random effect parameters.

In general, we will assume that $\{(\mathbf{y}_{ij}, \mathbf{u}_{ij})\}_{j\in\mathbb{Z}^+}$, is an ergodic and stationary autoregressive process of order n for each $i = 1, \ldots, N$ and satisfies the following linear difference equation induced by model (2.1) or (2.3)

$$\mathbf{y}_{ij} = h(\mathbf{y}_{i,j-1}^{j-n}, \mathbf{u}_i, x_j; \mathbf{T}, \Gamma_i) + \boldsymbol{\varepsilon}_{ij}, \quad j \in \mathbb{Z}^+ \tag{4.7}$$

where $n \in \mathbb{N}$ and $\boldsymbol{\varepsilon}_{ij}$ is a stationary and ergodic sequence of zero mean iid random vectors, with common distribution function μ_ε, that are independent of both $\{(\mathbf{y}_{11}, \mathbf{u}_{11}), \ldots, (\mathbf{y}_{NT}, \mathbf{u}_{NT})\}$ and $\{\Gamma_1, \ldots, \Gamma_N\}$. We also assumed previously that the autoregressive parameters satisfy the condition that the roots of the characteristic function (2.2) is never zero in the unit sphere so that dependence on initial values vanish exponentially fast.

Once the appropriate model order has been selected, the first step is to use data $\mathbf{Y}_{NT} = \{(\mathbf{y}_{11}, \mathbf{u}_{11}, x_1), \ldots, (\mathbf{y}_{NT}, \mathbf{u}_{NT}, x_T)\}$ to calculate a preliminary parameter estimate $\mathbf{T}_{NT} = (\hat{\Pi}^{\mathrm{D}*}, \hat{\Sigma})$ and the BLUP $\{\hat{\Gamma}_1, \ldots, \hat{\Gamma}_N\}$ for the random effect parameters which allows us to estimate the errors by the residuals $\hat{\varepsilon}_{ij} = \mathbf{y}_{ij} - h(\mathbf{y}_{i,j-1}^{j-n}, \mathbf{u}_i, x_j; \mathbf{T}_{NT}, \hat{\Gamma}_i)$. The BLUP's form an empirical distribution $\hat{\mu}_\Gamma$ from which we take a sample size r with replacement and call these $\{\hat{\Gamma}_1^*, \ldots, \hat{\Gamma}_r^*\}$. The residuals $\hat{\boldsymbol{\varepsilon}}_{ij}$, on the other hand, need to be centered $\tilde{\boldsymbol{\varepsilon}}_{ij} = \hat{\boldsymbol{\varepsilon}}_{ij} - \bar{\boldsymbol{\varepsilon}}$, with centering value $\bar{\boldsymbol{\varepsilon}} = (NT)^{-1} \sum_{i=1}^N \sum_{j=1}^T \hat{\boldsymbol{\varepsilon}}_{ij}$, to generate a valid approximation. Centering is important since it is presumed from the form of the bounding function requires that, componentwise, $\mathrm{E}\psi(\varepsilon) = 0$. If not done, it can cause random bias that does not vanish in the limit and renders the approximation useless (see, for example, Bickel and Freedman [13], Shorack [90], and Lahiri [58] that treats a similar bias phenomenon in regression problems). From these centered residuals, we can define the following centered empirical distribution function

$$\tilde{\mu}_{NT}(\mathbf{x}) = (NT)^{-1} \sum_{i=1}^N \sum_{j=1}^T \delta(\mathbf{x}) \quad \mathbf{x} \in \mathbb{R}^p$$

from which we draw a simple random sample $\tilde{\boldsymbol{\varepsilon}}_{i,n+1}^*, \ldots, \tilde{\boldsymbol{\varepsilon}}_{im}^*$ of size $m - n$ for each i with replacement from the collection of centered residuals $\{\tilde{\boldsymbol{\varepsilon}}_{1T}^1, \ldots, \tilde{\boldsymbol{\varepsilon}}_{NT}^1\}$. Alternatively, the residuals could be kept in groups by individuals, so that for each individual, a random i^* is

selected from $\{1, \ldots, r\}$, then a sample of size $m - n$ is taken with replacement from $\{\tilde{\boldsymbol{\varepsilon}}^1_{i^*T}\}$. Then we can we transform $\tilde{\boldsymbol{\varepsilon}}^*_{ij}$ or $\tilde{\boldsymbol{\varepsilon}}^*_{i^*j}$ recursively using (4.7) as

$$\begin{aligned} \mathbf{y}^*_{ij} &= \mathbf{y}_{ij} \quad \text{for} \quad j = 1, \ldots, n \quad \text{and} \\ \mathbf{y}^*_{ij} &= h(\mathbf{y}^{j-n*}_{i,j-1}, \mathbf{u}^*_{ij}, x_j; \mathbf{T}_{NT}, \hat{\Gamma}^*_i) \quad \text{for} \quad j = n+1, \ldots, m \end{aligned} \tag{4.8}$$

with initial values $\tilde{\boldsymbol{\varepsilon}}^*_{ij} = 0$ for $j = 1, \ldots, n$ to obtain the bootstrap sample $\mathbf{Y}^*_{rm} = \{(\mathbf{y}^*_{11}, \mathbf{u}^*_{11}, x_1), \ldots, (\mathbf{y}^*_{rm}, \mathbf{u}^*_{rm}, x_m)\}$. Then the bootstrap version of the random variable $\mathbf{T}_{NT}$ is given by $\mathbf{T}^*_{rm}$ which is the solution to the equation

$$\sum_{i=1}^{r} \sum_{j=n+1}^{m} \boldsymbol{\Psi}^*_{ij}(\mathbf{y}^{j-n+1*}_{ij}, \mathbf{u}^*_{ij}, x_j; \mathbf{T}^*_{rm}) = \mathbf{0}. \tag{4.9}$$

Consequently, the naive bootstrap approximation to the sampling distribution $\sqrt{NT}(\mathbf{T}_{NT} - \mathbf{T})$ is found by looking at the conditional distribution of $\mathbf{T}^*_{rm}$ given $\mathbf{Y}^*_{rm}$.

4.3 ASYMPTOTIC VALIDITY

Since the distribution of $\sqrt{NT}(\mathbf{T}_{NT} - \mathbf{T})$ is approximated by the conditional distribution given the original data $\mathbf{Y}_{NT}$ of a quantity $\mathbf{T}^*_{rm}$ which can be calculated from the bootstrap data $\mathbf{Y}^*_{rm}$ and the residuals $\tilde{\boldsymbol{\varepsilon}}^*_{ij}$, we have to justify that the purported conditional distribution is a reasonable approximation. Hence, it is important to deal with the problem of estimating the distributions of the centered residuals and the estimated random effects and making sure that their empirical distribution $\tilde{\mu}_{NT}$ and $\hat{\mu}_\Gamma$ conforms with the prescribed operating distribution μ_ε of the true residuals and μ_Γ of the true random effects in the suitable sense as $N, T \to \infty$, respectively. This result is essential since if they do not match, the bootstrap data can create a bias in the estimates that do not necessarily vanish even with large numbers of resampled observations.

Let $d_r (r \geq 1)$ be the Mallows metric [13] defined for probability measures μ_X on $\mathbb{R}^k$ with $\int \|\mathbf{x}\| d\mu_X < \infty$ such that for any two probability measures μ_X and μ_Y their distance d_r is given by

$$d_r(\mu_X, \mu_Y) = \inf(\mathrm{E}\|X - Y\|^r)^{1/r} \tag{4.10}$$

where the minimum is taken over all pairs (X, Y) with $X \sim \mu_X$ and $Y \sim \mu_Y$. Using this metric we have the following result similar to Theorem 3.1 in Kreiss and Franke [54] in the case of ARMA models.

Theorem 4. $d_2(\tilde{\mu}_{NT}, \mu_\varepsilon) \to 0$ in probability as $N, T \to \infty$.

Proof. When $\mu_{\varepsilon_{NT}}$ denotes the empirical distribution function of the unobservable residuals $\boldsymbol{\varepsilon}_{i1}, \ldots, \boldsymbol{\varepsilon}_{NT}$, Bickel and Freedman [13] (Lemma 8.4) showed that $d_2(\mu_{\varepsilon_{NT}}, \mu_\varepsilon) \to 0$ as $N, T \to \infty$ almost everywhere. Then it suffices to show that $d_2(\mu_{\varepsilon_{NT}}, \tilde{\mu}_{NT}) \to 0$.

Let I, J be uniformly distributed over the lattice with $I = \{1, \ldots, N\}$, $J = \{1, \ldots, T\}$ and define a random variable X_1 and Y_1 with marginals $\mu_{\varepsilon_{NT}}$ and $\tilde{\mu}_{NT}$, respectively, according to

$$X_1 = \boldsymbol{\varepsilon}_{IJ} \quad Y_1 = \hat{\boldsymbol{\varepsilon}}_{ij} - \bar{\boldsymbol{\varepsilon}} \tag{4.11}$$

and observe that

$$\begin{aligned}
\{d_2(\mu_{\varepsilon_{NT}}, \mu_\varepsilon)\}^2 &= \inf \mathrm{E}\|X - Y\|^2 \leq \mathrm{E}\|X_1 - Y_1\|^2 \\
&= \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \left\| \hat{\boldsymbol{\varepsilon}}_{ij} - \boldsymbol{\varepsilon}_{ij} - \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \hat{\boldsymbol{\varepsilon}}_{ij} \right\|^2 \\
&\leq \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \|\hat{\boldsymbol{\varepsilon}}_{ij} - \boldsymbol{\varepsilon}_{ij})\|^2 + \frac{1}{(NT)^2} \left\| \sum_{i=1}^{N} \sum_{j=1}^{T} \boldsymbol{\varepsilon}_{ij} \right\|^2 \\
&\leq \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} |\hat{\Pi}^{\mathrm{D}*} - \Pi^{\mathrm{D}*}|^2 \|\mathbf{z}_{ij}^*\|^2 + |\hat{\Gamma}_i - \Gamma_i|^2 \|\mathbf{w}_{ij})\|^2 + \\
&\qquad \frac{1}{(NT)^2} \left\| \sum_{i=1}^{N} \sum_{j=1}^{T} \boldsymbol{\varepsilon}_{ij} \right\|^2
\end{aligned}$$

Note that $|\mathbf{w}_j|$ is bounded by a constant and $\|\mathbf{z}_{ij}^*\|^2 = O_P(1)$ due to the stationarity of $\mathbf{y}_{ij}$. Then since $\sqrt{NT}(\mathbf{T}_{NT} - \mathbf{T}) = O_P(1)$ and $\sqrt{N}(\hat{\Gamma}_i - \Gamma_i) = O_P(1)$ [49] we expect that the first term is $o_P(1)$. The central limit theorem assures that $1/\sqrt{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \boldsymbol{\varepsilon}_{ij}$ is also $O_P(1)$ so the second term is $o_P(1)$. Thus we obtain the assertion of the theorem. □

The result that the empirical distribution of the predicted random effects converges suitably to its corresponding true distribution, i.e., $d_2(\hat{\mu}_\Gamma, \mu_\Gamma) \to 0$, is established in Jiang

[49]. Given these equivalence relations in distributions, we only need to show that the distribution of $\mathbf{T}^*_{rm}$ is "close" to the distribution of $\mathbf{T}_{NT}$.

In practice, one calculates $\mathbf{T}_{NT}$ in (3.3) by a finite iteration of the Newton's method starting with a $\sqrt{NT}$-consistent solution as an initial value, e.g. least squares estimates. This implies that any subsequent estimate update inherits the same $\sqrt{NT}$-consistency. In fact, using an arbitrary but $\sqrt{NT}$-consistent estimator for $\mathbf{T}$ as an initial estimator, it can be shown that the M-estimator $\mathbf{T}_{NT}$ which fulfills $\sum_{i=1}^{N}\sum_{j=n+1}^{T}\mathbf{\Psi}_{ij}(\mathbf{y}_{ij}^{j-n+1},\mathbf{u}_{ij},x_j;\mathbf{T}_{NT}) = O_p(\sqrt{NT})$ can be obtained by a one-step Newton iteration [54]. Hence, the definition of the M-estimator in (3.3) can actually be broadened by considering $\sqrt{NT}$-consistent estimators which solves the same equation only in an asymptotic sense, i.e., $\sum_{i=1}^{N}\sum_{j=n+1}^{T}\mathbf{\Psi}_{ij}(\mathbf{y}_{ij}^{j-n+1},\mathbf{u}_{ij},x_j;\mathbf{T}_{NT}) = O_p(\sqrt{NT})$, while preserving the asymptotic distribution theory of the estimates.

Using the expanded definition, assume $\mathbf{T}^*_{rm}$ is the M-estimator calculated from the bootstrap sample $\mathbf{Y}^*_{rm}$, i.e., $\mathbf{T}^*_{rm}$ is a function of $\mathbf{Y}^*_{rm}$ satisfying

$$\sum_{i=1}^{N}\sum_{j=n+1}^{T}\mathbf{\Psi}^*_{ij}(\mathbf{y}_{ij}^{j-n+1,*},\mathbf{u}^*_{ij},x_j;\mathbf{T}^*_{rm}) = O_{P^*}(\sqrt{NT}) \tag{4.12}$$

where $\mathbf{\Psi}^*_{ij}$ is analogous to that of $\mathbf{\Psi}_{ij}$. Since $\mathbf{T}_{NT}$ is the parameter of the process from which the bootstrap data is generated, it is not unreasonable to expect that $\mathbf{T}^*_{rm}$ is $\sqrt{NT}$ consistent, i.e.,

$$\sqrt{NT}(\mathbf{T}^*_{rm} - \mathbf{T}_{NT}) = O_{P^*}(1). \tag{4.13}$$

In fact, assuming compactness of the neighborhood around $\mathbf{T}_{NT}$, the application of the Mean Value theorem to $\mathbf{G}^*_{rm}(\mathbf{T}^*_{rm})$ yields

$$\begin{aligned} O_{P^*}(1) &= \sqrt{rm}\mathbf{G}^*_{rm}(\mathbf{T}^*_{rm}) \\ &= \sqrt{rm}\mathbf{G}^*_{rm}(\mathbf{T}^*_{NT}) + \sqrt{rm}\dot{\mathbf{G}}^*_{rm}(\tilde{\mathbf{T}})(\mathbf{T}_{NT} - \mathbf{T}^*_{rm}) \end{aligned}$$

where $\tilde{\mathbf{T}}$ lies in between $\mathbf{T}_{NT}$ and $\mathbf{T}^*_{rm}$. Since $\dot{\mathbf{G}}^*_{rm}(\tilde{\mathbf{T}})$ is bounded then if $rm/NT \to 1$ we have

$$\sqrt{NT}(\mathbf{T}_{NT} - \mathbf{T}^*_{rm}) = \left[-\dot{\mathbf{G}}^*_{rm}(\tilde{\mathbf{T}})\right]^{-1}\sqrt{NT}\mathbf{G}^*_{rm}(\mathbf{T}_{NT}) + O_{P^*}(1). \tag{4.14}$$

We then have the following result which establishes equivalence of bootstrap approximation with the fixed point limiting distribution of the estimates.

Theorem 5. Let $\rho : \mathbb{R} \to [0, \infty)$ satisfy (**A1**)-(**A3**) and let the conditions similar to Theorem 1 hold. If $\{\mathbf{T}.\} \subset \mathbf{\Theta}$ denotes a sequence of $\sqrt{NT}$-consistent estimators for (Π, Σ) and $\|\mathbf{T} - \tilde{\mathbf{T}}_1\| \leq \|\mathbf{T} - \tilde{\mathbf{T}}_{NT}\|$ and $\|\mathbf{T} - \tilde{\mathbf{T}}_2\| \leq \|\mathbf{T} - \tilde{\mathbf{T}}^*_{rm}\|$ then

$$d_2(\sqrt{NT}\dot{\mathbf{G}}_{NT}(\tilde{\mathbf{T}}_1), \sqrt{rm}\dot{\mathbf{G}}^*_{rm}(\tilde{\mathbf{T}}_2)) \to 0 \quad \text{in probability} \tag{4.15}$$

and

$$d_2(\sqrt{NT}\mathbf{G}_{NT}(\mathbf{T}_{NT}), \sqrt{rm}\mathbf{G}^*_{rm}(\mathbf{T}^*_{rm})) \to 0 \quad \text{in probability} \tag{4.16}$$

Recall that the M-estimator we calculated before, under certain regularity conditions, converges weakly to the normal law. The above expression along with (4.14) ensures that the conditional distribution of $\left[-\dot{\mathbf{G}}^*_{rm}(\tilde{\mathbf{t}})\right]^{-1} \sqrt{NT}\mathbf{G}^*_{rm}(\mathbf{T}_{NT})$ also converges weakly to the same normal law $\mathrm{N}(\mathbf{0}, \mathbf{M}^{-1}\Lambda\mathbf{M})$ which is the asymptotic distribution of the M-estimates and affirms the asymptotic validity of the bootstrap in this case.

Proof. We will prove the second assertion only with the first estimating equation. The proof for the other estimating equation and the first assertion is tedious but follows the same line of arguments.

The squared Mallows distance in (4.16)

$$\{d_2(\sqrt{NT}\mathbf{G}_{NT}(\mathbf{T}_{NT}), \sqrt{rm}\mathbf{G}^*_{rm}(\mathbf{T}^*_{rm}))\}^2 = \inf \mathrm{E}\|\sqrt{NT}\mathbf{G}_{NT}(\mathbf{T}_{NT}) - \sqrt{rm}\mathbf{G}^*_{rm}(\mathbf{T}^*_{rm}))\|^2. \tag{4.17}$$

Note that $\mathbf{G}$ or $\mathbf{G}^*$ is a sum of two functions defined on two distributions depending on whether $x_j = 1$ or $x_j = 0$. Hence the infimum has to be evaluated over $\{\mu_\varepsilon, \tilde{\mu}_{NT}\}$ and $\{\mu_\varepsilon + \mu_\Gamma, \tilde{\mu}_{NT} + \hat{\mu}_\Gamma\}$. It is clear that the right hand side of (4.17) has

$$\begin{aligned}
\leq \; & \mathrm{E}\left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{j=n+1}^{T} \mathbf{\Psi}_{ij}(\breve{\mathbf{y}}_{ij}^{j-n+1}, \mathbf{u}_{ij}, x_j; \mathbf{T}_{NT}) \right. \\
& \qquad \left. - \frac{1}{\sqrt{rm}} \sum_{i=1}^{r} \sum_{j=n+1}^{m} \mathbf{\Psi}^*_{ij}(\mathbf{y}_{ij}^{j-n+1*}, \mathbf{u}^*_{ij}, x_j; \mathbf{T}^*_{rm}) \right\|^2 \\
\leq \; & \frac{1}{\min\{NT, rm\}} \mathrm{E}\left\| \sum_{i=1}^{N} \sum_{j=n+1}^{T} u(\breve{d}_{ij})\mathbf{Q}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij} - \sum_{i=1}^{r} \sum_{j=n+1}^{m} u(d^*_{ij})\mathbf{Q}^{**}\Omega_j^{*-1}\mathbf{r}^*_{ij} \right\|^2.
\end{aligned}$$

Let $NT' = \min\{NT, rm\}$, $N' = \max\{N, r\}$, and $T' = \max\{T, m\}$ then the right hand side of the last equation is less than

$$\frac{1}{NT'}\mathrm{E}\left\|\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}(u(\breve{d}_{ij}) - u(d^*_{ij}))\mathbf{Q}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij}\right\|^2 + \frac{1}{NT'}\mathrm{E}\left\|\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}u(d^*_{ij})(\breve{\mathbf{Q}}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij} - \mathbf{Q}^{**}\Omega_j^{*-1}\mathbf{r}^*_{ij})\right\|^2 \quad (4.18)$$

Each term of (4.18) can be split into two parts conditional on x_j, for example the first term given in the following, so that the expectation of each part is evaluated over its respective distribution.

$$\frac{1}{NT'}\mathrm{E}\left\|\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}(u(\breve{d}_{ij}) - u(d^*_{ij}))\breve{\mathbf{Q}}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij}|x_j = 0\right\|^2 + \frac{1}{NT'}\mathrm{E}\left\|\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}(u(\breve{d}_{ij} - u(d^*_{ij}))\breve{\mathbf{Q}}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij}|x_j = 1\right\|^2 \quad (4.19)$$

Let $\breve{c}$ be the mean of $u(\breve{d}_{ij})$ under μ_ε and let c^* be the mean of $u(d^*_{ij})$ under $\tilde{\mu}_{NT}$. Since u has a bounded and continuous derivative the MVT and Schwartz's inequality immediately imply

$$\begin{aligned}\|c^* - \breve{c}\| &\leq \sup_x \|u'(x)\|\|d^*_{ij} - \breve{d}_{ij}\| \\ &\leq \sup_x \|u'(x)\| d_2(\tilde{\mu}_{NT}, \mu_\varepsilon) \end{aligned} \quad (4.20)$$

Since $\|u'(x)\|$ is bounded above and $d_2(\tilde{\mu}_{NT}, \mu_\varepsilon) \to 0$ in probability, then $c^* - c \to 0$ in probability as well. Similarly,

$$d_2(u(\breve{d}_{ij}), u(d^*_{ij})) \leq \sup_x \|x\| d_2(\mu_{\varepsilon_{NT}}, \tilde{\mu}_{NT}) \to 0 \quad \text{in probability} \quad (4.21)$$

Now consider the expansion of the first term on (4.19) dropping off the notation for its dependence on x_j

$$\frac{1}{NT'}\mathrm{E}\left[\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}(u(\breve{d}_{ij}) - u(d^*_{ij}) - c + c^*)\breve{\mathbf{Q}}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij}\right]^\top \times \left[\sum_{k=1}^{N'}\sum_{l=n+1}^{T'}(u(\breve{d}_{kl}) - u(d^*_{kl}) - c + c^*)\breve{\mathbf{Q}}^*\Omega_l^{-1}\breve{\mathbf{r}}_{kl}\right] \quad (4.22)$$

$$= \frac{1}{NT'} \sum_{i=1}^{N'} \sum_{j=n+1}^{T'} \mathrm{E}(u(\breve{d}_{ij}) - u(d^*_{ij}) - c + c^*)^2 \mathrm{E}\|\breve{\mathbf{Q}}^* \Omega_j^{-1} \breve{\mathbf{r}}_{ij}\|^2 \tag{4.23}$$

Since, conditional on previous $\mathbf{y}$ values, $\mathbf{r}_{ij}$ is independent of $\mathbf{r}_{kl}$ for $i \neq k$ and $j \neq l$ and the $\mathbf{w}_j$'s are independent, while the inequality is by the fact that $\mathrm{E}\|XY\| \leq \mathrm{E}\|X\|\mathrm{E}\|Y\|$. By (4.20-4.21) the first minuend of (4.23) goes to zero in probability while the second minuend is $O_P(1)$ since it can be seen as the estimating equation for ordinary least squares. In fact,

$$\begin{aligned} \mathrm{E}[\breve{\mathbf{r}}_{ij}^\top \Omega_j^{-1} \breve{\mathbf{Q}}^{*\top} \breve{\mathbf{Q}}^* \Omega_j^{-1} \breve{\mathbf{r}}_{ij}] &\leq \mathrm{E}[\breve{\mathbf{r}}_{ij}^\top \Omega_j^{-1} \Omega_j^{-1} \breve{\mathbf{r}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top} \breve{\mathbf{z}}_{ij}^*] \\ &= \mathrm{tr}(\Omega_j^{-1} \mathrm{E}[\breve{\mathbf{r}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top} \breve{\mathbf{z}}_{ij}^* \breve{\mathbf{r}}_{ij}^\top] \Omega_j^{-1}) \end{aligned} \tag{4.24}$$

and $\mathrm{E}[\breve{\mathbf{r}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top} \breve{\mathbf{z}}_{ij}^* \breve{\mathbf{r}}_{ij}^\top] = O_P(1)$. The former assertion implies that given some ϵ there is a positive number K such that when $NT' \geq K$, $|\mathrm{E}[\cdot]| \leq \epsilon$, so that (4.23) is bounded by $(NT)^{-1}\epsilon \sum_{i=1}^{N'} \sum_{j=n+1}^{T'} \mathrm{tr}(\Omega_j^{-1} \mathrm{E}[\breve{\mathbf{r}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top} \breve{\mathbf{z}}_{ij}^* \breve{\mathbf{r}}_{ij}^\top] \Omega_j^{-1})$ and it can be seen, by the boundedness of Ω_j^{-1}, that (4.23) goes to zero in probability. Similarly, the second term of (4.19) also goes to zero by the same arguments so that the first term of (4.18) goes to zero in probability.

Next we have to show that the second term of (4.18) also goes to zero in probability. To do this we will find an upper bound for $\|\breve{\mathbf{Q}}^* \Omega_j^{-1} \breve{\mathbf{r}}_{ij} - \mathbf{Q}^{**} \Omega_j^{*-1} \mathbf{r}_{ij}^*\|^2$. Since $\breve{\mathbf{Q}}^*$ is obtained from $\breve{\mathbf{z}}_{ij}^* \otimes I_p$ by deleting some rows, then

$$\begin{aligned} \|\breve{\mathbf{Q}}^* \Omega_j^{-1} \breve{\mathbf{r}}_{ij} - \mathbf{Q}^{**} \Omega_j^{*-1} \mathbf{r}_{ij}^*\|^2 &\leq \|(\breve{\mathbf{z}}_{ij}^* \otimes I_p) \Omega_j^{-1} \breve{\mathbf{r}}_{ij} - (\mathbf{z}_{ij}^{**} \otimes I_p) \Omega_j^{*-1} \mathbf{r}_{ij}^*\|^2 \\ &= \|\mathrm{vec}(\Omega_j^{-1} \breve{\mathbf{r}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top}) - \mathrm{vec}(\Omega_j^{*-1} \mathbf{r}_{ij}^* \mathbf{z}_{ij}^{**\top})\|^2 \\ &= \|\Omega_j^{-1} \breve{\mathbf{r}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top} - \Omega_j^{*-1} \mathbf{r}_{ij}^* \mathbf{z}_{ij}^{**\top}\|^2 \\ &\leq \|(\Omega_j^{-1} - \Omega_j^{*-1}) \mathbf{r}_{ij}^* \mathbf{z}_{ij}^{**\top}\|^2 \\ &\quad + \|\Omega_j^{-1} (\breve{\mathbf{r}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top} - \mathbf{r}_{ij}^* \mathbf{z}_{ij}^{**\top})\|^2 \end{aligned} \tag{4.25}$$

Since $\|\mathbf{r}_{ij}^* \mathbf{z}_{ij}^{**\top}\| = O_P(1)$ then the first term of (4.25) is bounded by $|\mathbf{T} - \mathbf{T}_{rm}^*| O_P(1)$. On the other hand,

$$\begin{aligned} \|\Omega_j^{-1} (\breve{\mathbf{r}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top} - \mathbf{r}_{ij}^* \mathbf{z}_{ij}^{**\top})\|^2 &\leq |\Omega_j^{-1}|^2 \|(\breve{\mathbf{y}}_{ij} \breve{\mathbf{z}}_{ij}^{*\top} - \mathbf{y}_{ij}^* \mathbf{z}_{ij}^{**\top})\|^2 \\ &\quad + |\Omega_j^{-1}|^2 \|\Pi^{\mathrm{D}*} \breve{\mathbf{z}}_{ij}^* \breve{\mathbf{z}}_{ij}^{*\top} - \hat{\Pi}^{\mathrm{D}*} \mathbf{z}_{ij}^{**} \mathbf{z}_{ij}^{**\top}\|^2 \end{aligned} \tag{4.26}$$

and

$$\|\Pi^{\mathrm{D}*}\breve{\mathbf{z}}^{*}_{ij}\breve{\mathbf{z}}^{*\top}_{ij} - \hat{\Pi}^{\mathrm{D}*}\mathbf{z}^{**}_{ij}\mathbf{z}^{**\top}_{ij}\|^2 \;\leq\; \|(\Pi^{\mathrm{D}*} - \hat{\Pi}^{\mathrm{D}*})\mathbf{z}^{**}_{ij}\mathbf{z}^{**\top}_{ij}\|^2 + \|\hat{\Pi}^{\mathrm{D}*}(\breve{\mathbf{z}}^{*}_{ij}\breve{\mathbf{z}}^{*\top}_{ij} - \mathbf{z}^{**}_{ij}\mathbf{z}^{**\top}_{ij})\|^2. \tag{4.27}$$

Note that $\mathbf{y}_{ij} = \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h\boldsymbol{\varepsilon}_{i,j-h} + \boldsymbol{\omega}_i$ for some ω_i so that

$$\mathbf{z}^{\top}_{ij} = \begin{bmatrix} 1 \\ \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h\boldsymbol{\varepsilon}_{i,j-1-h} + \boldsymbol{\omega}_i \\ \vdots \\ \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h\boldsymbol{\varepsilon}_{i,j-n-h} + \boldsymbol{\omega}_i \\ \mathbf{u}_i \\ x_j \end{bmatrix}. \tag{4.28}$$

Then $\|(\breve{\mathbf{y}}_{ij}\breve{\mathbf{z}}^{*\top}_{ij} - \mathbf{y}^{*}_{ij}\mathbf{z}^{**\top}_{ij})\|^2$ equals

$$\left\| \begin{matrix} \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h\breve{\boldsymbol{\varepsilon}}_{i,j-h} - \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})\boldsymbol{\varepsilon}^{*}_{i,j-h} + \boldsymbol{\omega}_i - \hat{\boldsymbol{\omega}}_i \\ \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h\breve{\boldsymbol{\varepsilon}}_{i,j-1-h}\breve{\boldsymbol{\varepsilon}}^{\top}_{i,j-1-h}\boldsymbol{\gamma}^{\top}_h - \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})\boldsymbol{\varepsilon}^{*}_{i,j-1-h}\boldsymbol{\varepsilon}^{*\top}_{i,j-1-h}\boldsymbol{\gamma}^{\top}_h(\mathbf{T}_{NT}) + \boldsymbol{\omega}_i\boldsymbol{\omega}^{\top}_i - \hat{\boldsymbol{\omega}}_i\hat{\boldsymbol{\omega}}^{\top}_i \\ \vdots \\ \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h\breve{\boldsymbol{\varepsilon}}_{i,j-n-h}\breve{\boldsymbol{\varepsilon}}^{\top}_{i,j-n-h}\boldsymbol{\gamma}^{\top}_h - \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})\boldsymbol{\varepsilon}^{*}_{i,j-n-h}\boldsymbol{\varepsilon}^{*\top}_{i,j-n-h}\boldsymbol{\gamma}^{\top}_h(\mathbf{T}_{NT}) + \boldsymbol{\omega}_i\boldsymbol{\omega}^{\top}_i - \hat{\boldsymbol{\omega}}_i\hat{\boldsymbol{\omega}}^{\top}_i \\ \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h\breve{\boldsymbol{\varepsilon}}_{i,j-h}\mathbf{u}_i - \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})\boldsymbol{\varepsilon}^{*}_{i,j-h}\mathbf{u}^{*}_i + \boldsymbol{\omega}_i\mathbf{u}_i - \hat{\boldsymbol{\omega}}_i\mathbf{u}^{*}_i \\ \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h\breve{\boldsymbol{\varepsilon}}_{i,j-h}x_j - \sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})\boldsymbol{\varepsilon}^{*}_{i,j-h}x_j + \boldsymbol{\omega}_i x_j - \hat{\boldsymbol{\omega}}_i x_j \end{matrix} \right\|^2 \tag{4.29}$$

which is less than

$$\left\| \begin{matrix} \sum_{h=0}^{\infty}(\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_h(\mathbf{T}_{NT}))\boldsymbol{\varepsilon}^{*}_{i,j-h} + \boldsymbol{\omega}_i - \hat{\boldsymbol{\omega}}_i \\ \sum_{h=0}^{\infty}(\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_h(\mathbf{T}_{NT}))\boldsymbol{\varepsilon}^{*}_{i,j-1-h}\boldsymbol{\varepsilon}^{*\top}_{i,j-1-h}(\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_h(\mathbf{T}_{NT}))^{\top} + \boldsymbol{\omega}_i\boldsymbol{\omega}^{\top}_i - \hat{\boldsymbol{\omega}}_i\hat{\boldsymbol{\omega}}^{\top}_i \\ \vdots \\ \sum_{h=0}^{\infty}(\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_h(\mathbf{T}_{NT}))\boldsymbol{\varepsilon}^{*}_{i,j-n-h}\boldsymbol{\varepsilon}^{*\top}_{i,j-n-h}(\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_h(\mathbf{T}_{NT}))^{\top} + \boldsymbol{\omega}_i\boldsymbol{\omega}^{\top}_i - \hat{\boldsymbol{\omega}}_i\hat{\boldsymbol{\omega}}^{\top}_i \\ \sum_{h=0}^{\infty}(\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_h(\mathbf{T}_{NT}))\boldsymbol{\varepsilon}^{*}_{i,j-h}\mathbf{u}_i + (\boldsymbol{\omega}_i - \hat{\boldsymbol{\omega}}_i)\mathbf{u}^{*}_i \\ \sum_{h=0}^{\infty}(\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_h(\mathbf{T}_{NT}))\boldsymbol{\varepsilon}^{*}_{i,j-h}x_j + (\boldsymbol{\omega}_i - \hat{\boldsymbol{\omega}}_i)x_j \end{matrix} \right\|^2$$

$$
+\left\|\begin{array}{c}
\sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})(\breve{\boldsymbol{\varepsilon}}_{i,j-h}-\boldsymbol{\varepsilon}^*_{i,j-h})+\boldsymbol{\omega}_i-\hat{\boldsymbol{\omega}}_i\\
\sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})(\breve{\boldsymbol{\varepsilon}}_{i,j-1-h}\breve{\boldsymbol{\varepsilon}}^\top_{i,j-1-h}-\boldsymbol{\varepsilon}^*_{i,j-1-h}\boldsymbol{\varepsilon}^{*\top}_{i,j-1-h})\boldsymbol{\gamma}_h(\mathbf{T}_{NT})+\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top-\hat{\boldsymbol{\omega}}_i\hat{\boldsymbol{\omega}}_i^\top\\
\vdots\\
\sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})(\breve{\boldsymbol{\varepsilon}}_{i,j-n-h}\breve{\boldsymbol{\varepsilon}}^\top_{i,j-n-h}-\boldsymbol{\varepsilon}^*_{i,j-n-h}\boldsymbol{\varepsilon}^{*\top}_{i,j-n-h})\boldsymbol{\gamma}_h(\mathbf{T}_{NT})+\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top-\hat{\boldsymbol{\omega}}_i\hat{\boldsymbol{\omega}}_i^\top\\
\sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})(\breve{\boldsymbol{\varepsilon}}_{i,j-h}\mathbf{u}_i-\boldsymbol{\varepsilon}^*_{i,j-h}\mathbf{u}^*_i)+\hat{\boldsymbol{\omega}}_i(\mathbf{u}_i-\mathbf{u}^*_i)\\
\sum_{h=0}^{\infty}\boldsymbol{\gamma}_h(\mathbf{T}_{NT})(\breve{\boldsymbol{\varepsilon}}_{i,j-h}-\boldsymbol{\varepsilon}^*_{i,j-h})x_j
\end{array}\right\|^2 \tag{4.30}
$$

Let C, κ, and ϑ be appropriately chosen constants. Then we have the following bounds

$$
|\boldsymbol{\gamma}_h| \leq C\kappa^\nu \quad \nu\geq 0 \tag{4.31}
$$

$$
|\boldsymbol{\gamma}_h-\boldsymbol{\gamma}_h(\mathbf{T}_{NT})| \leq C|\mathbf{T}-\mathbf{T}_{NT}|\vartheta^\nu \quad \nu\geq 0 \tag{4.32}
$$

$$
|\boldsymbol{\gamma}_h(\mathbf{T}_{NT})| \leq C|\mathbf{T}-\mathbf{T}_{NT}|\kappa^\nu \quad \nu\geq 0. \tag{4.33}
$$

Consequently,

$$
\begin{aligned}
\left\|\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}(\breve{\mathbf{y}}_{ij}\breve{\mathbf{z}}^{*\top}_{ij}-\mathbf{y}^*_{ij}\mathbf{z}^{**\top}_{ij})\right\|^2 &\leq C|\mathbf{T}-\mathbf{T}_{NT}|^2+N^{-1}C\sum_{i=1}^{N}|\Gamma_i-\hat{\Gamma}_i|^2\\
&\quad + C\|\breve{\boldsymbol{\varepsilon}}_{ij}-\boldsymbol{\varepsilon}^*_{ij}\|^2
\end{aligned} \tag{4.34}
$$

The bound for $\|\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}(\breve{\mathbf{z}}_{ij}\breve{\mathbf{z}}^{*\top}_{ij}-\mathbf{z}^*_{ij}\mathbf{z}^{**\top}_{ij})\|^2$ can be obtained similarly.

Now consider the expansion of the second term of (4.18)

$$
\begin{aligned}
&\frac{1}{NT'}\mathrm{E}\left\|\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}u(d^*_{ij})(\breve{\mathbf{Q}}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij}-\mathbf{Q}^{**}\Omega_j^{*-1}\mathbf{r}^*_{ij})\right\|^2\\
&\qquad\leq\frac{1}{NT'}\mathrm{E}\left\|\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}(u(d^*_{ij})-c+c^*)(\breve{\mathbf{Q}}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij}-\mathbf{Q}^{**}\Omega_j^{*-1}\mathbf{r}^*_{ij})\right\|^2\\
&\qquad\quad+\frac{1}{NT'}\mathrm{E}\left\|(c^*-c)\sum_{i=1}^{N'}\sum_{j=n+1}^{T'}(\breve{\mathbf{Q}}^*\Omega_j^{-1}\breve{\mathbf{r}}_{ij}-\mathbf{Q}^{**}\Omega_j^{*-1}\mathbf{r}^*_{ij})\right\|^2.
\end{aligned} \tag{4.35}
$$

Using the same arguments from (4.23-4.24), and by the results obtained in (4.25-4.27, 4.34) as well as $\|\breve{\boldsymbol{\varepsilon}}_{ij}-\boldsymbol{\varepsilon}^*_{ij}\|^2=d_2^2(\mu_\varepsilon,\tilde{\mu}_{NT})$ then it can be seen that the second term of (4.18) goes to zero in probability. □

The ensuing result implies that

$$\sqrt{NT}(\mathbf{T}^*_{rm} - \mathbf{T}_{NT}) \sim \mathrm{N}(\mathbf{0}, \mathbf{M}^{-1}\Lambda\mathbf{M}^{-1}). \tag{4.36}$$

Furthermore, since $\mathbf{T}^*_{rm}$ is essentially an M-estimate then $\sqrt{NT}(\mathbf{T}^*_{rm} - T) \sim \mathrm{N}(\mathbf{0}, \mathbf{M}^{-1}\Lambda\mathbf{M}^{-1})$. Therefore, if $\mathbf{V}$ denotes $\mathbf{M}^{-1}\Lambda\mathbf{M}^{-1}$ and since $\mathbf{V}_{NT} = \mathbf{M}^{-1}_{NT}\Lambda_{NT}\mathbf{M}^{-1}_{NT}$ converges weakly to $\mathbf{V}$ then we have

$$\sup_{\mathbf{x}} \|P^*\{\sqrt{NT}\mathbf{V}^{-1/2}_{NT}(\mathbf{T}^*_{rm} - \mathbf{T}) \leq \mathbf{x}\} - P\{\sqrt{NT}\mathbf{V}^{-1/2}_{NT}(\mathbf{T}_{NT} - \mathbf{T}) \leq \mathbf{x}\}\| \to 0 \quad \text{in probability.} \tag{4.37}$$

5.0 SIMULATIONS AND APPLICATION TO NEUROIMAGING DATA

Robust estimation in multivariate data models has been considered one of the most difficult problems in the robust literature [79]. This is even more challenging with multivariate time series models when autoregressive dependence in observations cannot be ignored. In this chapter, we apply the estimation procedure for the model specified in Chapter 2 both on real and simulated data. The simulated data considers two different types of outlier contamination to see how the estimation works under such circumstances. Some remarks are given initially with regards to some nuances in its computation.

5.1 NUMERICAL ISSUES

When minimizing (2.18), one must choose a function ρ satisfying the conditions set in Chapter 2. Although there is a broad class of functions one can choose from, we limit our choice to those whose derivatives redescend to zero when its argument is large. One popular choice is given by

$$\rho_{B,k}(x) = \begin{cases} \frac{k^2}{2}(1-(1-\frac{x^2}{k^2})^3), & |x| \leq k \\ \frac{k^2}{6}, & |x| > k \end{cases} \tag{5.1}$$

and its derivative is the Tukey biweight function $\psi_{B,k}(x) = k^2 x(1-\frac{x^2}{k^2})I(|x| \leq k)$. Note that (5.1) depends on the positive tuning parameter k which is actually associated with the asymptotic efficiency and breakdown point properties of the redescending M-estimator. For lack of information on its optimal value we will just assume a particular value based on the scalar multiple of the median of the squared residuals.

The choice of the ρ function also has implications both on the existence of the solution and to the numerical computation alike. For strong redescenders, such as the one given above where $\psi(x) = 0$ for large x, $|\Omega_j| \to -\infty$ and so the value of (2.18) also goes out of bounds. Therefore it may seem that no solution exists to the minimization problem or if such a solution exists it may not be quite desirable. A similar result can happen when one uses an initial estimate which is far from the global minimizer. The residuals obtained are usually large, thus are assigned zero weights by a strongly redescending ψ. Consequently, (2.19) and (2.20) are close to zero and the optimization terminates prematurely at an unsatisfactory solution.

To obtain good initial estimates, a resampling scheme can be adopted based on the the techniques used by Rousseeuw and Leroy [86] and Arslan *et al.* [4]. The strategy finds good initial points with non-zero gradients that catch the global minimum with good probability. Reasonable $\Pi_0^{\mathrm{D}*}$ values can be obtained by simplified fitting using a small subset of the $N(T-n)$ rows of the design matrix. If ps parameters need to be estimated, we can pick $N(ps+q)$ equations, i.e., $ps+q$ equations for every cross-section where q is some small number, at random each time generating a system to be solved for $\Pi^{\mathrm{D}*}$. The goal is to get at least some samples in each cross-section without outliers, giving good starting points from where the global minimum can be identified. However, the cross-sectional structure of the data may require a large number of generated samples to be assured of at least one "good sample". In fact, if we assumed that the fraction of outliers is 1% distributed uniformly among the cross-sections and the number of parameters to be estimated is 27, we may have to generate some 366 samples to be within 80% sure of getting at least one "good sample". If structure in the data is ignored, this number can be significantly reduced.

Once the initial estimates are set, the optimization can proceed by solving (2.19) and (2.20) jointly. We recommend, however, a two-stage iterative approach since convergence is more consistent. We have encountered examples when optimization on the whole parameter space failed to converge but was able to optimize in the partitioned approach with ease. In this approach, when Σ_ε and Σ_Γ are held fixed, we can simplify (2.19) to solve for $\Pi^{\mathrm{D}*}$ in the

following

$$\sum_{i=1}^{N}\sum_{j=n+1}^{T}\psi(d_{ij})\mathrm{vec}(\mathbf{Q}^{*}\Omega_{j}^{-1}\mathbf{y}_{ij}) = \sum_{i=1}^{N}\sum_{j=n+1}^{T}\psi(d_{ij})\mathrm{vec}(\mathbf{Q}^{*}\Omega_{j}^{-1}\Pi^{\mathrm{D}*}\mathbf{z}_{ij}) \tag{5.2}$$

$$= \sum_{i=1}^{N}\sum_{j=n+1}^{T}\psi(d_{ij})(\mathbf{z}_{ij}\otimes\mathbf{Q}^{*}\Omega_{j}^{-1})\mathrm{vec}(\Pi^{\mathrm{D}*}). \tag{5.3}$$

So that

$$\mathrm{vec}(\Pi^{\mathrm{D}*}) = \sum_{i=1}^{N}\sum_{j=n+1}^{T}\psi(d_{ij})\mathbf{R}^{-1}(\mathbf{z}_{ij}\otimes\mathbf{Q}^{*}\Omega_{j}^{-1})\mathbf{y}_{ij} \tag{5.4}$$

where $\mathbf{R} = \sum_{i=1}^{N}\sum_{j=n+1}^{T}\psi(d_{ij})(\mathbf{z}_{ij}\otimes\mathbf{Q}^{*}\Omega_{j}^{-1})$ We then substitute this estimate to (2.20) to obtain updated estimates of Σ_{ε} and Σ_{Γ}. The drawback to this procedure is that it can generally be slow. A multivariate analog of the one-dimensional line search method by Arslan *et al.* [4] for Σ_{ε} and Σ_{Γ} while minimizing for $\Pi^{\mathrm{D}*}$ can be adopted to accelerate it but the derivatives can be so complicated to be helpful.

5.2 SIMULATIONS

To illustrate the estimation procedure, we generated an artificial data set that is contaminated with two distinct outlier types: innovations outliers (IO) and additive outliers (AO). The observations were generated using the following replacement scheme: let the observed series be given by

$$\mathbf{z}_{ij} = \mathbf{y}_{ij} + \boldsymbol{\nu}_{ij} \tag{5.5}$$

so that autoregression is not perfectly observed and where

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\mu}_{g} + \mathbf{D}(\Pi + \Pi_{g})\mathbf{y}_{i,t-1} + \Gamma_{i}\mathbf{w}_{j} + \Upsilon\mathbf{u}_{i} + \boldsymbol{\varepsilon}_{ij} \tag{5.6}$$

and $\boldsymbol{\nu}_{ij}$ is an independent sequence of variables and independent of the sequence $\mathbf{y}_{ij}$. The innovation sequence $\boldsymbol{\varepsilon}_{ij}$ is independent and identically distributed with symmetric distribution G which, in our case, is a mixture of two concentric normal distributions given by

$$G = (1-\epsilon)N(0,\Sigma_{\varepsilon}) + \epsilon N(0, c\Sigma_{\varepsilon}). \tag{5.7}$$

The variables $\boldsymbol{\nu}_{ij}$, on the other hand, have distribution H, given by

$$H = (1-\epsilon)\delta_0 + \epsilon B, \tag{5.8}$$

where δ_0 is the atomic distribution that assigns probability 1 to the origin and B is an arbitrary distribution chosen to be a Bernoulli distribution that assigns probability π to a vector of constants $\mathbf{C}$ and $1-\pi$ to $-\mathbf{C}$. As for the other variables and fixed parameters in the model, we assume a univariate input covariate u_{ij} which has a value of 1 or 0 for some time duration so that its covariance is given by $\Sigma_\Gamma = \sigma^2 \otimes I_p$. This specification of the input covariate mimics the stimulus in the real data set. Finally, a matrix constraint is chosen arbitrarily.

The magnitude and proportion of outliers present in the data poses a difficult task for the redescending M-estimation (RM) procedure. Yet despite the situation, we can see from Table 1 that the RM-estimates are reasonably close to the true parameters compared to the maximum likelihood estimates whose estimates succumb to the bias caused by the contamination. In fact, the RM procedure still works well even in the presence of additive outliers which is generally considered more unwieldy than innovational outliers since small proportions of the former can already cause large bias. However, the estimates of the autoregressive parameters Π and Π_g tend to shrink toward zero, confirming the same observation noted by Denby and Martin [31, 67] within the univariate time series setting. This is because the contamination causes the data to "look like" multivariate white noise. The ML estimate of σ^2 is even negative which implies that in the presence of severe contamination the random effect becomes unidentifiable. On the other hand, the standard deviation (in parenthesis) obtained from RM estimates are, generally, larger than the ML estimates. This arises because the values of the Hessian matrix become smaller when observations are bounded than when they are not, hence its inverse will have larger values.

In figure 3 the trajectory of cross-section SH08's time series at three locations is shown. Note that the 1-0 stimulus is no longer distinguishable in all locations when contamination is present. In the figure, the predicted values using RM estimates are usually resistant against peaks and maintains a conservative prediction along the mean of the series. Although at first blush, one might have the impression that the prediction is too conservative since second

	True	ML	RM		True	ML	RM
μ_1	0	0.1632	-0.0160	γ_1	0.150	-0.0195	0.0058
		(0.1286)	(0.1563)			(0.0369)	(0.0436)
μ_2	0	0.0580	-0.0443	γ_2	0.100	0.1089	0.0142
		(0.1296)	(0.1497)			(0.0372)	(0.0432)
μ_3	0	0.1493	-0.0191	γ_3	-0.050	0.0051	-0.0211
		(0.1269)	(0.1461)			(0.0364)	(0.0430)
μ_{g1}	0.050	0.0831	0.0459	τ_1	0.005	0.0032	0.0052
		(0.0382)	(0.0447)			(0.0024)	(0.0029)
μ_{g2}	-0.010	0.0286	-0.0198	τ_2	-0.003	-0.0050	-0.0024
		(0.0375)	(0.0436)			(0.0024)	(0.0028)
μ_{g3}	0.030	0.0325	0.0428	τ_3	0.001	-0.0014	0.0018
		(0.0367)	(0.0434)			(0.0024)	(0.0027)
π_{11}	0.250	0.1979	0.2765	π_{g11}	0.100	0.0704	0.0640
		(0.0134)	(0.0187)			(0.0196)	(0.0245)
π_{22}	0.300	0.2135	0.2685	π_{g22}	-0.150	-0.1273	-0.1467
		(0.0142)	(0.0199)			(0.0201)	(0.0269)
π_{33}	0.350	0.3167	0.3398	π_{g33}	0.150	0.0513	0.1518
		(0.0128)	(0.0216)			(0.0192)	(0.0282)
ℓ_1	0.800	1.8078	0.7824	ℓ_4	0.800	1.7727	0.7824
		(0.0145)	——			(0.0153)	——
ℓ_2	0	0.4243	-0.0401	ℓ_5	0	0.3413	-0.0256
		(0.0185)	(0.0213)			(0.0172)	(0.0210)
ℓ_3	0	0.5533	-0.0120	ℓ_6	0.800	1.6629	0.7541
		(0.0177)	(0.0215)			(0.0137)	——
σ^2	0.300	-0.0433	0.2667				
		(0.0468)	——				

Table 1: True and estimated parameters of (5.6) using maximum likelihood (ML) and redescending-M (RM) estimation in the presence of innovation and additive outliers.

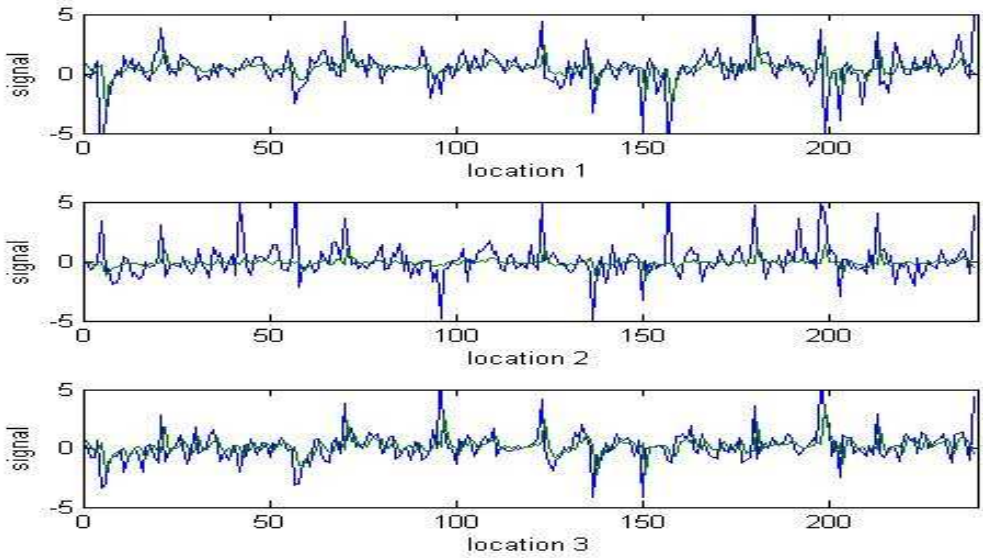

Figure 3: Time course of cross-section SH08 at three different locations.

location does not vary that much, it must be noted that this effect is actually attributed to the low magnitude of the linear combination of the observations when the parameter weights are negligible. On the other hand, figure 4 shows the residual time series from each location of the same cross-section. All the residual trajectories generally behave like white noise except for some spurious spikes. These spikes are obtained when the estimated model is fitted into the data wherein the robust procedure provides good prediction for "good data" and does not try to fit outliers. Hence the effect of the contamination does inflate prediction but remains in the heavy-tailed error distribution as shown by the quantile plot.

5.3 NEUROIMAGING DATA

In brief, the experiment was performed on a 3-T MRI scanner on 19 subjects. These subjects come from two groups: individuals who showed large cardiovascular reactions from an prior

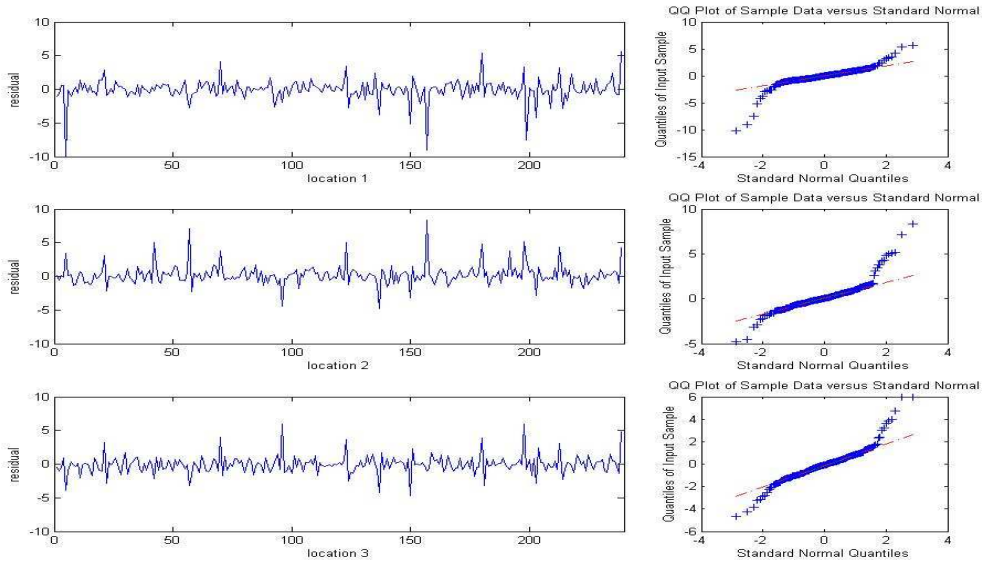

Figure 4: Residual time course and residual qqplot of cross-section SH08 at three different locations.

test battery, categorized as High Reactors and those who showed small cardiovascular reactions, categorized as Low Reactors. The stimulus involved was a modified Stroop color-word interference task where subjects were presented with a sequence of color words on a visual display and they are to determine the color in which the target word was shown as quickly and as accurately as possible. The task has two conditions: congruent and incongruent so that for all of the trials of the Congruent condition, the color of the target word and the identifier words are the same with the color in which the target word appears, while for all of the trials of the Incongruent condition, the color of the target word and the identifier words are incongruent with the color in which the target word appears (see figure 5) . The subjects complete eight 90-second blocks of each condition (8 Congruent, 8 Incongruent) in an alternating fixed order, beginning with the incongruent condition. This gives 60 images for each block for a total of 960 images for the whole experiment.

After image preprocessing, a two level mixed-effects parametric modulation analysis [20] was implemented using Statistical Parametric Mapping (SPM) to examine the correlation

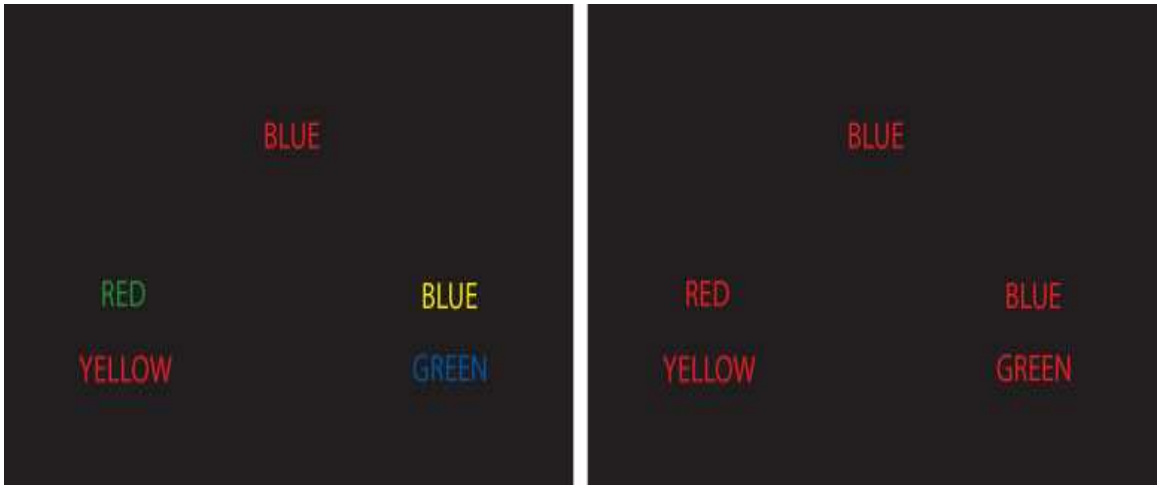

Figure 5: A Modified Stroop color-word interference task showing the incongruent and congruent condition.

between the amplitude of the fMRI BOLD hemodynamic response and the concurrent level of mean arterial pressure across the 8 congruent and 8 incongruent blocks. In the first level analysis, each subject's observed BOLD responses were correlated with the hemodynamic response convolved MAP. The resulting SPM(t) images from each subject were aggregated to obtain regions in which greater BOLD response amplitudes correlated with a concurrently high level of MAP. Some of the regions activated are the insular cortex, the anterior cingulate, and the perigenual cingulate shown in figure 6. A representative time series defined by the first eigenvariate was then extracted and processed from each of these regions for all individuals.

With these time series, we want to investigate the dynamic behavior and mutual predictability of the BOLD response at each location based on its recent past as well as how information from other regions are utilized to gain better prediction at that location. Based on initial investigations on an acceptable model order, we came up with $n = 3$ and $m = 1$ for the model in (2.2). The modified AIC, however, recommends a much higher autoregressive order but for the sake of parsimony and from correlation results on lagged residuals on some individuals (see for example figure 8, column 3), the chosen AR order maybe sufficient. A

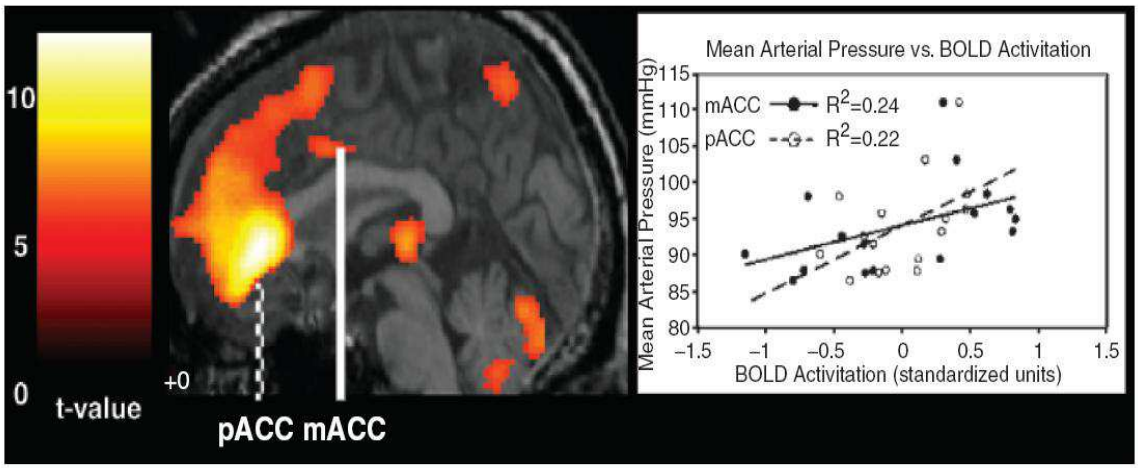

Figure 6: Sagittal slice of the brain showing regions in which greater hemodynamic BOLD response amplitudes correlate with concurrent greater levels of MAP during the Stroop color-word interference task.

more thorough investigation is perhaps needed with regards to this problem. On the other hand, the effect of the stimulus on the response becomes linear with longer time durations, such as in block designs, so that it is not necessary to use lagged terms on the stimulus part, i.e. $m = 1$ is satisfactory. Hence, the model we will be working on is given by

$$\mathbf{y}_{ij} = \sum_{h=1}^{3} \mathbf{D}_h(\Pi_h^{\text{diag}} + \Pi_{h(g)}^{\text{diag}})\mathbf{y}_{i,j-h} + \Gamma_i x_{ij} + \Upsilon u_i + \boldsymbol{\varepsilon}_{ij} \tag{5.9}$$

where the non-time varying u_i is chosen as the age of subject i. The mean of each subjects multiple time series has been removed so we shall no longer include an intercept term. The ML and RM estimates of this model are given in Table 2 where the constraint matrix used are the robust autocorrelations up to lag 3 given by

$$\mathbf{D}_1 = \begin{bmatrix} 1.0000 & 0.3269 & 0.1869 \\ 0.3472 & 1.0000 & 0.1943 \\ 0.1614 & 0.1618 & 1.0000 \end{bmatrix} \mathbf{D}_2 = \begin{bmatrix} 1.0000 & 0.2868 & 0.1586 \\ 0.2838 & 1.0000 & 0.1302 \\ 0.1458 & 0.1073 & 1.0000 \end{bmatrix}$$

$$\mathbf{D}_3 = \begin{bmatrix} 1.0000 & 0.2541 & 0.1606 \\ 0.2567 & 1.0000 & 0.1637 \\ 0.1197 & 0.1208 & 1.0000 \end{bmatrix}.$$

The predicting equations for each location up to grouping can be obtained using (5.9).

A plot of the predicted values, using RM estimates, for each location's time series course of subject H20 is shown in figure 7. By inspection, the predictions generally have good fit to the data and the model choice is appropriate since the residuals behave like white noise (see figure 8) and the lagged correlation between the residuals is almost zero. The latter remark can be verified by looking at the residual vs. lagged residual plot in figure 8 where the points concentrate elliptically about the origin. This implies that most of the serial information has been removed or accounted for in the model. The plots also show a feature of robust estimation that it exposes the outliers far from the bulk of the data.

Since the asymptotic distribution of the estimates have are complicated closed forms and are not accurate in the presence of contamination, bootstrap approximations that are applied directly to the RM-estimator were calculated over 2500 bootstrap samples. The quantile plots of the estimates are displayed in figures 9-10. The median of the bootstrap distribution for each parameter can be easily inferred from the plot. The overall shape of the plots are close to normal at the core and elongated at the tails. In fact, there is also some evidence of right skewness if one looks at the empirical histogram (not shown here) and supports the fact that the asymptotic distribution maybe unacceptable for use in inference which intrinsically relies on tail probabilities. On the other hand, bootstrap distribution for both the error and random effect variance are difficult to obtain since they easily get trapped at the initial estimate when it is close to the minimizer of the likelihood.

How do all these estimates relate to the question of temporal dynamics in the brain data? Temporal information in vector autoregressive models are obtained from the auto- and cross-covariance functions which are used to estimate the coefficients at different time lags. In this model, estimation of these coefficients are based on the constraint weighted auto- and cross-covariance functions and hence temporal information are also influenced by the constraint used. Temporal profile of the weighted coefficients characterize the temporal

	ML	RM		ML	RM
$\pi_{1,11}$	0.3017(0.0150)	0.3567(0.0571)	$\pi_{1,11(1)}$	-0.0378(0.0186)	0.0100(0.1136)
$\pi_{1,22}$	0.3026(0.0182)	0.3123(0.0435)	$\pi_{1,22(1)}$	-0.0608(0.0210)	0.1025(0.0660)
$\pi_{1,33}$	0.3437(0.0160)	0.2919(0.0409)	$\pi_{1,33(1)}$	0.0139(0.0203)	0.0635(0.0619)
$\pi_{2,11}$	0.1886(0.0155)	0.2133(0.0398)	$\pi_{2,11(1)}$	-0.0654(0.0190)	0.0018(0.0583)
$\pi_{2,22}$	0.2298(0.0185)	0.2355(0.0474)	$\pi_{2,22(1)}$	-0.1393(0.0213)	-0.0875(0.0612)
$\pi_{2,33}$	0.3224(0.0158)	0.4415(0.0450)	$\pi_{2,33(1)}$	-0.0750(0.0204)	-0.0128(0.0609)
$\pi_{3,11}$	0.0997(0.0148)	0.0426(0.0483)	$\pi_{3,11(1)}$	0.0580(0.0183)	0.0412(0.0560)
$\pi_{3,22}$	0.1358(0.0178)	0.0969(0.0465)	$\pi_{3,22(1)}$	0.1121(0.0206)	-0.0482(0.0478)
$\pi_{3,33}$	0.1287(0.0157)	0.0831(0.0353)	$\pi_{3,33(1)}$	-0.0064(0.0199)	0.0082(0.0436)
γ_1	0.0425(0.0124)	0.0467(0.0253)	τ_1	-0.0003(0.0001)	-0.0006(0.0003)
γ_2	0.0535(0.0182)	0.0946(0.0320)	τ_2	-0.0004(0.0002)	-0.0010(0.0004)
γ_3	0.1540(0.0192)	0.0659(0.0315)	τ_2	-0.0011(0.0002)	-0.0002(0.0004)
ℓ_1	0.5783(0.0049)	0.3181(——)	ℓ_4	0.6854(0.0061)	0.2616(——)
ℓ_2	0.5045(0.0084)	0.3903(0.0126)	ℓ_5	0.1480(0.0089)	-0.0463(0.0148)
ℓ_3	0.3435(0.0093)	0.1772(0.0138)	ℓ_6	-0.8163(0.0065)	-0.4370(——)
σ^2	0.0205(0.0052)	0.4926(——)			

Table 2: ML and RM estimates of the parameters of (5.9).

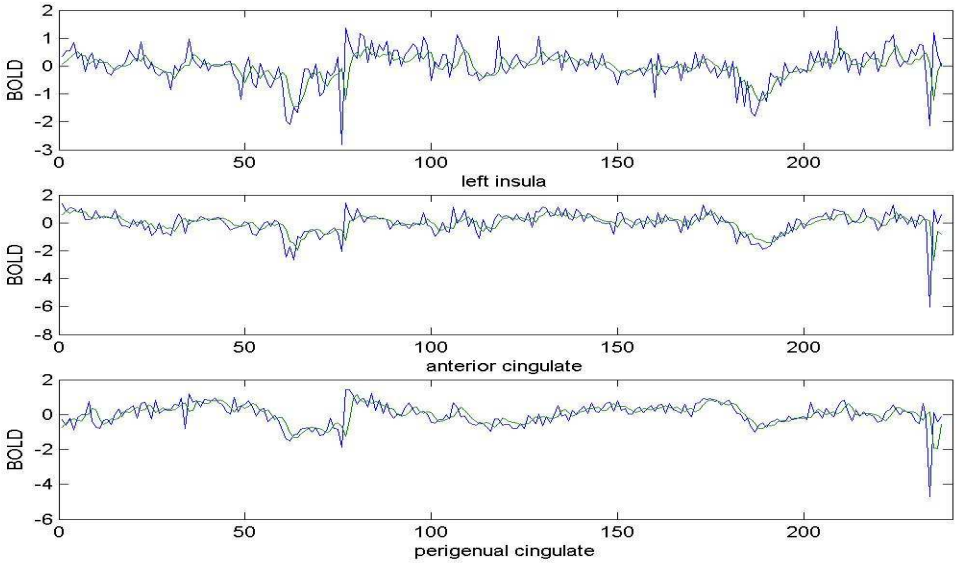

Figure 7: Hemodynamic BOLD response and predicted BOLD response of subject H20 at three different locations over time.

dependencies of the locations. When the constraint used is based on the correlation which is geometrically bounded over lags then one's influence over another cannot exceed its influence onto itself. Moreover, this influence exerted on another location is only proportional to its dependence on its past. Temporal decay of weighted coefficients happen much faster than in the classical vector autoregressive model. On the other hand, oscillations within location and between locations can happen in two ways, either due to the cross-correlation or the oscillatory movement of the influencing location itself.

Connectivity among the set of preselected regions depends on the constraint matrix used. The diagonal elements of Π_h are self connections or the dependence of one location on its past. Since the upper and lower diagonal elements are constrained to be zero then the feed-forward and the feed-backward influence of a location depends on the propagation of self connections through the constraint matrix. Therefore, if a spatially informed matrix is used then connectivity is directed if there is spatial connection between locations. In the case when the auto- and cross-correlation are used, connectivity can be assessed by the significance of

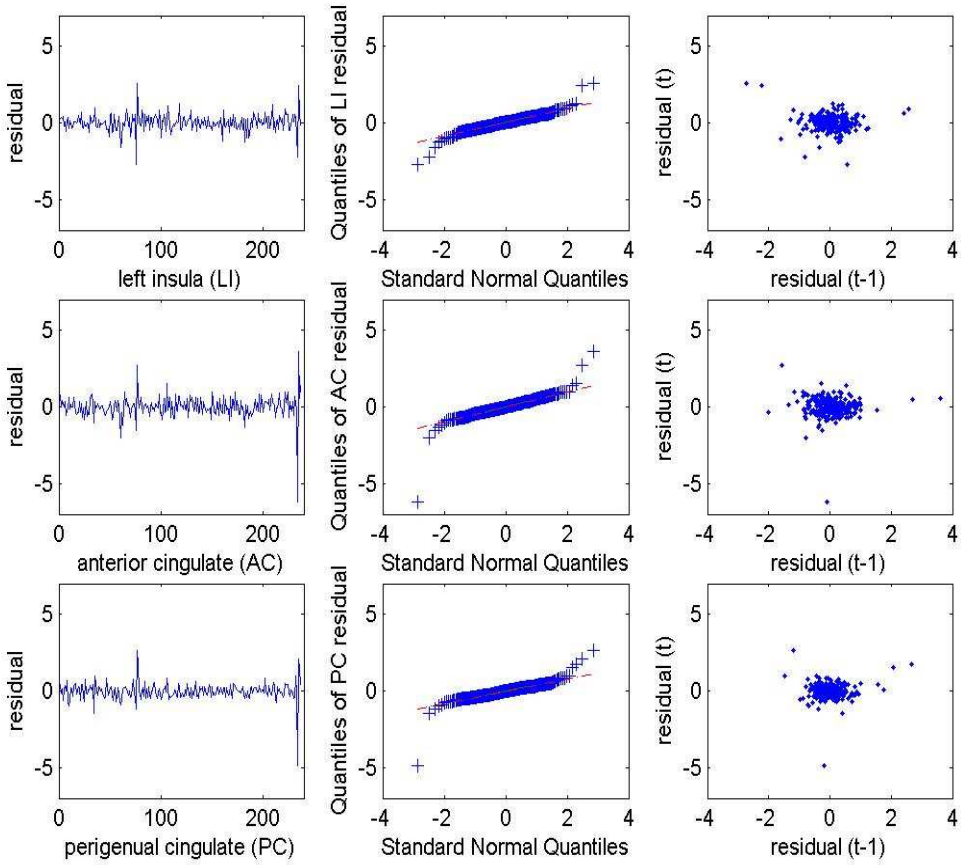

Figure 8: Residual diagnostic plots. First column: Univariate residual time course at three different locations of subject H20. Second column: qqplot or the residual series against the standard normal quantiles. Third column: residual (t) vs. lagged residual ($t-1$) plot.

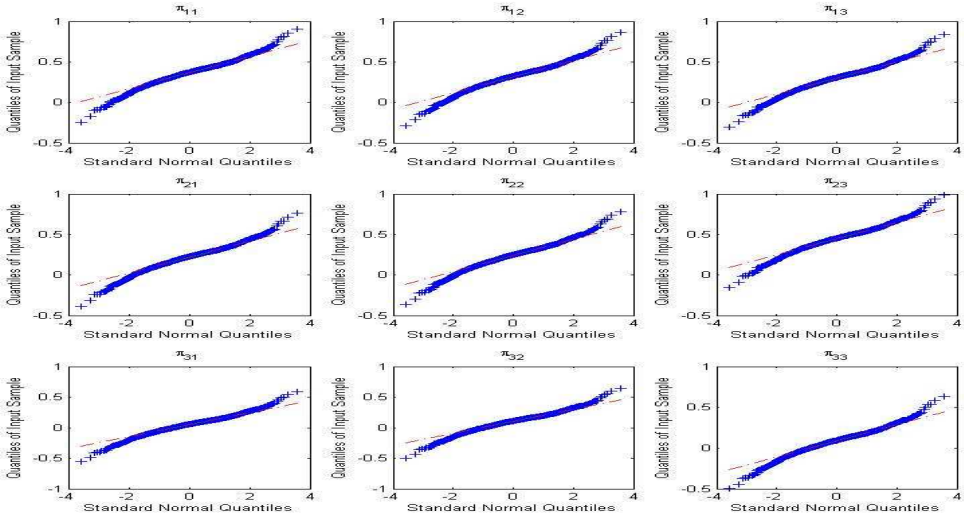

Figure 9: The QQ-plots of the autoregressive parameters.

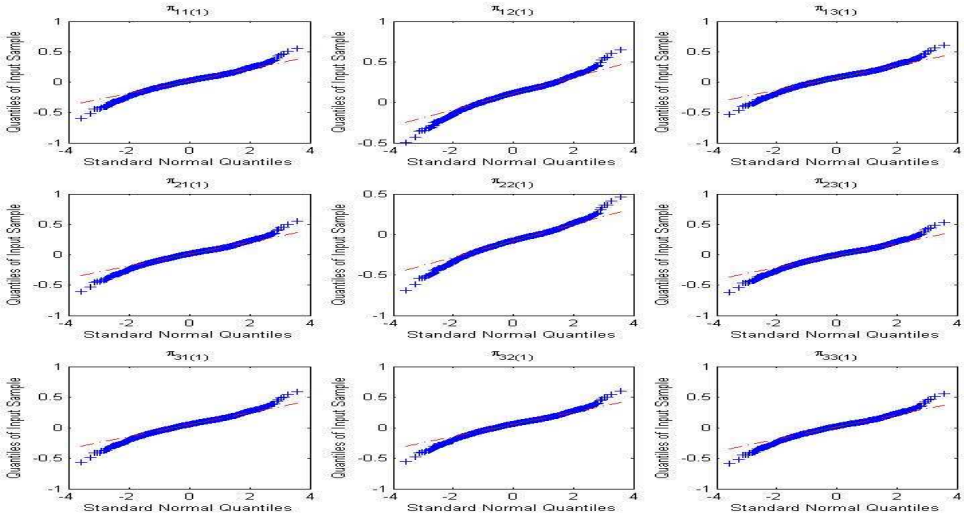

Figure 10: The QQ-plots of the group-related autoregressive parameters.

the linear combination of the coefficients across all time lags. Generally, the cross-correlation between location is still significant even if the weighted parameter is already negligible. On the other hand, feedback within location can be determined using the bootstrap distribution in figure 9. In the figure, one can infer that, on average, the strength of dependence of the left insula and the anterior cingulate on their respective past decays over time except for the perigenual cingulate where the strength of dependence reaches an optimal values at lag 2 (3 secs) and implies delay in the reaction and processing of information in this location.

Other nonlinear interactions such as modulatory interaction whereby one location affects the connection between two other locations can be modelled using bilinear terms [43], a term obtained by taking the product of the observations of the modulating and influencing location over time. Delay in the modulation is obtained by taking the product of the lagged observation of the influencing location with the current value of the modulating location. This technique is, however, limited to such types of interaction. There are many other situations where it can be rendered useless, e.g. inhibiting feedback. In this setting, suppose locations A influences B but but B reciprocates an effect to A. Working individually A and B influences location C but the controlling or inhibiting effect of B on A cancels their joint effect on C. It is obvious that this cannot be modelled through bilinear terms. This requires a more elaborate technique which will be a topic for future investigations.

Changes in the connectivity induced by group differences failed to attain desired significance levels. For High Reactors, the anterior cingulate tends (but not significant) to have a stronger dependence on itself, which is transmitted also on all the other locations, than for Low Reactors.

6.0 CONCLUDING DISCUSSION

In the preceding chapters, we showed the Constrained Mixed-VARX model in terms of its specification, estimation, asymptotic theory, and bootstrap inference and applied it to a neuroimaging data. In this chapter, we will summarize the results presented and elucidate some research problems that can be pursued in the future.

6.1 SUMMARY

Clinical data usually consists of observations or signals recorded from several subjects nested within naturally occurring groups. One important statistical task is to combine information obtained from each subject's data within these groups. The usual problem, pointed earlier in this manuscript, is that data across individuals or panels is usually not homogeneous so that imposing the homogeneity assumption can be too restrictive and is likely to be violated in practice. For example, each individual adopts varying cognitive strategies or the brain adopts degenerative solutions to perform the same controlled stimulus which may be reflected in differences in their hemodynamic BOLD response associated to that task. Heterogeneity, in this situation, might be a bane but this information can actually be helpful because it reveals how certain factors affect a cross-section specifically and, in some cases, might even reveal an irregularity. For instance, in the study of growth curves of certain quantities relating to pregnancy [64] it is important to account for individual effects to see if there are deviating quantities as this maybe a sign of a pathologic condition. These examples point out that accounting for individual heterogeneity is an important question in the analysis and aggregation of cross-sectional data.

Despite prevailing heterogeneity, it is logical to ask if it is still possible to determine a consistent model across subjects and groups? In the same way, is it possible to obtain a consistent network influences in the brain while still allowing individual variation? Numerous approaches have been used to answer this question, e.g. characterizing the network by analyzing each subject separately. However, this approach sometimes lead to inconsistent results [35, 72] and clinical and demographic variables cannot be included. Moreover, aggregation using the ordinary arithmetic average among resulting separate regression estimates can be heavily biased with just one corrupt individual estimate.

In the application described herein, we have used an aggregation scheme using techniques in random effects analysis to accommodate individual heterogeneity. The random effect specification reduces the number of parameters to be estimated substantially in comparison to individual regressions while still allowing the coefficients to differ from cross-section to cross-section. Hence, the model is equivalent to postulating a separate regression for each cross-sectional unit but that some of the free coefficients are assumed to have come from a certain hypothesized distribution. On the other hand, this stipulation also provides some method for modelling the cross-sectional units as a group and gives consistent patterns of influences while allowing one to relate clinical and demographic variables to the response. Prediction of subject specific variation is obtained via empirical Bayes based upon consistent estimates of group parameters and with the assumption that the observations are conditionally independent given their past. These random effects predictions are shrinkage estimators in the sense that their variance diminishes in the presence of data and moreover, they are the best among the class of linear unbiased predictors.

Another important devise used in this application pertains to modelling general dynamic behavior of the vector series where the components of the vector series are spatially informed or at least there is an *a priori* physical correlation other than that which is associated with temporal sequencing. This complicates correlativeness of observations since it is not only due to adjacency in time but also because of some other physical phenomenon be it known or unknown. The method used here asks for an informative constraint to the autoregressive parameter to provide an alternative approach to describing interregional dependence among multiple time series cross-sectional data. It is based on space-time ARMAX (STARMAX)

by Stoffer [93] who successfully used it in spatial time series from a geophysical standpoint. The technique, whereby off-diagonal autoregressive parameters are set to zero, essentially removes redundant off-diagonal parameter components and effectively reduces the number of parameters to be estimated. As a result, the model is more flexible in the sense that higher order lags can be explored for longer dependence or the vector observation can be augmented through bilinear terms to model some nonlinear interaction can be pursued.

The constraining approach actually follows an interesting heuristic argument. In general the autoregressive parameters in an ARMAX model are derived through the auto- and cross-covariance functions. Since these quantities can be estimated prior to estimating the model, it is helpful if one uses separate regressions among the vector components and let them correlate through the auto and cross-covariance functions or by any matrix that provides information on how these components are related. Some types of constraints were described in the course of the discussion but we focused on the autocorrelation since this is the same quantity the autoregressive estimates are derived from. In addition, the auto- and cross-correlation incorporates both temporal and spatial dependence among locations, i.e., both types of dependence are collapsed into this single quantity.

While one could use usual estimation techniques for the parameters of the constrained model, we believe upon inspection of the marginal univariate time series from each location that a robust procedure is advantageous as the data is contaminated by sporadic outliers. In this study, we showed that the redescending M-estimation is a valuable tool in estimating the parameters of multiple time series because is resistant to the effects of aberrant observations. From the simulation that was done, the method performed very well compared to the classic maximum likelihood estimation when the data is contaminated with both additive and innovational outliers. Additive outliers are generally more unwieldy since it inflates a sequence of innovations, and more alarmingly, may create a pattern into the residuals. Moreover, the method exceeded expectations that it would breakdown easily because of the inherent cross-sectional structure of the data albeit more investigations are needed to determine the amount of contamination this method will eventually fail.

From a theoretical standpoint, we have shown that the general estimation procedure, assuming stationarity, ergodicity and equivariance, can actually be obtained through M-

functionals. The M-functional is a function whose expectation is zero at the solution. In this problem, the M-functional has a compound form due to an alteration on the distribution depending on which condition the observation happened. Depending on the type of bounding function, this functional may admit multiple solutions, yet a desirable solution can actually be determined and in fact this solution is consistent as long as the variance does not become degenerate. Moreover, assuming consistency and uniqueness of the functional at a certain neighborhood, the infinitesimal contribution of a single observation on to the estimates through the M-functional can be obtained by differentiating the statistical functional with respect to the contamination. It was shown that this functional, both true and empirical, are actually bounded when a redescending bounding function is used both for the fixed and variance parameters. Thus, we expect that the effect of a single observation on the estimates remains bounded. The same influence functional can be used to obtain the asymptotic distribution of the estimates.

Inference on the parameter estimates were obtained through bootstrap approximation which gives their limiting distribution conditional on the data. Many papers on robust estimation use this method instead of the limiting distribution because it gives a better approximation of the true distribution in the presence of contamination. In this approach, we adopted the conditional *iid* resampling scheme for innovations and generated data by reconstructing the series through these sampled innovations. Innovations were grouped by subject instead of pooling them altogether and the M-estimate was obtained from each bootstrap sample. We showed that this scheme actually follows the bootstrap principle that the empirical distribution of the innovations and the random effects are good approximations of their corresponding true distributions using the Mallow's metric. In this way any bootstrap sample does not cause a bias that does not vanish even in large samples. We also showed that the limiting distribution of some functions of the estimating equations actually converge to the derived limiting distribution asymptotically. Unfortunately, this technique is generally slow. In large samples, alternative methods can be used through bootstrapping the fixed point representation of the influence functional, i.e., reconstructing the distribution not by naive estimation but from bootstrap approximations of the"estimating equations".

Lastly, we also proposed an alternative model selection scheme since the bootstrap dis-

tribution requires that the correct specification of the model is used so that the distribution is not biased. This proposition still needs to be improved and will be discussed later in this chapter.

6.2 FUTURE WORK

6.2.1 Aggregative Methods

In most imaging studies, a typical estimate of group activity is obtained by averaging individual subject estimates. This could have been an easy and effective approach if all the estimates are "nice", i.e. it follows desired distributional assumptions, but since this rarely happens in practical clinical applications, the resulting estimate can be be greatly distorted with just one corrupt estimate. One can overcome this problem by using a more robust estimator but one can also use a data-adaptive strategy which give, as a group estimate, a linear combination of the individual estimates, where the coefficients are chosen adaptively from the data. This approach follows an easy procedure: (1) initial subject-specific estimates are computed, e.g. data fusion models with spatiotemporal structure such as Dale [30]; then (2) group estimate is computed as a weighted average of subject estimates with data dependent weights. Bunea, Ombao and Auguste [21] argued that using the weighted average of the curves, rather than the arithmetic average, reduces the bias introduced by denoising (or smoothing) in the subject estimates and at the same time avoids the potential problem of distortion that can be caused by some corrupted individual estimates. In fact, they showed that the data-adaptive estimate of a group mean function is superior (in minimax error sense) to the arithmetic average.

We can reconstruct the problem we presented in the previous chapter in the following general context while avoiding robust estimation. Given N data sets $\{\mathbf{Y}_i\}_{i=1}^N$ such that each subject $\mathbf{Y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$, a matrix of size $p \times T$, and can be either an observed signal or brain activity at each voxel in a brain surface or volume over time. Suppose that $\mathbf{Y}_i$ follows the general model $\mathbf{Y}_i = \mathbf{f}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{f}_i = \mathbf{g} + \mathbf{h}_i$ denotes the subject specific mean, $\mathbf{g}$ the

group specific fixed effect, $\mathbf{h}_i$ the random effect which is associated with the deviation of the individual from the group-specific function, and $\boldsymbol{\varepsilon}_i$ is the error which can be correlated within subject. Then, the first step of the data-adaptive procedure gives N smoothing estimates $\{\hat{\mathbf{f}}_i\}_{i=1}^N$. The second step, accomplished by finding weights in an adaptive way, uses the following solution: use data set from the N-th subject and fit the linear model with response $\mathbf{Y}_N$ and covariates $\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_{N-1}$. This can be repeated N times using the leave-one-out strategy so that the aggregate is not dependent on the particular choice of data. Each partial aggregate is defined as $\tilde{\mathbf{f}}^i = \sum_{\ell \neq i} \hat{w}_\ell \hat{\mathbf{f}}_\ell$, where $\hat{w}_\ell$ is the minimizer over w of the complexity penalized least squares

$$\frac{1}{T}\sum_{j=1}^{T}\left(\mathbf{y}_{ij} - \sum_{\ell \neq i} w_\ell \hat{\mathbf{f}}_\ell\right)^\top \left(\mathbf{y}_{ij} - \sum_{\ell \neq i} w_\ell \hat{\mathbf{f}}_\ell\right) + \mathrm{pen}(w). \tag{6.1}$$

Then $\hat{g} = \frac{1}{N}\sum_{i=1}^N \hat{\mathbf{f}}_i$ and the subject specific deviation will be predicted by subtracting the aggregated group estimate from the individual estimates. Variance of the parameters can be estimated using standard mixed effects procedures. The penalty term in (6.1) has some theoretical motivations, e.g. LASSO-type, which results in an aggregate having the highest level of accuracy among other possible aggregates [22] in the fixed effects model.

Then we have the following goals: (1) compare the results of the current approach with this data adaptive algorithm; (2) explore different penalty functions such as LARS and investigate their associated accuracy particularly in the mixed case; and (3) incorporate constraints on coefficients and develop a way of using the same aggregating scheme on signals coming from different modalities.

6.2.2 Space-time Modelling

The emergence of a wide range of physical processes that involves variability over space and time generated a tremendous growth in developing statistical models and techniques to analyze spatio-temporal data. Examples of spatio-temporal processes include applications in modelling space-time patterns of disease and disease risk [60], population dynamics in ecology [32]; monitoring office unit price distributions [97], characterizing changes in space-time pattern of brain signals [100], among others. In all of these applications, the models

attempt to provide a probabilistic framework for the analysis and prediction that is built on the joint spatial and temporal dependence of the observations.

For the spatio-temporal data adopted in this study, the model is created through a spatial augmentation of a model initially developed for temporal distributions. The special characteristic of this model is that it does not try to create a continuous space over which the time series are observed. In some applications, such as electroencephalogram (EEG) recordings or extracted blood oxygenated level dependent (BOLD) responses from several brain regions, interpolating information or function at an unknown location does not make any sense. However in more general applications, e.g. determination of space-time trends in the deposition of pollutants [87, 73], monitoring ozone concentrations [39], characterization of space-time variability of temperatures [41] etc., where the goal is to make predictions at some unknown location given a network of observation stations, the model we presented may not be helpful at all. To date, analytic tools that support both temporal and spatial analysis of data over continuous space and discrete time are still sorely lacking. The following is just an attempt to model space-time data in the cross-sectional case.

Consider the model for one subject given by (2.1). Using the state-space representation, this model can be written in the following *observation* equation and *state* equation

$$\mathbf{z}_{i,j+1} = \begin{bmatrix} \mathrm{D}_1\Pi_1 & \dots & \mathrm{D}_h\Pi_h \\ I_{(h-1)p\times(h-1)p} & & \mathbf{0}_{(h-1)p\times p} \end{bmatrix} \mathbf{z}_{ij} + \begin{bmatrix} \mathrm{D}\Pi_1 \\ \vdots \\ \mathrm{D}_h\Pi_h \end{bmatrix} \boldsymbol{\varepsilon}_{ij}, \tag{6.2}$$

$$\mathbf{y}_{ij} = [\underbrace{I,\dots,I}_{p\times(h-1)p},\mathbf{0}]\mathbf{z}_{ij} + \sum_{k=0}^{m}\Gamma_{ik}x_{j-k} + \Upsilon\mathbf{u}_i + \boldsymbol{\varepsilon}_{ij}, \tag{6.3}$$

which can be written in the general form

$$\mathbf{z}_{i,j+1} = \mathbf{P}\mathbf{z}_{ij} + \mathbf{K}\boldsymbol{\varepsilon}_{ij} \tag{6.4}$$

$$\mathbf{y}_{ij} = \mathbf{H}^\top\mathbf{z}_{ij} + \Gamma_i\mathbf{x}_j + \Upsilon\mathbf{u}_i + \boldsymbol{\varepsilon}_{ij} \tag{6.5}$$

where $\mathbf{H}$ and Υ are fixed parameter vectors, Γ_i is the random effect parameter, $\mathbf{P}$ is the transition matrix, $\mathbf{K}$ is the innovation coefficient matrix, $\mathbf{z}_{ij}$ is the state vector, $\mathbf{y}_{ij}$ is the observation vector, and $\boldsymbol{\varepsilon}$ is the observation error. This general framework can be augmented

by featuring some dominant components of variation in (6.5) as seen in the models seen in Wikle and Cressie [98] and Mardia *et al.* [65] which we will use and modify to include individual subject variation. In particular, consider the following equations:

$$\begin{aligned}
\mathbf{y}_{ij}(\mathbf{s}) &= \mathbf{z}_{ij}(\mathbf{s}) + \boldsymbol{\varepsilon}_{ij}(\mathbf{s}) && (6.6)\\
\mathbf{z}_{ij}(\mathbf{s}) &= \mu_j(\mathbf{s}) + \zeta_i(\mathbf{s}) + \Upsilon\mathbf{u}_i + \nu_{ij}(\mathbf{s}) && (6.7)\\
\mu_j(\mathbf{s}) &= \int \omega_s(\mathbf{q})\mu_{j-1}(\mathbf{q})d\mathbf{q} + \eta_j(\mathbf{s}). && (6.8)
\end{aligned}$$

Note that (6.6)-(6.7) corresponds to the observation equation (6.5), while (6.8) corresponds to the state equation (6.4). In the above model, we suppose that the observational process has a component of measurement error and $\mathbf{z}_{ij}(\mathbf{s})$ can be thought of as an unobservable process 'smoother' than $\mathbf{y}_{ij}(\mathbf{s})$. While in (6.6), the error term $\boldsymbol{\varepsilon}_{ij}(\mathbf{s})$ is white noise an represents observation error, the error term in (6.7), $\nu_{ij}(\mathbf{s})$, represents a spatial structure that is independent of time and cross-section and does not have temporally dynamic behavior. By contrast, the component $\mathbf{z}_{ij}(\mathbf{s})$ is assumed to evolve according to the state equation in (6.8) where $\eta_j(\mathbf{s})$ is spatially colored process. Furthermore, the former can be decomposed into dominant components from a set of deterministic basis functions that are complete and normal. For more details on this topic one can refer to the discussion in Wikle and Cressie [98], Mardia *et al.* [65], and Sanso and Guenni [91].

The estimation of the model given by equations (6.6)-(6.8) can proceed by specifying the joint posterior distribution of the parameters which can be obtained as the product of the loglikelihood distribution of the hierarchical model and prior distribution of the parameters. In the future, we will investigate the identifiability, estimability, model selection and covariance estimation, implementation issues and model diagnostics.

6.2.3 Robust Estimation

In a remark made earlier in Chapter 5, we often encounter situations when there seems to be no solution to the minimization problem (2.18) because the variance parameters Σ_ε and Σ_Γ become degenerate, i.e., $\log|\Omega_j| \to -\infty$. This leads us to search other robust procedures that try to bound the variance parameter away from degeneracy.

In the mixed model literature, most of the robust procedures focus on univariate responses (see for example [78, 101, 27, 28] among others). Recently, high breakdown methods such as constrained S [27] and MM-estimators [28] were introduced. These types of estimators are known to have good global properties as indicated by their high tolerance for the proportion of corrupt observations. Unfortunately, these methods have poor local properties and low efficiency, even for the fixed regression case. Although global properties are desirable, in practice, we do not expect conspicuous size of outliers in the data (see for example Rocke and Woodruff [79]). Rather, we usually expect a small fraction of outliers whose effect on an estimator can be tremendous. Therefore, it is important to consider estimation procedures that have good local properties.

In the robust literature, a general measure of local stability is the *contamination sensitivity of order* q [102] given by

$$\gamma_{\mathbf{T}}^{q} = \limsup_{\epsilon \downarrow 0} \frac{b_{\mathbf{T}}(\epsilon)}{\epsilon^{q}} \tag{6.9}$$

where $0 \leq q$ and $b_{\mathbf{T}}(\epsilon)$ is the maximum bias curve over the neighborhood of ϵ contaminated distribution μ^n given by

$$b_{\mathbf{T}}(\epsilon) = \sup_{\mu^n \in N_\epsilon(\mu^{n,\epsilon})} (\mathbf{T}(\mu^n) - \boldsymbol{\pi})^\top \mathbf{C}_0(H_0)(\mathbf{T}(\mu^n) - \boldsymbol{\pi}) \tag{6.10}$$

with $\mathbf{C}_0(H_0)$ a suitable positive definite scatter matrix functional satisfying some equivariance property (see [102] for details of this measurement). Then an estimator $\mathbf{T}$ is said to be *locally stable of order* q if $\gamma_{\mathbf{T}}^{q} < \infty$. He [44] proved that the above mentioned estimators, i.e. S- and MM-estimators, along with the LMS-estimator, $\frac{1}{2}$LTS-estimator and the τ-estimators have $\gamma_{\mathbf{T}}^{q} = \infty$ for $q > 0.5$ and therefore these estimators are not locally robust.

In the future, we will investigate robust procedures geared toward estimation of multivariate mixed models subject to the bound on the local sensitivity of order q while having the maximum achievable breakdown point possible. Examples of such estimation procedure are the *Constrained M-estimation* [52] which follows the same objective function as (2.18) with an added constraint that $\sum_{i=1}^{N} \sum_{j=1}^{T-n} \rho(\mathbf{r}_{ij}^\top \Omega_j^{-1} \mathbf{r}_{ij}) \leq b$ and the *GM-estimator* [102]. These procedures have known good local and global robustness properties in the multivariate location and scatter case [102]. It would be worthwhile to investigate these methods in our

setting while providing modified testing procedures. e.g. Wald test and the F-test. We will also compare their results with some established ones.

6.2.4 Bootstrap Approximation

In this study we explored naive bootstrap methods to approximate the limiting distribution of the the parameters by estimating the parameter of the model repeatedly on several bootstrap samples of reconstructed time series based on the conditional independence of the innovations. We said that the procedure is slow and convergence of the estimate of the variance components is often problematic. Furthermore, even if the estimation method is robust, the proportion of outlying innovations through the resampling scheme can be high enough to breakdown the the estimator for the particular resample. To overcome the first concern, we can bootstrap the fixed point representation of the parameter estimate (see Salibian-Barrera and Zamar [88] for the MM-estmators of univariate regression and Kreiss and Franke [54] for univariate ARMA models) instead, i.e., using the linear approximation of limiting value of the estimate, say $\mathbf{T} = (\boldsymbol{\pi}, \Sigma)$ given by

$$\mathbf{T}_{NT} = \mathbf{G}_{NT}(\mathbf{T}) + \dot{\mathbf{G}}_{NT}(\mathbf{T})(\mathbf{T}_{NT} - \mathbf{T}) + \mathbf{R} \tag{6.11}$$

where $\mathbf{R}$ is a remainder term. When this remainder term in small, (6.12) can be written as

$$\sqrt{NT}(\mathbf{T}_{NT} - \mathbf{T}) \approx [I - \dot{\mathbf{G}}_{NT}(\mathbf{T})]^{-1}\sqrt{NT}(\mathbf{G}_{NT}(\mathbf{T}) - \mathbf{T}). \tag{6.12}$$

Then the bootstrap equivalents of both sides is obtained by estimating $[I - \dot{\mathbf{G}}_{NT}(\mathbf{T})]^{-1}$ by $[I - \dot{\mathbf{G}}_{NT}(\mathbf{T}_{NT})]^{-1}$ and by

$$\sqrt{NT}(\mathbf{T}^*_{NT} - \mathbf{T}_{NT}) \approx [I - \dot{\mathbf{G}}_{NT}(\mathbf{T}_{NT})]^{-1}\sqrt{NT}(\mathbf{G}^*_{NT}(\mathbf{T}_{NT}) - \mathbf{T}_{NT}). \tag{6.13}$$

Therefore, instead of calculating the left hand side of (6.14), we can calculate its right hand side, i.e., we actually approximate the estimator for each sample by computing the function $\mathbf{G}^*_{NT}$ in $\mathbf{T}_{NT}$.

On a different note, instead of reconstructing the time series by approximating independence of innovations we can use *moving block bootstrap* (MBB) [56, 62]which resamples *blocks*

of consecutive observations at a time. As a result, the dependence structure of the original observations is preserved within each block while reducing computational time in reconstructing observations. Moreover, this technique is helpful when the experimenter does not have enough prior information to specify appropriate models or any parametric assumption.

In the future, we will investigate the properties of above technique and establish its use in time-series cross-sectional data.

6.2.5 Model Selection

One issue that we encountered in model selection is the tendency of the procedure to choose higher model order even if there is "not enough" significant information left in the innovations to warrant the use of extra terms. From experience, the value of either (4.5) or (4.6) regardless of the penalty function is dependent on the resulting estimator, the order used, and the value of the constant k in the bounding function ρ. When higher orders are used, say n, the value of the Ω_j becomes very small so that even if the effect of a single outlier reverberates up to the next n consecutive observations and the bounding function assigns a maximum weight on the observation the upset is still negligible compared to the value of the former. Hence, the selection criteria always chooses the model with a higher order.

One procedure which could probably overcome this is to use cross-validation methods [83]. The basic idea of the selection procedure is to split the sample of size NT into a construction sample of size NT_c and a validation sample of size $N(T-T_c)$ so that we use the construction sample to fit the model and use the validation sample to evaluate the prediction error of the model. The splitting of observations can be done using the MBB method described above to obtain blocks of observations which are approximately independent and preserves the structure of the time series. Then the final model is chosen as the one which gives the smallest average of prediction error over different validation samples.

APPENDIX A

Definition 3 (**Regression Equivariance**)**.** Let $\mathbf{v}$ be any $p \times 1$ vector. The estimator $\hat{\Theta}$ is regression equivariant if

$$\hat{\boldsymbol{\pi}}(\mathbf{z}, \mathbf{y} + \mathbf{z}\mathbf{v}) = \mathbf{T}(\mathbf{z}, \mathbf{y} + \mathbf{z}\mathbf{v}) = \mathbf{T}(\mathbf{z}, \mathbf{y}) + \mathbf{v} = \hat{\boldsymbol{\pi}} + \mathbf{u}. \tag{A.1}$$

So regression equivariance means that we can assume without loss of generality that $\boldsymbol{\pi} = \mathbf{0}$.

Definition 4 (**Affine Equivariance**)**.** Let $\mathbf{A}$ be any $s \times s$ nonsingular matrix. Then $\hat{\boldsymbol{\pi}}$ is affine equivariant if

$$\hat{\boldsymbol{\pi}}(\mathbf{z}\mathbf{A}, \mathbf{y}) = \mathbf{T}(\mathbf{z}\mathbf{A}, \mathbf{y}) = \mathbf{A}^{-1}\mathbf{T}(\mathbf{z}, \mathbf{y}) = \mathbf{A}^{-1}\hat{\boldsymbol{\pi}}. \tag{A.2}$$

Theorem 6 (**Implicit Function Theorem, [Lopuhaa, 1989]**)**.** Let $\mathscr{Q}$ be a metric space, $(h_0, \mathbf{t}_0) \in \boldsymbol{\Omega} \subset \mathscr{Q} \times \mathbb{R}^{s+\frac{1}{2}p(p+1)}$, $\boldsymbol{\Omega}$ open. When $\mathbf{W} : \mathscr{Q} \times \mathbb{R}^{s+\frac{1}{2}p(p+1)} \to \mathbb{R}^{s+\frac{1}{2}p(p+1)}$, with $\mathbf{W}(h_0, \mathbf{t}_0) = \mathbf{0}$ is such that

1. $\mathbf{W}$ is continuous on $\boldsymbol{\Omega}$,
2. $\partial\mathbf{W}/\partial\mathbf{t}$ is continuous on $\boldsymbol{\Omega}$
3. $\partial\mathbf{W}/\partial\mathbf{t}$ is nonsingular at $(h_0, \mathbf{t}_0)$,

then there exists a neighborhood $B_1 \times B_2$ of $(h_0, \mathbf{t}_0)$ on which a function $\mathbf{t}(\cdot) : B_1 \to B_2$ exists such that $\mathbf{W}(h_0, \mathbf{t}_0(h)) = \mathbf{0}$. Moreover it holds that:

1. If $(\tilde{h}, \tilde{\mathbf{t}}) \in B_1 \times B_2$ with $\mathbf{W}(\tilde{h}, \tilde{\mathbf{t}}) = \mathbf{0}$, then $\tilde{\mathbf{t}} = \mathbf{t}(\tilde{h})$.
2. $\mathbf{t}(\cdot)$ is continuous on B_1

APPENDIX B

Matlab codes for the robust parameter estimation and bootstrap inference of the constrained Mixed-VARX Model.

```
function [XP, XS, logL, hessian1, hessian2]=rolex(XP, XS, data,
cov, group, stim, W, n)

%iterative algorithm to compute for the parameters of the no-intercept
%model

%OUTPUT
%XP gives the autoregressive and fixed parameters
%XS gives the scale parameters
%logL is the loglikelihood
%hessian1 and hessian2 are the hessian matrices associated to XP and XS, resp.

%INPUT
%XP and XS are the initial values corresponding to the autoregressive and
%variance parameters, respectively.
%data is the data matrix
%cov is the non-time varying covariate, e.g. age.
%group is the vector of 1s and 0s corresponding to group membership of
%subjects
%W is the constraint matrix
%stim is the time-varying covariate, e.g. stimulus
%n is the autoregressive order

options1=optimset('Display', 'final', 'GradObj', 'on',
'LargeScale', 'on'); tol=1e-3; iter=0; iterlim=30;

iter=0;
while(iter==0 | norm(XP-xp0)>tol | norm(XS-xs0)>tol)
    iter=iter+1;
    if (iter>iterlim), warning('Iteration limit reached.'); break; end;
    xs0=XS;
    xp0=XP;
    [XP,logL, exitflag1, output1, grad1, hessian1]=fminunc(@(XP)
        Phi(XS, XP, Kz((j-1)*s+1:j*s,:), Y((j-1)*p+1:j*p,:),
        stim((j-1)*m+1:j*m,:), W, n), xp0, options1);
```

```
    [XS,logL, exitflag2, output2, grad2, hessian2]=fminunc(@(XS)
        LSS(XS, XP, Kz((j-1)*s+1:j*s,:), Y((j-1)*p+1:j*p,:),
        stim((j-1)*m+1:j*m,:), W, n), xs0, options1);
end;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [logL, glogLp] = Phi(XP, XS, data, cov, group, stim, W, n)

%function for solving unknown autoregressive parameters
%OUTPUT: loglikelihood and its corresponding gradient

%some constants
p=3; m=1; s=2*n*p+m+1; sd=2*n+m+1; [T,Np]=size(data); N=Np/p;
pm=p*m; ps=p*s; psd=p*sd; r1=0.5*p*(p+1); r2=0.5*pm*(pm+1);
r3=ps+r1+r2; r3d=psd+r1+r2; C=(N*(T-n))^(-1);

%to reshape vectorized parameters
ematS=mat(p);
ematSgma=mat(pm);

%partitioning initial parameters and reshaping vecp, S and Sgma
vecp=XP;
Psi_hat=zeros(p,s);
for k=1:n;
    Psi_hat(:,(k-1)*p+1:k*p)=W(:,(k-1)*p+1:k*p)*diag(vecp((k-1)*p+1:k*p));
    Psi_hat(:,n*p+(k-1)*p+1:n*p+k*p)=W(:,(k-1)*p+1:k*p)
           *diag(vecp(n*p+(k-1)*p+1:n*p+k*p));
end;
Psi_hat(:,2*n*p+1:s-1)=vecp(2*n*p+1:p*2*n+m*p);
Psi_hat(:,s)=vecp(psd-2:psd);
Phi=Psi_hat;
vecp=Phi(:);

%forming the variance matrices
vechS=XS(1:r1, 1);
LSgma=XS(r1+1, 1);
lS=ematS'*vechS;
LS=reshape(lS,p,p);
S=LS*LS';
Sgma=LSgma*eye(pm);

%response and design matrix
Y=response(data,n);
Kz=design(data, cov, stim, group, n, m);

Wstim=zeros(m,T-n);
for j=1:T-n
    for k=1:m
        Wstim(k,j)=stim(j+n-k,1)';
    end;
end;

L1=0;
for j=1:T-n
    A2=kron(Wstim(:,j)', eye(p));
    A3=A2*Sgma*A2'+S;
    A4=log(det(A3));
    L1=L1+A4;
```

```
end;
logL1=N*L1;

D=zeros(N*(T-n),1);
for i=1:N
    for j=1:T-n
        A1=kron(Kz(:,(i-1)*(T-n)+j), eye(p));
        A2=kron(Wstim(:,j)', eye(p));
        A3=A2*Sgma*A2'+S;
        iOmegaj=inv(A3);
        R=Y(:,(i-1)*(T-n)+j)-A1'*vecp;
        D((i-1)*(T-n)+j,1)=R'*iOmegaj*R;
    end;
end;
c=max(4.835,(1/0.6748)*median(D));

L2=zeros(1,N);
for i=1:N
    l2=0;
    for j=1:T-n
        A1=kron(Kz(:,(i-1)*(T-n)+j), eye(p));
        A2=kron(Wstim(:,j)', eye(p));
        A3=A2*Sgma*A2'+S;
        iOmegaj=inv(A3);
        R=Y(:,(i-1)*(T-n)+j)-A1'*vecp;
        Dij=R'*iOmegaj*R;
        h1=tukeyc(Dij,c,0);
        l2=l2+h1;
    end;
    L2(:,i)=l2;
end;
logL2=sum(L2,2);

logL=logL1+logL2;

if nargout > 1
    H1=derivLp(data, cov, stim, group, W, n, c, vecp, LS, LSgma);
    glogLp=H1;
end;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [logL, glogLS] = LSS(XS, XP, data, cov, group, stim, W, n)

%function for solving unknown scale parameters
%OUTPUT: loglikelihood and its corresponding gradient

%some constants
p=3; m=1; s=2*n*p+m+1; sd=2*n+m+1; [T,Np]=size(data); N=Np/p;
pm=p*m; ps=p*s; psd=p*sd; r1=0.5*p*(p+1); r2=0.5*pm*(pm+1);
r3=ps+r1+r2; r3d=psd+r1+r2; C=(N*(T-n))^(-1);

%to reshape vectorized parameters
ematS=mat(p);
ematSgma=mat(pm);

%partitioning initial parameters and reshaping vecp, S and Sgma
vecp=XP;
```

```
Psi_hat=zeros(p,s);
for k=1:n;
    Psi_hat(:,(k-1)*p+1:k*p)=W(:,(k-1)*p+1:k*p)*diag(vecp((k-1)*p+1:k*p));
    Psi_hat(:,n*p+(k-1)*p+1:n*p+k*p)=W(:,(k-1)*p+1:k*p)
            *diag(vecp(n*p+(k-1)*p+1:n*p+k*p));
end;
Psi_hat(:,2*n*p+1:s-1)=vecp(2*n*p+1:p*2*n+m*p);
Psi_hat(:,s)=vecp(psd-2:psd);
Phi=Psi_hat;
vecp=Phi(:);

%forming the variance matrices
vechS=XS(1:r1, 1);
LSgma=XS(r1+1, 1);
lS=ematS'*vechS;
LS=reshape(lS,p,p);
S=LS*LS';
Sgma=LSgma*eye(pm);

%response and design matrix
Y=response(data,n);
Kz=design(data, cov, stim, group, n, m);

Wstim=zeros(m,T-n);
for j=1:T-n
    for k=1:m
        Wstim(k,j)=stim(j+n-k,1)';
    end;
end;

L1=0;
for j=1:T-n
    A2=kron(Wstim(:,j)', eye(p));
    A3=A2*Sgma*A2'+S;
    A4=log(det(A3));
    L1=L1+A4;
end;
logL1=N*L1;

D=zeros(N*(T-n),1);
for i=1:N
    for j=1:T-n
        A1=kron(Kz(:,(i-1)*(T-n)+j), eye(p));
        A2=kron(Wstim(:,j)', eye(p));
        A3=A2*Sgma*A2'+S;
        iOmegaj=inv(A3);
        R=Y(:,(i-1)*(T-n)+j)-A1'*vecp;
        D((i-1)*(T-n)+j,1)=R'*iOmegaj*R;
    end;
end;
c=max(7,(1/0.6748)*median(D));

L2=zeros(1,N);
for i=1:N
    l2=0;
    for j=1:T-n
        A1=kron(Kz(:,(i-1)*(T-n)+j), eye(p));
        A2=kron(Wstim(:,j)', eye(p));
```

```
        A3=A2*Sgma*A2'+S;
        iOmegaj=inv(A3);
        R=Y(:,(i-1)*(T-n)+j)-A1'*vecp;
        Dij=R'*iOmegaj*R;
        h1=tukeyc(Dij,c,0);
        l2=l2+h1;
    end;
    L2(:,i)=l2;
end;
logL2=sum(L2,2);

logL=logL1+logL2;

if nargout > 1
    [H2,H3]=derivLs(data, cov, stim, group, W, n, c, vecp, LS, LSgma);
    glogLS=[H2; H3];
end;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function Y = response(data, n)

%gives the appropriate response vector

[T,Np]=size(data);
p=3;
N=Np/p;

%removing the first n observations for the response vector
data_trunc=data(n+1:T,:);

data_cat=zeros(N*(T-n),p);
for i=1:N;
    data_cat((i-1)*(T-n)+1:i*(T-n),1:p)=data_trunc(:,(i-1)*p+1:i*p);
    %[y_11,...,y_1T,..., y_N1,...,y_NT]'
end;

Y=data_cat';
%ycat=Y(:); %concatenates data_cat'

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function Kz = design(data, cov, stim, group, n, m)

%design matrix for the no-intercept model

[T,Np]=size(data);
p=3;
N=Np/p;

%creating the design matrix
%Kz=[Z1,...,ZN]
Kz=zeros(2*n*p+m+1,N*(T-n)); %first row is mean of the process

for i=1:N
    for j=1:T-n
        for k=1:n
        Kz((k-1)*p+1:k*p,(i-1)*(T-n)+j)=data(j+n-k,(i-1)*p+1:i*p )';
        Kz(p*n+(k-1)*p+1:p*n+k*p,(i-1)*(T-n)+j)=group(1,i)
                *data(j+n-k,(i-1)*p+1:i*p )';
```

```
        end;
    end;
end;
for i=1:N
    for j=1:T-n
        for k=1:m
            Kz(2*n*p+k, (i-1)*(T-n)+j)=stim(j+n+1-k,1);
        end;
    end;
end;

for i=1:N
    Kz(2*n*p+m+1,(i-1)*(T-n)+1:i*(T-n))=cov(n+1:T,i)';
    %adding the covariates(age)
end;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function W=crosscorr(D, n)

%gives the constraint matrix based on auto- and cross-correlations
%D is data matrix

[T,Np]=size(D);
p=3;
N=Np/p;
mean_data=zeros(1,Np);
for i=1:Np
    mean_data(i)=mean(D(:,i));
end;

num=zeros(p,n*p);
den=zeros(p,n*p);
W=zeros(p,n*p);
for k=1:n
    WS=zeros(p,p);
    for i=1:N
        R=(D(1:T-k,(i-1)*p+1:i*p)-ones(T-k,1)*mean_data(1,(i-1)*p+1:i*p));
        S=(D(1+k:T,(i-1)*p+1:i*p)-ones(T-k,1)*mean_data(1,(i-1)*p+1:i*p));
        RSC=zeros(T-k,1);
        SSC=zeros(T-k,1);
        for j=1:T-k
            rr=sqrt(R(j,1:p)*R(j,1:p)');
            RSC(j)=tukeyc(rr,4.18,1)/rr;
            ss=S(j,1:p)*S(j,1:p)';
            SSC(j)=tukeyc(ss,4,1)/ss;
        end;
        RR=diag(RSC)*R;
        SS=diag(SSC)*S;
        Q=diag(SS'*SS)*diag(SS'*SS)';
        den(:,(k-1)*p+1:k*p)=Q.^(-0.5);
        num(:,(k-1)*p+1:k*p)=RR'*SS;
        WS=WS+den(:,(k-1)*p+1:k*p).* num(:,(k-1)*p+1:k*p);
    end;
    W(:,(k-1)*p+1:k*p)=N^(-1)*WS;
end

for k=1:n
```

```
    W(:,(k-1)*p+1:k*p)=W(:,(k-1)*p+1:k*p)-diag(diag(W(:,(k-1)*p+1:k*p)))+eye(p);
end;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function v = tukeyc(d,c,diff)

% TUKEY The Tukey rho-function.
% r is a vector quantity which is usually the mahalanobis distances
% of the residuals.
% c is the tuning parameter
% diff is the order of differentiation

sid = find(abs(d)<=c);
lid = find(abs(d)>c);

v = zeros(size(d));
switch diff
case 0
  v(sid) = (c^2)/6*(1 - (1 - (1/c)*d(sid)).^3);
  v(lid) = c^2/6*ones(size(lid));
case 1
  v(sid) = sqrt(d(sid)).*(1 - (1/c)*d(sid)).^2;
case 2
  d2 = (1/c)*d(sid);
  v(sid) = (-2/c)*(1 - d2);
otherwise
  error('Illegal order of differentiation')
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function H1=derivLp(data, cov, stim, group, W, n, c, vecp, LS, LSgma)

%gives the derivatives with respect to Phi in the no-intercept model
%vecp is the vectorized matrix parameter
%LS is the Cholesky factor of S
%LSgma is the Cholesky factor of Sgma
%c is the tuning parameter

m=1; Y=response(data,n);
Kz=design(data,cov, stim, group, n, m);

p=3; s=2*n*p+m+1; [T,Np]=size(data); N=Np/p; pm=p*m; ps=p*s;
r1=0.5*p*(p+1); r2=0.5*pm*(pm+1); r3=ps+r1+r2;

%positions of variance components to be used in derivatives
lS=tril(ones(p));
[iS,jS,vS]=find(lS);
S=LS*LS';
Sgma=LSgma*eye(pm);

%stimulus design matrix
Wstim=zeros(m,T-n);
for j=1:T-n
    for k=1:m
        Wstim(k,j)=stim(j+n-k,1)';
    end;
end;
```

```
%first derivative broken down into three components
HH1=zeros(2*p*n,1);
HH2=zeros(p,1);
HH3=zeros(pm,1);
Hh1=zeros(2*p*n,N);
Hh2=zeros(p,N);
Hh3=zeros(pm,N);
for i=1:N
    Hh1(:,i)=zeros(2*p*n,1);
    Hh2(1:p,i)=zeros(p,1);
    Hh3(:,i)=zeros(pm,1);
    for j=1:T-n
        A1=kron(Kz(:,(i-1)*(T-n)+j), eye(p));
        A2=kron(Wstim(:,j)', eye(p));
        A3=A2*Sgma*A2'+S;
        iOmegaj=inv(A3);
        R=Y(:,(i-1)*(T-n)+j)-A1'*vecp;
        Dij=R'*iOmegaj*R;
        for k=1:n
            h1d=tukeyc(Dij,c,1)*kron(Kz((k-1)*p+1:k*p,(i-1)*(T-n)+j),
                    W(:,(k-1)*p+1:k*p)')*iOmegaj*R;
            h1=[h1d(1);h1d(p+2); h1d(3*p)];
            h11d=tukeyc(Dij,c,1)*kron(Kz(n*p+(k-1)*p+1:n*p+k*p,(i-1)*(T-n)+j),
                    W(:,(k-1)*p+1:k*p)')*iOmegaj*R;
            h11=[h11d(1);h11d(p+2); h11d(3*p)];
            Hh1((k-1)*p+1:k*p,i)=Hh1((k-1)*p+1:k*p,i)+h1;
            Hh1(n*p+(k-1)*p+1:n*p+k*p,i)=Hh1(n*p+(k-1)*p+1:n*p+k*p,i)+h11;
        end;
        h2=tukeyc(Dij,c,1)*iOmegaj*R * Kz(s,(i-1)*(T-n)+j)';
        Hh2(1:p,i)=Hh2(1:p,i)+h2;
        for km=1:m
             h3=tukeyc(Dij,c,1)*iOmegaj*R * Kz(2*n*p+km,(i-1)*(T-n)+j)';
             Hh3((km-1)*p+1:km*p,i)=Hh3((km-1)*p+1:km*p,i)+h3;
        end;
    end;
end;
HH1=-2*sum(Hh1,2);
HH2=-2*sum(Hh2,2);
HH3=-2*sum(Hh3,2);
H1=[HH1; HH3; HH2];

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [H2,H3]=derivLs(data, cov, stim, group, W, n, c, vecp, LS, LSgma)

%derivatives of the scale parameters: H2 for LS and H3 for LSgma

m=1;
Y=response(data,n);
Kz=design(data,cov, stim, group, n, m);

p=3; s=2*n*p+m+1; [T,Np]=size(data); N=Np/p; pm=p*m; ps=p*s;
r1=0.5*p*(p+1); r2=0.5*pm*(pm+1); r3=ps+r1+r2;

%positions of variance components to be used in derivatives
lS=tril(ones(p));
lSgma=tril(ones(pm));
[iS,jS,vS]=find(lS);
```

```
S=LS*LS';
Sgma=LSgma*eye(pm);

%stimulus design matrix
Wstim=zeros(m,T-n);
for j=1:T-n
    for k=1:m
        Wstim(k,j)=stim(j+n-k,1)';
    end;
end;

%first derivative broken down into three components

H2=zeros(r1,1);
H3=zeros(1,1);
for i=1:N
    for j=1:T-n
        A1=kron(Kz(:,(i-1)*(T-n)+j), eye(p));
        A2=kron(Wstim(:,j)', eye(p));
        A3=A2*Sgma*A2'+S;
        iOmegaj=inv(A3);
        R=Y(:,(i-1)*(T-n)+j)-A1'*vecp;
        Dij=R'*iOmegaj*R;
        for k=1:length(iS)
            Sdot=zeros(p,p);
            Sdot(iS(k),jS(k))=1;
            A4=-tukeyc(Dij,c,1)*trace(iOmegaj*R*R'*iOmegaj*(LS*Sdot'
                    + Sdot*LS'))+trace(iOmegaj*(LS*Sdot'+Sdot*LS'));
            H2(k,1)=H2(k,1)+A4;
        end;
        A4=-tukeyc(Dij,c,1)*trace(iOmegaj*R*R'*iOmegaj*A2*A2')
                +trace(iOmegaj*A2*A2');
        H3=H3+A4;
    end;
end;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [FXP, FXS, FlogL, FXPH, FXSH]=rolexbstrap(data, cov,
stim, group, W, n, c, xp, xs)

%OUTPUT
%FXP a set of bootstrap estimates for the autoregressive parameters
%FXS a set of bootstrap estimates for the scale parameters
%FlogL corresponding loglikehood evaluated at each column of FXP and FXS
%FXPH and FXSH diagonal of the hessian matices at
%each column of FXP and FXS

%INPUT
%xp initial value for autoregressive parameters
%xs initial value for scale parameters

n=1; m=1; q=10; p=3; s=2*n*p+m+1; sd=2*n+m+1; [T,Np]=size(data);
N=Np/p; pm=p*m; ps=p*s; psd=p*sd; r1=0.5*p*(p+1);
r2=0.5*pm*(pm+1); r3d=psd+r1+r2; rps=r3d+2;

Y=response(data, n); Kz=design(data, cov, stim, group, n, m);
```

```
%generating bootstrap data
RqN=zeros(q*(T-n), N); iRqN=zeros(size(RqN));
for j=1:q
    for i=1:N;
        RqN((j-1)*(T-n)+1:j*(T-n), i)=randperm(T-n)';
        [RqN((j-1)*(T-n)+1:j*(T-n), i), iRqN((j-1)*(T-n)+1:j*(T-n), i)]=
            sort(RqN((j-1)*(T-n)+1:j*(T-n), i));
    end;
end;

RKz=zeros(q*s, N*rps); for j=1:q;
    for i=1:N
        iKz=iRqN((j-1)*(T-n)+1:j*(T-n), i);
        RKz((j-1)*s+1:j*s,(i-1)*rps+1:i*rps)=Kz(:, iKz(1:rps));
    end;
end;

RY=zeros(q*p, N*rps); for j=1:q;
    for i=1:N
        iKz=iRqN((j-1)*(T-n)+1:j*(T-n), i);
        RY((j-1)*p+1:j*p,(i-1)*rps+1:i*rps)=Y(:, iKz(1:rps));
    end;
end;

Wstim=zeros(m,T-n); for j=1:T-n
    for k=1:m
        Wstim(k,j)=stim(j+n-k,1)';
    end;
end;

Rstim=zeros(q*m, N*rps);
for j=1:q;
    for i=1:N
        iKz=iRqN((j-1)*(T-n)+1:j*(T-n), i);
        Rstim((j-1)*m+1:j*m,(i-1)*rps+1:i*rps)=Wstim(:, iKz(1:rps));
    end;
end;

RXP=zeros(psd,q); RXS=zeros(r1+r2, q); xpr=repmat(xp, 1, q);
xsr=repmat(xs, 1, q);

options1=optimset('Display', 'final', 'GradObj', 'on',
'LargeScale', 'on'); tol=1e-3; iter=0; iterlim=30;
for j=1:10
    XP=xpr(:, j);
    XS=xsr(:,j);
    iter=0;
    while(iter==0 | norm(XP-xp0)>tol | norm(XS-xs0)>tol)
        iter=iter+1;
        if (iter>iterlim), warning('Iteration limit reached.'); break; end;
        xs0=XS;
        xp0=XP;
        [XP,logL]=fminunc(@(XP) Phi2(XS, XP, RKz((j-1)*s+1:j*s,:),
            RY((j-1)*p+1:j*p,:), Rstim((j-1)*m+1:j*m,:), W, n), xp0, options1);
        [XS,logL]=fminunc(@(XS) LSS2(XS, XP, RKz((j-1)*s+1:j*s,:),
            RY((j-1)*p+1:j*p,:), Rstim((j-1)*m+1:j*m,:), W, n), xs0, options1);
    end;
```

```
    RXP(:,j)=XP;
    RXS(:,j)=XS;
end;

%NOTE: An alternative way would be to use a set of initial values obtained by
%perturbing the estimate. In addition since the variance parameters usually get
%stucked when the initial estimate is already close to the minimizer at that
%bootstrap data, it is practical to just perform the bootstrap over the
%autoregressive parameters since these estimates are assymptotically
%inddependent.

options2=optimset('Display', 'iter', 'GradObj', 'on',
'LargeScale', 'off');

FXP=zeros(psd,q); FXS=zeros(r1+r2, q); FXPG=zeros(psd,q);
FXSG=zeros(r1+r2, q); FXPH=zeros(psd,q); FXSH=zeros(r1+r2,q);
FlogL=zeros(1,q);
for j=1:5
    XS=RXS(:,j);
    XP=RXP(:,j);
    iter=0;
    while(iter==0 | norm(XP-xp0)>tol | norm(XS-xs0)>tol)
        iter=iter+1;
        if (iter>iterlim), warning('Iteration limit reached.'); break; end;
        xs0=XS;
        xp0=XP;
        [XP,logL, exitflag, output, grad1, hessian1]=fminunc(@(XP)
            Phi(XP, XS, data, cov, group, stim, W, n), xp0, options2);
        [XS,logL, exitflag, output, grad2, hessian2]=fminunc(@(XS)
            LSS(XS, XP, data, cov, group, stim, W, n), xs0, options2);
    end;
    FXP(:,j)=XP;
    FXS(:,j)=XS;
    FlogL(j)=logL;
    FXPG(:,j)=grad1;
    FXSG(:,j)=grad2;
    FXPH(:,j)=diag(inv(hessian1));
    FXSH(:,j)=diag(inv(hessian2));
end;
```

BIBLIOGRAPHY

[1] Aertsen, A. and Preiβl H. (1991). Dynamics of activity and connectivity in physiological neuronal networks. In *Nonlinear Dynamics and Neuronal Networks.* (Schuster, H.G. eds.), VCH Publishers, New York, 281-302.

[2] Allen, M. and Datta, S. (1999). A note on bootstrapping M-estimators in ARMA models. *Journal of Time Series Analysis* **20**, 365-379.

[3] Anderson, T.W. (1978). Repeated measurements on autoregressive processes. *Journal of the American Statistical Association* **73**, 371-378.

[4] Arslan, O., Edlund, O., Ekblom, H. (2002). Algorithms to compute CM- and S-estimates for regression. *Metrika* **55**, 37-51.

[5] Basawa, L.V., A.K.Mallik, W.P. McCormick, and R.L. Taylor (1989). Bootstrapping explosive autoregressive processes. *Annals of Statistics*, **17**, 1479- 1486.

[6] Bedrick, E. and Tsai, C.L. (1994). Model selection for multivariate regression in small samples. *Biometrics* **50**, 226-231.

[7] Belliveau, J.W., Rosen, B.R., Kantor, H.L., Rzedzian, R.R., Kennedy, D.N., McKinstry, R.C., Vevea, J.M., Cohen, M.S., Pykett, I.L., and Brady, T.J. (1990) Functional cerebral imaging by susceptibility-contrast NMR. *Magnetic Resonanace Medicine* **14**, 538-546.

[8] Belliveau, J.W., Kennedy, D.N., McKinstry, R.C., Buchbinder, B.R., Weisskoff, R.M., Cohen, M.S., Vevea, J.M., Brady, T.J., and Rosen, B.R. (1991) Functional mapping of the human visual cortex by magnetic resonance imaging. *Science* **254**, 716-719.

[9] Ben, M.G., Martinez, E.J., and Yohai, V.J. (1998). Robust Estimation in vector autoregressive moving-average models. *Journal of Time Series Analysis* **20**, 381-399.

[10] Behrens, J. (1991). *Robuste ordnungswahl fur autoregressive prozesse.* Ph.D. Thesis, University of Kaiserslautern, Germany.

[11] Besag, J.S. (1974). Spatial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society Ser. B* **36**, 197-242.

[12] Bhansali, R.J. and Downham, D.Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* **64**, 547-551.

[13] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for bootstrap. *Annals of Statistics* **9**, 1196-1217.

[14] Boos, D.D. and Serfling, R.J. (1980). A note on differentials and the CLT and LIL for statistical functions, with application to M-estimates. *Annals of Statistics* **8**, 618-624.

[15] Bose, A. (1988), Edgeworth correction by bootstrap in autoregressions, *Annals of Statistics*, **16**, 1709-1722.

[16] Bose, A. (1990). Bootstrap in moving average models. *Annals of the Institute of Statistical Mathematics*, **42**, 753-768.

[17] Box, G.E.P., and Jenkins, J.M. (1970). *Time Series Analysis-Forecasting and Control.* Holden-Day, San Francisco

[18] Boos, D.D. (1977). *The differential approach in statistical theory and robust inference.* Ph.D. Dissertation, Florida State University.

[19] Buchel, C. and Friston, K.J. (1997). Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modeling and fMRI. *Cerebral Cortex* **7**, 768-778.

[20] Buchel,C., Holmes,A.P., Rees, G. and Friston, K.J. (1998) Characterizing stimulusresponse functions using nonlinear regressors in parametric fMRI experiments. *Neuroimage* **8**, 140-148.

[21] Bunea, F., Ombao, H. and Auguste, A. (2005). Minimax adaptive estimation from an ensemble of signals. *IEEE Transactions on Signal Processing*, in press.

[22] Bunea, F., Tsybakov, A. and Wegkamp, M.H.(2004). Aggregation for regression learning. *Available online at arXiv:math.ST/0410214.*

[23] Bustos, O., Fraiman, R. and Yohai, V. (1986). Robust estimates for ARMA models. *Journal of the American Statistical Society* **81**, 155-168.

[24] Chatterjee, S. (1986). Bootstrapping ARMA Models: Some Simulations. *IEEE Transactions Systems, Man and Cybernetics*, SMC-16, 294-299.

[25] Cliff A. and Ord J.K. (1981). *Spatial Processes, Models and Applications.* Pion, London.

[26] Collins, J.R. (1976). Robust estimation of a location parameter in the presence of asymmetry. *Annals of Statistics* **4**, 68-85.

[27] Copt, S. and Victoria-Feser, M.P. (2005). High breakdown inference for mixed linear models. *Journal of the American Statistical Association*, **101**, 292-3000.

[28] Copt, S. and Heitier, S. (2006). Robust MM-estimation and inference in mixed linear models. *Journal of the American Statistical Association* to appear.

[29] Cressie, N.A. (1993). *Statistics for Spatial Data.* Wiley, New York

[30] Dale, A. and Halgren, E. (2001). Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. *Current Opinions in Neurobiology*, **11**, 202-208.

[31] Denby, L. and Martin, R.D.. (1979). Robust estimation of the first-order autoregressive parameter. *Journal of the American Statistical Association* **74**, 140-146.

[32] Fewster, R.M. (2003) A spatiotemporal stochastic process model for species spread. Biometrics, 59:640-649.

[33] Gianaros, P.J., Derbyshire, S.W.G., May, J.C., Siegle, G.J., Gamalo, M.A., and Jennings, J.R. (2005) Cortical and subcortical regulation of blood pressure reactivity to a behavioral stressor. *Psychophysiology* (in press).

[34] Geweke, J.F. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association* **77**, 304-324.

[35] Goncalves, M.S., Hall, D.A., Johnsrude, I.S. and Haggard, M.P. (2001). Can meaningful effective connectivities be obtained between auditory cortical regions? *Neuroimage* **14**, 1353-1360.

[36] Goodrich, R.L. and Caines, P.E. (1979). Linear system identification from nonstationary cross-sectional data. *IEEE Transactions on Automatic Control*, **24**, 403-411.

[37] Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424-438.

[38] Granger, C.W.J. (1980). Testing for causality: a personal viewpoint. *Journal of Econometric Dynamic Control* **2**, 329-352.

[39] Guttorp, P., Meiring, W, and Sampson, P. (1994). A space-time analysis of ground-level ozone data. *Environmetrics* **5**, 241-254.

[40] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

[41] Handock, M. and Wallis, J. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association* **89**, 368-390.

[42] Hannan, H.J., Dunsmuir, W.T.M. and Deistler, M. (1980). Estimation of vector ARMAX models, *Journal of Multivariate Analysis* **10**, 275-295.

[43] Harrison, L., Penny, W., and Friston, K. (2003). Multivariate autoregressive modelling of fMRI time series. *Neuroimage* **19**, 1477-1491.

[44] He, X. (1991). A local breakdown property of robust tests in linear regression. *Journal Multivariate Analysis* **38**, 294-305.

[45] Huber, P.J. (1964). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability* **1**, 221-233. University of California Press.

[46] Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73-101.

[47] Huggins, R.M. (1993a). A robust approach to repeated measures. *Biometrics* **49**, 715-720.

[48] Huggins, R.M. (1993b). A robust analysis of variance component models for pedigree data. *Australian Journal of Statistics* **36**, 271-286.

[49] Jiang, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in Mixed linea models. *Statistica Sinica* **8**, 861-885.

[50] Journel, A. G. and Huijbregts, C. J. (1978) *Mining Geostatistics*. Academic Press, London.

[51] Kent, J. and Tyler, D. (1991). Redescending M-estimates of multivariate location and scatter. *Annals of Statistics* **19**, 2102-2119.

[52] Kent, J. and Tyler, D. (1996). Constrained M-estimation for multivariate location and scatter. *Annals of Statistics* **24**, 13461370.

[53] Kilian, L. (1996). Impulse response analysis in vector autoregressions with unknown lag order, manuscript, Department of Economics, University of Pennsylvania.

[54] Kreiss, J.P., and J. Franke (1992). Bootstrapping stationary autoregressive moving-average models, *Journal of Time Series Analysis*, **13**, 287-317.

[55] Kunsch, H. (1984). Infinitesimal robustness for autoregressive processes. *Annals of Statistics* **12**, 843-863.

[56] Kunsch, H.R. (1989). The jacknife and the bootstrap for general stationary observations. *Annals of Statistics* **17**, 1217-1261.

[57] Kwong,K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskoff, R.M., Poncelet, B.P., Kennedy, D.N., Hoppel, B.E., Cohen, M.S., and Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Science USA* **89**, 56755679.

[58] Lahiri, S.N. (1992). Edgeworth Correction by Moving Block Bootstrap for Stationary and Nonstationary Data. In *Exploring the Limits of Bootstrap*. (R. LePage and L. Billard, eds.) John Wiley k Sons, New York.

[59] Laird, N. M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963974.

[60] Lawson, LA and Clark, A (2002) Spatial mixture relative risk models applied to disease mapping, *Statistics in Medicine* **21**, 359-370.

[61] Li, W.K. and Hui, Y.V. (1989). Robust multiple time series modelling. *Biometrika* **76**, 309-315.

[62] Liu, R.Y. and Singh, K. (1992). Moving blocks jacknife and bootstrap capture weak dependence. In it Exploring the Limits of the Bootstrap. (R. Lepage and L. Billard, eds) Wiley, New YOrk, pp. 225-248.

[63] Lopuhaa, H. P (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics* **17**, 1622-1683.

[64] Lundbye-Christensen, S. (1991). A multivariate growth curve model for pregnancy. *Biometrics*, **47**, 637-657.

[65] Mardia, K. V., Goodall, C., Redfern, E. J. and Alonso, F. J. (1998) The Kriged Kalman filter (with discussion). *Test*, **7**, 217-252.

[66] Maronna, R., Bustos, O.H. and Yohai, V.J. (1979) Bias and efficiency Robustness of general M-estimators for regression with random carriers. In *Soothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt eds.), Springer, New York.

[67] Martin, R.D. (1980). Robust estimation of autoregressive models (with discussion). In *Directions in Time Series* (D.R. Brillinger and G.C. Tiao, eds.) IMS, Hayward, Calif.

[68] Martin, R.D. and Yohai, V.J. (1986). Influence functionals for time series. *Annals of Statistics* **14**, 781-818.

[69] Masarotto, G. (1987). Robust and consistent estimates of autoregressive-moving average parameters. *Biometrika* **74**, 791-797.

[70] Matheron, G. (1963). Principles of geostatistics. *Economic Geology* **58**, 1246-1266.

[71] McIntosh, A.R. and Gonzalez-Lima, F. (1994). Structural equation modelling and its application to network analysis in functional brain imaging. *Human Brain Mapping* **2**, 2-22.

[72] Mechelli, A., Penny, W., Price, J., Gitelman, D. and Friston, K.J. (2002). Effective connectivity and intersubject variability: using a multisubject network to test differences and commonalities. *Neuroimage* **17**, 1459-1469.

[73] Oehlert, G.W. (1993). Regional trends in sulfate wet deposition. *Journal of the American Statistical Association* **88**, 390-399.

[74] Ogawa, S., Lee, T.M, Kay, A.R., and Tank, D.W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Science* **87**, 9868-9872.

[75] Pfeiffer, P.E. and Deutsch, S.J. (1980). A three-stage iterative procedure for space-time modeling. *Technometrics* **22**, 35-47.

[76] Portnoy, S.L. (1977). Robust estimation in dependent situations. *Annals of Statistics* **5**, 22-43.

[77] Rahiala, M. (1999). Random coefficient autoregressive models for longitudinal data. *Biometrika* **86**, 718-722.

[78] Richardson, A.M. and Welsh, A.H. (1995). Robust restricted maximum likelihood in mixed linear models. *Biometrics* **51**, 1429-1439.

[79] Rocke, D.M. and Woodruff, D.L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association* **91**, 1047-1061.

[80] Roebroeck, A., Formisano, E., and Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage* **25**, 230-242.

[81] Ronchetti, E. (1985). Robust model selection in regression. *Statistics and Probability Letters* **3**, 21-23.

[82] Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica* **7**, 327-338.

[83] Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association* **92**, 1017-1023.

[84] Rosenberg, B. (1973). Linear Regression with Randomly Dispersed Parameters. *Biometrika* **60**, 61-75.

[85] Rousseeuw, P.J. and Yohai, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (Franke, J., Hardle, W., and Martin, R.D. eds.) New York, Springer-Verlag, 256-272.

[86] Rousseeuw, P, J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

[87] Sahu, S.K. and Mardia, K.V. (2005). A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. *Applied Statistics* **54**, 224-244.

[88] Salibian-Barrera, M. and Zamar, R. (2002). Bootstrapping robust estimates of regression. *Annals of Statistics* **30**, 556-582.

[89] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

[90] Shorack, G. (1982). Bootstrapping robust regression. *Communications in Statistics: Theory and Methods*, **11**, 961-972.

[91] Sanso, B. and Guenni, L. (1999) Venezuelan rainfall data analysed by using a Bayesian spacetime model. *Applied Statistics*, **48**, 345362.

[92] Stahel, W.A. and Welsh, A.H. (1992). Robust estimation of variance components. *Research Report* **69**, ETH, Zurich.

[93] Stoffer, D.S. (1986). Estimation and identification of space-time ARMAX models in the presence of missing data. *Journal of the American Statistical Association* **81**, 762-772.

[94] Stoffer, D.S. (1999). Detecting common signals in multiple time series using the spectral envelope. *Journal of the American Statistical Association* **94**, 1341-1356.

[95] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B* **39**, 44-47.

[96] Tsay, R. S., Pena, D. and Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika* **87**, 789-804.

[97] Tu, Shi-Ming Yu and Hua Sun (2004). Transaction-Based Office Price Indexes: A Spatiotemporal Modeling Approach. *Real Estate Economics* **32**, 297-328.

[98] Wikle, C. K. and Cressie, N. (1999) A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815829.

[99] Vaida, F. and Blanchard S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351-370.

[100] Valdez-Sosa, P. (2004). Spatio-temporal autoregressive models defined over brain manifolds *Neuroinformatics* **2**, 239-250.

[101] Welsh, A.H. and Richardson, A.M. (1997). Approaches to the Robust Estimation of Mixed Models. In *Handbook of Statistics* Vol. 15, Elsevier Science B.V.

[102] Yohai, V.J. (1997) Local and global robustness of regression estimators. *Journal of Statistical Planning and Inference*, **57**, 73-92.