A Function Space Approach to the Generalized Nonlinear Model with Applications to Frequency Domain Spectral Estimation

by

Keith N. Selander

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Statistics

APPROVED:

Robert V. Foutz, Chairman

Raymond H. Myers

Donald R. Jensen

George R. Terrell

Marion R. Reynolds, Jr.

February 6, 1992

Blacksburg, Virginia

A Function Space Approach to the Generalized Nonlinear Model with Applications to Frequency Domain Spectral Estimation

by

Keith N. Selander

Robert V. Foutz, Chairman

Statistics

(ABSTRACT)

Peter McCullagh (1983) outlined the theory of quasi-likelihood estimation in generalized linear models. Chiu (1988) showed that an iterated, reweighted least squares procedure applied to the periodogram produces estimates of spectral density model parameters for Gaussian univariate time series which have the same asymptotic variance as those produced by maximizing the true likelihood. In this dissertation, McCullagh's theory is combined with a functional analysis approach and extended to parametric estimation of the spectral density matrix components of a non-Gaussian bivariate time series. An asymptotic optimality theorem is given, which shows optimality of an iterated, reweighted least squares procedure within a class of procedures. However, the principal application of the theory is for parametric spectral estimation in the case of an observed "contaminated" Gaussian series X(t)+N(t), where the noise series is uncorrelated with the X series, and it is desired to estimate the spectrum of the X series. Previous literature suggests removing contaminated bands of the periodogram prior to analysis, but the results of the dissertation may be used to unbiasedly estimate the spectrum of f without a precise knowledge of which bands are contaminated.

Acknowledgements

First, I would like to thank my parents for their continuing support and encouragement during the years of both my graduate and undergraduate degrees. Their help, both emotional and financial, and during often difficult times, has been greatly appreciated.

I would like to thank my wife, Sindee Sutherland, not only for her love and support, especially during the times I felt most discouraged, but also her practical help in debugging computer programs, the more efficient use of word processors, and proofreading. She was undoubtedly a major factor in my completion of this degree.

I would like to express my appreciation to all my professors at Virginia Tech, in both the math and statistics departments. In the math department, I would especially like to thank Peter Fletcher and Rebecca Crittenden, who introduced me to theoretical mathematics. The years I spent studying math have been an invaluable background for my current research. In the statistics department, I would like to thank all the members of my committee, but also express special appreciation to three members whose contributions have been greatest. To my advisor, Robert Foutz, for introducing me to time series, for being easy to get along with, for always being encouraging and enthusiastic about whatever I was doing, for allowing me to pursue what I was interested in for a dissertation topic, and for suggesting changes during the preparation of the manuscript. To Ray Myers, who is one of the best teachers in the statistics department and who has been a major influence in helping me intuitively understand applied statistics. This dissertation is essentially an extension of the viewpoint taught in his classes

and written in his papers, expressed in the language of theoretical math. Finally, to Don Jensen, for his careful reading of the dissertation and many suggestions of changes and constructive criticisms. Embarrassingly, he discovered how little I retained from past English and writing classes. All mistakes remaining in the dissertation are solely the fault of the author.

Table of Contents

List of Figures

Chapter I
Introduction

This dissertation will be concerned with a relationship which does not seem to be commonly recognized between two apparently different areas of statistics: frequency domain spectral estimation and the "generalized linear/nonlinear model". Because of this lack of recognition, there has been little development of generalized model theory as specifically applied to time series. This is surprising in a way, since many concepts and ideas which have only intuitive or "rough" sketches of proof in the context of the generalized model may be made rigorous in the context of spectral estimation.

The key to this relationship for Gaussian univariate time series lies in the asymptotic distribution of the periodogram as independently distributed exponential random variables whose means are the spectrum, the spectrum being a deterministic function which completely describes the covariance structure of the series. Peter McCullagh (1983) outlined the theory of "quasi-likelihood estimation" in generalized linear models. Briefly put, in the generalized linear model we have some observations y_i from an exponential family distribution, and wish to model the means of these observations as $s(\mathbf{x}_i'\boldsymbol{\beta})$, a function of the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ for some regressors $\mathbf{x}_i$. The obvious way to obtain an estimate $\widehat{\boldsymbol{\beta}}$ of the unknown parameter vector $\boldsymbol{\beta}_0$ is to maximize the likelihood function. For example, if the y_i have an exponential distribution, we maximize $-\sum \log s(\mathbf{x}_i'\boldsymbol{\beta}) + y_i/s(\mathbf{x}_i'\boldsymbol{\beta})$. McCullagh points out that if the observations y_i are not exponential, *but have the same (up to) second order moment structure as exponential random variables*, then obtaining $\widehat{\boldsymbol{\beta}}$ by maximizing the exponential likelihood

function rather than the true likelihood function which may be much more complicated, results in an estimate which is optimal with respect to asymptotic variance (is "asymptotically BLUE"). In such a situation, we say that the exponential likelihood function is a "quasi likelihood function", as it is not the true likelihood. Recall that exponential family distributions have their *variances a function of their means.* If a variance (or the covariance structure) can be so expressed, we may solve the QL equations $\partial l(\mu, y)/\partial\mu = V^{-}(\mu)(y - \mu)$ for the QL function, where y is the observation vector and $V^{-}(\mu)$ is a generalized inverse of the variance covariance matrix. McCullagh says that "The most interesting (case) involves uncorrelated observations" (McCullagh and Nelder (1983), p. 169), so that the variance matrix has diagonal form, and this corresponds to the cases we are most familiar with, such as independent observations from an exponential family (second) moment structure distribution. The central premise of this dissertation is that the uses of "non independent" QL functions are unrecognized.

Frequency domain spectral estimation is really an attempt to model a one dimensional generalized model response surface. One method of obtaining parametric estimates is to choose the θ maximizing

$$-\sum \log f_\theta(\lambda_i) + I_n(\lambda_i)/f_\theta(\lambda_i) \qquad [1.1]$$

where $\{f_\theta\}$ is the spectral model and $I_n(\lambda)$ is the periodogram (see section 3.3 for definitions). Notice that [1.1] is the log likelihood function for independent, exponential random variables, and its use is obviously motivated by the asymptotic distribution of the periodogram. Such an idea is quite old, dating to Whittle's work in the 1950's, e.g Whittle (1951), (1953). It turns out that the estimate $\widehat{\theta}$ from this procedure is asymptotically equivalent to that obtained by maximizing the true likelihood function *if the time series is Gaussian.* If the series is not

Gaussian, $\widehat{\theta}$ is *consistent but not optimal with respect to asymptotic variance.* One application of the results in this dissertation is an extension of the work of Chiu (1988) to "optimal" spectral estimation in non-Gaussian (univariate) time series. Chiu showed that an iterated, reweighted least squares procedure applied to the periodogram produces estimates of spectral density model parameters for Gaussian univariate time series which have the same asymptotic variance as those produced by maximizing the true likelihood. For non-Gaussian univariate series, we cannot consider the periodogram to be "asymptotically independent" in a sense to be described later, but the dissertation will show how to obtain "asymptotically BLUE" estimates of parameters for density models. In the case of multivariate time series, we will also give an asymptotic optimality theorem which indicates a new way in which splines or other parametric functions may be fit to the cross spectrum. The only previous methods of multivariate time series analysis are "smoothing" of the periodogram matrix, a non-parametric approach, or multivariate ARMA models, which are difficult to fit due to their complexity and abundance of parameters. The results presented here point towards the development of new parametric multivariate models with practical fitting methods as an alternative to both of these.

A main theme of the dissertation is how to fit spectral models to "contaminated" series of the form X(t)+N(t), where N(t) is uncorrelated noise which may severely damage parametric spectral estimates. It turns out that maximizing [1.1] where the periodogram has been calculated from a contaminated series will result in a biased estimate (Taniguchi (1979), Hosoya and Taniguchi (1982)). A way out of this unfortunate situation is the use of a "misspecified" QL function. The misspecification is in the covariance structure: we incorrectly assume covariance in the observations which does not exist in order to lessen bias. If this is correctly done, *and if in fact no contamination existed in the observed series, the asymptotic variance of our parametric estimates will be little damaged.* A connection exists here to non-

Gaussian spectral estimation, in which there *is* (asymptotic) covariance in the periodogram. This is apparently unrecognized in the literature.

The central viewpoint of the dissertation is *that the observation vector (i.e. the periodogram) is really a function in* $L^2[-\pi, \pi]$*, the space of square integrable functions on* $[-\pi, \pi]$. For example, if the periodogram is viewed as only being defined at the Fourier frequencies, then we can extend it in an obvious way to being defined on $[-\pi, \pi]$. The asymptotic means function is then the true spectrum $f(\lambda)$. The asymptotic variance at a frequency λ is $f^2(\lambda)$, *and the variance function is viewed as a multiplication operator on* L^2, i.e. $M_{f^2}[g]=f^2g$ for $g \in L^2$. This corresponds to the "independent observations case" in GLIM where the design space is the interval $[-\pi, \pi]$ and the variance matrix is of diagonal form, so that Vx multiplies each x_i by d_i, the d_i being the diagonal. There are many different linear operators on L^2; for example, one which will play an important role in the sequel is the kernel operator which maps a function $f(\lambda)$ to the function $\int k(\lambda, x) f(x)\, dx$. The main difference between Gaussian and non Gaussian processes is that *the variance operator for a non Gaussian process is not a multiplication operator.* This introduces a new concept into "response surface GLIM": thinking of the "variance" as an operator on an infinite dimensional Hilbert space rather than a matrix, which is an operator on a finite dimensional Hilbert space. If the data is contaminated, using a non multiplication operator (which will sometimes be referred to as "non Gaussian variance operators" for obvious reasons) as a variance operator may result in reduced bias if correctly chosen. Previously, it has been suggested to eliminate periodogram ordinates corresponding to "contaminated" frequency bands. The use of a non-Gaussian variance operator can achieve the same results *if it is not precisely known which bands are contaminated, which is perhaps a more realistic assumption in practice.*

The dissertation is divided into 11 chapters. Chapters 2 and 3 describe in further

detail what time series and generalized linear/nonlinear models are, including literature reviews of relevant material. Chapter 4 shows the link between maximum likelihood estimation in time series and the generalized model response surface. Chapter 5 describes the problem of frequency domain cospectral estimation in multivariate time series. The essence of the dissertation is chapters 6-9. Chapter 6 gives a new definition of quasi-likelihood, in which estimation is shown to be a problem in nonlinear functional analysis, and gives general definitions involved in the setup of QL type problems as here defined. Chapter 7 states and proves an optimality theorem for estimates obtained by minimizing a QL function as defined in chapter 6, essentially giving a version of the Gauss Markov theorem. Chapter 8 shows that the definitions of chapter 6 are satisfied in the framework of spectral estimation, so that the optimality theorems of chapter 7 apply. Chapter 9 discusses the problem of spectral estimation in contaminated series as a main application of the theory in chapters 6 and 7. Chapter 10 extends the theory of chapter 6, using the limitations of chapter 6's framework which become apparent in chapter 9 as motivation. Finally, chapter 11 gives some conclusions and directions for future research.

Chapter II
The Generalized Linear/Nonlinear Model

2.1 Introduction to Generalized Models

The usual regression problem is to find a model for the means of independent, normally distributed random variables as a function of explanatory variables $x_1, x_2, \ldots, x_p$, i.e. we model

$$y_i = f(\beta, \mathbf{x}) + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \qquad \mathbf{x} = (x_1, x_2, \ldots, x_p)$$

where β is an unknown column vector of parameters to be estimated by the method of least squares. If $f(\beta, \mathbf{x}) = \mathbf{x}'\beta$, then we have a (multiple) linear regression problem, and otherwise a nonlinear least squares problem which may be solved by the Gauss-Newton method. But suppose the model does not have this "classical" form; it may be that the means of the y_i are a function of some explanatory variables $\mathbf{x}$, but the variances of the y_i are not constant. These variances may actually be a function of the means, i.e. the variance of y_i is $V(\mu_i)$ for some positive function $V(\mu)$. We are still interested in modeling the means of $y_i = f(\beta, \mathbf{x}_i)$ where f is not necessarily linear, but it must be taken into consideration when β is estimated that if μ_i=$f(\beta, \mathbf{x}_i)$, then the variance of y_i is $V(f(\beta, \mathbf{x}_i))$. This is known as a **generalized nonlinear model (GNM)**, and as one might suspect it is probably not a good idea to estimate β by minimizing

$$\sum (y_i - f(\beta, x_i))^2$$

as one would do if the variances were constant, but instead follow some other procedure. One example of such a procedure is called **iterated reweighted least squares (IRWLS)** and is given

by the following algorithm (which is from Carrol and Rupert, (1988), p. 10).

Step 1 Start with a preliminary estimator $\hat{\beta}_*$.

Step 2 Form the estimated weights $w_i=1/g^2(\mu_i(\hat{\beta}_*), z_i, \theta)$, where $var(y_i) = g^2(\mu_i(\hat{\beta}_*), z_i, \theta)$ may depend on $\mu_i=E(y_i)$, some other (known or estimable) parameter θ, and some known vectors z_i (whose components "may or may not include some or all of the predictors x_i").

Step 3 Let $\hat{\beta}_G$ be the weighted least squares estimate using the estimated weights in step 2.

Step 4 Update the preliminary estimator by setting $\hat{\beta}_*= \hat{\beta}_G$, and update the weights as in step 2.

Step 5 Repeat steps 3 and 4 $C-1$ more times, where C is the number of cycles chosen by the experimenter.

If $\mu_i(\beta) = x'\beta$ (i.e. the means are a linear function of the regressors), then according to the Gauss Markov Theorem, the BLUE estimate of the parameter β would be given doing weighted least squares using the weights $w_i=1/g^2(\mu_i(\beta), z_i, \theta)$. Asymptotically, the algorithm given above will yield an estimator with the same variance. If $\mu_i(\beta)$ is not a linear function of the regressors, the Gauss Markov theorem and a linearization argument yields the same result. Thus, we may consider the estimator given by the IRWLS algorithm "asymptotically BLUE".

One model where the mean is a function of the variance occurs when the y_i are members of an exponential family; i.e. y_i has a Poisson, Gamma, Binomial, etc. distribution, and the parameter needed to completely specify the distribution is a function of the regressors. We may write the distribution as

$$f(y_i, \theta_i) = \exp\{y_i\theta_i + g(\theta_i) + d(y_i, \theta_i)\} \qquad [2.1.1]$$

so that the joint log likelihood function of the observation vector **y** may be written

$$l(\theta, \mathbf{y}) = \sum_{i=1}^{N} y_i\theta_i + g(\theta_i) + d(y_i, \theta_i). \qquad [2.1.2]$$

It turns out that if the model is correctly specified, the estimator $\hat{\beta}$ obtained by maximizing the likelihood equations, and that obtained by using IRWLS are asymptotically equivalent (have the same asymptotic variance matrix).

In the **generalized linear model (GLIM)**, the relationship between the means, the regressors, and the parameters has the special form $\mu_i = s(\mathbf{x}_i\beta)$ for some monotone, twice differentiable function s(.) called the link function. Each exponential family function has a special link function, known as the natural link, which occurs when $\theta_i = \mathbf{x}_i\,\beta$ (i.e. $s^{-1}(\mu_i)=\theta_i$). Common link functions include the identity ($\mu_i=\mathbf{x}_i\beta$), the log ($\mu_i=\log(\mathbf{x}_i\beta)$), and the exponential ($\mu_i=\exp(\mathbf{x}_i\beta)$). Generalized linear models have become more popular in recent years due to a computer package called GLIM which allows the user to choose a distributional form and a link function to fit these models to data.

The main results of the dissertation fall under the domain of generalized *nonlinear* models, although obviously any result which holds for a generalized nonlinear model would also hold for a generalized linear model. Most of the published literature has to do with generalized *linear* models, since these are simpler to work with in practice. For example, it can be shown that IRWLS is equivalent to Fisher's "scoring method" where the second derivative matrix is replaced by its expected value (see, e.g. Green (1984)). If the natural or canonical link is used in a generalized linear model, then the second derivative matrix and its expected value are the same (McCullagh and Nelder (1983), p. 33).

Peter McCullagh, who was involved in the development of generalized linear models, observed in his 1983 paper that "likelihood" type methods could be extended to models where the full distributional structure need not be specified, but only the relationship between the

mean and the variance. If $\mathbf{y}$ is an observation vector which has mean vector $\boldsymbol{\mu}$ and variance covariance matrix $\mathbf{V}(\boldsymbol{\mu})$, he defined what is called a **quasi-likelihood function** by solving the equation

$$\frac{\partial l(\boldsymbol{\mu}, \mathbf{y})}{\partial \boldsymbol{\mu}} = \mathbf{V}^{-}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}) \qquad [2.1.3]$$

for $l(\boldsymbol{\mu}, \mathbf{y})$, and gave a rough sketch of an argument that the estimator $\hat{\boldsymbol{\beta}}$ obtained by maximizing the quasi-likelihood function is "asymptotically BLUE", and thus has the same asymptotic variance matrix as the IRWLS estimator (now obtained by iteratively minimizing the quadratic form $(\mathbf{y} - \boldsymbol{\mu})'\mathbf{V}^{-}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})$) or the estimator obtained by maximizing the true likelihood function. Thus, if you use a Poisson likelihood function to estimate $\boldsymbol{\beta}$ in a generalized nonlinear model, assuming the y_i's have the same relationship between the mean and variance as do Poisson random variables, your IRWLS estimators or ML estimators will have an asymptotic optimality property. This is true even if higher order cumulants of the y_i's are not the same as higher order Poisson cumulants.

Notice that only the zeros of the derivative $\partial l(\boldsymbol{\mu}(\boldsymbol{\beta}), \mathbf{y})/\partial \boldsymbol{\beta}$ are needed (McCullagh, (1983)); therefore, it is not necessary to explicitly solve [2.1.3] in order to use a QL function in practice (and of course, in the IRWLS procedure, only $\mathbf{V}(\boldsymbol{\mu})$ needs to be specified). If the means vector is written as a function of the unknown $\boldsymbol{\beta}$ vector, we may write the "generalized least squares" equations as

$$\mathbf{D}^T\mathbf{V}^{-}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))=0 \qquad [2.1.4]$$

where $\mathbf{D} = \frac{d\boldsymbol{\mu}(\boldsymbol{\beta})}{d\boldsymbol{\beta}}$ is an $N \times p$ matrix, and the solution $\hat{\boldsymbol{\beta}}$ to these equations is our QL estimate.

All of the above discussion about "asymptotic optimality" applies only if the means model has been correctly specified. One main result of the dissertation will show what happens

under model misspecification in generalized nonlinear model response surfaces (i.e. what is $\widehat{\beta}$ from solving the QL equations estimating if the model is not correct?). It will be seen that if the data $y(x_i)$ is "contaminated" in the sense that $E(y(x_i))=f_{\beta_0}(\mathbf{x}_i) + f_N(\mathbf{x}_i)$, where $\mathbf{x}_i$ is in a design space, $f_{\beta_0}(\mathbf{x})$ is the true means function it is desired to estimate, $f_N(\mathbf{x})$ is due to "contamination" in the data, and the $y(x_i)$ are independent, solving the QL equations formed by using the correct variance function $V(\mu)$ does not consistently estimate β_0. Solving an *alternative* set of QL equations obtained by *assuming covariance in the data which does not exist* will yield consistent estimates of β_0. If the alternative variance matrix is correctly chosen, asymptotic variance of parameter estimates will be little damaged if the model were in fact correct (i.e. if $f_N(x)=0$).

2.2 Literature Review of Generalized Linear/Nonlinear Models

Generalized linear models were introduced in 1972 by Nelder and Wedderburn. Their paper discussed the GLM as applied to different exponential family distributions and noted the equivalence of IRWLS to solving the likelihood equations. It also introduced the concept of "deviance" as a diagnostic in the fitting process (as an analog to the SSE or MSE in the usual regression setting). Wedderburn's (1974) paper recognized the fact that to define a likelihood requires a full specification of the form of the distribution, but to define a quasi-likelihood requires only the specification of the relation between the mean and variance. Various properties of QL functions are discussed, and the equivalence of IRWLS and using the Newton-Raphson method with expected second derivative of the QL function was observed. Wedderburn continued his GLIM research with his 1976 paper, which explored existence and uniqueness properties for maximum likelihood type estimates in the generalized linear model.

In the 1980's, generalized linear/nonlinear models became a "hot topic" for research,

and the number of papers published in this period increased dramatically. Much of the research was geared towards model diagnostics and the application of likelihood type theory to "smoothing spline" (nonparametric) estimation in GLIM (mostly one dimensional) response surfaces. As most of this research does not have a direct bearing on the results of the dissertation, we will not review the voluminous literature but instead concentrate on the papers which are most closely related.

In 1983, McCullagh's paper reiterated "the connection between quasi likelihood functions, exponential family models and nonlinear weighted least squares", and has been previously mentioned is a major background paper for the dissertation. Jorgensen (1983) began to examine GLIM with correlated observations and error distributions not of the exponential family form. Green's (1984) paper was in part a survey article, but also gave a discussion of such topics as specific numerical methods to carry out IRWLS and convergence in the IRWLS procedure. It was also an attempt to examine "resistant" reweighted least squares estimation techniques, which mainly involved downweighting "outliers" in the residuals during the process of IRWLS.

The papers discussed in the previous paragraph are of a "general theory" nature. Another main area of GLIM research apparently beginning in the mid 1980's is the "GLIM response surface". One of the first papers in this area was O'Sullivan, Yandell, and Raynor (1986), in which the problem of obtaining a nonparametric estimate of a regression function f(t) is considered. The method used is the "penalized log likelihood" concept, first introduced by Good and Gaskins (1971), where the nonparametric estimate $\hat{f}(t)$ ($t \in T$) is chosen from (some subset of) the continuous functions on T to minimize $l_{n\lambda}(f) = \sum_{i=1}^{n} l_i(y_i;\, f(t_i)) + n\,\lambda J(f)$. Here, l_i is the likelihood function, and J(f) is a "penalty functional" designed to "incorporate prior notions, such as smoothness, about the behavior of f" (O'Sullivan et. al. (1986), p. 96).

For example, if T = [a, b], then $\hat{f}(t)$ might be chosen from the set of continuously differentiable functions and J(f) might be $\int_a^b [f'(t)]^2\, dt$. λ is a "smoothing parameter" which controls the relative weighing of the penalty function in estimating f, and needs to be estimated from the data. The theme of "nonparametric GLIM response surfaces" is continued by Tibshirani and Hastie (1987), which discusses fitting "local" linear models (e.g. $\beta_0+\beta_1 x$) by maximizing the likelihood function in a neighborhood of each point to obtain an estimate of the mean response at that point (e.g. to obtain $\hat{f}(x_0)$, let N be a neighborhood of x_0 and fit a linear model using only the data corresponding to regressors which fall in N. Then $\hat{f}(x_0)=\hat{\beta}_0+\hat{\beta}_1 x_0$). This idea is expanded upon in Staniswalis (1989), who examined the notion of maximizing a "weighted likelihood" at each x_0 of the form

$$W(\lambda)=\sum_i W\left(\frac{x_0 - x_i}{b}\right) \log f(y_i, \lambda).$$

The kernel bandwidth b is chosen (by cross validation), and then the $\hat{\lambda}$ maximizing $W(\lambda)$ is the estimate of the regression function at x_0. Azzalini, Bowman and Hardle (1989) discuss using nonparametric regression as a diagnostic tool by comparing nonparametric and parametric estimates of the response function, an idea which is again elaborated on by Staniswalis and Severini (1991). The papers involving a combination of nonparametric and parametric regression are mentioned because they somewhat relate to the material in chapter 9 of the dissertation, which in part has to do with fitting a parametric model to a nonparametric estimate of a regression function.

Several important books were also published on generalized linear models in the 1980's, the most important being McCullagh and Nelder's "Generalized Linear Models" (1983), and Carrol and Rupert's "Transforming and Weighing in Regression" (1988). The

McCullagh/Nelder book was essentially an expansion of the McCullagh/Nelder papers (some of which are rather terse), attempting to give a more comprehensive view but not containing any essential new material. Carrol and Rupert also gave a practical treatment of "how to" modeling using GLIM. One important discussion in their book is the concept of "pseudo likelihood", in which the variance function does not depend upon the mean, but instead upon some other unknown parameters (as in the IRWLS algorithm of section 2.1). This will be similar to the cross spectral estimation situation in chapter 5.

Chapter III
Time Series

3.1 Introduction

When one thinks of a "statistical model", perhaps the first thing to come to mind might be the areas of multiple regression, response surfaces, or the "generalized linear model" as described in the preceding chapter; that is, modeling the means of independent random variables as functions of "x values" or "regressors". However in time series analysis, the random variables being analyzed are all assumed to have mean 0, and the problem is finding a model to describe the covariance structure of these variables. In practice, one would usually have to remove a trend line by some means (such as regression, for example), and the time series part of the statistical analysis would consist of modeling the covariance structure of the residuals. For example, if one considers the yearly GNP from 1850 to the present, a curve might be fit to the data using regression, but it probably would not be valid to consider the residuals from the fit to be independent random variables.

There are essentially two basic types of time series, continuous and discrete. An example of a continuous time series might be the recording from a seismograph or an electroencephalograph; that is, a numerical reading for each point in a continuous interval of time. Some examples of discrete time series include the monthly production of steel, automobiles, etc., the weekly rainfall, the daily Dow Jones Industrial Average, and the daily value of a certain stock when the market closes. We might expect the last two of these series to be related (i.e. correlated), and taken together they form a bivariate time series. This

dissertation will also be concerned with modeling the joint covariance structure of two such series, but that comes later. For now, let us just state that the dissertation will be concerned solely with discrete series. Of course, one can always create a discrete series from a continuous one by taking observations every δ units of time (how to choose δ has been studied, but we will not go into this here), and for purposes of computer analysis this must often be done. Here, we will concentrate solely on how to model the covariance structure of a discrete, mean 0, stationary (to be defined in the next section) time series. There is a connection between modeling the covariance structure and a topic perhaps more familiar to statisticians known as "generalized linear models", and the main purpose of the dissertation is to develop this connection. But first we must establish some notation and describe some of the key concepts involved in time series analysis.

3.2 The Covariance Function and the Spectral Density Function

Let us establish some notation. We will define a **weakly stationary time series** to be an infinite sequence of square integrable random variables X(t) (t=0, ± 1, ± 2, . . .) having mean 0 and covariance function B(t, s)=E(X(t)X(s)) so that B(t, s) depends only on the difference $t-s$. In other words, we can write $B(t, s) = \gamma(t-s)$ for some function γ on the integers. A **strictly stationary time series** has the property that the vector $(X(t_1+t), \ldots X(t_k+t))$ has the same distribution for all t. Whether the series is weakly or strictly stationary, the central problem we are concerned with is how to estimate $\gamma(\ .\ )$. At first glance this would be a seemingly impossible problem (since you only observe a finite number of observations), but fortunately we have the following theorem due to Herglotz (see Brockwell and Davis (1987), p. 115):

Theorem 4.3.1 A complex-valued function $\gamma(h)$ defined on the integers is nonnegative definite iff

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda}\, dF(\lambda)$$

where $F(\lambda)$ is a right continuous, non-decreasing, bounded function on $[-\pi, \pi]$ called the spectral distribution function. If

$$F(\lambda) = \int_{-\pi}^{\lambda} f(\nu)\, d\nu$$

then f is called the **spectral density** or **spectrum** of $\gamma(\,.\,)$.

Note that for real valued processes, in order to make the covariances real the spectral density f must be symmetric about the origin. The time series we will be concerned with will all have a spectral density function (spectrum), so if we can succeed in estimating the spectrum, we will have completely specified the covariance structure of the series and will thus consider ourselves to have analyzed the time series as far as possible.

There is also a multivariate analog to the spectral density for stationary multivariate processes. A vector valued process $\mathbf{X}=(X_1(t), \ldots X_n(t))'$ is said to be weakly stationary if the joint covariance matrix $B(t, s)= \mathbf{E}\, \mathbf{X}(t)\, \mathbf{X}'(s)$ depends only on the difference $t-s$, and strictly stationary if the distribution of the vector $(X_{a_1}(t_1+t), \ldots\ X_{a_k}(t_k+t))$ does not depend on t. In this case, instead of the spectral density function, there is a spectral density *matrix* of functions to describe the covariance structure of the process. This matrix must be nonnegative definite at each frequency, and the Fourier transforms of the components give the covariance and cross covariance functions, i.e.

$$\gamma_{ij}(h) = \int_{-\pi}^{\pi} e^{i\lambda h}\, f_{ij}(\lambda)\, d\lambda$$

where $\gamma_{ij}(h)$ = cov $X_i(t+h)$ $\overline{X_j(t)}$ for all t (which is constant for all t by stationarity). If $(X_1(t), X_2(t))$ is a bivariate process, then $f_{12}(\lambda)$ is called the **cross spectrum.** It is complex valued, and if we define $c_{12}(\lambda)=\mathrm{Re}\{f_{12}(\lambda)\}$, $q_{12}(\lambda)=-\mathrm{Im}\{f_{12}\}$, then c_{12} and q_{12} are called the **cospectrum** and **quadrature spectrum**, respectively.

In the case of strictly stationary processes, there are higher order analogues of the spectrum known as **cumulant spectra of order k**, the spectrum being a "cumulant spectra of order 2". Given the strictly stationary r-vector valued time series $\mathbf{X}(t)$, $t=0, \pm 1, \pm 2, \ldots$ with components $\mathbf{X}_a(t)$, $a=1, \ldots r$, if

$$c_{a_1,\ldots,a_k}(t_1, \ldots t_k) \equiv \mathrm{cum}\{X_{a_1}(t_1), \ldots, X_{a_k}(t_k)\},$$

the **kth order cumulant spectrum** , $f_{a_1,\ldots,a_k}(\lambda_1, \ldots \lambda_k)$ is defined by

$$c_{a_1,\ldots,a_k}(t_1, \ldots t_k)=\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \exp\left\{ i \sum_{1}^{k} \lambda_j u_j \right\} f_{a_1,\ldots,a_k}(\lambda_1, \ldots \lambda_k)\, d\lambda_1 \ldots d\lambda_k.$$

Recall that if $X_1, \ldots, X_k$ are random variables, and $\phi(\alpha_1, \ldots \alpha_k) = E[\exp(i(\alpha_1 X_1 + \ldots + \alpha_k X_k))]$, then

$$\mathrm{cum}\,(X_1, \ldots, X_k) = \frac{\partial^k}{\partial\alpha_1 \ldots \partial\alpha_k} \log\,[\phi(\alpha_1, \ldots \alpha_k)]$$

where the partial derivative is evaluated at $(0, 0, \ldots 0)$.

For the results of this dissertation to hold, *it is assumed that the fourth order cumulant spectrum exists, but stationarity beyond fourth order cumulants is not required.*

3.3 The Periodogram: Motivation From Deterministic Time Series

One way of looking at the covariance function $\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\, d\lambda$ is observing that $\gamma(h)$ is simply the hth Fourier coefficient of $f(\lambda)$. Reversing this, we can write

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma(h).$$

Suppose now one had a function X(t), where $t \in \mathbb{Z}$ (integers) which could be written as

$$X(t) = \int_{-\infty}^{\infty} e^{it\lambda} f(\lambda)\, d\lambda.$$

Since X(t) is defined only on the integers, the analog of the Fourier transform $\int_{-\infty}^{\infty} e^{-it\lambda} x(t)\, dt$ would be $\frac{2\pi}{\sqrt{n}} d^{(n)}(\lambda)$, where

$$d^{(n)}(\lambda) \equiv \sum_{t=0}^{n-1} e^{-it\lambda} X(t) \qquad [3.3.1]$$

and it turns out by a central limit theorem that if X(t) is a stationary process, $d^{(n)}(\lambda)$ has an asymptotic (complex) normal distribution. But the mean of this random variable is 0, since the means of the X(t)'s are all 0. Nonetheless there is an important relationship between $d^{(n)}(\lambda)$ and the spectral density, which is that asymptotically variance($d^{(n)}(\lambda)$)=$f(\lambda)$. Hence if we consider the statistic

$$I_n(\lambda) = \frac{1}{2\pi n} |\, d^{(n)}(\lambda)|^2 \qquad [3.3.2]$$

which is known as the second order periodogram, $\lim_{n\to\infty} E(I_n(\lambda)) = f(\lambda)$. Furthermore, we

have the following theorem describing one aspect of its asymptotic behavior (Brockwell and Davis (1987), p. 337):

Theorem 10.3.2

If $f(\lambda)>0$ for all $\lambda \in [-\pi,\pi]$ and if $0<\lambda_1<\lambda_2< \ldots <\lambda_m<\pi$, then the random vector $(I_n(\lambda_1), I_n(\lambda_2), \ldots, I_n(\lambda_m))$ converges in distribution to a vector of independent and exponentially distributed random variables, the ith component of which has mean $f(\lambda_i)$, i=1, . . . , m .

It is this theorem which motivates the following: Define $F_n = \{j \in \mathbb{Z}: -\pi<\omega_j \equiv 2\pi j/n \leq \pi\} = \{-[(n-1)/2, \ldots, [n/2]\}$, where [x] denotes the integer part of x and $\mathbb{Z}$ denotes the positive and negative integers. Then $\{\omega_j: j \in F_n\}$ are called the **Fourier frequencies**. We may consider the set of random variables $\{I_n(\omega_j): j \in F_n\}$ to be independent, exponentially distributed random variables with $E(I_n(\omega_j)) = f(\omega_j)$. The sense in which this is true will be described later. But for now, let us simply observe that the periodogram is not a consistent estimator of the spectrum, and so we will have to find some other method of estimating it.

3.4 Kernel Estimates of the Spectral Density

If the spectrum is sufficiently smooth, the periodogram ordinates are approximately uncorrelated with means which shouldn't be too different, provided they are restricted to a sufficiently small interval. For any given interval, as n increases there will be an increasing number of Fourier frequencies in the interval. Hence one might expect to construct a consistent estimator for $f(\lambda)$ by taking some kind of (possibly weighted) average of the periodogram ordinates for the Fourier frequencies falling in the interval $[\lambda-\epsilon, \lambda+\epsilon]$ for some suitably chosen ϵ. Actually, it is possible to do better than this by making the estimator have

the form

$$\hat{f}(\omega_j) = \sum_{|k| < m_n} W_n(k)\, I_n(\omega_{j+k})$$

where $\{m_n\}$ is a sequence of positive integers (which presumably increases with n) and $W_n(\,.\,)$ is a sequence of weight functions. These are called "Kernel Estimators", and the asymptotic properties of such estimators are well known. In this dissertation we will seek to find parametric models for the spectrum, but as will be seen in chapter 9, kernel smoothing will play a role in a robust parametric estimation procedure. Before we can describe the parametric approach further, we must review some other properties of the spectrum.

3.5 The Wald Decomposition and the Innovation Variance: Kolmogorov's Formula

The random variables X(t), t=0, ± 1, ± 2, . . . are all square integrable, and so generate a Hilbert space H (the closure of all finite linear combinations). Let us define H_t to be the closed linear subspace of H generated by $\{X(s)\colon s \leq t\}$. From the standpoint of prediction, we are mainly interested in those stationary processes X(t) for which $X(t+1) \notin H_t$ (since otherwise, the projection of X(t+1) onto H_t will be itself and there is no statistical problem with prediction). A process which satisfies this condition is known as a **regular process**, and otherwise the process is known as a **singular process**. Every regular process has what is known as its **Wald decomposition**, that is, it can be written as

$$X(t) = \sum_{s=1}^{\infty} c(t-s)\, Z(s)$$

where Z(s) are uncorrelated elements of H (Rozanov (1967), p. 56).

The question arises, "Is there any restriction on the spectral density function $f(\lambda)$ so

that a process X(t) which has spectrum $f(\lambda)$ will be a regular process?". The answer is "yes", and the condition is given by the following.

Theorem 5.1 (Rozanov (1967), p. 64)

In order that the stationary process ξ be linearly regular, it is necessary and sufficient that it have an almost everywhere positive spectral density $f(\lambda)$ such that

$$\int_{-\pi}^{\pi} \log f(\lambda)\, d(\lambda) > -\infty.$$

If we have a regular process with spectrum $f(\lambda)$, the innovation variance (i.e. the variance of $X(t+1) - \hat{X}(t+1)$, where $\hat{X}(t+1)$ is the projection of $X(t+1)$ onto H_t) is given by what is known as "Kolmogorov's Formula", which is

$$\inf_P \int_{-\pi}^{\pi} |1+P(e^{-i\lambda})|^2 f(\lambda)\, d(\lambda) = 2\pi \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\lambda)\, d(\lambda)\right\}$$

(Rozanov (1967), p. 66). The infimum is taken over all polynomials P(z) such that P(0)=0. This shows the "frequency domain" squared distance from the constant function 1 to the subspace spanned by $\{e^{-i\lambda}, e^{-2i\lambda}, e^{-3i\lambda}, \ldots\}$ is given by the expression on the right side of the equation. By the isomorphism between the frequency domain space and the process space, this squared distance is also the variance of $X(t+1) - \hat{X}(t+1)$.

3.6 ARMA Models and their Motivation

Because most of the parametric spectral estimation research has been directed towards a special type of model, known as the "Autoregressive/moving average" or ARMA models, it is

necessary to explain what these models are and why they are so important.

An ARMA model for a (univariate) spectral density is simply

$$\sigma^2 \left| \frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})} \right|^2$$

where σ^2 is the innovation variance, and $\theta(x) = 1+\theta_1 x+\theta_2 x^2+ \ldots + \theta_q x^q$, $\phi(x) = 1-\phi_1 x-\phi_2 x^2- \ldots -\phi_p x^p$ are polynomials having no roots in the unit disk of the complex plane. Thus, the ARMA model is a squared rational polynomial on the unit circle in the complex plane. This definition may seem strange, and there is another interpretation of ARMA models in the "time domain" which is as follows. We can write an expression involving the time series and its Wald decomposition as

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = Z_t+\theta_1 Z_{t-1}+\theta_2 Z_{t-2}+ \ldots + \theta_q Z_{t-q}.$$

If the polynomial $\theta(x) \equiv 1$, this expression involves only the time series under observation and Z_t, and we can write $X_t=\phi_1 X_{t-1}+\phi_2 X_{t-2}+ \ldots +\phi_p X_{t-p}+ Z_t$. This shows the observation at time t may be viewed as a linear combination of the p previous observations plus a "random shock" Z_t. If the polynomial $\phi(x) \equiv 1$, we can write $X_t=Z_t+\theta_1 Z_{t-1}+\theta_2 Z_{t-2}+ \ldots + \theta_q Z_{t-q}$, and X_t may be viewed as a finite linear combination of uncorrelated random variables, which is known as a "moving average".

It is a fact (Brockwell and Davis (1987), p. 125) that squared rational polynomials are dense in the space of spectral densities corresponding to regular processes. Actually, rational polynomials of the form

$$\left| \frac{1}{\phi(e^{-i\lambda})} \right|^2 \text{ or } \left| \theta(e^{-i\lambda}) \right|^2$$

are also dense in the space of regular spectral densities, and hence theoretically if you put

enough coefficients in an autoregressive, moving average, or ARMA model, you should be able to fit to within sufficient accuracy any spectral density. Of course in practice, you want the model you fit to contain as few parameters as possible, and one would hope to choose an autoregressive, moving average, or ARMA model which does this and yet still fits the data reasonably well.

Another nice interpretation of ARMA models is to look at where the roots of the polynomials $\phi(x)$ and $\theta(x)$ lie. If $\phi(x)$ has a root near the unit circle (say near $e^{-i\lambda}$), then the spectral density will have a peak at λ. If $\theta(x)$ has a root near the unit circle (say near $e^{-i\lambda}$), then the spectral density will have a valley at λ. Unfortunately during the fitting process this knowledge is of no help, since the parametrization is done in terms of the polynomial coefficients and not the (complex) roots.

The time domain interpretation, however, will be of help in the fitting process. As pointed out by Box and Jenkins (1976), one can compare ACF's and PACF's calculated from the data to theoretical ACF's and PACF's of autoregressive, MA or ARMA processes in the model screening stage (see chapter 3 of Brockwell and Davis (1987) for definitions of these).

For multivariate processes, there is an analogous class of models known as **multivariate ARMA models**. The time domain representation of an m-dimensional ARMA process is

$$\mathbf{X}(t) - \Phi_1\mathbf{X}(t-1) - \ \ldots\ - \Phi_p\mathbf{X}(t-p) = \mathbf{Z}(t) + \Theta_1\mathbf{Z}(t-1) + \ \ldots\ + \Theta_q\mathbf{Z}(t-q)$$

where the Φ's and the Θ's are $m \times m$ matrices. There is a noniterative time domain method for obtaining parameter estimates in the autoregressive case (Yule-Walker equations), but in general due to the large number of parameters in these models likelihood type solutions are difficult to obtain.

Disadvantages of ARMA models include the fact they are "global" models; you cannot say how a change in a parameter will affect the shape of the fitted curve. ARMA models,

especially multivariate ARMA models, can be difficult to fit since you must have starting values for the parameters in order to do Newton-Raphson iterations. There may be many local maxima in the likelihood surface, so it is important to have consistent starting estimates. In the multivariate case, even assuming you can obtain these starting values the model is very complicated and difficult to work with from the practical viewpoint. Restrictions must be placed on the multivariate ARMA model to make the model unique within the parameter space.

3.7 Problems in Time Series Analysis

There are essentially two main problems in time series analysis, (1) estimation of the spectral density for its own sake, i.e. to see where "important" frequency ranges lie, and (2) prediction of future values of the time series. Let us discuss each of these.

In the univariate case, the spectrum tells us where most of the "power" in the time series is coming from (i.e. low or high frequencies). Knowledge such as this may be important in its own right, such as in the case where the series represents "noise" and an engineer would try to broadcast a radio signal on a frequency as far as possible from where most of the noise spectrum is concentrated. The case of multivariate series becomes more complicated, because there are other functions of the spectral density matrix which have nice interpretations. For example, three of these are the the phase spectrum ($\phi_{12}(\lambda)$=arg($f_{12}(\lambda)$), the group delay ($\phi'_{12}(\lambda)$, i.e. the derivative of the phase), and the squared coherency spectrum ($|f_{12}(\lambda)|/\sqrt{f_1(\lambda)f_2(\lambda)}$). Functions of the spectral density matrix are usually estimated by substituting a smoothed version of the periodogram for the true spectrum in the appropriate formulas. Of course, if we have a parametric model for the spectrum of the bivariate series, then we may use this model to estimate the functions of the spectral density matrix in a

similar way.

The second main problem in time series analysis is the prediction of future values. If the spectrum is known, then to predict X_t having observed $\{X_{t-1}, X_{t-2}, \ldots\}$, we simply construct the L^2 projection $\hat{X}_{t+1}$ of X_{t+1} onto H_t. Of course, if we have a model for the spectrum, then we will simply act as if it were the real spectrum when finding the projection $\hat{X}_{t+1}$. Sometimes, two univariate series, say $X_1(s)$ and $X_2(s)$ may be observed, which are the input and output of the transfer function model

$$X_2(s) = \sum_{j=0}^{\infty} t_j X_1(t-j) + N(t)$$

where $T=\{t_j, j= 0, 1, 2....\}$ is a linear filter and $N(t)$ is a stationary process uncorrelated with the input process $X_1(t)$. It turns out that the transfer function $T(e^{-i\lambda}) = \sum_{j=0}^{\infty} t_j e^{-ij\lambda}$ can be expressed as $T(e^{-i\lambda}) = \frac{f_{12}(\lambda)}{f_{11}(\lambda)}$, and if one is interested in predicting $X_1(t)$ having observed both series until time $t-1$, a better estimate can be obtained by projecting $X_1(t)$ onto the L^2 closure of the linear span of $\{X_1(t-1), X_1(t-2), \ldots\} \cup \{X_2(t-1), X_2(t-2), \ldots\}$. This projection can be done in practice if one has an estimate of the bivariate spectrum of both processes.

Chapter IV
Maximum Likelihood Estimation of Time Series Parameters

4.1 Three Asymptotically Equivalent Ways of Estimating Parameters

We would like to distinguish between two basic types of spectral models and discuss the differences in maximum likelihood estimation for the two types.

The two basic categories of spectral models are (1) those which have the innovation variance as a parameter, and (2) those which do not. This dissertation is concerned mainly with the second category, but theory developed for these models may also be applied to the first category. Thus it is necessary to see how a separately parametrized innovation variance simplifies things in some cases, and results in differences in others.

Using as motivation the previously mentioned theorem from Brockwell and Davis (1987) which gives the asymptotic distribution of the periodogram (the periodogram ordinate at λ becoming exponentially distributed with mean $f(\lambda)$, the spectral density function), we estimate θ where the model is $\{f_\theta\}$ by minimizing

$$\sum \log(f_\theta(\lambda_i)) + \frac{I(\lambda_i)}{f_\theta(\lambda_i)} \qquad [4.1.1]$$

summing over the Fourier frequencies. Of course, this function is the negative log likelihood function assuming the "observations" $I(\lambda_i)$ are exponentially distributed (which is the link

between spectral estimation and GLIM). As previously mentioned, the idea of using the exponential likelihood function in this manner instead of the true likelihood of the series is quite old, dating back to Whittle's work in the 1950's. It is further suggested that one might view frequency domain spectral estimation as *estimating a one dimensional generalized model response surface over the design space [0,π]* (or a subset of [0,π], see the next section). This "design" aspect of spectral estimation will be crucial in obtaining rigorous proofs of results which do not exist in the general GLIM context.

On the other hand, the Gaussian likelihood of the vector of observations $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)'$ is given by

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} |G_n(\beta)|^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2} \mathbf{X}_n' G_n^{-1}(\beta) \mathbf{X}_n\} \qquad [4.1.2]$$

where $G_n(\beta) = \sigma^{-2} \Gamma_n(\beta)$ and $\Gamma_n(\beta)$ is the covariance matrix of $\mathbf{X}_n$. Three asymptotically equivalent estimators are found as follows:

(a) $\hat{\beta}_n$ minimizes

$$l(\beta) = \frac{\mathbf{X}_n' G_n^{-1}(\beta) \mathbf{X}_n}{n} + n^{-1} \ln \det(G_n(\beta)) \qquad [4.1.3]$$

and then

$$\hat{\sigma}^2(\hat{\beta}_n) = \frac{\mathbf{X}_n' G_n^{-1}(\hat{\beta}) \mathbf{X}_n}{n} \qquad [4.1.4]$$

which is full ML estimation of parameters.

(b) $\tilde{\beta}$ minimizes

$$\tilde{\sigma}^2(\beta) = \frac{\mathbf{X}_n' \, G_n^{-1}(\hat{\beta}) \, \mathbf{X}_n}{n} \qquad [4.1.5]$$

which is then taken to be the innovation variance. This is called the "least squares" estimator for obvious reasons.

(c) $\bar{\beta}$ minimizes

$$\tilde{\sigma}^2(\beta) = n^{-1} \sum \frac{I_n(\lambda_i)}{g(\lambda_i, \beta)} \qquad [4.1.6]$$

which is then taken to be the innovation variance, where

$$g(\lambda, \beta) = \frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2} \, . \qquad [4.1.7]$$

Chapter 10, section 8 of Brockwell and Davis (1987) shows the consistency and asymptotic normality of these three estimates, and also shows that they are asymptotically equivalent.

Since [4.1.6] looks like half of [4.1.1], let's concentrate on this estimator. Why does the first term of [4.1.1] disappear when the innovation variance is separately parametrized? Why can't you make $\tilde{\sigma}^2(\beta)$ as small as you want by making $g(\lambda, \beta)$ large? The answer lies in the way the model is parametrized. The model is

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q} \qquad [4.1.8]$$

where $Z_t \sim \text{IID}(0, \sigma^2)$, $\beta = (\phi_1, \ldots \phi_p, \theta_1, \ldots, \theta_q)$, and the parameter set is $\{\beta \varepsilon \mathbf{R}^{p+q}$: $\phi(z)\theta(z) \neq 0$ for $|z| \leq 1$ and $\phi(\,.\,)$, $\theta(\,.\,)$ have no common zeros$\}$. As Brockwell and Davis (1987) observe (p. 366), β can be expressed as a continuous function $\beta(a_1, \ldots a_p, b_1, \ldots b_q)$ of

the zeros $a_1, \ldots a_p$ of $\phi(\ .\ )$ and $b_1, \ldots, b_q$ of $\theta(.\ )$. Since these functions have no roots in the unit circle, $|a_i|>1$ and $|b_i|>1$. Hence to make $\sigma^2(\beta)$ small, you would need to, say, make $|\theta(z)| = |\prod(1 - \frac{1}{a_i} z)|$ $(z=e^{-i\lambda})$ large, which would be done by making the $|a_i|$'s small. . . but they've got to be bigger than 1! This is a rather informal argument, but a more rigorous one can be constructed from a theorem known as the "Toeplitz Asymptotic Homomorphism" which can be stated as follows: If the spectral density f is strictly positive, then the one step prediction error σ_f^2 is given by

$$\log \sigma_f^2 = \lim_{n\to\infty} \frac{1}{n} \log \det T_n(2\pi f) = \frac{1}{2\pi} \int \log 2\pi\ f(\lambda)\ d\lambda$$

where $T_n(g)$ is the covariance matrix calculated using g as the spectral density (Azencott and Dacunha-Castelle (1986), p113). So, if the model is separately parametrized, the first term in [4.1.1] would (asymptotically) be replaced by σ^2, which then plays no role when the partial derivatives are taken with respect to the other parameters. This means only the second term is really used in the determination of parameters.

4.2 Some Examples of GLIM Parametrizations

The most commonly used spectral density models, the ARMA models, unfortunately do not fall in the GLIM category and instead must be viewed as generalized models. There are, however, two models of lesser popularity which are GLIM models. These are known as Kolmogorov's turbulence model and Bloomfield's exponential model (Bloomfield (1973)).

According to Kolmogorov (1941), the spectral density for turbulence in a fluid may be written as $f_c(\omega) = c\ \omega^{-5/3}$. In practice, however, the ideal conditions required for this formula to hold are rarely met. Instead, a model must be fit which views the power as a parameter;

i.e. the model is $f_{\alpha, c}(\omega) = c\ \omega^{\alpha}$. Rewrite this as exp $\{\log c + \alpha \log \omega\}$. This may be fit with the GLIM package by letting the x variate be the frequencies, using the log link (Cameron and Turner (1988)).

In 1973, Bloomfield proposed the model

$$2\pi f(\omega; \theta) = \sigma^2 \exp\left\{2\sum_{j=1}^{p} \theta_j \cos j\omega\right\}.$$

Letting $\theta_0 = 1/2 \log \frac{\sigma^2}{2\pi}$, we can rewrite this as

$$f(\omega; \theta) = \exp\left\{2\sum_{j=0}^{p} \theta_j \cos j\omega\right\}$$

and estimates can be obtained using gamma error with log link and performing the regression on the p variables $2 \cos k\omega_j$, $k = 1 \ldots p$.

There are also other possibilities for GLIM spectral models, such as regression splines (Smith (1979)), which will not be discussed here but might be applicable to spectral analysis.

4.3 Why Fit Models Over Frequency Bands?

A central motivational idea for this dissertation which differentiates it from most of the time series parameter estimation literature is *fitting spectral density models over frequency bands.* This is hinted at in the preceding section, where the relationship between the actual log likelihood of the series and the log likelihood of exponentially distributed random variables with "design matrix" a vector of Fourier frequencies is pointed out. If it is desired to fit a model where some periodogram ordinates have been excluded from the analysis, all that is necessary is to remove the appropriate frequencies from the sum in the log exponential

likelihood (of course, you'll have a more difficult time doing this using the *true* likelihood rather than the *approximate* log exponential likelihood). This fact seems to have been discussed for the first time by John Rice (1979). One might naturally be inclined to ask the following question: "Since in the final analysis, you need an estimate of the entire spectrum (i.e. to make predictions, etc.), why would anyone be interested in fitting a model to just a subset of the spectrum?" There are two main points to be made in response.

First, in the "exploratory" stage of an analysis it might be desirable to fit simple parametric models (such as splines, for example) to segments of the periodogram, rather than immediately trying to fit a very complex model to the entire periodogram. This is perhaps even more relevant in the case of a multivariate time series, where even "simple" parametric models probably aren't. A preliminary "segmented" approach would perhaps give an intuitive idea of what is going on, using parametric GLIM type model diagnostics and screening rather than nonparametric smoothing techniques, before attempting to fit the ultimately desired spectral model.

Second, there is the "contaminated series" problem, an example of which appears in Cameron and Turner (1987). The data consist of 300 observations taken in one second of the EEG record obtained from a steer. It was desired to fit an autoregressive model to the data, but it was known from graphing the periodogram that a spike existed at 50 hz *due to the oscillation of this frequency in the electrical supply and its influence on the recording equipment.* Although not stated in Cameron and Turner's paper, one might view the model as being $Y(t)=X(t)+N(t)$, where $X(t)$ and $N(t)$ are uncorrelated processes, $Y(t)$ is the observed process, $X(t)$ is the "true" process, and $N(t)$ is the "noise" process due to the power supply having a spectrum concentrated in a narrow band around 50 hz. It does not make sense to ignore the fact that $Y(t)$ is observed, not $X(t)$, and the way to correct for this in the analysis is

to eliminate the periodogram ordinates in a small frequency band around 50 hz, even though the spectral density model (for X(t)) is to be used for the entire spectrum. It should be pointed out that this raises an interesting "design" question: the recording equipment could probably be altered so that the power frequency translates into a data contaminant at *any* frequency. Given that some frequency band must be contaminated, what would be the "best" frequency to contaminate so as to "best" estimate the parameters of the spectral density model? Such questions will not be considered here, but are a topic for future research. Other references to the "contamination problem" occurring in practice are given in Rice (1979).

4.4 Taniguchi, Chiu, Kulperger

There are five papers in the time series field which most closely relate to this dissertation, written by Taniguchi, Chiu, and Kulperger. As such, we would like to outline the relevant parts of these results.

Masanobu Taniguchi, a Japanese applied mathematician, published two related papers; the first in 1979 and the second (with Hosoya) in 1982. The main purpose of the 1982 paper was to extend some of the 1979 results to the non Gaussian and multivariate cases. The non Gaussian results are complicated by the fourth cumulant spectrum, see chapter 5 of the dissertation.

One of the central ideas in these papers was to describe what $\widehat{\theta}$ obtained by "maximum likelihood" is estimating asymptotically if the model $\{f_\theta\}_{\theta \in \Theta}$ *does not necessarily include the true spectral density function g.* He observed that in this case, "maximum likelihood" may not be the "best" way to estimate parameters, i.e. there are alternative parameter estimation techniques which may result in a smaller MSE if bias is also taken into consideration, and went on to describe one such technique (in chapter 9, we will give another

technique). Although not observed by Taniguchi, his work is trivially extended to the case where we are fitting a model over frequency bands instead of $[-\pi, \pi]$.

Before we give specific results from Taniguchi, let us observe that the periodogram may be regarded as a function on $[-\pi, \pi]$ because it may be extended as follows (see Brockwell and Davis (1987), p. 333).

Method 1

For any $\lambda \in [-\pi, \pi]$, define

$$I_n(\lambda) = \begin{cases} I_n(\lambda_k) & \text{if } \lambda_k - \pi/n < \lambda \leq \lambda_k + \pi/n \text{ and } 0 \leq \lambda \leq \pi, \\ I_n(-\lambda) & \text{if } \lambda \in [-\pi, 0) \end{cases} \qquad [4.4.1]$$

With the periodogram viewed as a function, we can then "step functionize" the (continuous) spectral density f in a similar way, and write

$$\int_{-\pi}^{\pi} \log f(\lambda) + \frac{I_n(\lambda)}{f(\lambda)} \, d\lambda \quad \text{instead of } \frac{2\pi}{n} \sum \log f(\lambda_i) + \frac{I_n(\lambda_i)}{f(\lambda_i)} .$$

Heuristically, the reason this is so is that the later expression may be viewed as a "Riemann sum" of the former integral. On the other hand, the "natural extension" is:

Method 2

Define $I_n(\lambda)$ as being

$$I_n(\lambda) \equiv \frac{1}{n} \left| \sum X_i \, e^{-i\lambda} \right|^2 \qquad [4.4.2]$$

for all $\lambda \in [-\pi, \pi]$. Then the "likelihood" function is exactly the integral above.

In the statement of his theorems, Taniguchi extends his periodograms to $[-\pi, \pi]$ using

method 2, but the proofs of these theorems are given for the extension done the other way. Actually, the way in which the periodogram is extended is irrelevant, as the main results hold in both cases. Let us conclude this short discussion of the periodogram by saying that viewing the periodogram as a function will be an extremely important concept in the sequel, but it doesn't really matter how the extension is obtained.

Taniguchi (1979) defines a "distance measure" D(. , .) on the space of spectral density functions as follows. For two spectral densities f and g (positive functions defined on [-π, π]) let

$$D(f, g) = \int_{-\pi}^{\pi} \log f(\lambda) + \frac{g(\lambda)}{f(\lambda)}\, d\lambda.$$

Note that this function is not a metric or semimetric on the space of densities as it is not symmetric and does not satisfy the triangle inequality.

If we have a model $\{f_\theta\}$, for each function g, define T(g) to be the value of θ minimizing $D(f_\theta, g)$. Then using this definition, $\hat{\theta}_n = T(I_n)$, where $\hat{\theta}_n$ is the "maximum likelihood" estimator. If the model contains the true spectral density g (which equals, say, f_{θ_0}), then θ_0 minimizes $D(f_\theta, g)$ and $\hat{\theta}$ is estimating θ_0. If the model does not contain the true spectral density g, then $\hat{\theta} = T(I)$ asymptotically is estimating $\theta_0 = T(g)$ (i.e. the "closest" density in the model space to g according to the "distance" D(. , .)). Taniguchi goes on to define "weak convergence" in the space of spectral densities as follows: If, for every continuous function $\psi(x)$,

$$\int_{-\pi}^{\pi} \psi(x)\, g_n(x)\, dx \rightarrow \int_{-\pi}^{\pi} \psi(x)\, g(x)\, dx$$

then we say that g_n **converges to g weakly**, denoting this by $g_n \xrightarrow{w} g$. Viewed in light of this definition, a previously existing theorem of Brillinger says that the periodogram converges

weakly in probability to g. Taniguchi then gives an asymptotic "linear representation theorem", which can be stated as follows.

Theorem 2 (Taniguchi (1979), p. 577)

Suppose T(g) exists uniquely and lies in int(Θ), and

$$\int_{-\pi}^{\pi} \left. \frac{\partial^2 f_\theta^{-1}}{\partial\theta\partial\theta'} g(x) + \frac{\partial^2 \log f_\theta(x)}{\partial\theta\partial\theta'} \right|_{\theta = T_1(g)} dx$$

is a nonsingular matrix. Then for every sequence of spectral densities $\{g_n\}$ satisfying $g_n \xrightarrow{w} g$, we have

$$T(g_n) = T(g) + \int_{-\pi}^{\pi} \rho_g(x)(g_n(x) - g(x))\, dx + a_n \int_{-\pi}^{\pi} \left. \frac{\partial f_\theta^{-1}}{\partial\theta} \right|_{\theta = T_1(g)} (g_n(x) - g(x))\, dx$$

where $\rho_g(x) = \left[\int_{-\pi}^{\pi} \left. \frac{\partial^2 f_\theta^{-1}}{\partial\theta\partial\theta'} g(x) + \frac{\partial^2 \log f_\theta(x)}{\partial\theta\partial\theta'} \right|_{\theta = T_1(g)} dx \right]^{-1} \left. \frac{\partial f_\theta^{-1}}{\partial\theta} \right|_{\theta = T_1(g)}$

and a_n is a real $p \times p$ matrix which goes in probability to 0 as $n \to \infty$.

This theorem yields that the vector $\sqrt{n}\,(T(I_n) - T(g))$ converges in distribution to a multivariate normal random variable, and the vector $(T(I_n) - T(g))$ converges in probability to 0. It should be compared with the following comment from McCullagh (1983, p62): "It can be shown that among all estimators of β for which the influence function is linear, i.e. estimators satisfying $\hat{\beta} - \beta_0 = L_\mu(Y - \mu) + o_p(N^{-1/2})$, where L_μ is a $p \times N$ matrix of influences, quasi likelihood estimators have minimum asymptotic variance".

There are two central concepts in this paper we will extend. The first is viewing

"maximum likelihood" as minimizing a distance function with the true means function (the spectral density) replaced with the "observations" (periodogram), and being able to identify what maximum likelihood is estimating when the true spectral density is not in the model in "geometric" terms as the closest function to the spectral density. The "Taniguchi distance" $D(f_{\theta_0}, g)$ is really a measure of bias, and the deviance is really estimating a function of $D(f_{\theta_0}, g) - D(g, g)$. The second main idea to be extended is being able to represent $\hat{\theta}_n - \theta_0$ as a linear function of the periodogram minus the true spectral density.

Taniguchi does not derive any "optimality" results, but simply finds the asymptotic variance of his estimator and points out that in the Gaussian case it is the same as that obtained by maximizing the true Gaussian likelihood.

Chiu (1988) is the first to give a theorem concerning IRWLS estimates of the periodogram. The following appears in his 1988 paper concerning the estimate $\hat{\theta}$ minimizing

$$\sum \frac{1}{f^2(\hat{\eta}, \lambda_i)} (f(\theta, \lambda_i) - I(\lambda_i))^2 \qquad [4.4.3]$$

where $f(\theta, \lambda)$ is a model containing the true spectrum.

Theorem 7 (Chiu (1988), p. 1321)

Let $\phi(\lambda) = \psi(\lambda)/f^2(\lambda, \eta_0)$, and suppose $\eta_T \rightarrow \eta_0$ a.s., and $\sqrt{T}(\hat{\eta}_T - \eta_0)$ converges in law. Then (under some standard regularity assumptions) $\hat{\theta}_T$ is strongly consistent and $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is asymptotically normal with mean 0 and covariance matrix $2A^{-1}BA^{-1} + A^{-1}DA^{-1}$, the elements (j, k) of A, B, and D being

$$a_{jk}(\theta_0) = (2\pi)^{-1} \int \phi(\lambda)\, g_j(\lambda, \theta_0)\, g_k(\lambda, \theta_0)\, d\lambda$$

$$b_{jk}=(2\pi)^{-1}\int \phi^2(\lambda)\, f^2(\lambda, \theta_0)\, g_j(\lambda, \theta_0) g_k(\lambda, \theta_0)\, d\lambda$$

$$d_{jk}= (2\pi)^{-1}\int \phi(\lambda)\phi(\mu)\, f_4(\lambda, -\lambda, \mu)\, g_j(\lambda, \theta_0)\, g_k(\lambda, \theta_0)\, d\lambda\, d\mu \qquad [4.4.4]$$

respectively, where $g_j(\lambda, \theta)= \frac{\partial f(\lambda,\theta)}{\partial \theta_j}$ and $f_4(\lambda, -\lambda, \mu)$ is the fourth cumulant spectrum of the process.

Similar to Taniguchi, Chiu does not give any optimality theorem proofs. He simply observes that if the time series is Gaussian and the correct weighting function (i.e $1/f^2(\theta_0, \lambda)$) is used, then $D \equiv 0$ and the variance matrix is asymptotically the same as that obtained by maximizing the true likelihood ($2A^{-1}$). Chiu's theorem shows that IRWLS yields this same optimal variance, since one may first obtain a consistent, but not "optimal variance" estimate $\hat{\eta}$ of θ_0 by minimizing $\sum \phi(\lambda_i)\, (f_\theta(\lambda_i) - I(\lambda_i))^2$ for an arbitrary function $\phi(\lambda)$, and then using this preliminary estimate to find the $\hat{\theta}$ minimizing [4.4.3].

Chiu (1990) also considers the problem of "sensitivity" of parametric estimates to perturbed periodic components of a time series. In this paper, he examines series of the form $Y(t)=\mu+S(t)+X(t)$, where $X(t)$ is a stationary, Gaussian time series with spectrum $f(\lambda, \theta_0)$, and

$$S(t)=\sum_{k=1}^{K} R_k(t)\, \cos(\omega_k t+\phi_k(t)).$$

In this expression, there are K periodic components, each at a frequency ω_k. $R_k(t)$ and $\phi_k(t)$ are functions that "change slowly over time t".

Chiu (1990) essentially proposes using classical techniques on a modified periodogram $\tilde{I}(\lambda)$, where $\tilde{I}(\lambda) \equiv \rho\ g\{\tilde{I}(\lambda)/f(\lambda, \tilde{\theta})\}\ f(\lambda, \tilde{\theta})$, $\tilde{\theta}$ is an estimate of θ_0, $g \geq 0$ and satisfies a

smoothness and integrability condition, and ρ is a constant to adjust the expectation of $\tilde{I}(\lambda)$ approximately to $f(\lambda)$. For example, theorem 4 (Chiu, (1990)) shows the asymptotic variance of the $\hat{\theta}$ minimizing

$$Q_T(\theta)=1/T \sum \psi(\lambda) (\tilde{I}(\lambda) - f(\lambda, \theta))^2$$

where $\psi(\lambda)$ is a weighting function. Section 4 shows if "the 'energy' of the periodic components does not spread away from the peaks too much, then the estimates based on $\tilde{I}_Y(\lambda)$ have the same asymptotic properties (e.g. variance) as (those based on $\tilde{I}_X(\lambda)$) in theorem 4". An IRWLS procedure is proposed as in the 1988 paper, the justification of which is based on theorem 4.

The problem discussed in Chiu (1990) is similar to the "contamination" problem of chapter 9, but the proposed remedies of the paper and this dissertation are quite different. Interestingly, Chiu notes that "though the problem of estimating the parameters for the case $S(t)=0$ has been studied extensively, there are relatively little researches concerning the situation of $S(t) \neq 0$". While this dissertation does not base itself on the problem as posed by Chiu (e.g. we will have a "noise series" $X_N(t)$ which has a spectrum in place of Chiu's "S(t)"), it does offer an alternative look at the same basic type of problem.

Kulperger (1985) gives an "asymptotic optimality" result as follows: He considers the class of functions of the form

$$S_N(\theta) = \frac{1}{N} \sum_{j=1}^{N-1} h_1(\lambda_j, \theta)+h(\lambda_j, \theta)\, I_n(\lambda_j)$$

where h_1 and h are twice continuously differentiable with respect to θ and $\frac{\partial h_1}{\partial \theta} = -2\pi f \frac{\partial h}{\partial \theta}$. His main theorem is:

Theorem 3.2 (Kulperger, (1985)) For a stationary ARMA(p, q) process, the best Gaussian estimate is when $h(\lambda, \theta) = -(2\pi f(\lambda, \theta))^{-1} \Rightarrow h_1(\lambda, \theta) = -\log f(\lambda, \theta)$, in the sense that if $\hat{\theta}_N$ is the asymptotic likelihood estimate, and $\hat{\hat{\theta}}_N$ is any other Gaussian estimate, then

$$\lim_{N\to\infty} N\,(\mathrm{var}(\hat{\hat{\theta}}_N) - \mathrm{var}(\hat{\theta}_N)) \text{ is nonnegative definite.}$$

Note that Kulperger doesn't 1) give an optimality theorem for non-ARMA processes 2) give an optimality theorem over frequency bands 3) say what happens in the case of model misspecification 4) seem to be aware of the connection to McCullagh's (1983) QL functions, and doesn't use the same viewpoint as McCullagh. For example, he does not separate the model from the "variance" function as does McCullagh, or identify a "variance" function as such. A theorem to be given in chapter 6 of this dissertation will essentially extend the class of estimating functions considered by Kulperger (1985) (i.e. what he calls "Gaussian" estimating functions) and give a procedure yielding optimal parameter estimates within the new class. For non-Gaussian processes, it will be seen that using "Gaussian" estimating functions is inefficient.

4.5 Discussion and Goals of Dissertation: Why Generalized Models?

Most of the literature on parametric estimation in time series has concentrated on fitting "separately parametrized innovation variance" (e.g ARMA) models for the spectrum, assuming the model has been correctly specified. The Generalized Models approach to spectral analysis may be viewed as trying to find a one dimensional response surface when the "observations" (periodogram ordinates) are exponentially distributed and there are

observations at a uniform grid of points (the Fourier Frequencies) over the model space. As more observations are taken of the time series, the Fourier frequencies become closer together yielding more information about the surface. Because of this special situation, we are able to give new interpretations which combine ideas in both the GLIM and response surface areas.

One of the main ideas in the dissertation is that the periodogram should be viewed as a function in L^2 regardless of how defined, as a step function with steps at the Fourier frequencies, or by the natural definition. This is implicit in Taniguchi's papers, e.g. when he defines the "distance" $D_1(f, g)= \int \log(f)+g/f \, d\lambda$, but its ramifications have not been developed. One major consequence of the "functional" approach to spectral estimation is that instead of having a "variance matrix" for the observation vector, we will have a "variance operator" on L^2. If the time series is Gaussian, the variance operator will correspond to a diagonal matrix. Otherwise, it will take on a different form. Chapter 6 will elaborate on this theme and gives a new definition of QL function. Specifically, a "quasi likelihood function" will be defined as a function $D(\, . \, , \, . \,)$ with certain properties which maps two functions f and g to a real number $D(f, g)$. D_1 defined above will be but one example of a QL function under the extended definition. If the observed series is Gaussian and the model is correctly specified, obtaining parametric estimates by minimizing $D_1(f_\theta, I_n)$ is optimal with respect to asymptotic variance. If the model is not correctly specified due to "contamination", estimates obtained by minimizing D_1 will be asymptotically biased, which leads us to attempt to define what is meant by "contamination".

As has been already observed (e.g. Chiu (1988)), the generalized model approach offers the advantage of being able to estimate the spectrum when some frequency ranges are "contaminated" in the sense that the model is not valid there. The dissertation alters the conditions on a "contamination" problem: we observe the series $Y(t)=X(t)+N(t)$, but are

interested in estimating the spectrum of X(t). Chiu (1988) shows his IRWLS procedure works if the spectrum of Y(t) satisfies a rather strong "smoothness" condition: $\sum |t|\ |\gamma(t)| < \infty$ (a sufficient condition is that the spectrum have a continuous derivative satisfying a Lipschitz condition), and we exclude the "contaminated" bands (due to N(t)) from the analysis. It seems reasonable that "noise", even if limited to certain frequency bands, would not necessarily have a continuously differentiable spectrum. It also seems reasonable that we might not have a clear idea of exactly which frequency bands are affected by the noise series, or maybe the influence of N(t) can't be restricted to "bands". The dissertation will instead assume that X(t) has a covariance function satisfying $\sum |t|\ |\gamma(t)| < \infty$, while the spectrum of N(t) is only of bounded variation. Under these conditions, an IRWLS procedure is given in which it is not necessary to precisely know which frequency bands are contaminated. Traditional IRWLS requires the specification of the variance matrix for an observation vector, our version requires the specification of the variance operator. Traditional IRWLS also assumes the model is correct, and estimates the true parameters with the same variance as maximizing the QL function corresponding to the variance matrix of the observation vector. Our version of IRWLS shows that to eliminate the bias due to contamination, one may deliberately use the "wrong" QL function, i.e one which would not give optimal variance were the model correct, and hence the "wrong" variance operator, in order to eliminate bias. If correctly chosen, even assuming the model were correct the variance of parametric estimates should be close to optimal.

4.6 Literature Review of Time Series and General Parametric Estimation

As in the area of generalized linear models, there is a voluminous literature on time

series, so that only the papers which are most closely related to the dissertation will be mentioned. The first results on parametric estimation of time series parameters were given by Whittle (1951, 1953), which discussed hypothesis testing and parametric estimation in time series analysis, but did not tend to give rigorous proofs. Walker (1964) attempted to correct this situation by showing the asymptotic properties (consistency, normality) of the "least squares" estimators. Ibragomov (1966) gave results about properties of maximum likelihood type parametric estimates similar to those of Walker. Dzhaparidze (1974) gave a variant of a "reweighted least squares" method for solving the Quasi likelihood equations

$$\int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_j} \left[\log f(\lambda, \theta) + \frac{I_n(\lambda)}{f(\lambda, \theta)} \right] d\lambda = 0$$

and showed its "optimality" among solutions obtained by using "Gaussian" type likelihood approximations even when the series is not Gaussian. Kulperger's paper is in some ways an extension of Dzhaparidze's. Another of the earlier "optimality" papers is Davies (1973), in which the author discussed asymptotic inference in stationary Gaussian time series, and proved the asymptotic optimality of parametric estimates for Gaussian stationary processes using the full likelihood function. This was done by showing time series parametric estimation satisfies the assumptions of Le Cam's (1969) asymptotic optimality theory. Most authors (e.g. Chiu, Taniguchi) prove the optimality of their procedures by finding their asymptotic variance matrix and showing it is the same as Davies'.

The 1970's saw a refinement of the earlier theory, relaxing conditions on the series and showing the usual inferential theory continues to hold. Along these lines are Hannan (1973), which proved a CLT for univariate linear time series, Dunsmuir and Hannan (1976), and Dunsmuir (1979). The latter two papers essentially extend Hannan's (1973) work to vector

processes. Dunsmuir's (1979) paper removes the separately parametrized innovation variance assumption in the 1976 paper.

In the late 1970's, papers utilizing the "GLIM viewpoint" (although the authors usually did not recognize it as such) began to appear. For example, Robinson (1978) seems to be one of the first to discuss fitting univariate spectral models where the innovation variance is not separately parametrized, giving the asymptotic variance properties of the "ML" estimate of θ minimizing $\int \log f_\theta(\lambda) + I(\lambda)/f_\theta(\lambda)\, d\lambda$. Rice's (1979) is among the first papers published in a statistical journal which discuss different methods of parametric spectral estimation over frequency bands (of not necessarily separately parametrized innovation variance models) when the observed series is "contaminated".

The Taniguchi papers (1979) and Hosoya and Taniguchi (1982) have already been mentioned (and will be again in chapter 5) as main background for the dissertation. They are a periodogram based approach to parametric spectral estimation, and the second paper is essentially an extension of Taniguchi's 1979 results to vector valued and non Gaussian processes. Another interesting paper which examines periodogram based (rather than true likelihood) parametric estimation procedures is Dahlhaus' (1988). Dahlhaus' results suggest tapering the series before calculating the periodogram as a method of reducing bias, although the dissertation will not consider this topic.

Explicit links to generalized (linear/nonlinear) models in the time series literature do not appear until the late 1980's with Cameron and Turner (1987), an applied paper showing details of applying IRWLS to ARMA processes, but giving no theorems or proofs, and Chiu (1988, 1990) which have been previously discussed. Kulperger (1985) might also be considered as falling in this category, as it examines the consequences in terms of parameter variances of fitting correct spectral models using an incorrect QL function (analogous to using the wrong

likelihood function in GLIM). As previously mentioned, this idea is one of the main themes of the dissertation to be developed in chapter 9.

Chapter V
Estimation of Cospectra in Multivariate Time Series

5.1 Introduction

In their 1982 paper, Hosoya and Taniguchi extended the univariate theory of Taniguchi (1979) as described in chapter 4 to the multivariate case. In this chapter, we will take a closer look at the "function oriented" theory for multivariate processes.

For any two pointwise positive definite matrices of functions $f(\omega)$ and $g(\omega)$, define

$$D(f, g) \equiv \int_{-\pi}^{\pi} \log \det f(\omega) + \text{trace } (f^{-1}(\omega)\, g(\omega)\,)\, d\omega. \qquad [5.1.1]$$

Let $\{f_\theta\}_{\theta \in \Theta}$ be a family of positive definite spectral density matrices. Define $T(f)$ as a natural extension of the way it was previously defined, that is, for any positive definite matrix valued function g, $T(f)$ is the θ minimizing $D(f_\theta, g)$. If the true spectral density matrix is g and $\hat{\theta}_n = T(I_n)$, then $\hat{\theta}_n$ is estimating $\theta_0 = T(f)$ and $\hat{\theta} - \theta_0 \rightarrow N(0, W)$ where the periodogram matrix I_n is extended to a matrix of functions on $[-\pi, \pi]$ as in [4.4.1]. W has the form $M_f^{-1}\tilde{V}\, M_f^{-1}$, where

$$M_f = \int_{-\pi}^{\pi} \left[\frac{\partial^2}{\partial\theta\partial\theta'} \text{tr}\{f_\theta(\omega)^{-1} f(\omega)\} + \frac{\partial^2}{\partial\theta\partial\theta'} \log \det f_\theta(\omega) \right]_{\theta = T(f)} d\omega \qquad [5.1.2]$$

$$\tilde{V}_{jl} = 4\pi \int_{-\pi}^{\pi} \text{tr} \left[f(\omega) \frac{\partial}{\partial\theta_j} \{f_\theta^{-1}(\omega)\} f(\omega) \frac{\partial}{\partial\theta_l} \{f_\theta^{-1}(\omega)\} \right]_{\theta = T(f)} d\omega +$$

$$2\pi \sum_{r,t,u,v=1}^{8} \int\int_{-\pi}^{\pi} \left\{ \frac{\partial}{\partial\theta_j} f_\theta^{(r,t)}(\omega_1) \frac{\partial}{\partial\theta_l} f_\theta^{(u,v)}(\omega_2) \right\}_{\theta = T(f)}$$

$$\times k_{ra}(-\omega_1) k_{tb}(\omega_1) k_{uc}(-\omega_2) k_{cd}(\omega_2) \tilde{Q}^e_{abcd}(-\omega_1, \omega_2, -\omega_2)\, d\omega_1\, d\omega_2. \qquad [5.1.3]$$

The observed series Z(t) is a linear process of the form $Z(t)=\sum_{j=0}^{\infty} G(j)\, e(t-j)$, where t is an integer, the Z(t)'s are s dimensional vectors for each t, and the e(t)'s are p dimensional vectors for each t having the properties that E(e(t))=0 and $E(e(t)\, e(t)') = \delta(m, n)\, K$. Here, $\delta(m, n) = 0$ for $m \neq n$ and 1 otherwise, and K is a nonsingular $p \times p$ matrix. The G(j)'s are $s \times p$ matrices, and it is assumed that $\sum_{j=0}^{\infty} \text{tr}\, G(j)KG(j)' < \infty$. Under these conditions, {Z(t)} is a second order stationary process having a spectral density matrix $f(\omega)$ which is representable as $f(\omega) = \frac{1}{2\pi} k(\omega)\, K\, k(\omega)'$, $-\pi \leq \omega \leq \pi$, where $k(\omega)=\sum_{j=0}^{\infty} G(j)e^{i\omega j}$. The $k_{ij}(\omega)$'s in [5.1.3] are components of $k(\omega)$. $\tilde{Q}^e_{abcd}$ in [5.1.3] is the fourth order cumulant spectrum of {e(n)}, see section 3.2 for definition.

Chapter 7 will give a more enlightening form of the matrices [5.1.2] and [5.1.3], separating first and second derivatives, among other things, in the context of a more general theorem and definition of multivariate QL functions so that expressions such as these may be more effectively studied in the case of model misspecification.

Hosoya and Taniguchi (1982) argue that the parametric estimate $\hat{\theta}$ given by minimizing $D(f_\theta, I_n)$, where I_n is the multivariate periodogram, is asymptotically optimal if the model is correct. They do this by showing that it has the same asymptotic variance as the estimator obtained from minimizing the true likelihood function.

We would like to pose four basic spectral estimation problems, problem three being the main focus of this chapter. The last problem will be discussed in chapter 6, which will also give a general theorem answering all of the questions raised here.

1) For the Gaussian ARMA case, Kulperger (1985) has shown the asymptotic optimality property of minimizing $D(f_\theta, I_n) = \int_{-\pi}^{\pi} \log f_\theta + \frac{I_n}{f_\theta}\, d\lambda$. What happens with regard to optimality of parametric estimates if the integral is restricted to the frequencies in $\Lambda \subset [-\pi,\pi]$ (where Λ is a finite collection of intervals), the model isn't ARMA, but is defined on $[-\pi, \pi]$? If the model itself is only defined on Λ, do we still have optimality?

2) Is the multivariate spectral density estimator given above (for Gaussian series) optimal if the integral is taken over Λ? Is there an IRWLS procedure for multivariate series similar to that of Chiu (1988) in the univariate case which is optimal in the multivariate case?

3) For Gaussian series, if the quadrature or co-spectra have a separately parametrized model, is there an optimal procedure for estimating the co/quad spectrum "one curve at a time", i.e. without using the full multivariate structure of the spectrum for simplicity when "model screening"?

4) If the time series is non-Gaussian, are the answers to 1-3 above the same? If not, what is an optimal procedure?

5.2 Separately Parametrized Cross Spectral Estimation

Multivariate ARMA models may be difficult to work with, i.e. obtaining parametric estimates, due to the complex structure and large numbers of parameters. But suppose one had a model in which there were separate parametrizations of families of functions modeling f_1, f_2, c_{12}, and q_{12}, where c_{12} and q_{12} are the co and quad spectra, respectively. We would like to create a theory which can be used for estimating at most one curve at a time. We already know how to estimate f_1 and f_2 separately. What about c_{12} and q_{12}? Before answering this question, let's review some of the basic properties of multivariate spectral densities. Defining $\gamma_{xy}(\tau) = E(X(t+\tau)\ \overline{Y(t)})$ we have

$$\begin{aligned}\gamma_{yx}(\tau) &= E(Y(t+\tau)\overline{X(t)}) \\ &= E(\overline{X(t)}Y(t+\tau)) \\ &= E(\overline{X(t-\tau)}Y(t)) \quad \text{(by stationarity)} \\ &= \overline{\gamma_{xy}(-\tau)}\end{aligned}$$

Hence

$$\gamma_{xy}(\tau) = \overline{\gamma_{yx}(-\tau)} \qquad [5.2.1]$$

Note $\gamma_{xy}(\tau)$ needn't be symmetric. The cross spectral density $f_{xy}(\lambda)$ is the function whose Fourier coefficients are $\gamma_{xy}(\tau)$. If X(t), Y(t) are real valued, $\gamma_{xy}(\tau)$ must be real valued. So

$$\begin{aligned}\gamma_{xy}(\tau) &= \int_{-\pi}^{\pi} f_{xy}(\lambda)\, e^{i\lambda\tau}\, d\lambda \\ &= \int_{-\pi}^{\pi} \overline{f_{xy}(\lambda)}\, e^{-i\lambda\tau}\, d\lambda \\ &= \int_{-\pi}^{\pi} \overline{f_{xy}(-\lambda)}\, e^{i\lambda\tau}\, d\lambda \qquad \text{(change of variable)}\end{aligned}$$

Thus we see

$$f_{xy}(\lambda) = \overline{f_{xy}(-\lambda)} \qquad [5.2.2]$$

Also,

$$\begin{aligned}\gamma_{yx}(\tau) &= \overline{\gamma_{xy}(-\tau)} \\ &= \overline{\int_{-\pi}^{\pi} f_{xy}(\lambda)\, e^{-i\lambda\tau}\, d\lambda} \\ &= \int_{-\pi}^{\pi} \overline{f_{xy}(\lambda)}\, e^{i\lambda\tau}\, d\lambda.\end{aligned}$$

Hence

$$f_{yx}(\lambda) = \overline{f_{xy}(\lambda)}. \qquad [5.2.3]$$

Define

$$c_{xy}(\lambda) = \text{Re } f_{xy}(\lambda) \; ; \quad q_{xy}(\lambda) = -\text{Im } f_{xy}(\lambda).$$

So $f_{xy}(\lambda) = c_{xy}(\lambda) - i\, q_{xy}(\lambda)$. Using the relation [5.2.3] it is seen that $c_{xy}(\lambda) - i\, q_{xy}(\lambda) = c_{xy}(-\lambda) + i\, q_{xy}(-\lambda)$, which implies

$$c_{xy}(\lambda) = c_{xy}(-\lambda)$$

$$q_{xy}(\lambda) = -q_{xy}(-\lambda). \qquad [5.2.4]$$

From [5.2.4], it is sufficient to define $c_{xy}(\lambda)$ and $q_{xy}(\lambda)$ on $[0,\pi]$. We need the spectral density matrix

$$f_\theta(\lambda) = \begin{bmatrix} f_{1\theta} & c_\theta - i\, q_\theta \\ c_\theta + i q_\theta & f_{2\theta} \end{bmatrix}$$

to be positive definite. Thus, $\det f_\theta(\lambda) = f_{1\theta}(\lambda)\, f_{2\theta}(\lambda) - [\, c^2_\theta(\lambda) + q^2_\theta(\lambda)] > 0$ for all $\lambda \in [-\pi,\pi]$. Note this is really the only restriction on the model. If we make "separate parametrizations" of c_θ and q_θ, the model parameter space must be restricted so that this is satisfied. To simplify the notation we will write $\{c_\theta\}$, $\{q_\theta\}$ to denote the models for the co and quad spectra, respectively, even though these models do not depend on the same parameters. One might regard θ as a column vector with the first n_1 rows containing the parameters for $\{c_{\theta_1}\}$, rows n_1+1 to n_2 containing the parameters for $\{q_{\theta_2}\}$, etc. where θ_1 and θ_2 are in different parameter spaces.

5.3 Cross Spectral Models

Now that some elementary results about bivariate processes have been reviewed, we would like to return to the original problem of "one curve at a time" spectral estimation. First let us suppose that we have a bivariate time series and are in the position of knowing exactly the spectral densities f_x and f_y for the two series, and we know also the quad spectrum q but not the co-spectrum c. Is there an analogous "likelihood" function we can minimize which will yield a "Taniguchi" type theory? If such a function exists, in what sense is it "best", i.e. resulting in better asymptotic variances than minimizing other similar functions? The answer to the first question is "yes" and the sense in which it is asymptotically optimal will be described later. But first, let's make an attempt at constructing one such "likelihood".

We begin by examining the function Hosoya and Taniguchi (1982) want to minimize in the bivariate case; let's write out $D(f_\theta, I)$ more explicitly:

$$f_\theta = \begin{bmatrix} f_{1\theta} & c_\theta - i\, q_\theta \\ c_\theta + iq_\theta & f_{2\theta} \end{bmatrix} \quad ; \quad f_\theta^{-1} = \frac{1}{f_1 f_2 - (c^2+q^2)} \begin{bmatrix} f_2 & -c_\theta + iq_\theta \\ -c_\theta - i\, q_\theta & f_1 \end{bmatrix}$$

$$\text{So } f_\theta^{-1} I = \frac{1}{f_{1\theta}\, f_{2\theta} - (c_\theta^2 + q_\theta^2)}$$

$$\times \begin{bmatrix} f_{2\theta}\, I_x + (-c_\theta + iq_\theta)(\hat{c} + i\hat{q}) & f_{2\theta}(\hat{c} - i\hat{q}) + I_y(-c_\theta + iq_\theta) \\ -I_x(c_\theta + iq_\theta) + f_{1\theta}(\hat{c} + i\hat{q}) & (c_\theta + iq_\theta)(\hat{c} - i\hat{q}) + f_{1\theta} I_y \end{bmatrix}$$

$$\text{and } \operatorname{trace}(f_\theta^{-1} I) = \frac{1}{f_{1\theta} f_{2\theta} - (c_\theta^2 + q_\theta^2)} \left[f_{2\theta} I_x - 2\, c_\theta\, \hat{c} - 2\, q_\theta\, \hat{q} + f_{1\theta}\, I_y \right]$$

where

$$I = \begin{bmatrix} I_x & \hat{c} - i\hat{q} \\ \hat{c}+i\hat{q} & I_y \end{bmatrix}$$

is the periodogram matrix. So Taniguchi wants to minimize

$$\int_{-\pi}^{\pi} \log(f_{1\theta}f_{2\theta} - (c_\theta^2+q_\theta^2)\ ) +$$

$$\frac{1}{f_{1\theta}f_{2\theta} - (c_\theta^2+q_\theta^2)} [f_{2\theta}I_x - 2\, c_\theta\hat{c} - 2\, q_\theta\, \hat{q} + f_{1\theta}\, I_y]\, d\lambda. \qquad [5.3.1]$$

If one knew f_x and f_y and was attempting to estimate c and q, it might seem reasonable to adjust the above function and minimize instead

$$\int_{-\pi}^{\pi} \log(f_{1\theta}\, f_{2\theta} - (c_\theta^2 + q_\theta^2)\ ) +$$

$$\frac{1}{f_{1\theta}f_{2\theta} - (c_\theta^2+q_\theta^2)} [2f_{1\theta}f_{2\theta} - 2\, (c_\theta\, \hat{c} + q_\theta\, \hat{q})\ ]\, d\lambda. \qquad [5.3.2]$$

If one knew f_x, f_y and a function $\phi(\lambda)$ which has the property that $f_x(\lambda)f_y(\lambda) - (c^2(\lambda)+\phi^2(\lambda)) > 0$ for every λ (for example, $\phi(\lambda) \equiv 0$ is one such function), then it might be reasonable to choose θ to minimize (in the general case, as described in the introduction)

$$D^A_{(f_x,f_y,\phi)}(c_\theta,\hat{c}) \equiv \int_\Lambda \log(f_x f_y - (c_\theta^2+\phi^2)\ ) +$$

$$\frac{2}{f_x f_y - (\, c_\theta^2+\phi^2)} [\, f_x\, f_y - c_\theta\hat{c} - \phi^2]\, d\lambda. \qquad [5.3.3]$$

Defining $L_A(x,y,\lambda) \equiv \log(k(\lambda)-x^2) + \frac{2}{k-x^2}[k(\lambda) - xy]$ where x, y $\in$ **R**, $\lambda \in [-\pi,\pi]$, note that

$$\frac{\partial L_A}{\partial x} = \frac{-2x}{k-x^2} + 2\left(\frac{(k-x^2)(-y) - (k-xy)(-2x)}{(k-x^2)^2}\right)$$

$$= 2\left(\frac{-x(k-x^2) + (k-x^2)(-y) - (k-xy)(-2x)}{(k-x^2)^2}\right)$$

$$= 2\left(\frac{-xk+x^3 - yk + x^2y + 2kx - 2x^2y}{(k-x^2)^2}\right)$$

$$= 2\left(\frac{xk - x^2y - yk + x^3}{(k-x^2)^2}\right)$$

$$= 2\left(\frac{k(x-y) + x^2(x-y)}{(k-x^2)^2}\right)$$

$$= 2\left(\frac{(k+x^2)(x-y)}{(k-x^2)^2}\right)$$

$$= \frac{x-y}{\left(\frac{(k-x^2)^2}{2(k+x^2)}\right)}$$

Recalling Taniguchi's proof, the likelihood function possessing a partial derivative of this form will be crucial in showing both (1) If the model contains c, $D^A_{f_x,f_y,\phi}(c_\theta, c)$ is minimized at θ_0, where $c_{\theta_0} = c$, and (2) $\hat{\theta}$ minimizing $D^A_{f_x,f_y,\phi}(c_\theta, \hat{c})$ consistently estimates θ_0 minimizing $D^A{}_{f_x,f_y,\phi}(c_\theta,c)$.

There is another nice property of $D^A_{f_x,f_y,\phi}(\,.\,,\,.\,)$, and that is if one has separate parametrizations of f_1 , f_2, and c, the parametrization of c_θ must be restricted so that $f_{1\theta}f_{2\theta} - (c_\theta^2 + q_\theta^2) > 0$. Notice that for a fixed λ as c gets larger,

$$\log (k - c^2) + \left(\frac{2}{k - c^2}\right)(k - c\,\hat{c})$$

goes to ∞ because of the second term. Hence given a starting value for θ which satisfies the positive definite requirement, minimizing

$$\sum_{\lambda} \log(k(\lambda) - c_\theta^2) + \left(\frac{2}{k(\lambda) - c_\theta^2(\lambda)}\right)(k(\lambda) - c_\theta(\lambda)\,\hat{c}(\lambda)) \qquad [5.3.4]$$

cannot result in a $\hat{\theta}$ which doesn't satisfy the positive definite requirement. The question which naturally arises is this: If the model is correct, in the i.i.d. case ML estimates are "asymptotically optimal" in the sense of asymptotically satisfying the Cramér-Rao lower bounds. Our "observations" $\hat{c}(\lambda)$ are only "asymptotically independent", but aren't "asymptotically identically distributed". Does the estimate minimizing $D^1_{f_x, f_y, \phi}(c_\theta, \hat{c})$ possess any "asymptotic optimal" (with regard to variance) property in case $\{c_\theta\}$ is the correct model and $\phi(\lambda)=q(\lambda)$? The answer is NO, but to see this we will have to develop the theory of asymptotically BLUE estimators, as suggested by McCullagh (1983). Recall from section 2.1 that the QL function $l(\mu, y)$ may be obtained by solving

$$\frac{\partial l(\mu, y)}{\partial \mu} = \frac{y - \mu}{V(\mu)} \qquad [5.3.5]$$

where $V(\mu)$ is the variance. From Brockwell and Davis (1987), we know that asymptotically at each $\lambda \in [-\pi, \pi]$,

$$\mathrm{Var}\begin{pmatrix}\hat{c} \\ \hat{q}\end{pmatrix} = \begin{bmatrix} \frac{1}{2}(f_1 f_2 + c^2 - q^2) & cq \\ c\,q & \frac{1}{2}(f_1 f_2 + q^2 - c^2) \end{bmatrix}. \qquad [5.3.6]$$

Notice that unfortunately var($\hat{c}$) depends on q, but if we substitute an arbitrary function ϕ (satisfying positive definite conditions) for q, using μ=c, V(c)=($f_1 f_2 - \phi^2$)+c^2 and solve [5.3.5] while ignoring the dependence of $f_1 f_2 - \phi^2$ on λ by pretending it's a constant, we get

$$L_B(c,y,\lambda) \equiv \frac{y}{\sqrt{f_1(\lambda)f_2(\lambda) - q^2}} \tan^{-1}\left(\frac{c}{\sqrt{f_1(\lambda)f_2(\lambda) - q(\lambda)}}\right) -$$

$$\frac{1}{2} \ln(f_1(\lambda)f_2(\lambda) - q^2(\lambda) + c^2). \qquad [5.3.7]$$

Here, $L_B(c, y,\lambda)$ is a function **R**x**R**x$[-\pi,\pi]\rightarrow$**R**. Maximizing the above is equivalent to minimizing its negative, so now, following Hosoya and Taniguchi's (1982) idea in the univariate spectral estimation case define

$$D^B_{f_x,f_y,\phi}(h(\lambda)\ ,\ g(\lambda)) \equiv \int_\Lambda \frac{1}{2} \ln(f_x f_y - \phi^2 + h^2) -$$

$$\frac{g}{\sqrt{f_x f_y - \phi^2}} \tan^{-1}\frac{h}{\sqrt{f_x f_y - \phi^2}}\, d\lambda. \qquad [5.3.8]$$

Now we've got two "Taniguchi" functions, D^A and D^B. Notice they are both of the same form in that

$$D^A_{f_x f_y \phi}(h(\lambda),\ g(\lambda)) = \int_\Lambda L_A(h,\ g,\ \lambda)\ d\lambda$$

$$D^B_{f_x f_y \phi}(h(\lambda),\ g(\lambda)) = \int_\Lambda L_B(h,\ g,\ \lambda)\ d\lambda \qquad [5.3.9]$$

where $L_A(c_1,\ c_2,\lambda)$ and $L_B(c_1,\ c_2,\ \lambda)$ are functions from **R**x**R**x$[-\pi,\pi]\rightarrow$**R** such that

$$\frac{\partial L_A(c_1,c_2,\lambda)}{\partial c_1} = \frac{c_1 - c_2}{V_A(c_1,\lambda)}$$

$$\frac{\partial L_B(c_1,c_2,\lambda)}{\partial c_1} = \frac{c_1 - c_2}{V_B(c_1,\lambda)}$$

$$V_A(c_1,\lambda) = \frac{(k(\lambda) - c_1^2)^2}{2(k(\lambda) + c_1^2)}$$

$$V_B(c_1,\lambda) = \tfrac{1}{2}(f_x(\lambda)f_y(\lambda) - \phi^2(\lambda)) + c_1^2 . \qquad [5.3.10]$$

At this point it might be noted that one way in which our theory will differ from that of McCullagh (1983) is that our $V(\mu)$ is dependent upon both the mean c and the frequency λ, whereas his depends solely upon the mean. So in "response surface" terminology, our variance may be associated with the *location* in the model space together with the *mean at that location.* As previously mentioned (section 2.1), this idea is described by Carrol and Rupert (1988) (see chapter 3), but not made into a formal definition. One very important consequence of incorporating this into a general definition of "quasi-likelihood function" (to be made in the next chapter) is that the "least squares" function $\int (y_n(\lambda) - f_\theta(\lambda))^2/V(\lambda)\, d\lambda$ *will be considered a QL function under the new definition* (but not under McCullagh's).

As will be shown in the next chapter, minimizing $D^B_{f_1 f_2 q}(c_\theta, \hat{c})$ results in an "optimal" estimate of $\hat{\theta}$. Unfortunately, we don't know f_1, f_2, and q exactly. So how would the asymptotic variance of the estimator obtained by the following two procedures compare with the "optimal" estimate? Are the procedures consistent?

Procedure 1

(1) We don't have an estimate of q, but it's not unreasonable to assume we can find a function ϕ satisfying $f_1(\lambda)\, f_2(\lambda) - (c^2(\lambda)+\phi^2(\lambda)) > 0$. Such a function might be obtained by smoothing,

or if ϕ is not obvious, $\phi \equiv 0$ can always be used.

(2) Use as variance function $V(c(\lambda), \lambda) = f_{\hat{\theta}_1}(\lambda)\, f_{\hat{\theta}_2}(\lambda) - (c^2(\lambda)+\phi^2(\lambda))$ in the QL function D^B, and obtain a parametric estimate $c_{\hat{\theta}}$ of c.

(3) Use the $c_{\hat{\theta}}$ from step 2 for the "ϕ" function when repeating steps 1 and 2 in an attempt to estimate q. You now have an estimate $q_{\hat{\theta}}$.

(4) Re-estimate $c_{\hat{\theta}}$, using $q_{\hat{\theta}}$ from step 3 as "ϕ".

Procedure 2

Exactly the same as procedure 1, except use the least squares distance measure with variance function V(.) as stated in step 2 instead of D^B.

Note that these procedures are essentially Caroll and Rupert's (1988) algorithm for generalized least squares on p. 69 (see also the discussion in section 3.2 of Carrol and Rupert (1988)).

We will delay until the next chapter a theorem which settles the question of cross spectral estimators, because the theorem needed here can be stated as a corollary of a more general theorem which solves spectral estimation problems for non-Gaussian time series, whose periodograms lose the property of "asymptotic independence". However, now is a good time to give a discussion of where the co and quad spectral variances come from.

5.4 Background For Variances

There are two main variance theorems, stated in Brillinger (1981) and Hosoya and

Taniguchi (1982), which we will use for variance and covariance type results. These are as follows.

Theorem 7.6.1 (Brillinger (1981))

Let X(t), t=0, ± 1, . . be an r vector valued series satisfying

$$\sum_{u_1,u_2,u_3=-\infty}^{\infty} |c_{a_1a_2a_3}(u_1,u_2,u_3)| < \infty$$

where $c_{a_1a_2a_3}(u_1,u_2,u_3) = \text{cum}\{X_{a_1}(t_1),\ X_{a_2}(t_2),\ X_{a_3}(t_3)\}$ (see section 3.2 for definition of cumulant and cumulant spectrum). Defining

$$J^{(T)}_{a_1b_1}(A) = \frac{2\pi}{T} \sum_{s=1}^{T-1} A(\tfrac{2\pi s}{T})\ I^{(T)}_{a_1b_1}(\tfrac{2\pi s}{T}) \qquad [5.4.1]$$

where $A(\lambda)$ is a function of bounded variation on $[0,\ 2\pi]$, then if $A_1(\lambda)$ and $A_2(\lambda)$ are any two such functions, then

$$\text{cov}\ \{\ J^{(T)}_{a_1b_1}(A_1)\ ,\ J^{(T)}_{a_2b_2}(A_2)\ \} = \frac{2\pi}{T} \int_0^{2\pi} A_1(\alpha)\ \overline{A_2(\alpha)}\ f_{a_1a_2}(\alpha)\ f_{b_1b_2}(-\alpha)\ d\alpha$$

$$+ \frac{2\pi}{T} \int_0^{2\pi} A_1(\alpha)\ \overline{A_2(2\pi-\alpha)}\ f_{a_1b_2}(\alpha)\ f_{b_1a_2}(-\alpha)\ d\alpha$$

$$+ \frac{2\pi}{T} \int_0^{2\pi}\int_0^{2\pi} A_1(\alpha)\ \overline{A_2(\beta)}\ f_{a_1b_1a_2b_2}(\alpha,\ -\alpha,\ -\beta)\ d\alpha\ d\beta + o(T^{-1}). \qquad [5.4.2]$$

Lemma A2.2 (Hosoya and Taniguchi (1982))

If A_1 and A_2 are any square integrable functions on $[-\pi,\ \pi]$, and the process satisfies the property that each component of its spectral matrix has an absolutely summable Fourier series, then

$$\lim_{N\to\infty} N \operatorname{cov} \left\{ \int_{-\pi}^{\pi} A_1(\alpha)\, I_{a_1a_2}(\alpha)\, d\alpha,\ \int_{-\pi}^{\pi} A_2(\alpha)\, I_{a_3a_4}(\alpha)\, d\alpha \right\} =$$

$$2\pi \int_{-\pi}^{\pi} A_1(\alpha)\, \overline{A_2(\alpha)}\, f_{a_1a_3}(\alpha)\, \overline{f_{a_2a_4}(\alpha)}\, d\alpha$$

$$+2\pi \int_{-\pi}^{\pi} A_1(\alpha)\, A_2(-\alpha)\, f_{a_1a_4}(\alpha)\, \overline{f_{a_2a_3}(\alpha)}\, d\alpha$$

$$+2\pi \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} A_1(\alpha)\, A_2(-\beta)\, f_{a_1a_2a_3a_4}(\alpha, \beta, -\beta)\, d\alpha\, d\beta. \qquad [5.4.3]$$

Notice there must be some conditions on either the components of the spectral matrix or on the functions A_1 and A_2 for this theorem to hold, because if not, a contradiction results in the univariate case if the spectrum $f(\lambda) \in L^2[-\pi, \pi]$, but $f^2(\lambda) \notin L^2[-\pi, \pi]$. Notice that if $A_1(\lambda)$ and $A_2(\lambda)$ are both chosen to equal $f(\lambda)$, [5.4.3] is not defined. Chapter 8 will essentially show this theorem holds if the process has a spectrum with components of bounded variation.

Brillinger's (1981) theorem essentially handles the case where the periodogram is regarded as a step function with steps at each Fourier frequency, whereas Hosoya and Taniguchi's (1982) theorem uses the continuous "natural" definition of the periodogram. The statement of Brillinger's theorem is for the periodogram defined on $[0, 2\pi]$, and the reason for doing this lies in simplification of proofs so that sums may be taken between 1 and N rather than between $-N/2$ and $N/2$. However, using the (2π) periodicity of the periodogram and assuming that the functions A_1 and A_2 are periodic with period 2π, it is easily seen that the two expressions are equivalent. To do this, first use a common subscript notation in both expressions. Then observe that the integral of a periodic function over any interval whose length is the period is the same to convert Brillinger's integral to one between $-\pi$ and π, and

use the fact that $f_{a_1 a_2}(-\lambda) = \overline{f_{a_1 a_2}(\lambda)}$, i.e. relation [5.2.2].

Adapted to the cross spectral case for a bivariate process, [5.4.3] may be rewritten as

$$\lim_{N\to\infty} N \operatorname{cov} \left\{ \int_{-\pi}^{\pi} A_1(\alpha)\, I_{xy}(\alpha)\, d\alpha,\ \int_{-\pi}^{\pi} A_2(\alpha)\, I_{xy}(\alpha)\, d\alpha \right\} =$$

$$2\pi \int_{-\pi}^{\pi} A_1(\alpha)\, \overline{A_2(\alpha)}\, f_{xx}(\alpha)\, \overline{f_{yy}(\alpha)}\, d\alpha$$

$$+2\pi \int_{-\pi}^{\pi} A_1(\alpha)\, A_2(-\alpha)\, f_{xy}(\alpha)\, \overline{f_{yx}(\alpha)}\, d\alpha$$

$$+2\pi \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} A_1(\alpha)\, A_2(-\beta)\, f_{xyxy}(\alpha, \beta, -\beta)\, d\alpha\, d\beta. \qquad [5.4.4]$$

We may use this to obtain the "variance function" for the co and quad spectra as follows.

First, suppose $A_1(\alpha)$ and $A_2(\alpha)$ are both real valued even functions (i.e. $A_i(\lambda) = A_i(-\lambda)$), and use these in [5.4.4]. Recalling $f_{yx}(\lambda) = \overline{f_{xy}(\lambda)}$ (i.e. relation [5.2.3]), so the second integral contains the term $f_{xy}^2(\lambda) = (c(\lambda) - iq(\lambda))^2 = c^2(\lambda) - iq^2(\lambda) - 2ic(\lambda)q(\lambda)$. As $A_2(-\alpha) = A_2(\alpha)$ and $c(\lambda)$ are even, $A_1(\lambda)\, A_2(\lambda)\, c(\lambda)\, q(\lambda)$ must be odd since $q(\lambda)$ is odd. Hence this term integrates to 0. Except for the final term involving a fourth cumulant spectrum, we are left with

$$\operatorname{cov} \{ J_{xy}^{(T)}(A_1),\ J_{xy}^{(T)}(A_2) \} =$$

$$\frac{2\pi}{T} \int_{-\pi}^{\pi} A_1(\alpha)\, A_2(\alpha)\, f_{xx}(\alpha)\, f_{yy}(\alpha)\, d\alpha + \frac{2\pi}{T} \int_{-\pi}^{\pi} A_1(\alpha)\, A_2(\alpha)\, (c^2(\alpha) - q^2(\alpha))\, d\alpha =$$

$$\frac{2\pi}{T} \int_{-\pi}^{\pi} A_1(\alpha)\, A_2(\alpha)\, (f_{xx}(\alpha)\, f_{yy}(\alpha) + c^2(\alpha) - q^2(\alpha))\, d\alpha .$$

On the other hand, if $A_1(\alpha)$ and $A_2(\alpha)$ are both real valued odd functions (i.e. $A_1(\alpha) = -A_2(-\alpha)$), $A_1(\lambda)A_2(\lambda)c(\lambda)q(\lambda)$ is still odd and integrates to 0. Note $A_1(\lambda)A_2(-\lambda) = -A_1(\lambda)A_2(\lambda)$, so cov $\{ J_{xy}^{(T)}(A_1), J_{xy}^{(T)}(A_2) \} =$

$$\frac{2\pi}{T} \int_{-\pi}^{\pi} A_1(\alpha) A_2(\alpha) f_{xx}(\alpha) f_{yy}(\alpha) d\alpha - \frac{2\pi}{T} \int_{-\pi}^{\pi} A_1(\alpha) A_2(\alpha) (c^2(\alpha) - q^2(\alpha)) d\alpha =$$

$$\frac{2\pi}{T} \int_{-\pi}^{\pi} A_1(\alpha) A_2(\alpha) (f_{xx}(\alpha) f_{yy}(\alpha) + q^2(\alpha) - c^2(\alpha)) d\alpha.$$

Recall $\hat{c}(\lambda)$ and $\hat{q}(\lambda)$ are also even and odd, respectively. Hence if $A(\lambda)$ is even, $J_{xy}^{T}(A) = \sum A(\lambda_i)(\hat{c}(\lambda_i) - i\hat{q}(\lambda_i)) = \sum A(\lambda_i) \hat{c}(\lambda_i)$, and if $A(\lambda)$ is odd, $J_{xy}^{T}(A) = -i \sum A(\lambda_i) \hat{q}(\lambda_i)$. So if A_1 and A_2 are even, cov $(\sum A_1(\lambda_i) \hat{c}(\lambda_i), \sum A_2(\lambda_i) \hat{c}(\lambda_i)) =$ cov $(\sum A_1(\lambda_i) I_{xy}(\lambda_i), \sum A_2(\lambda_i) I_{xy}((\lambda_i))$, and if A_1 and A_2 are odd, cov $(\sum A_1(\lambda_i) \hat{q}(\lambda_i), \sum A_2(\lambda_i) \hat{q}(\lambda_i)) =$ cov $(-i \sum A_1(\lambda_i) \hat{q}(\lambda_i), -i \sum A_2(\lambda_i) \hat{q}(\lambda_i)) =$ cov $(\sum A_1(\lambda_i) I_{xy}(\lambda_i), \sum A_2(\lambda_i) I_{xy}((\lambda_i))$. Also observe that $E\left[\frac{2\pi}{T} \sum A_1(\lambda_i) I_{xy}(\lambda_i) \right] = \int_{-\pi}^{\pi} A(\lambda) f_{xy}(\lambda) d\lambda = \int_{-\pi}^{\pi} A(\lambda) c(\lambda) d\lambda$ if A is even, and $= -i \int_{-\pi}^{\pi} A(\lambda) q(\lambda) d\lambda$ if A is odd.

Of course, the same result holds if the sum is replaced by an integral in the "natural" definition of the cross periodogram.

This concludes our overview of cross spectra. The purpose of this chapter was to examine some multivariate spectral estimation problems and point towards some possible approaches to their solution. The expression for the parametric variance matrix [5.1.2] and [5.1.3] does not appear to give insight into what is happening in the case of model misspecification. Chapter 6 will begin to define a framework into which all parametric spectral estimation problems, univariate or multivariate, may be studied. Viewed in light of the definitions in chapter 6 and the main theorems in chapter 7, Hosoya and Taniguchi's (1982)

variance matrix will be expressed in a more comprehensible form. Chapters 9 and 10 are concerned mainly with misspecified univariate series, but the ideas there regarding how to deal with "contaminated" series also apply in the multivariate case.

Chapter VI
Quasi-Likelihood for Non-Gaussian Processes

6.1 Introduction

We have stated that for univariate Gaussian stationary time series, the periodogram may asymptotically be regarded as independent, exponentially distributed random variables whose mean at frequency λ is $f(\lambda)$, the spectral density. To see the sense in which this is true, it is not sufficient to simply show that the joint asymptotic distribution for a finite number of fixed Fourier frequencies approaches that of independent exponential random variables (i.e. Brockwell and Davis' (1987) theorem 10.3.2 quoted in section 3.3). Instead, we must look at how the periodogram behaves *as a function* when n is increased, e.g. see the theorems in section 5.4. To do this, let's take our motivation from the familiar regression models where the observations have a common variance.

Suppose y_i i=1..n have a common variance σ^2, and **a** and **b** are two orthogonal vectors in $\mathbf{R}^n$. Then we know that the random variables $\mathbf{y} \bullet \mathbf{a}$ (or $\mathbf{y}'\mathbf{a}$) and $\mathbf{y} \bullet \mathbf{b}$, where $\bullet$ is the Euclidian inner product $\mathbf{a} \bullet \mathbf{b} = \sum a_i b_i$, are uncorrelated. The analog of this for periodograms is the following. Suppose $A(\lambda)$ and $B(\lambda)$ are two even functions of bounded variation such that $\int A(\lambda)B(\lambda)\, f^2(\lambda)\, d\lambda = 0$. If the spectrum is continuous, then J(A) and J(B) are asymptotically uncorrelated, where

$$J(Q) \equiv \frac{2\pi}{T} \sum_{s=1}^{T-1} Q(\tfrac{2\pi s}{T})\, I(\tfrac{2\pi s}{T}) \qquad [6.1.1]$$

for any function $Q(\lambda)$ of bounded variation. If the periodogram is taken from a (univariate)

Gaussian time series, this is in fact the case. But if the time series is not Gaussian, the covariance of J(A) and J(B) is instead

$$2\pi \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} A(\lambda)\, B(\mu)\, f_4(\lambda, -\lambda, \mu)\, d\lambda\, d\mu \qquad [6.1.2]$$

where $f_4(\lambda, -\lambda, \mu)$ is the fourth cumulant spectrum (e.g. Brillinger's (1981) theorem 7.6.1, see also section 5.4).

From another angle, we know that if the series is Gaussian and $\hat{\theta}$ is chosen to minimize the weighted least squares function $\sum \phi(\lambda)\ (f(\theta,\ \lambda) - I_n(\lambda))^2$ with the correct weighting $\phi(\lambda)=1/f^2(\lambda)$, then the variance matrix of $\sqrt{n}\ (\hat{\theta} - \theta_0)$ is given by $2A^{-1}$. But if the series is *not* Gaussian, the variance matrix is $2A^{-1} + A^{-1}DA^{-1}$ where *D is nonzero* (see section 4.4). So if the series isn't Gaussian, the parametric estimates do not have the same asymptotic variance, and the asymptotically BLUE methods discussed in this dissertation aren't necessarily BLUE anymore. We need a new, more general theory to cover the non Gaussian case. Of course, we would expect the theory developed for Gaussian series to be a special case of the more general theory.

To motivate this more general theory, let's consider the question of what should correspond to doing weighted least squares using the correct weighting function, which is one over the square of the unknown spectral density. Since apparently the problem for non-Gaussian series is that the periodogram isn't asymptotically independent in some sense, the answer would be to minimize an expression of the form

$$(I_n(\lambda_i) - f_\theta(\lambda_i))' \mathbf{V}_n^{-1} (I_n(\lambda_i) - f_\theta(\lambda_i)) \qquad [6.1.3]$$

where $\mathbf{V}_n$ is a "covariance matrix" to take into account the dependencies in the periodogram. But $\mathbf{V}_n$ has to change with n, i.e. its size must increase. We need to identify an "object" that

represents what $\mathbf{V}_n$ is "converging" to ideally as $n \to \infty$. Using as motivation our previous concept of regarding the periodogram as a function, a natural "function space" generalization of finite dimensional Euclidian space with its concepts of vectors and perpendicularity is $L^2[-\pi, \pi]$, the set of all square (Lebesgue) integrable functions with the inner product $f \bullet g = \int fg \, d\lambda$. One might suspect that the "ideal" $\mathbf{V}$ is in fact an *operator on* L^2, and the $\mathbf{V}_n$ can be regarded as *approximations to this operator which become better as* $n \to \infty$. Recall that we are viewing the periodogram as a function in L^2 by defining $I_n(\lambda) = I_n(\lambda_k)$ if $\omega_k - 2\pi/n < \omega \leq \omega_k$, so it makes sense that the variance is an operator on L^2. This will in fact be shown in the sequel. But before doing this, we must consider how to create an analog of the "quasi-likelihood", i.e. among other things we need a class of procedures containing the weighted least squares procedure so that we can show minimizing the "weighted least squares" above is optimal in this class. We would also hope that the Gaussian likelihood function may be considered a non-Gaussian quasi-likelihood, so that the weighted least squares results in estimators which asymptotically have a smaller variance if the series is non-Gaussian. As will be seen, this is in fact true.

In the independent observations case, we defined a QL function at each x, μ to be a function $L(x, \mu, \lambda)$ satisfying

$$\frac{\partial L(x, \mu, \lambda)}{\partial \mu} = \frac{x - \mu}{W(\mu, \lambda)} . \qquad [6.1.4]$$

Notice that the QL function is *independent of the parametrization of the spectral density.* From the QL function, we create the "distance" function $D(\,.\,,\,.\,)$ as

$$D(x(\lambda), y(\lambda)) = \int_{-\pi}^{\pi} L(x(\lambda), y(\lambda), \lambda) \, d\lambda. \qquad [6.1.5]$$

What if the observations aren't independent? In this case, an "L" function as such doesn't exist, but we can *take the definition from the "distance measure"* instead of the "likelihood" at each point. Note that this gives a definition which does not depend upon the parametrization of the spectral density. So for univariate series, let's define the "QL distance" to be a function $D^W(x(\lambda),\ y(\lambda))$: $M \times PL^2[\Lambda] \rightarrow \mathbf{R}$ where the "model space" M is contained in the positive continuous functions and $PL^2[\Lambda]$ is the set of positive square integrable functions on the finite collection Λ of intervals, so that the partial derivative with respect to the vector (function) $x(\lambda)$ satisfies

$$\frac{\partial}{\partial[x(\lambda)]}\ D^W(x(\lambda),\ y(\lambda)) = W(x(\lambda))[\ x(\lambda) - y(\lambda)] \qquad [6.1.6]$$

where $W(x(\lambda))[\ .\ ]$ is *a linear operator defined on L^2 for each $x(\lambda) \in M$.* This definition will be refined and formalized later, as obviously some continuity conditions on D will be needed. What it means to take the derivative or partial derivative with respect to a function will be discussed in the next section, but let us for now say that the partial derivative of $D(x(\lambda),\ y(\lambda))$ evaluated at $(x_0(\lambda),\ y_0(\lambda))$ is a function in L^2 which will be regarded as a linear functional on the space of spectral densities.

6.2. Derivatives in Normed Spaces

Does the rough definition we've made for $D^W(x(\lambda),\ y(\lambda))$ include the previous definition of a distance derived from a QL function as a special case? To answer this question, the definition needs to be made more precise. Specifically, we must state what we mean by

$$\frac{\partial}{\partial[x(\lambda)]}\ D(x(\lambda),\ y(\lambda)). \qquad [6.2.1]$$

If $y(\lambda)$ is fixed, $D(x(\lambda), y(\lambda))$ as a function of $x(\lambda)$ is a nonlinear functional on the space of spectral densities, which is a subset of a linear space. Therefore, it is necessary to discuss what it means to take a derivative of a mapping between a normed linear space and $\mathbf{R}$. More generally, we will find it necessary to take the derivative of a nonlinear mapping between two normed linear spaces. Let us first review the more familiar notion of derivatives of mappings from $\mathbf{R}^p \to \mathbf{R}^k$.

If $f:\mathbf{R}\to\mathbf{R}$ is differentiable, the derivative of f at a point c is defined to be $f'(c)$ where

$$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c}$$

But the derivative may be thought of in another way, that of being a *linear mapping.* If we make a Taylor expansion of f about c, we may write $f(x) \approx f(c) + f'(c)\,(x - c)$, where the closer x is to c, the better the approximation holds. So the function f has been approximated by a translate of a linear mapping, the mapping $L(x) = f'(c)\,x$. It is this aspect of derivative, rather than the interpretation as a rate of change, which is used when making generalizations. Furthermore, we could take the definition of "derivative" to be the unique linear mapping $L(x)$ = ax satisfying: for each $\epsilon>0$, $\exists\ \delta>0$ so that if $|x - c|<\delta$, then $|\ f(x) - [f(c) + L(x - c)]\ | < \epsilon\ |x - c|$.

If $f:\mathbf{R}^p\to\mathbf{R}$ is differentiable at the vector $\mathbf{c}$, then we define the derivative, which is also known as the gradient, to be the vector of functions $\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots \frac{\partial f}{\partial x_p}\right)'$, where the partials are evaluated at $\mathbf{c}$. But a vector in $\mathbf{R}^p$ may also be regarded as a linear mapping $\mathbf{R}^p\to\mathbf{R}$: if $\mathbf{x} \in \mathbf{R}^p$, define $F_{\mathbf{x}}(\mathbf{y}) = \mathbf{x} \bullet \mathbf{y}$, where $\bullet$ represents the usual inner product in $\mathbf{R}^p$. As in the one dimensional case, the function f may be approximated in a neighborhood of a vector $\mathbf{c}$ by a translate of the linear mapping represented by its derivative at $\mathbf{c}$, i.e. $f(\mathbf{x}) = f(\mathbf{c})$

$+ \left.\frac{\partial f}{\partial x}\right|_c \bullet (\mathbf{x}-\mathbf{c})$, or using more familiar notation, $f(\mathbf{x}) = f(\mathbf{c}) + \left.\frac{\partial f}{\partial \mathbf{x}'}\right|_c (\mathbf{x}-\mathbf{c})$. As before, the definition of "derivative" could be taken as the unique linear mapping $L(\mathbf{x}) = \mathbf{a}'\mathbf{x}$ satisfying for each $\epsilon>0$, $\exists\ \delta>0$ so that if $\| \mathbf{x}-\mathbf{c} \|<\delta$, then $\| f(\mathbf{x}) - [f(\mathbf{c}) + L(\mathbf{x}-\mathbf{c})] \| < \epsilon \|\mathbf{x}-\mathbf{c}\|$, where $\| \cdot \|$ is the Euclidian norm.

If $f:\mathbf{R}^p \rightarrow \mathbf{R}^k$ $(= (f_1, f_2, \ldots f_k)$, where each f_i is a real valued function on $\mathbf{R}^p)$ is differentiable at the vector $\mathbf{c}$, then we define the derivative at $\mathbf{c}$ to be the matrix

$$M = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdot & \cdot & \cdot & \frac{\partial f_1}{\partial x_p} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \frac{\partial f_k}{\partial x_1} & \cdot & \cdot & \cdot & \frac{\partial f_k}{\partial x_p} \end{bmatrix}$$

where the partials are evaluated at $\mathbf{c}$. This matrix may be regarded as a linear mapping from $\mathbf{R}^p \rightarrow \mathbf{R}^k$ which maps $\mathbf{x} \in \mathbf{R}^p$ to $M\mathbf{x}$. As before, in a neighborhood of $\mathbf{c}$, f may be approximated by a translate of the linear mapping M, i.e. $f(\mathbf{x}) = f(\mathbf{c}) + M(\mathbf{x}-\mathbf{c})$. Again, the definition of "derivative" could be taken as the unique linear mapping $L(\mathbf{x}) = M\mathbf{x}$ satisfying for each $\epsilon>0$, $\exists$ $\delta>0$ so that if $\| \mathbf{x}-\mathbf{c} \|<\delta$, then $\| f(\mathbf{x}) - [f(\mathbf{c}) + M(\mathbf{x}-\mathbf{c})] \| < \epsilon \|\mathbf{x}-\mathbf{c}\|$, where $\| \cdot \|$ is the Euclidian norm.

What this suggests is that the concept of "derivative" as a linear mapping used to approximate a function (which maps one finite dimensional vector space into another) at a point may be extended to functions mapping two normed linear spaces, because the definition essentially depends only upon the norms on the spaces.

Suppose U is an open subset of a normed linear space X and F is contained in a

normed linear space Y. We define a mapping f:U→F to be **differentiable** at $\xi \in U$ if there exists a bounded (i.e. continuous) linear operator A_ξ such that

$$\lim_{x \to \xi} \frac{|| f(x) - f(\xi) - A_\xi(x - \xi) ||}{||x - \xi||} = 0. \qquad [6.2.2]$$

If there is such a linear operator A_ξ, then it is unique and will be denoted by $Df(\xi)$. A good reference for material on calculus extended to normed spaces is Chae (1985), from which this definition was taken (p. 54).

There are also other notations used in the dissertation for derivatives between normed spaces. As previously mentioned, we will often be concerned with partial derivatives on vector functions, such as $\frac{\partial}{\partial[x(\lambda)]} D(x(\lambda), y(\lambda))$, which refers to the derivative of the nonlinear functional $D(x(\lambda), y(\lambda))$ when $y(\lambda)$ is a fixed function, *evaluated at x(λ).* Another example is the mapping between the parameter space Θ and the model f_θ. The derivative of this mapping (evaluated at θ), viewed as a mapping between a subset of $\mathbf{R}^p$ and C[a, b] (the continuous functions on the interval [a, b]), will be denoted as $\frac{\partial f_\theta}{\partial \theta}$. The reason for this is that the "normed spaces derivative" and the ordinary partial derivative of the function $f(\theta, \lambda)$ are identical, see the discussion in the second paragraph of section 7.1. Sometimes functions on product vector spaces $X \times Y$ will have their partial derivatives given by $D_1 f(x, y)$, which is the same as $\frac{\partial f(x, y)}{\partial x}$. See, for example, proposition 7.2.2.

Most important theorems from the ordinary calculus have normed linear space versions. Two of these which we will need are the following.

1) The Mean Value Theorem: Suppose that U contains the line segment [a, b]. If f:U→**R** is a differentiable map, then there exists $c \in (a, b)$ such that $f(b) - f(a) = Df(c)(b - a)$ (Chae (1985), p. 69).

2) The Chain Rule: Let E, F, G be normed spaces, U an open subset of E, and V an open subset of F. Let f:U→V and g: V→G. If f is differentiable at $\xi \in U$ and g is differentiable at $f(\xi) \in V$, then $g \circ f$: U→G is differentiable at ξ and $D(g \circ f)(\xi) = Dg(f(\xi)) \circ Df(\xi)$ (Chae (1985), p. 58).

In the previous chapters, we have always considered the "model" to be a set of vector valued continuous or continuously differentiable functions $\{f_\theta\}$ indexed by a parameter θ. As such, the model could be regarded as a subset of $\Pi_k C[\Lambda] \equiv C[\Lambda] \times C[\Lambda] \times \ldots \times C[\Lambda]$, the product of C[Λ] k times, written as column vectors of continuous functions. There is a natural norm imposed on spaces of continuous functions, the "sup" norm $\| \cdot \|_\infty$; i.e.

$$\| f \|_\infty = \sup_{x \in \Lambda} |f(x)|.$$

The notation $C_\infty[\Lambda]$ will be used to denote C[Λ] equipped with the sup norm, and $\Pi_k C_\infty$ to denote $\Pi_k C$ with the product norm given by $\|(f_1, f_2, \ldots f_k)\| = \sup \{\|f_1\|_\infty, \|f_2\|_\infty, \ldots \|f_k\|_\infty\}$.

We will now give a technical result which is needed later. Suppose g(x, λ) is defined on $U \times \Lambda$ for some $U \subset \mathbf{R}$ and has a continuous partial derivative with respect to x. Then a nonlinear operator $C_\infty[\Lambda] \to C_\infty[\Lambda]$ is defined by $F_g[z(\lambda)] = g(z(\lambda), \lambda)$ for $z \in C_\infty[\Lambda]$. The following proposition gives the derivative $D_z F_g$, and is needed in the sequel to help show the new definition of "Quasi-likelihood" to be made in this section includes the old definition as a special case. Its proof is an example of how to work with derivatives of mappings on normed spaces.

Proposition 6.2.1

If $F_g[h] = g(h(\lambda), \lambda)$ for h in an open subset of $C_\infty[\Lambda]$, $\lambda \in \Lambda$ (Λ is a finite union of intervals), and g(x, λ) has a continuous partial derivative with respect to x (on an appropriate open

rectangle to make F_g well defined), then F_g is differentiable at each h and

$$F'_g(h)\,[k] = \frac{\partial g(h(\lambda),\,\lambda)}{\partial x}\;k(\lambda)$$

Remark: This is simply the product of the two functions $\frac{\partial g(h(\lambda),\,\lambda)}{\partial x}$ and $k(\lambda)$.

Before giving the proof, a preliminary result is needed. As noted in exercise 27.M p. 199 of Bartle (1976), if a function f is continuously differentiable, then for every $\epsilon>0$ there exists a $\delta(\epsilon) > 0$ such that if $0<|x-y|<\delta(\epsilon)$, $x, y \in [a, b]$, then

$$\left|\frac{f(x)-f(y)}{x-y} - f'(y)\right| < \epsilon \quad (\text{or } |f(x)-[f(y)+f'(y)(x-y)]| < \epsilon\,|x-y|\,).$$

We can extend this as follows.

Proposition 6.2.2

Let K be a compact set in $\mathbf{R}^n$ and let θ denote a vector in $\mathbf{R}^n$. If $f(x, \theta)$ is continuous in a rectangle $[a, b] \times K$ and has a continuous partial derivative with respect to x, then for every $\epsilon>0$ $\exists$ a $\delta(\epsilon) > 0$ such that if $0<|x-y|<\delta(\epsilon)$, $x, y \in [a, b]$, $\theta \in K$, then

$$\left|\frac{f(x,\,\theta)-f(y,\,\theta)}{x-y} - \frac{\partial f(y,\,\theta)}{\partial x}\right| < \epsilon \quad \left(\text{or } \left|f(x,\,\theta)-[f(y,\,\theta)+\frac{\partial f(y,\,\theta)}{\partial x}(x-y)]\right| < \epsilon\,|x-y|\,\right).$$

proof

Let $\epsilon>0$. By the continuity of $\frac{\partial f(x,\,\theta)}{\partial x}$, choose δ so if $||(x_1,\,\theta_1)-(x_2,\,\theta_2)||<\delta$, then $\left|\frac{\partial f(x_1,\,\theta_1)}{\partial x} - \frac{\partial f(x_2,\,\theta_2)}{\partial x}\right|<\epsilon$. Now suppose $|x-y|<\delta$ for some $x, y \in \mathbf{R}$. For each θ the one variable mean value theorem gives a c_θ between x and y so that $\frac{f(x,\,\theta)-f(y,\,\theta)}{x-y} = \frac{\partial f(c_\theta,\,\theta)}{\partial x}$.

Since c_θ is within δ of y, the continuity of the partial yields $\left|\frac{\partial f(c_\theta, \theta)}{\partial x} - \frac{\partial f(y, \theta)}{\partial x}\right| < \epsilon$. The result follows. □

proof of proposition 6.2.1

To show that $F_g'(h)\,[k] = g'(h(\lambda), \lambda)\,k$, it suffices to show that for every $\epsilon > 0$ $\exists$ a $\delta(\epsilon) > 0$ such that if $||x - h||_\infty < \delta(\epsilon)$, where x and h are functions in $C[\Lambda]$, then

$$||F_g[x] - [F_g[h] + g'(h(\lambda), \lambda) \cdot (x - h)]\,||_\infty < \epsilon\, ||x - h||_\infty$$

or equivalently, that

$$||\, g(x(\lambda), \lambda) - [g(h(\lambda), \lambda) + g'(h(\lambda), \lambda) \cdot (x(\lambda) - h(\lambda))]\,||_\infty < \epsilon\, ||x(\lambda) - h(\lambda)||_\infty.$$

First, note that the functions in a ball about h will have their ranges in some finite interval [c, d]. So we may view g as being restricted to a rectangle. By the proposition applied to g, there exists a $\delta(\epsilon)$ such that the conclusion of the proposition holds if $|x - y| < \delta(\epsilon)$. So if $||x(\lambda) - h(\lambda)||_\infty < \delta(\epsilon)$, for each λ the conclusion of the proposition holds at g evaluated at $h(\lambda)$ and we can put the sup norm on both sides of the inequality. □

We may now show the function $\int_\Lambda \log(f) + g/f\, d\lambda$ (and more generally [6.1.5]) is indeed a QL function. In doing so, a simple use of the chain rule is illustrated which will be often referred to in the sequel.

Proposition 6.2.3

Suppose Λ is a finite collection of intervals. Let the domain of the function

$$D(f, g) = \int_\Lambda \log(f) + g/f\, d\lambda$$

be $C[a, b] \times L^2[a, b]$ (or $C[\Lambda] \times L^2[\Lambda]$, where $\Lambda \subset [a, b]$). Then D satisfies [6.1.6].

Note two results are given in the proposition: the "design space" can be either Λ or [a, b]. If the design space is Λ, this QL function will be seen to result in optimal estimates (and

otherwise not).

proof

Define functions I: C[a, b]→L^2[a, b] (or C[Λ]→L^2[Λ]) as the identity mapping (i.e. I(f)=f), F_1: C[a, b]→C[a, b] (or C[Λ]→C[Λ]) by $F_1(f)=\log(f)$ (where of course the domain is really positive continuous functions), F_2: C[a, b]→C[a, b] (or C[Λ]→C[Λ]) by $F_2(f)=1/f$ (where the domain is again positive functions), F_3:C[a, b]→L^2[a, b] (or C[Λ]→L^2[Λ]) by $F_3(f)=fg$ (ordinary function multiplication for some fixed $g \in L^2$), F_4: L^2[a, b]→**R** (or L^2[Λ]→**R**) by $F_4(f)=\int_\Lambda f\, d\lambda$. Notice that for g fixed, D may be regarded as a function from C[a, b]→**R** (or C[Λ]→**R**) and may be written as $F_4 \circ (I \circ F_1+F_3 \circ F_2)[f]$. One application of the chain rule says the derivative of this mapping *evaluated at f and applied to q* is

$$F_4'\big|_{(I \circ F_1 + F_3 \circ F_2)[f]} \circ (I \circ F_1+F_3 \circ F_2)'\big|_f [q].$$

Another application gives the derivative as

$$F_4'\big|_{(I \circ F_1 + F_3 \circ F_2)[f]} \circ \left(I'\big|_{F_1(f)} \circ F_1'\big|_f + F_3'\big|_{F_2(f)} \circ F_2'\big|_f \right) [q].$$

Observe that the derivative of the mapping x→M[x] is M for any linear mapping M (i.e. corresponding to the usual d/dx (mx)=m for m, $x \in \mathbf{R}$) regardless of where it's evaluated, so $I'=I$, $F_3'=F_3$, $F_4'=F_4$. Proposition 6.2.1 gives $F_1'(f)[q]=q/f$ and $F_2'(f)[q]=-f^{-2}q$. Hence the expression simplifies to $F_4 \circ (q/f - (g/f^2)\, q)$. F_4 is just the integral, so this simplifies to

$$\int_\Lambda \left(\frac{1}{f}-\frac{g}{f^2}\right) q\, d\lambda = \int_\Lambda \frac{f-g}{f^2}\, q\, d\lambda.$$

Hence the "derivative" is $I_\Lambda(\lambda)\,\frac{f(\lambda)-g(\lambda)}{f^2(\lambda)}$ (where I_Λ is the indicator function), which is of the correct form. □

More generally, note that any function of the form $L(\mu, y, \lambda)$ where $\partial L/\partial\mu=(y-\mu)/W(\mu, \lambda)$ can be expressed as $\int \frac{\mu}{W(\mu, \lambda)}\,d\mu + y \int \frac{1}{W(\mu, \lambda)}\,d\mu + \phi(y)$ $(=l(\mu, \lambda) + y\, r(\mu, \lambda) + \phi(y))$. The above proof can essentially be applied to the two indefinite integrals to yield [6.1.5] is a QL function (see theorem 8.5.1).

The following corollary (of proposition 6.2.1) is a consequence of propositions 6.2.1 and 7.2.2:

Corollary 6.2.1

Suppose $g(\mathbf{x}, \lambda)$: $\mathbf{R}^k \times \Lambda \rightarrow \mathbf{R}^p$ has continuous partial derivatives with respect to the components of $\mathbf{x}$. Define $F_g[h(\lambda)] = g(h(\lambda), \lambda)$ for $h:\Lambda\rightarrow\mathbf{R}^k$. Then the derivative of F_g, evaluated at $h(\lambda) \in \Pi_k C$ and applied to $p(\lambda) \in \Pi_k C$ is given by

$$F'_g(h(\lambda))[p(\lambda)]= \frac{\partial g(h(\lambda), \lambda)}{\partial \mathbf{x}}\ (p_1(\lambda), p_2(\lambda), \ldots p_k(\lambda))'.$$

Remark: $\frac{\partial g(h(\lambda), \lambda)}{\partial \mathbf{x}}$ is a matrix of functions, do the usual matrix multiplication pointwise and the result is a function.

Now we may give a rigorous definition of the quasi-likelihood. The definition is part of a more general scenario, so that the theory to be developed may cover spectral estimation over bands in both the univariate series case as described in chapter 4, and the cospectral and multivariate spectral estimation case as described in chapter 5.

In the following, note that $\Pi_k L^2$ denotes the k fold product of L^2 with itself, which is a Hilbert space with the inner product $(f_1, f_2, \ldots f_k) \bullet (g_1, g_2, \ldots g_k) = f_1 \bullet g_1 + f_2 \bullet g_2 + \ldots +$

$f_k \bullet g_k$.

6.3. Definitions

In this section we give some definitions and establish notation to be needed in the sequel.

Definition 6.3.1 A continuous function on $[-\pi, \pi]$ with a Fourier series satisfying $\sum |k| |\tilde{f}(k)| < \infty$ (where $\tilde{f}(k)$ is the kth Fourier coefficient of the function $f(\lambda)$), is said to satisfy the **basic continuity condition (BCC).** A sufficient condition on f for this to occur is for f to have a continuous derivative satisfying a Lipschitz condition of order $\alpha > 0$ (e.g. see Chiu's (1988) theorem 1).

Definition 6.3.2 A **random L^2 sequence** is a 5 tuple $(\langle f_n \rangle, f, M, SL^2, V)$.

(a) Λ is a finite set of disjoint closed intervals.

(b) There exists a probability space $(\Omega, \mathcal{P})$ so that for each $\omega \in \Omega$ and each n a positive integer, $f_n(\lambda, \omega)$ as a function of λ is in $SL^2 \subset \Pi_k L^2[\Lambda]$ (column vectors of L^2 functions). The set of finite linear combinations of elements in SL^2 is assumed to be dense in $\Pi_k L^2$. f is a (nonrandom) function in SL^2 called the **means function**, or **limiting function.** In the future, we shall suppress the ω and simply write $f_n(\lambda)$ for $f_n(\lambda, \omega)$.

(c) M is the **model space**, an open subset of $\Pi_k C$ which is also contained in SL^2.

(d) V is the **variance operator**, and is a nonnegative definite, invertible self adjoint operator such that $\lim_{n \to \infty} n \operatorname{cov}(\psi_1 \bullet f_n, \psi_2 \bullet f_n) = \psi_1 \bullet V[\psi_2]$ for any $\psi_1, \psi_2 \in \Pi_k L^2$.

(e) $\exists K_1 > 0$ so $n |\operatorname{cov}(\psi_1 \bullet f_n, \psi_2 \bullet f_n)| \leq K_1 ||\psi_1||_2 ||\psi_2||_2$ for all $n \in \mathbf{N}$, $\psi_1, \psi_2 \in \Pi_k L^2$.

(f) $\exists K_2 > 0$ so $|E(\psi \bullet f_n)| \leq K_2 ||\psi||_2$ for all $n \in \mathbf{N}$, $\psi \in \Pi_k L^2$.

(g) $\lim_{n \to \infty} E(\psi \bullet f_n) = \psi \bullet f$.

Definition 6.3.3 The **model** is a collection of functions $\{f_\theta\}_{\theta \in \Theta}$. Θ is assumed to be a convex, compact subset of $\mathbf{R}^p$. The functions f_θ are contained in M and are twice continuously differentiable with respect to θ.

Definition 6.3.4 The **quasi likelihood distance** or **quasi likelihood function** is a function $D_\psi(.\ ,\ .)$ on $\Psi \times M \times SL^2$ satisfying:

(a) Ψ is a "parameter" space with some topology.

(b) $\frac{\partial}{\partial[x(\lambda)]} D_\psi(x(\lambda),\ y(\lambda)) = W_\psi(x(\lambda))[x(\lambda) - y(\lambda)]$ for some not necessarily self adjoint or invertible operator W_ψ, see remark 5 below.

(c) (i) Given $\epsilon>0$, $y \in \Pi_k L^2$, $x_0 \in M$, and ψ_0 a limit point of Ψ, then there exists a neighborhood N of $(\psi_0,\ x_0)$ in $\Psi \times M$ so that if $(\psi,\ x) \in N$, then $||W^*_\psi(x(\lambda))[y] - W^*_{\psi_0}(x_0(\lambda))[y]||_2<\epsilon$.

(ii) If K is a bounded subset of M with the sup norm, then $\exists$ a positive R so that $||W^*_\psi(x)||<R$ for all $x \in K$, $\psi \in \Psi$.

(d) Define the mapping $\Psi \times \Pi_k C \rightarrow B(\Pi_k L^2)$ by $F(\psi, x(\lambda))=W_\psi(x(\lambda))[\ .\ ]$.

(i) F is partially differentiable with respect to $x(\lambda)$, i.e. for each $(\psi, x(\lambda))$, the derivative is in $B(\Pi_k C, B(\Pi_k L^2))$.

Let $L_{\psi,f} \in B(\Pi_k C, B(\Pi_k L^2))$ denote the derivative evaluated at (ψ, f) (so for each $g \in \Pi_k C$, $L_{\psi,f}(g) \in B(\Pi_k L^2)$).

(ii) If $\epsilon>0$, $f_0 \in M$, $g_0 \in \Pi_k C$, $y \in L^2$ and ψ_0 a limit point of Ψ, then there exists a neighborhood N of (ψ_0, f_0, g_0) so that if $(\psi, f, g) \in N$, then $||L^*_{\psi,f}(g)[y] - L^*_{\psi_0,f_0}(g)[y]||_2<\epsilon$.

(iii) If K_1 is a bounded subset of M with the sup norm and K_2 is a bounded subset of $\Pi_k C$,

then $\exists$ a positive R so that $||L_{\psi,f}(g)|| \leq R$ (operator norm, see remark (2) below) for all $f \in K_1$, $g \in K_2$, $\psi \in \Psi$.

If $W_\psi(f)[\,.\,]$ satisfies (c) and (d), then it is a **quasi likelihood operator.**

Definition 6.3.5 Suppose $\{f_\theta\}$ is a specific model, $\Theta \subset \mathbf{R}^p$, $W_\psi(f)[\,.\,]$ is a QL operator, and $(\langle f_n \rangle, f, M, SL^2, V)$ is a random L^2 sequence. Then we say that the **link condition** holds if

$$\lim_{n\to\infty} \sqrt{n}\, E\left[\int_\Lambda W_\psi^*(f_\theta)\left[\frac{\partial f_\theta}{\partial \theta_j}\right] f_n^i d\lambda - \int_\Lambda W_\psi^*(f_\theta)\left[\frac{\partial f_\theta}{\partial \theta_j}\right] f^i\, d\lambda\right] = 0$$

for all fixed θ and ψ, every component f_n^i, f^i of f_n, f, respectively, j=1 to p.

Remarks:

1) All function spaces (e.g. $\Pi_k L^2$, etc.) and integrals are over Λ.

2) $B(\Pi_k L^2)$ is the space of all continuous operators on $\Pi_k L^2$. Whenever an operator is given a norm, unless specified otherwise it is the operator norm defined by $||W|| = \sup_{|x|=1} ||W[x]||_2$.

3) Ψ is a topological space of "parameters" which might be a subset of $\mathbf{R}^q$ or some sort of function space, for example. This allows the "variance" operator to depend upon some additional parameters besides the (unknown) means function.

4) M and SL^2 will depend upon the specific problem one is addressing. For example, in the univariate spectral estimation case one would take SL^2 to be the nonnegative L^2 functions and M to be the subset of strictly positive continuous functions on Λ.

5) By the Riesz representation theorem (see, e.g. theorem 3.4, p. 12 of Conway (1985)), if F is a linear functional on the Hilbert space H, then there exists an element $f \in H$ so that for any $g \in H$, $F(g) = f \bullet g$. Fix $y(\lambda)$ and regard $D(x(\lambda), y(\lambda))$ as a nonlinear functional on $\Pi_k L^2$. If $D(x(\lambda), y(\lambda))$ has a partial derivative with respect to $x(\lambda)$, then the partial derivative

(evaluated at a particular $x(\lambda)$) is a linear functional, which can be represented as an element of H. To satisfy definition 6.3.4 (b), this element must be $W_{\psi}(x(\lambda))[x(\lambda) - y(\lambda)]$.

6) The conditions in (c) and (d) of QL distance are a weakening of continuity (It isn't necessary to have continuity when the range space has the operator norm). The "y" in (i) of (c) or (ii) of (d) may come from a compact set and the convergence is uniform (see, e.g. proposition 6.5.3).

7) The Link Condition is the only "link" or connection between the QL distance or operator and the specific model used. More will be said about this in chapter 8.

8) For an operator W, we use the notation W^* to denote the adjoint of W. See Conway (1985), theorem 2.2.2 p. 31 for a theorem concerning existence of adjoints. Briefly put, the adjoint is the operator satisfying $h \bullet W[k] = W^*[h] \bullet k$. In Euclidian space, if matrices are viewed as operators, the adjoint of a matrix is simply its conjugate transpose.

The scope of this definition is unclear and is a topic for future research. For example, consider the spectral estimation problem for univariate series. Does there exist an "exponential family" of time series, of which the Gaussian is a special case, for which maximizing the likelihood of the finite number of observed random variables asymptotically is equivalent to minimizing the correct D function, where $W(x(\lambda))$ is a nontrivial function of $x(\lambda)$? Knowing this "D" function would be useful in model fitting in the same sense as "deviance" is used in GLIM. Chapter 9 will give examples of nontrivial QL functions which may be useful when it is suspected there may be certain types of misspecification in the model, and chapter 10 will suggest another way in which QL operators could be defined.

6.4. Examples of Random L^2 Sequences and QL Distances

Without proof (proofs will be given in chapter 8), we show how the preceding

definitions capture the spectral parametric estimation problems so far discussed. For all of the following examples, we assume the process under consideration has a spectrum whose components satisfy the BCC. As will be seen in chapter 8, for time series applications this is sufficient for the "link condition" to hold regardless of what QL function or operator is used.

Example 1: Usual Gaussian (univariate) case.

Let $f_n(\lambda)$ be the periodogram, restricted to $\Lambda \subset [-\pi, \pi]$ and extended to all frequencies in Λ either as a step function or by the natural extension, and let $f(\lambda)$ be the true spectrum. Then V is the multiplication operator on $L^2[\Lambda]$ i.e. $V[x(\lambda)] = 2\pi f^2(\lambda)x(\lambda)$. The model space M might be all positive, continuous functions on Λ, and SL^2 is all nonnegative, square integrable functions. One example of a QL distance would be

$$D(x(\lambda), y(\lambda))=\int_{\Lambda} L(x(\lambda), y(\lambda))\, d\lambda \qquad [6.4.1]$$

where $\frac{\partial L(x, y)}{\partial x} = \frac{x-y}{W(x)}$ for a positive function $W(x)$ (e.g. if $L(x, y)=\log(x) + y/x$, $W(x)=x^2$). See the discussion following the proof of corollary 6.2.1.

Example 2: Cross spectral case.

$(X_1(t), X_2(t))$ is a bivariate Gaussian process, f_n is the co (or quad) spectrum restricted to Λ, and f is the true co (or quad) spectrum. The model space M is all continuous functions $c(\lambda)$ (or $q(\lambda)$) satisfying $f_1(\lambda)f_2(\lambda) - (c^2(\lambda)+q^2(\lambda))>0$, where f_1, f_2 are the true spectra of the two processes, and $q(\lambda)$ is the true quad spectrum (or $c(\lambda)$ is the true co spectrum). Ψ is the function space $\{f_1\} \times \{f_2\} \times \{q\}$ (or $\{f_1\} \times \{f_2\} \times \{c\}$). V_ψ is the multiplication operator on Λ defined by $V_\psi(c(\lambda))[x(\lambda)] = 2\pi(f_1(\lambda)f_2(\lambda)+c^2(\lambda) - q^2(\lambda))\ x(\lambda)$ (or $V_\psi(q(\lambda))[x(\lambda)] = 2\pi(f_1(\lambda)f_2(\lambda)+q^2(\lambda) - c^2(\lambda))\ x(\lambda)$). The functions $D_A(\ .\ ,\ .\ )$ and $D_B(\ .\ ,\ .\ )$ given in chapter 5

are examples of QL distances.

Example 3: Weighted least squares.

Let W be any positive, self adjoint invertible operator. Then the "weighted least squares" distance $D(x(\lambda), y(\lambda)) = (x(\lambda) - y(\lambda)) \bullet W[x(\lambda) - y(\lambda))]$ is a QL distance for any spectral random L^2 sequence.

Example 4: Bivariate Gaussian process.

Let $(X_1(t), X_2(t))$ be a bivariate Gaussian process. For a column vector $x=(x_1, x_2, x_3, x_4)'$, define x^m ("m" for "matrix") to be the 2×2 matrix defined by

$$x^m = \begin{bmatrix} x_1 & x_3 - ix_4 \\ x_3+ix_4 & x_2 \end{bmatrix}. \qquad [6.4.2]$$

(I_n, f_0, M, SL^2, V) is a random L^2 sequence where

1) $I_n = (I_1, I_2, \hat{c}, \hat{q})$ restricted to Λ.

2) f_0 is the spectrum of the process (written as (f_{11}, f_{22}, c, q)).

3) M is the subset of $\Pi_4 C$ satisfying $\det f^m > 0$.

4) SL^2 is the subset of $\Pi_4 L^2$ functions $x(\lambda)$ satisfying $\det x^m(\lambda) > 0$ for almost all λ.

5) V is the operator on $\Pi_4 L^2$ defined by $V[x(\lambda)] = M(f_0)\, x(\lambda)$ for $x(\lambda) \in \Pi_4 L^2$, where $(1/2\pi)M(f_{11}, f_{22}, c, q)$ is the matrix valued function

$$\begin{bmatrix} f_{11}^2 & |f_{12}|^2 & f_{11}c_{12} & f_{11}q_{12} \\ |f_{12}|^2 & f_{22}^2 & f_{22}c_{12} & f_{22}q_{12} \\ f_{11}c_{12} & f_{22}c_{12} & \frac{1}{2}(f_{11}f_{22}+c_{12}^2 - q_{12}^2) & c_{12}q_{12} \\ f_{11}q_{12} & f_{22}q_{12} & c_{12}q_{12} & \frac{1}{2}(f_{11}f_{22}+ q_{12}^2 - c_{12}^2) \end{bmatrix}. \qquad [6.4.3]$$

The multiplication $M(f_0)\ x(\lambda)$ is done pointwise at each λ, and the result is a function in $\Pi_4 L^2$. Let $f_\theta = (f_{1\theta}, f_{2\theta}, c_\theta, q_\theta)'$ and $g=(g_1, g_2, g_3, g_4)'$. Then the bivariate distance function defined by

$$D(f_\theta, g) \equiv \int_\Lambda \log \det f_\theta^m(\omega) + \text{trace}\ ([f_\theta^m(\omega)]^{-1}\ g^m(\omega)\ )\ d\omega \qquad [6.4.4]$$

is a QL distance.

Each of the examples above can actually be viewed in two different ways. For example, suppose we do not restrict the periodogram (or model) to Λ, but define D the same way (i.e. as integrating over Λ). Then D will be a QL distance *which does not result in optimal parametric estimates.* If the periodogram (and model) is viewed as being restricted to Λ, then D does result in optimal parametric estimates.

Example 5: Non-Gaussian process.

X(t) is a "filtered white noise process" of the form $X(t)=\sum a_s\ Z(t-s)$, where Z(t) is not assumed Gaussian. Then the operator $W(f(\lambda), \kappa)$ defined by $W(f(\lambda), \kappa)[x(\lambda)] = 2\pi f^2(\lambda)\ x(\lambda) + 2\pi\kappa f(\lambda) \int f(\psi)x(\psi)\ d\psi$ is the QL operator which is the variance operator for the random L^2 sequence (see p. 47 Rosenblatt (1985)). The QL distance (if it exists) associated with this operator is not known.

6.5 Theorems about QL distances

Chapter 8 will give proofs that the periodogram (defined continuously or as a step function, under various smoothness conditions on the spectrum) is a random L^2 sequence, and proofs that the examples described in the preceding section are indeed QL distances. Here, however, we will give some general propositions and theorems needed in the next chapter about QL distances which follow directly from the definition.

Proposition 6.5.1

If there exists an operator $W_\psi(x(\lambda))[\ .\ ]$ so $\frac{\partial}{\partial[x(\lambda)]} D_\psi(x(\lambda), y(\lambda)) = W_\psi(x(\lambda))[x(\lambda) - y(\lambda)]$, then $W(x(\lambda))$ is unique.

proof

SL^2 contains a set whose span is dense in $\Pi_k L^2$ (indicator functions of open intervals, for example). Any two continuous operators having the same values on a dense set must be the same. □

Proposition 6.5.2

Let (y_n, y_0, M, SL^2, V) be a random L^2 sequence.

a) For any $\epsilon, \delta > 0$, $\exists$ K so that if $||\psi||_2 \leq K$, $P\{|\psi \bullet y_n| > \epsilon\} < \delta$.

b) For any K, $\epsilon, \delta > 0$, $\exists$ N so that if $n \geq N$ and $||\psi||_2 \leq K$, $P\{|\psi \bullet (y_n - y_0)| > \epsilon\} < \delta$.

proof

(a) Choose K so $\frac{K_1 K^2}{(\epsilon/2)^2} \leq \delta$ and $K_2 K < \epsilon/2$, where K_1 and K_2 are from (e) and (f) of definition 6.3.2 (let $\psi_1 = \psi_2 = \psi$ in part (e) to choose K_1). Suppose $||\psi||_2 \leq K$. Then $|E(\psi \bullet y_n)| \leq K_2 K \leq \epsilon/2$. By Chebychev's inequality,

$$P\{|\psi \bullet y_n - E(\psi \bullet y_n)| > \epsilon/2\} \leq \frac{||\psi||_2^2 K_1}{(\epsilon/2)^2} \leq \frac{K_1 K^2}{(\epsilon/2)^2} \leq \delta.$$

(b) Choose N_1 by definition 6.3.2 (g) so $n \geq N_1 \Rightarrow |E(\psi \bullet y_n) - \psi \bullet y_0| \leq \epsilon/2$. Choose N_2 so $n \geq N_2 \Rightarrow \frac{K_1 K^2}{(\epsilon/2)^2 n} < \delta$. Let $N = \max\{N_1, N_2\}$. Then as in (a), $n \geq N_2 \Rightarrow P\{|\psi \bullet y_n - E(\psi \bullet y_n)| > \epsilon/2\} \leq \delta$, by Chebychev's inequality. But $n \geq N_1$ also implies

$\{|\psi \bullet y_n - \psi \bullet y_0)|>\epsilon\} \subset \{|\psi \bullet y_n - E(\psi \bullet y_n)|>\epsilon/2\}$, and so we are done. □

Proposition 6.5.3

Suppose $W_\psi(x)[\ .\ ]$ satisfies (c) in the definition of QL distance, and (y_n, y_0, M, SL^2, V) is a random L^2 sequence. Then

(a) If S_1 is a compact subset of M, S_2 is a compact subset of $\Pi_k L^2$, $\psi_0 \in \Psi$, and $\epsilon>0$, then there exists a neighborhood U of ψ_0 so that $||W^*_\psi(x)[y] - W^*_{\psi_0}(x)[y]||_2<\epsilon$ for $x \in S_1$, $y \in S_2$.

(b) If S_1 is a compact subset of M, S_2 is a compact subset of $\Pi_k L^2$, $\epsilon>0$, $\delta>0$, and $\psi_n \xrightarrow{P} \psi_0$, then $\exists N$ so $n \geq N \Rightarrow |y_n \bullet W^*_{\psi_n}(k_1)[k_2] - y_0 \bullet W^*_{\psi_0}(k_1)[k_2]|<\epsilon$ for all $k_1 \in S_1$, $k_2 \in S_2$ off a set with probability less than δ.

proof

(a) For each $(x, y) \in S_1 \times S_2$, choose a neighborhood $U_{(x,y)}$ of ψ_0 and neighborhoods B_x of x, B_y of y so $(\psi, x', y') \in U_{(x,y)} \times B_x \times B_y$ implies $||W^*_\psi(x')[y'] - W^*_{\psi_0}(x)[y]||_2<\epsilon/2$ (by condition (c) of definition 6.3.4). $B_x \times B_y$ is a cover of $S_1 \times S_2$, so there exists a finite set $\{(x_i, y_i)\}$ with $\{B_{x_i} \times B_{y_i}\}$ a finite subcover. Let $U = \bigcap_i U_{(x_i,y_i)}$. If $(\psi, x, y) \in U \times K_1 \times K_2$, choose i so $(x, y) \in B_{x_i} \times B_{y_i}$. Then

$$||W^*_\psi(x)[y] - W^*_{\psi_0}(x)[y]||_2 \leq ||W^*_\psi(x)[y] - W^*_{\psi_0}(x_i)[y_i]||_2 + ||W^*_{\psi_0}(x_i)[y_i] - W^*_{\psi_0}(x)[y]||_2$$

Each of these is less than $\epsilon/2$, so we are done.

(b) By part (a), Choose U a neighborhood of ψ_0 so $||W^*_\psi(x)[y] - W^*_{\psi_0}(x)[y]||_2<\epsilon/2$ and $||W^*_\psi(x)[y] - W^*_{\psi_0}(x)[y]||_2\ ||y_0||_2<\epsilon/2$ for $x \in S_1$, $y \in S_2$, $\psi \in U$. By proposition 6.5.2 choose N_1 so $n \geq N_1 \Rightarrow P\{|\psi \bullet (y_n - y_0)|>\epsilon/2\}<\delta/2$ for K satisfying $||W^*_\psi(k_1)[k_2]||<K$ for $k_1 \in S_1$, $k_2 \in S_2$ (K exists by definition 6.3.4 (c)). Choose N_2 so $n \geq N_2 \Rightarrow P\{\hat{\psi}_n \notin U\}<\delta/2$. Then if $N=\max\{N_1, N_2\}$, off a set with probability less than δ we have

$|y_n \bullet W^*_{\psi_n}(k_1)[k_2] - y_0 \bullet W^*_{\psi_0}(k_1)[k_2]| \leq$

$|y_n \bullet W^*_{\psi_n}(k_1)[k_2] - y_0 \bullet W^*_{\psi_n}(k_1)[k_2]| + |y_0 \bullet W^*_{\psi_n}(k_1)[k_2] - y_0 \bullet W^*_{\psi_0}(k_1)[k_2]| \leq \epsilon$

(by proposition 6.5.2 (b)). □

Theorem 6.5.1

Suppose

1) (y_n, y_0, M, SL^2, V) is a random L^2 sequence.

2) $\psi_n \xrightarrow{P} \psi_0$.

3) $\{f_\theta\}_{\theta \in \Theta_0}$ is a model.

4) $\theta_0 \in \Theta$.

Then

(a) $D_{\psi_n}(f_\theta(\lambda), y_n(\lambda)) - D_{\psi_n}(f_{\theta_0}(\lambda), y_n(\lambda)) \rightarrow D_{\psi_0}(f_\theta(\lambda), y_0(\lambda)) - D_{\psi_0}(f_{\theta_0}(\lambda), y_0(\lambda))$

uniformly in probability for $\theta \in \Theta$.

(b) If $W_\psi(x(\lambda))$ satisfies (c) and (d) in the definition of QL distance, then $\frac{\partial f_\theta}{\partial \theta_i} \bullet W_{\psi_n}(f_\theta)[f_\theta - y_n] \rightarrow \frac{\partial f_\theta}{\partial \theta_i} \bullet W_{\psi_0}(f_\theta)[f_\theta - y_0]$ uniformly in probability for $\theta \in \Theta$.

proof

(a) Let $\epsilon > 0$ and $\delta > 0$. For each n, by the mean value theorem (Chae (1985), p. 69) applied to the function $G(\theta) = D_{\psi_0}(f_\theta, y_0(\lambda)) - D_{\psi_n}(f_\theta, y_n(\lambda))$ we may write

$[D_{\psi_0}(f_{\theta_0}(\lambda), y_0(\lambda)) - D_{\psi_n}(f_{\theta_0}(\lambda), y_n(\lambda))] - [D_{\psi_0}(f_{\theta_1}(\lambda), y_0(\lambda)) - D_{\psi_n}(f_{\theta_1}(\lambda), y_n(\lambda))] =$

$[W_{\psi_0}(f_{\theta_n})[f_{\theta_n} - y_0] - W_{\psi_n}(f_{\theta_n})[f_{\theta_n} - y_n] \bullet \left(\left[\frac{\partial f_{\theta_n}}{\partial \theta'} \right] (\theta_1 - \theta_0) \right)$ for some θ_n between θ_0 and θ_1.

Rearrange the left side of the equality to read

$$[\, D_{\psi_n}(f_{\theta_1}(\lambda),\, y_n(\lambda)\, - D_{\psi_n}(f_{\theta_0}(\lambda),\, y_n(\lambda))\,] - [\, D_{\psi_0}(f_{\theta_1}(\lambda),\, y_0(\lambda)) - D_{\psi_0}(f_{\theta_0}(\lambda),\, y_0(\lambda))\,]$$

$$= [W_{\psi_0}(f_{\theta_n})[f_{\theta_n} - y_0] - W_{\psi_n}(f_{\theta_n})[f_{\theta_n} - y_n]] \bullet \left(\left[\frac{\partial f_{\theta_n}}{\partial \theta'} \right] (\theta_1 - \theta_0) \right).$$

Note that

$$[W_{\psi_0}(f_{\theta_n})[f_{\theta_n} - y_0] - W_{\psi_n}(f_{\theta_n})[f_{\theta_n} - y_n]] \bullet \left(\left[\frac{\partial f_{\theta_n}}{\partial \theta'} \right] (\theta_1 - \theta_0) \right) =$$

$$[W_{\psi_0}(f_{\theta_n})[f_{\theta_n}] - W_{\psi_n}(f_{\theta_n})[f_{\theta_n}]] \bullet \left(\left[\frac{\partial f_{\theta_n}}{\partial \theta'} \right] (\theta_1 - \theta_0) \right) +$$

$$[W_{\psi_n}(f_{\theta_n})[y_n] - W_{\psi_0}(f_{\theta_n})[y_0] \bullet \left(\left[\frac{\partial f_{\theta_n}}{\partial \theta'} \right] (\theta_1 - \theta_0) \right)$$

$$= [W^*_{\psi_0}(f_{\theta_n}) - W^*_{\psi_n}(f_{\theta_n})] \left(\left[\frac{\partial f_{\theta_n}}{\partial \theta'} \right] (\theta_1 - \theta_0) \right) \bullet [f_{\theta_n}] +$$

$$y_n \bullet W^*_{\psi_n}(f_{\theta_n}) \left(\left[\frac{\partial f_{\theta_n}}{\partial \theta'} \right] (\theta_1 - \theta_0) \right) - y_0 \bullet W^*_{\psi_0}(f_{\theta_n}) \left(\left[\frac{\partial f_{\theta_n}}{\partial \theta'} \right] (\theta_1 - \theta_0) \right).$$

Define $F: \mathbf{R}^p \times \mathbf{R}^p \to \Pi_k L^2$ by $F(\theta_1,\, \theta_2) = \left[\frac{\partial f_{\theta_1}}{\partial \theta'} \right] \theta_2$. Assuming the map $\theta \to \frac{\partial f_\theta}{\partial \theta_i}$ $(\mathbf{R}^p \to L^2)$ is continuous (i=1, . . , p), F will be continuous, and hence the image of the compact set $\Theta \times \Theta$ is compact in $\Pi_k L^2$. Of course, $\{f_\theta\}$ is compact in $\Pi_k L^2$ by definition. Because of this, proposition 6.5.3 yields that each of the two pieces converge uniformly in probability to 0.

(b) follows directly from proposition 6.5.3 □

Corollary 6.5.1 (to theorem 6.5.1)

(a) If $\hat{\theta}_n$ minimizes $D_{\psi_n}(f_\theta,\, f_n)$ and θ_0 minimizes $D_{\psi_0}(f_\theta,\, f)$, then assuming θ_0 is unique,

$\widehat{\theta}_n \xrightarrow{P} \theta_0$ under the conditions of theorem 6.5.1.

(b) If $\frac{\partial f_\theta}{\partial \theta} \odot W(f_\theta)[f_\theta - y_0]$ has a unique zero at θ_0, and $\widehat{\theta}_n$ is a zero of $\frac{\partial f_\theta}{\partial \theta} \odot W(f_\theta)[f_\theta - y_n]$, then $\widehat{\theta}_n \xrightarrow{P} \theta_0$ under the conditions of theorem 6.5.1 (b) (Note: see definition 7.1.1 for "$\odot$").

proof

(a) We elaborate on the argument in the proof of Taniguchi's (1979) theorem 1 (ii). Specifically, fix $\theta_0 \in \Theta$ and define $h_n(t)=D_{\psi_n}(f_t, f_n) - D_{\psi_n}(f_{\theta_0}, f_n)$, $h(t)=D_{\psi_0}(f_t, f) - D_{\psi_0}(f_{\theta_0}, f)$. Of course, $\widehat{\theta}_n$ also minimizes h_n and θ_0 minimizes h. Given $\epsilon>0$, there must exist a $\delta>0$ such that $|h(t) - h(\theta_0)|<\delta \Rightarrow ||t - \theta_0||<\epsilon$. (if not, choose t_n so $||t_n - \theta_0||>\epsilon$ and $|h(t_n) - h(\theta_0)|<1/n$. As Θ is compact, $\{t_n\}$ must have a cluster point t_0, implying there is a subsequence t_{n_k} converging to t_0. Apparently $|h(t_0) - h(\theta_0)|=0$, a contradiction) Suppose ϵ and δ are given positive numbers. Choose ϵ_2 so $|h(t) - h(\theta_0)|<\epsilon_2 \Rightarrow ||t - \theta_0||<\epsilon$. By theorem 6.5.6 (a) there exists N so $n \geq N \Rightarrow |h_n(t) - h(t)|<\epsilon_2$ except on a set S with probability less than δ. Hence off S (and for $n \geq N$), $|h_n(\theta_n) - h(\theta_0)|<\epsilon_2$ also (because if the surfaces are always within ϵ_2 of each other, so must be their minimums), and so $||\widehat{\theta}_n - \theta_0||<\epsilon$.

(b) Define $h_n(t) = \frac{\partial f_t}{\partial t} \odot W_{\psi_n}(f_t)[f_t - f_n]$, $h(t) = \frac{\partial f_t}{\partial t} \odot W_{\psi_0}(f_t)[f_t - f_0]$. Given $\epsilon>0$, there must exist a $\delta>0$ such that $||h(t)||_\infty<\delta \Rightarrow ||t - \theta_0||<\epsilon$, where the norm on $h(\lambda)$ is the supremum of the entries in the (real valued) matrix h(t) (prove by contradiction as above). Suppose ϵ and δ are given positive numbers. Choose ϵ_2 so $||h(t)||_\infty<\epsilon_2 \Rightarrow ||t - \theta_0||<\epsilon$. By theorem 6.5.6 (b) there exists N so $n \geq N \Rightarrow ||h_n(t) - h(t)||_\infty<\epsilon_2$ except on a set S with probability less than δ (use the theorem for each component of the matrix). Hence off S (and for $n \geq N$), we have $||h_n(\widehat{\theta}_n) - h(\widehat{\theta}_n)||_\infty<\epsilon_2$, so $||h(\widehat{\theta}_n)|_\infty|<\epsilon_2 \Rightarrow ||\widehat{\theta}_n - \theta_0||<\epsilon$.

□

The following proposition 6.5.4, theorem 6.5.2 and corollary 6.5.2 will be used in chapter 7, but resemble material in this section (needed to prove corollary 6.5.1).

Proposition 6.5.4

Suppose condition (c) of the definition of QL distance holds, and $\psi_n \to \psi_0 \in \Psi$, $x_n \to x_0 \in \Pi_k C$, $x_n, x_0 \in M$, $g_n \to g_0 \in \Pi_k C$. If either

(a) (y_n, y_0, M, V, SL^2) is a random L^2 sequence, or

(b) $y_n \to y_0 \in \Pi_k L^2$, where convergence is in the L^2 norm

then $g_n \bullet W_{\psi_n}(x_n(\lambda))[y_n] \to g \bullet W_{\psi_0}(x_0(\lambda))[y_0]$. If the convergences (e.g. $\psi_n \to \psi_0$, $x_n \to x_0$, $g_n \to g_0$, $y_n \to y_0$) hold "in probability", then the conclusion holds "in probability".

proof

$$| g_n \bullet W_{\psi_n}(x_n(\lambda))[y_n] - g \bullet W_{\psi_0}(x_0(\lambda))[y_0]| \leq$$

$$|g_n \bullet W_{\psi_n}(x_n(\lambda))[y_n] - g \bullet W_{\psi_n}(x_n(\lambda))[y_n]| + |g \bullet W_{\psi_n}(x_n(\lambda))[y_n] - g \bullet W_{\psi_0}(x_0(\lambda))[y_0]|$$

$$\leq |W^*_{\psi_n}(x_n(\lambda))[g_n - g] \bullet y_n| + |W^*_{\psi_n}(x_n(\lambda))[g] \bullet y_n - W^*_{\psi_0}(x_0(\lambda))[g] \bullet y_0|$$

Under the conditions in (a), the first piece goes to 0 by proposition 6.5.2 (a) and the boundedness of $||W^*_{\psi_n}(x_n(\lambda))||$ (definition 6.3.4 (c)). The second piece goes to 0 by proposition 6.5.2 (b) and definition 6.3.4 (c).

Under the conditions in (b), the first piece goes to 0 because $|W^*_{\psi_n}(x_n(\lambda))[g_n - g] \bullet y_n$ $\leq$ $||W^*_{\psi_n}(x_n(\lambda))||$ $||g_n - g||$ $||y_n||$, and $||W^*_{\psi_n}(x_n(\lambda))||$, $||y_n||$ are bounded.

If we assume the convergences are in probability, then the limits and bounds in the above proof will hold in probability, and so must the conclusion. □

Theorem 6.5.2

Suppose the mapping from $\Psi \times \Pi_k C \to B(\Pi_k L^2)$ defined by $F(\psi, x(\lambda)) = W_\psi(x(\lambda))[\ .\ ]$ satisfies condition (d) in the definition of QL distance, $\psi_n \to \psi_0 \in \Psi$, $f_n \to f_0 \in \Pi_k C$, $g_{in} \to g_{i0} \in \Pi_k C$ for i=1, 2, and g_{in} , $g_{i0} \in M$. If either

(a) (y_n, y_0, M, V, SL^2) is a random L^2 sequence , or

(b) $y_n \to y_0 \in \Pi_k L^2$, where convergence is in the L^2 norm

then $g_{1n} \bullet L_{\psi_n f_n}(g_{2n})\ [y_n] \to g_{10} \bullet L_{\psi_0 f_0}(g_{20})\ [y_0]$. If the convergences (e.g. $\psi_n \to \psi_0$, $f_n \to f_0$, $g_{in} \to g_{i0}$, $y_n \to y_0$) hold "in probability", then the conclusion holds "in probability".

proof

Observe $g_{1n} \bullet L_{\psi_n f_n}(g_{2n})[y_n] = \left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}] \bullet y_n$, and so it suffices to show that

$$\left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}] \bullet y_n \to \left(L_{\psi_0 f_0}(g_{20})\right)^*[g_{10}] \bullet y_0.$$

(a) $\left|\left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}] \bullet y_n - \left(L_{\psi_0 f_0}(g_{20})\right)^*[g_{10}] \bullet y_0\right| \leq$

$$\left|\left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}] \bullet y_n - \left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}] \bullet y_0\right| +$$
$$\left|\left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}] \bullet y_0 - \left(L_{\psi_0 f_0}(g_{20})\right)^*[g_{10}] \bullet y_0\right|.$$

The first piece goes to 0 by the boundedness of $\left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}]$ and proposition 6.5.2 (a). The second piece goes to 0 by definition 6.3.4 (d) and the Cauchy Schwartz inequality.

(b) The difference breaks up the same as in part (a). The first piece is bounded by $\left\|\left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}]\right\| \; ||y_n - y_0||$, which goes to 0 by definition 6.3.4 (d) (boundedness condition (iii)). The second piece is bounded by

$$\left\|\left(L_{\psi_n f_n}(g_{2n})\right)^*[g_{1n}] - \left(L_{\psi_0 f_0}(g_{20})\right)^*[g_{10}]\right\| \; ||y_0||_2$$

which again goes to 0 by definition 6.3.4 (d). □

Corollary 6.5.2 (to theorem 6.5.2)

If ψ_n, ψ_0, g_{in}, (i=1,2), g_{10}, g_{20}, f_n, f_0, y_n, y_0 are exactly as in theorem 6.5.2, defining $Q_\psi(x(\lambda), y(\lambda)) \equiv W_\psi(x(\lambda))[y(\lambda)]$ (for each fixed $y(\lambda) \in \Pi_k L^2$, a mapping $\Pi_k C \to \Pi_k L^2$), we have

$$g_{1n} \bullet \left[\frac{\partial Q_{\psi_n}(f_n(\lambda),\, y_n(\lambda))}{\partial x(\lambda)}\,[g_{2n}]\right] \to g_1 \bullet \left[\frac{\partial Q_{\psi_0}(f_0(\lambda),\, y_0(\lambda))}{\partial x(\lambda)}\,[g_2]\right].$$

proof

It suffices to verify that $\dfrac{\partial Q_{\psi_n}(f_n(\lambda),\, y_n(\lambda))}{\partial x(\lambda)}\,[g_{2n}] = L_{\psi_n f_n}(g_{2n})\,[y_n]$. This follows from the chain rule: regard the mapping $x(\lambda) \to W(x(\lambda))[y(\lambda)]$ from $\Pi_k C \to \Pi_k L^2$ as a composition of the maps F_1: $\Pi_k C \to B(\Pi_k L^2)$ defined by $x(\lambda) \to W(x(\lambda))[\ .\ ]$ and F_2: $B(\Pi_k L^2) \to \Pi_k L^2$ defined by $W \to W[y(\lambda)]$. So the derivative of the composition, evaluated at (ψ_n, f_n) and applied to g_{2n} is $F_2'\big|_{(\psi_n, F_1(f_n))} \circ F_1'(f_n)\,[g_{2n}]$. Note that F_2 is linear, and so is its own derivative (i.e evaluation of an operator at y_n). $F_1'(f_n)[g_{2n}]$ is by definition the operator $L_{\psi_n f_n}(g_{2n})\,[\ .\ ]$.

□

6.6 Conclusions

In this chapter, we have observed that non Gaussian time series have periodograms which do not "act" like independent, exponential random variables as do the periodograms of Gaussian series. We have formalized the problem of "generalized nonlinear models" in possibly multivariate function spaces in terms of linear functionals and linear operators. Section 6.2 reviewed "calculus" on normed linear spaces as a prelude to our applications in section 6.3. Section 6.3 defined "QL distances" and "QL operators" (definition 6.3.4), viewing the QL distance as a nonlinear functional on the "model space", and gave the general setup for the types of estimation problems to be considered in the dissertation. Section 6.4 showed how examples from spectral estimation fit the framework of section 6.3. Section 6.5, specifically, corollary 6.5.1, gave consistency results of parametric estimates obtained from minimization of the QL distance or solution of the QL equations to be defined in chapter 7.

Now that the basic ideas of the dissertation have been introduced, some specific theorems regarding the "optimality" of parametric estimates obtained by using the QL operators of chapter 6 must be established. This will be done in the following chapter.

Chapter VII
Generalized Optimality

7.1 Introduction

Before stating the main results of this section, it is necessary to first establish some notation and technical results.

First, note that the model $\{f_\theta\}$ can be regarded as a (nonlinear) mapping $\Theta \rightarrow \Pi_k C$, where Θ consists of column vectors in $\mathbf{R}^p$. As such, the derivative with respect to θ is a linear mapping $\Theta \rightarrow \Pi_k C$, and hence $\frac{\partial f_\theta}{\partial \theta}$ can be written as the column vector of $\Pi_k C$ functions

$$\left[\frac{\partial f_\theta}{\partial \theta_1}, \frac{\partial f_\theta}{\partial \theta_2}, \cdots \frac{\partial f_\theta}{\partial \theta_p} \right]'.$$

The derivative of f_θ, evaluated at θ_0 and applied to θ_1 is $\frac{\partial f_{\theta_0}}{\partial \theta'} \theta_1$. Apply the row vector of functions to the column vector θ_1 of scalars pointwise to get a function. This is because of propositions 6.2.2 and 7.2.2. Propositions 6.2.2 and 7.2.2 say that if $\Theta \subset \mathbf{R}$, then the ordinary partial derivative of $f(\theta, \lambda)$ with respect to θ (a vector in $\Pi_k C$) is the derivative of the nonlinear mapping between Θ and $\Pi_k C[a, b]$. Proposition 7.2.2 shows the form of the derivative if θ is a vector in $\mathbf{R}^p$.

We will define the derivative with respect to a column vector as a column vector, and use the notation $\frac{\partial f_\theta}{\partial \theta'}$ to denote row vectors. If $a(\theta)$ is a column vector, we use the notation $\frac{\partial a(\theta)}{\partial \theta'}$ to denote the matrix whose rows are the derivatives of the components of $a(\theta)$. Hence,

$\frac{\partial^2 f_\theta}{\partial\theta'\partial\theta}$ can be represented as the $p \times p$ matrix with entries $\Pi_k C$ functions

$$\begin{bmatrix} \frac{\partial^2 f_\theta}{\partial\theta_1\partial\theta_1} & \frac{\partial^2 f_\theta}{\partial\theta_2\partial\theta_1} & \cdot & \cdot & \cdot & \frac{\partial^2 f_\theta}{\partial\theta_p\partial\theta_1} \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \frac{\partial^2 f_\theta}{\partial\theta_1\partial\theta_p} & \frac{\partial^2 f_\theta}{\partial\theta_2\partial\theta_p} & \cdot & \cdot & \cdot & \frac{\partial^2 f_\theta}{\partial\theta_p\partial\theta_p} \end{bmatrix}$$

Definition 7.1.1

Suppose that A is an $n \times m$ matrix with entries from a Hilbert space H, and B is an $m \times k$ matrix with entries from H. Then we define $A \odot B$ to be the matrix with real valued entries defined by performing the usual matrix multiplication as if the matrices A and B had real entries, but replacing scalar multiplication with the Hilbert space inner product. If A consists of one element, then define $A \odot B$ to be the matrix obtained by replacing each element of B by the inner product of A with that element of B.

Proposition 7.1.1 (product rule)

Let H be a Hilbert space, and suppose $a(\theta) \in H$, $b(\theta) = (b_1(\theta), b_2(\theta), \ldots b_p(\theta))'$, where $b_i(\theta) \in H$ $i = 1, \ldots p$. Then $D_\theta \left[a(\theta) \odot b(\theta) \right] = a(\theta) \odot \frac{\partial b(\theta)}{\partial\theta'} + b(\theta) \odot \frac{\partial a(\theta)}{\partial\theta'}$.

proof

First observe that if $a(\theta)$ and $b(\theta)$ are both H valued functions and $\theta \in \mathbf{R}$, then

$$\frac{d}{d\theta}\left[a(\theta) \bullet b(\theta) \right] = a(\theta) \bullet b'(\theta) + b(\theta) \bullet a'(\theta)$$

(See remark following the proof of proposition 7.2.1, and use the chain rule on the mapping $\theta \rightarrow (a(\theta), b(\theta)) \rightarrow a(\theta) \bullet b(\theta)$). By proposition 7.2.2 (b), the derivative of the vector $a(\theta) \odot b(\theta)$

with respect to θ is

$$\begin{bmatrix} a(\theta)\bullet\frac{\partial b_1}{\partial\theta_1}+b_1(\theta)\bullet\frac{\partial a}{\partial\theta_1} & \cdot & \cdot & a(\theta)\bullet\frac{\partial b_1}{\partial\theta_p}+b_1(\theta)\bullet\frac{\partial a}{\partial\theta_p} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ a(\theta)\bullet\frac{\partial b_p}{\partial\theta_1}+b_p(\theta)\bullet\frac{\partial a}{\partial\theta_1} & \cdot & \cdot & a(\theta)\bullet\frac{\partial b_p}{\partial\theta_p}+b_p(\theta)\bullet\frac{\partial a}{\partial\theta_p} \end{bmatrix}$$

Observe that this is the matrix $a(\theta)\odot\frac{\partial b(\theta)}{\partial\theta}+b(\theta)\odot\frac{\partial a(\theta)}{\partial\theta'}$. □

Definition 7.1.2

Suppose A is a linear mapping $\mathbf{R}^p\to H$, and $W:H\to H$ is a linear operator on H. Then the composition $W\circ A$ may be represented as the row vector of H elements $[W(f_1), W(f_2), \ldots W(f_p)]$ with respect to the basis $e_1, e_2, \ldots e_p$ of $\mathbf{R}^p$, where $A(e_i)=f_i$, $i = 1, \ldots, p$. So we will define "$*$" to denote a representation of the composition $W\circ A$ as $W*[f_1, f_2, \ldots f_p] \equiv [W(f_1), W(f_2), \ldots W(f_p)]$.

Definition 7.1.2 will be used to obtain representations of expressions such as $D_\theta F(f_\theta)$, where $F:\Pi_k C\to\Pi_k C$. For example, by the chain rule, $D_\theta F(f_\theta)\big|_{\theta_0} = D_f F(f)\big|_{f_{\theta_0}} \circ \frac{\partial f_\theta}{\partial\theta}\Big|_{\theta_0}$, which can be represented as $D_f F(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial\theta}$ if $\frac{\partial f_{\theta_0}}{\partial\theta}$ is represented as a column vector of functions as discussed in the second paragraph of this section. When used as a superscript on a matrix or operator, "$*$" will, without confusion, denote the adjoint (i.e. for an operator A, $A[h]\bullet k = h\bullet A^*[k]$).

One final definition is necessary before stating the main results of this chapter.

Definition 7.1.3

Suppose (y_n, y, M, V, SL^2) is a random L^2 sequence, and $W(f)[\,.\,]$ is a QL operator.

(a) $\Phi_\psi(f, g) \equiv W_\psi(f)[g]$ for $f \in M$, $g \in \Pi_k L^2$.

(b) $M_W(\psi, \theta) \equiv [\Phi_\psi(f_\theta, f) - \Phi_\psi(f_\theta, f_\theta)] \odot \frac{\partial^2 f_\theta}{\partial\theta'\partial\theta} +$

$$\frac{\partial f_\theta}{\partial\theta} \odot \left(\left[W_\psi(f_\theta) + \frac{\partial\Phi_\psi(f_\theta, f_\theta)}{\partial x} - \frac{\partial\Phi_\psi(f_\theta, f)}{\partial x} \right] * \frac{\partial f_\theta}{\partial\theta'} \right) \qquad [7.1.1]$$

(a $p \times p$ matrix).

(c) $Q_W(\psi, \theta) \equiv \frac{\partial f_\theta}{\partial\theta} \odot \left(W_\psi(f_\theta) \ V \ W^*_\psi(f_\theta) * \frac{\partial f_\theta}{\partial\theta'} \right)$ [7.1.2]

(a $p \times p$ matrix).

(d) $M^0_W \equiv M_W(\psi_0, \theta_0)$, $Q^0_W \equiv Q_W(\psi_0, \theta_0)$.

(e) $M_V \equiv \frac{\partial f_\theta}{\partial\theta} \odot \left(V^{-1} * \frac{\partial f_\theta}{\partial\theta'} \right)$.

Theorem 7.1.1 (Consistency and Representation theorem)

Suppose

(1) (f_n, f, M, SL^2, V) is a random L^2 sequence.

(2) $\{f_\theta\}_{\theta \in \Theta}$ is a model for f, not assumed to contain f.

(3) $\hat{\psi}_n \xrightarrow{P} \psi_0$.

(4) $D^W_{\psi_0}(f_\theta, f)$ has a unique minimum for $\theta = \theta_0$.

(5) $\hat{\theta}_n$ is obtained by minimizing the QL function $D^W_{\hat{\psi}}(\,.\,,\,.\,)$.

(6) M^0_W is invertible.

Then $\hat{\theta}_n$ is a consistent estimator of θ_0. Furthermore, we may write $\hat{\theta}_n - \theta_0 = L^W[f_n - f] + \epsilon_n$, where L^W: $\Pi_k L^2 \rightarrow \mathbf{R}^p$ is a linear operator defined by

$$L^W(h) = (M_W^0)^{-1}\left[\left(W^*_{\psi_0}(f_{\theta_0}) * \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta} \right] \right) \odot h(\lambda) \right] \qquad [7.1.3]$$

and $\sqrt{n}\epsilon_n \xrightarrow{P} 0$.

Note that [7.1.3] is the reason for the "link condition" in the definitions of the preceding chapter. The link condition creates asymptotic unbiasedness in $\hat{\theta}$, as it is needed for $\sqrt{n}$ $L^W[f_n - f]$ to have an asymptotic mean of 0.

Corollary 7.1.1

The conclusion of theorem 7.1.1 holds if $W_\psi(\ )[\ .\ ]$ is a QL operator and (4) and (5) are replaced with

(4′) The QL equations

$$W_\psi(f_\theta)[f_\theta - g] \odot \frac{\partial f_\theta}{\partial \theta} = 0 \qquad [7.1.4]$$

have a unique solution θ_0.

(5′) $\hat{\theta}_n$ is obtained as a solution to $W_{\hat{\psi}}(f_\theta)[f_\theta - f_n] \odot \frac{\partial f_\theta}{\partial \theta} = 0$.

Corollary 7.1.1 says it is only necessary to specify the relationship between the mean and variance operator in order to have a QL estimator, and that it is not necessary to define an actual QL distance as such.

Theorem 7.1.2 (Optimality Theorem)

If $\hat{\theta} - \theta_0$ can be expressed in the form of theorem 7.1.1 (5), then asymptotically Var $\sqrt{n}(\hat{\theta}_n - \theta_0) = (M_W^0)^{-1} Q_W^0 \left[(M_W^0)^{-1}\right]'$. The minimum this can attain in the sense of any linear combination of the parametric estimates having asymptotically smallest variance is M_V^{-1}. If the model contains the limiting function, a sufficient condition for the variance matrix to attain this minimum is if the operator $W^*(f_{\theta_0}) = cV^{-1}$ (where c is a positive real number) on span $\{\partial f_{\theta_0}(\lambda)/\partial\theta_i\}$ (For any finite collection $\{f_i\}$ of functions in $\Pi_k L^2$, we define span $\{f_i\}$ to be the finite dimensional vector space consisting of all finite linear combinations of elements of $\{f_i\}$).

A major focus of chapters 9 and 10 will be how to approximate the optimal variance matrix when the model does not contain the limiting function. There we shall give sufficient conditions for this to occur.

Corollary 7.1.2 (IRWLS corollary)

Suppose $\{f_\theta\}$ is a model for the random L^2 sequence (f_n, f, M, SL^2, V), and $W_\psi(f_\theta)[\ .\ ]$ is a self adjoint operator satisfying (c) in the definition of QL distance. Suppose

1) $\hat{\theta}_1 \xrightarrow{P} \theta_1$, $\hat{\psi} \xrightarrow{P} \psi_0$.

2) $(f_\theta - f) \bullet W_\psi(f_{\theta_1})[f_\theta - f]$ has a minimum at θ_0.

3) Defining

$$M_W^1(\psi, \theta) \equiv \left[[\Phi_\psi(f_\theta, f) - \Phi_\psi(f_\theta, f_\theta)] \odot \frac{\partial^2 f_\theta}{\partial\theta'\partial\theta}\right] + \left[\frac{\partial f_\theta}{\partial\theta} \odot \left(W_\psi(f_\theta) * \frac{\partial f_\theta}{\partial\theta'}\right)\right]$$

$M_W^1(\psi, \theta)$ is an invertible matrix for all θ, ψ.

Under these three assumptions, the $\hat{\theta}$ minimizing $(f_\theta - f_n) \bullet W_{\hat{\psi}}(f_{\hat{\theta}_1})[f_\theta - f_n]$ converges in probability to θ_0, and has an asymptotic variance matrix

$$[M_W^1(\psi_0, \theta_1)]^{-1} Q_W(\psi_0, \theta_1) [M_W^1(\psi_0, \theta_1)]^{-1}. \qquad [7.1.5]$$

Corollary 7.1.2 is analogous to Chiu's (1988) theorem 7. Its proof follows immediately

from theorem 7.1.1, defining "Ψ" to be $\{f_\theta\} \times \Psi$ and "$W_\psi[\ .\ ]$" to be $W_{\psi,\ f_\theta}[\ .\ ]$. Notice that the stronger conditions on the second derivative of W(f)[.] are not needed. In fact, W needn't even be defined off the model. Some examples of how this might occur will be given in chapter 9. Also observe the stronger conditions on the QL operator $W_\psi(\ .\ )$; we are assuming it to be self adjoint, a condition not assumed in theorems 7.1.1, 7.1.2 or corollary 7.1.1. Chapters 9 and 10 will discuss how to do IRWLS with a non self adjoint "inverse variance operator".

It should be noted that the IRWLS corollary, in the case of model misspecification, gives the asymptotic variance matrix for *one iteration.* If the model is misspecified, one step IRWLS and the solution of the QL equations *are not consistently estimating the same* θ_0. This is also reflected in the differences between M_W^1 and [7.1.1]: [7.1.1] has an extra "second derivative part" (of the QL function). Fully iterated IRWLS, i.e. until convergence, yields a solution to the QL equations (Green (1984), Carrol and Rupert (1988)). So in practice, for a misspecified model it will be necessary to study the "complete" expression [7.1.1] to obtain the correct asymptotic variance.

To prove these theorems, we need consistency and asymptotic optimality. Consistency follows from corollary 6.5.1. Assuming consistency, we will complete the proof of theorem 7.1.1 by making Taylor expansions (as did Taniguchi (1979)) to obtain the asymptotic linear representation of $\sqrt{n}(\widehat{\theta} - \theta_0)$ where $\widehat{\theta}$ is obtained by using any QL distance or operator as defined above. Before doing this, it will be helpful to obtain a series of technical propositions and corollaries needed in the proof of the main theorem. In these preliminary results, we shall suppress the "ψ" to simplify the notation.

The following two propositions are needed in the proof of the optimality theorem 7.1.2. They are essentially from Taniguchi (1979) in the univariate case, and are included here since they follow quickly from the definition of "random L^2 sequence".

Proposition 7.1.2

Let (f_n, f, M, SL^2, V) be a random L^2 sequence. If ψ_1 and ψ_2 are vectors with components in $\Pi_k L^2$, then

$$\lim_{n\to\infty} n \operatorname{cov}(\psi_1 \odot f_n, \psi_2 \odot f_n) = \psi_1' \odot [V * \psi_2].$$

The proof is straightforward by the definition of V (It's really just a matter of notation).

Proposition 7.1.3

Let (f_n, f, M, SL^2, V) be a random L^2 sequence, W a bounded operator on $\Pi_k L^2$, and ψ_1 and ψ_2 vectors with components in $\Pi_k L^2$. Then

$$\lim_{n\to\infty} n \operatorname{cov}(\psi_1(\lambda) \odot W[f_n], \psi_2(\lambda) \odot W[f_n]) = \psi_1(x) \odot (W \, V \, W^* * \psi_2'(x)).$$

proof

$\psi_1(\lambda) \odot W[f_n]$ and $\psi_2(\lambda) \odot W[f_n]$ may be rewritten as

$$[W^* * \psi_1(x)] \odot f_n \text{ and } [W^* * \psi_2(x)] \odot f_n$$

which have asymptotic covariance

$$[W^* * \psi_1(x)] \odot \left(V * [W^* * \psi_2(x)]' \right) \quad \text{(by proposition 7.1.2)}$$

$$= \psi_1(x) \odot \left(W * \left[V * [W^* * \psi_2(x)]' \right]' \right) \text{(by the definition of adjoint)}$$

$$= \psi_1(x) \odot [W \, V \, W^* * \psi_2'(x)]. \qquad \square$$

7.2 Proof of Representation Theorem

Let θ_0 be the true value of the parameter which minimizes $D^W_{\psi_0}(f_\theta, f)$.

Let $\hat{\theta}_n$ minimize $D^W_{\hat{\psi}}(f_\theta, f_n)$.

Proposition 7.2.1

$$\frac{\partial}{\partial\theta} D^W(f_\theta, f_n) = W(f_\theta)[f_\theta] \odot \frac{\partial f_\theta}{\partial\theta} - W(f_\theta)[f_n] \odot \frac{\partial f_\theta}{\partial\theta}$$

Note that $\frac{\partial}{\partial\theta} D^W(f_\theta, f_n)$ is a vector in $\mathbf{R}^p$, and using previous notation the right side of the equation may be written as $[\Phi(f_\theta, f_\theta) - \Phi(f_\theta, f_n)] \odot \frac{\partial f_\theta}{\partial\theta}$. $W(f_\theta))[f_\theta]$ and $W(f_\theta)[f_n]$ are viewed as functionals on $\Pi_k L^2$. They are applied to each of the p entries of the column vector $\frac{\partial f_\theta}{\partial\theta}$ to get a column of scalars, a scalar in place of each row. This is easily expressed using the "$\odot$" notation, as explained in the proof below.

proof

Recalling $\frac{\partial}{\partial[x(\lambda)]} D(x(\lambda), y(\lambda)) = \Phi(x(\lambda), x(\lambda)) - \Phi(x(\lambda), y(\lambda))$ (which is a linear functional), we have by the chain rule $\frac{\partial}{\partial\theta} D^W(f_\theta, f_n) = [W(f_\theta))[f_\theta] - W(f_\theta)[f_n]] \circ \frac{\partial f_\theta}{\partial\theta}$, where $\frac{\partial f_\theta}{\partial\theta}$ is the derivative of the mapping between Θ and $\Pi_k C$ defined by $\theta \to f_\theta$. Here, as in many other similar expressions throughout the dissertation, we are slightly abusing notation for simplicity in that we are taking the derivative with respect to θ and evaluating the derivative at θ. Using definition 7.1.2 and the proper "representation" as a column vector of functions for $\frac{\partial f_\theta}{\partial\theta}$, "$\circ$" becomes "$\odot$". □

We will follow Taniguchi (1979) and make a Taylor expansion of each term in

proposition 7.2.1. But to do this some preliminary theorems are needed.

As noted in Chae (1985) p. 95, if $f:U_1 \times U_2 \rightarrow F$ is differentiable at $(a,b) \in U_1 \times U_2$, then $D_1f(a, b)$ and $D_2f(a, b)$ exist and $Df(a, b)(x, y) = D_1f(a, b)(x) + D_2f(a, b)(y) = [D_1f(a, b), D_2f(a, b)] \begin{bmatrix} x \\ y \end{bmatrix}$. This is easy to see by definition of the derivative, taking the limit along the coordinate axes and using the uniqueness of the derivative. There is a converse to this theorem which would be stated as follows.

Proposition 7.2.2

(a) Suppose X and Y are normed linear spaces, U_1 is open in X and U_2 is open in Y. If $D_1f(a, b)$ and $D_2f(a, b)$ exist and are continuous at (a, b), then $f: U_1 \times U_2 \rightarrow F$ is differentiable at $(a,b) \in U_1 \times U_2$ with derivative given by $Df(a, b)(x, y) = D_1f(a, b)(x) + D_2f(a, b)(y)$.

(b) Let U_i and V_j be subsets of normed linear spaces X_i, Y_j, i=1. . p, j=1 . . q, and suppose c is an interior point of ΠU_i. Let $f: \Pi U_i \rightarrow \Pi V_j$ be a (nonlinear) mapping so that the partial derivatives $D_i f_j(c)$ (a linear operator $X_i \rightarrow Y_j$ for each c) exist and are continuous at c. Then f is differentiable at c, and the linear operator $Df(c) : \Pi X_i \rightarrow \Pi Y_j$ may be represented by the operator matrix

$$\begin{bmatrix} D_1f_1(c) & D_2f_1(c) & \dots & D_pf_1(c) \\ D_1f_2(c) & D_2f_2(c) & \dots & D_pf_2(c) \\ \cdot & \cdot & \cdot & \cdot \\ D_1f_q(c) & D_2f_q(c) & \cdot & D_pf_q(c) \end{bmatrix}$$

in the sense that a vector $(x_1, x_2, \dots x_p)'$ in ΠX_i is mapped to the vector obtained by applying the operator matrix to it.

proof

This proposition is well known in the case of functions $f:\mathbf{R}^p \to \mathbf{R}^q$. In fact, a proof is given in Bartle (1976) p. 355 which applies word for word to prove proposition 7.2.2. □

By the chain rule, regarding the domain of Φ as being the product space $\Pi_k C \times \Pi_k L^2$ with product of sup norm and Hilbert space norm, and the range of Φ as being $\Pi_k L^2$ with the Hilbert norm, we have the following proposition.

Proposition 7.2.3

$$D_x\Phi(x(\lambda), x(\lambda))\Big|_{x_0(\lambda)} = \frac{\partial\Phi(x(\lambda), y(\lambda))}{\partial[x(\lambda)]}\Bigg|_{(x_0(\lambda),x_0(\lambda))} + \frac{\partial\Phi(x(\lambda), x(\lambda))}{\partial[y(\lambda)]}\Bigg|_{(x_0(\lambda), x_0(\lambda))}$$

Note: Here, we are taking the derivative of the mapping $\Pi_k C \to \Pi_k L^2$ defined by $x(\lambda) \to \Phi(x(\lambda), x(\lambda))$ and evaluating it at the function $x_0(\lambda)$. This derivative (evaluated at $x_0(\lambda)$) must be a linear mapping $\Pi_k C \to \Pi_k L^2$. In fact, each piece of the expression on the right side of the equality is a linear mapping $\Pi_k C \to \Pi_k L^2$. The first piece is the derivative of the mapping $x(\lambda) \to \Phi(x(\lambda), x_0(\lambda))$ (where the second coordinate is fixed at $x_0(\lambda)$ before taking the derivative) evaluated at $x_0(\lambda)$, and the second piece is the derivative of the mapping $x(\lambda) \to \Phi(x_0(\lambda), x(\lambda))$ (where the first coordinate is fixed at $x_0(\lambda)$ before taking the derivative), evaluated at $x_0(\lambda)$.

proof

View $\Phi(x(\lambda), x(\lambda))$ as the composition $\Phi \circ F$, where $F[x(\lambda)] = [x(\lambda), x(\lambda)]'$ $(\Pi_k C \to \Pi_k C \times \Pi_k L^2)$. So by the chain rule and proposition 7.2.2, the derivative evaluated at $z(\lambda)$ and applied to $h(\lambda)$ is $D\Phi(x(\lambda), y(\lambda))\big|_{(z(\lambda), z(\lambda))}$ $[h(\lambda), h(\lambda)]'$ (as F is linear), which equals

$$\left[\left.\frac{\partial\Phi(x(\lambda),\, y(\lambda))}{\partial[x(\lambda)]}\right|_{(z(\lambda),\, z(\lambda))},\ \left.\frac{\partial\Phi(x(\lambda),\, y(\lambda)}{\partial[y(\lambda)]}\right|_{(z(\lambda),\, z(\lambda))}\right]\begin{bmatrix}h(\lambda)\\ h(\lambda)\end{bmatrix}$$

With some abuse of notation for simplification in writing the expression, this is equal to

$$\frac{\partial\Phi(z(\lambda),\, z(\lambda))}{\partial[x(\lambda)]}\,[h(\lambda)] + \frac{\partial\Phi(z(\lambda),\, z(\lambda))}{\partial[y(\lambda)]}\,[h(\lambda)]$$

$$= \left[\frac{\partial\Phi(z(\lambda),\, z(\lambda))}{\partial[x(\lambda)]} + \frac{\partial\Phi(z(\lambda),\, z(\lambda))}{\partial[y(\lambda)]}\right][h(\lambda)]. \qquad \square$$

Remark: For $h \in \Pi_k C$, $\frac{\partial\Phi(a(\lambda),\ b(\lambda))}{\partial[x(\lambda)]}$ [h] (the partial derivative evaluated at $(a(\lambda), b(\lambda)) \in \Pi_k C \times \Pi_k L^2$ and applied to h), equals $((D_x W(a(\lambda))\ [h(\lambda)])\ [b(\lambda)]$ by the chain rule. $D_x W(a(\lambda))$ is a linear mapping from $\Pi_k C \to B(\Pi_k L^2)$, so $D_x W(a(\lambda))[h(\lambda)] \in B(\Pi_k L^2)$. The operator is applied to $b(\lambda) \in \Pi_k L^2$ to obtain another element of $\Pi_k L^2$.

Corollary 7.2.1 (to proposition 7.2.3)

$$\frac{\partial}{\partial\theta}\,\Phi(f_\theta,\, f_\theta) = \left[\frac{\partial\Phi(f_\theta,\, f_\theta)}{\partial[x(\lambda)]} + \frac{\partial\Phi(f_\theta,\, f_\theta)}{\partial[y(\lambda)]}\right] * \frac{\partial f}{\partial\theta}$$

Remark: $\theta \to \Phi(f_\theta,\, f_\theta)$ is a nonlinear mapping between $\Theta \subset \mathbf{R}^p$ and $\Pi_k L^2$, so its derivative evaluated at θ_0 is a linear mapping $\mathbf{R}^p \to \Pi_k L^2$. In the above expression, we are taking the derivative with respect to θ and evaluating the derivative at θ.

proof: The proof follows immediately from proposition 7.2.3 and the chain rule. $\square$

As $\frac{\partial}{\partial\theta} D(f_\theta,\, f_n) = \frac{\partial}{\partial[x(\lambda)]} D(f_\theta,\, f_n) \odot \frac{\partial f_\theta}{\partial\theta}$, in order to do the Taylor expansion (i.e. in order

to find $\frac{\partial^2}{\partial\theta'\partial\theta} D(f_\theta, f_n)$), we need to find the matrices $\frac{\partial}{\partial\theta'}\left[\Phi(f_\theta, f_\theta) \odot \frac{\partial f_\theta}{\partial\theta}\right]$ and

$\frac{\partial}{\partial\theta'}\left[\Phi(f_\theta, f_n) \odot \frac{\partial f_\theta}{\partial\theta}\right]$, which will be done in the following two corollaries.

Corollary 7.2.2 (to proposition 7.2.3)

$$\frac{\partial}{\partial\theta'}\left[\Phi(f_\theta, f_\theta) \odot \frac{\partial f_\theta}{\partial\theta}\right]$$

$$= \Phi(f_\theta, f_\theta) \odot \frac{\partial^2 f_\theta}{\partial\theta'\partial\theta} + \frac{\partial f}{\partial\theta} \odot \left(\left[\frac{\partial\Phi(f_\theta, f_\theta)}{\partial x(\lambda)} + \frac{\partial\Phi(f_\theta, f_\theta)}{\partial y(\lambda)} \right] * \frac{\partial f}{\partial\theta'} \right)$$

Remark: $\Phi(f_\theta, f_\theta) \odot \frac{\partial f_\theta}{\partial\theta}$ is a column vector in $\mathbf{R}^p$, so the derivative must be a $p \times p$ matrix.

proof

Corollary 7.2.1 gives the derivative of $\Phi(f_\theta, f_\theta)$, so the product rule proposition 7.1.1 gives the derivative of $\Phi(f_\theta, f_\theta) \odot \frac{\partial f_\theta}{\partial\theta}$. □

Using the same reasoning as above, we can also find the other derivative.

Corollary 7.2.3 (to proposition 7.2.3)

$$\frac{\partial}{\partial\theta'}\left[\Phi(f_\theta, f_n) \odot \frac{\partial f_\theta}{\partial\theta}\right] = \Phi(f_\theta, f_n) \odot \frac{\partial^2 f_\theta}{\partial\theta'\partial\theta} + \frac{\partial f_\theta}{\partial\theta} \odot \left(\left[\frac{\partial\Phi(f_\theta, f_n)}{\partial x} * \frac{\partial f_\theta}{\partial\theta'} \right] \right)$$

Note that this is again a $p \times p$ matrix.

We now have all of the pieces necessary to prove the representation theorem, the proof of which is below.

proof of theorem 7.1.1

From the definition of $\hat{\theta}$,

$$\frac{\partial}{\partial\theta} D_{\psi}(f_{\theta}, f_{n}) \Big|_{\hat{\theta},\hat{\psi}} = 0.$$

But $\frac{\partial}{\partial\theta} D_{\psi}(f_{\theta}, f_{n}) = \frac{\partial}{\partial[x(\lambda)]} D_{\psi}(f_{\theta}, f_{n}) \odot \frac{\partial f_{\theta}}{\partial\theta}$

$$= \left(W_{\psi}(f_{\theta}(\lambda))[\, f_{\theta}(\lambda)\,] - W_{\psi}(f_{\theta}(\lambda))[f_{n}(\lambda)\,]\right) \odot \frac{\partial f_{\theta}}{\partial\theta}. \qquad [7.2.1]$$

We do a Taylor expansion of each piece around θ_0.

$$W_{\psi}(f_{\theta}(\lambda))[\, f_{\theta}(\lambda)] \odot \frac{\partial f_{\theta}}{\partial\theta} \Big|_{\theta_0,\hat{\psi}} +$$

$$\Phi_{\psi}(f_{\theta}, f_{\theta}) \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta} + \frac{\partial f}{\partial\theta} \odot \left(\left[\frac{\partial\Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial x(\lambda)} + \frac{\partial\Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial y(\lambda)} \right] * \frac{\partial f}{\partial\theta'} \right) \Big|_{\theta_1^*,\hat{\psi}} (\hat{\theta} - \theta_0) -$$

$$W_{\psi}(f_{\theta}(\lambda))[\, f_{n}(\lambda)] \odot \frac{\partial f_{\theta}}{\partial\theta} \Big|_{\theta_0,\hat{\psi}} -$$

$$\Phi_{\psi}(f_{\theta}, f_{n}) \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta}}{\partial\theta} \odot \left[\frac{\partial\Phi_{\psi}(f_{\theta}, f_{n})}{\partial x} * \frac{\partial f_{\theta}}{\partial\theta'} \right] \Big|_{\theta_2^*,\hat{\psi}} (\hat{\theta} - \theta_0) = 0 \qquad [7.2.2]$$

This implies

$$W_{\psi}(f_{\theta}(\lambda))[\, f_{\theta}(\lambda) - f_{n}] \odot \frac{\partial f_{\theta}}{\partial\theta} \Big|_{\theta_0,\hat{\psi}} +$$

$$\left\{ \Phi_{\psi}(f_{\theta}, f_{\theta}) \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta}}{\partial\theta} \odot \left(\left[\frac{\partial\Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial x(\lambda)} + \frac{\partial\Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial y(\lambda)} \right] * \frac{\partial f}{\partial\theta'} \right) \Big|_{\theta_0,\hat{\psi}} -$$

$$\Phi_{\psi}(f_{\theta}, f_{\theta}) \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta}}{\partial\theta} \odot \left[\frac{\partial \Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial x} * \frac{\partial f_{\theta}}{\partial\theta'} \right] \Bigg|_{\theta_0, \hat{\psi}} \Bigg\} (\hat{\theta} - \theta_0) +$$

$$\left\{ \Phi_{\psi}(f_{\theta}, f_{\theta}) \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta}}{\partial\theta} \odot \left(\left[\frac{\partial \Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial x(\lambda)} + \frac{\partial \Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial y(\lambda)} \right] * \frac{\partial f_{\theta}}{\partial\theta'} \right) \Bigg|_{\theta_1^*, \hat{\psi}} d\lambda \quad -$$

$$\Phi_{\psi}(f_{\theta}, f_{\theta}) \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta}}{\partial\theta} \odot \left[\frac{\partial \Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial x(\lambda)} + \frac{\partial \Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial y(\lambda)} \right] * \frac{\partial f_{\theta}}{\partial\theta'} \Bigg|_{\theta_0, \hat{\psi}} \Bigg\} (\hat{\theta} - \theta_0) +$$

$$\left\{ \Phi_{\psi}(f_{\theta}, f_{n}) \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta}}{\partial\theta} \odot \left[\frac{\partial \Phi_{\psi}(f_{\theta}, f_{n})}{\partial x} * \frac{\partial f_{\theta}}{\partial\theta'} \right] \Bigg|_{\theta_2^*, \hat{\psi}} \quad - \right.$$

$$\Phi_{\psi}(f_{\theta}, f_{\theta}) \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta}}{\partial\theta} \odot \left[\frac{\partial \Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial x} * \frac{\partial f_{\theta}}{\partial\theta'} \right] \Bigg|_{\theta_0, \hat{\psi}} \Bigg\} (\hat{\theta} - \theta_0) = 0. \qquad [7.2.3]$$

The matrices in the next to last term in parenthesis of [7.2.3] will go in probability to 0 because of proposition 6.5.4 and corollary 6.5.2, and because we know $\hat{\theta}$, $\theta_i^* \rightarrow \theta_0$, $\hat{\psi} \rightarrow \psi_0$. Note that $\frac{\partial W(x)[y]}{\partial y} = W(x)[\ .\ ]$, regardless of where the derivative is evaluated. The first term in parenthesis goes in probability to

$$\frac{\partial f_{\theta}}{\partial\theta} \odot \left[W_{\psi}(f_{\theta}) * \frac{\partial f_{\theta}}{\partial\theta'} \right] \Bigg|_{\theta_0, \psi_0}. \qquad [7.2.4]$$

The last term in parenthesis goes (again by proposition 6.5.4 and corollary 6.5.2) in probability to

$$[\Phi_{\psi}(f_{\theta}, f) - \Phi_{\psi}(f_{\theta}, f_{\theta})] \odot \frac{\partial^2 f_{\theta}}{\partial\theta'\partial\theta}\Big|_{\theta_0,\psi_0} +$$

$$\frac{\partial f_{\theta}}{\partial\theta} \odot \left[\frac{\partial\Phi_{\psi}(f_{\theta}, f_{\theta})}{\partial x} - \frac{\partial\Phi_{\psi}(f_{\theta}, f)}{\partial x}\right] * \frac{\partial f_{\theta}}{\partial\theta'}\Big|_{\theta_0,\psi_0} . \qquad [7.2.5]$$

Notice (1) the sum of these terms is the previously defined $M_W(\theta_0, \psi_0)$. (2) If $f_{\theta_0} = f$, this reduces to [7.2.4].

We have now demonstrated that $\sqrt{n}\ M_n(\hat{\theta}_n - \theta_0)$ and $W_{\hat{\psi}}(f_{\theta})[f_{\theta} - f_n]\Big|_{\theta=\theta_0}$ have the same distribution, where $M_n \xrightarrow{P} M_W(\theta_0, \psi_0)$ and M_n is defined to be the coefficient matrix of $\hat{\theta} - \theta_0$ in [7.2.3]. The proof of the representation theorem is completed by showing the following proposition.

Proposition 7.2.4

$\sqrt{n}\ (\hat{\theta} - \theta_0)$ and $\sqrt{n}\ \ [M_W(\theta, \psi)]^{-1}\ W_{\psi}(f_{\theta}(\lambda))[\ f_n - f_{\theta}(\lambda)]\ \odot \frac{\partial f_{\theta}}{\partial\theta}\Big|_{\theta_0,\psi_0}$ have the same asymptotic distribution.

proof

By definition of $\odot$, $\sqrt{n}\ M_n^{-1} W_{\hat{\psi}}(f_{\theta_0})[f_n - f_{\theta_0}] \odot \frac{\partial f_{\theta_0}}{\partial\theta}$ equals

$$\sqrt{n}\ [M_n]^{-1} \begin{bmatrix} W_{\hat{\psi}}(f_{\theta_0}(\lambda))[\ f_n - f_{\theta_0}(\lambda)] \bullet \frac{\partial f_{\theta_0}}{\partial\theta_1} \\ \vdots \\ W_{\hat{\psi}}(f_{\theta_0}(\lambda))[\ f_n - f_{\theta_0}(\lambda)] \bullet \frac{\partial f_{\theta_0}}{\partial\theta_p} \end{bmatrix} \qquad [7.2.6]$$

Notice that for each i, $W_{\widehat{\psi}}(f_{\theta_0}(\lambda))[\, f_n - f_{\theta_0}(\lambda)] \bullet \frac{\partial f_{\theta_0}}{\partial \theta_i} =$

$$(f_n - f_{\theta_0}(\lambda)) \bullet W^*_{\widehat{\psi}}(f_{\theta_0}(\lambda)) \left[\frac{\partial f_{\theta_0}}{\partial \theta_i} \right]. \qquad [7.2.7]$$

But by condition (c) of the definition of QL distance (definition 6.3.4) we know that

$$W^*_{\widehat{\psi}}(f_{\theta_0}(\lambda)) \left[\frac{\partial f_{\theta_0}}{\partial \theta_i} \right] \xrightarrow{L^2} W^*_{\psi_0}(f_{\theta_0}(\lambda)) \left[\frac{\partial f_{\theta_0}}{\partial \theta_i} \right]. \qquad [7.2.8]$$

Condition (e) in the definition of random L^2 sequence (definition 6.3.2), together with Chebychev's inequality, now gives us that

$$\sqrt{n}\, (f_n - f_{\theta_0}(\lambda)) \bullet \left(\left\{ W^*_{\widehat{\psi}}(f_{\theta_0}(\lambda)) - W^*_{\psi_0}(f_{\theta_0}(\lambda)) \right\} \left[\frac{\partial f_{\theta_0}}{\partial \theta_i} \right] \right) \xrightarrow{P} 0. \qquad [7.2.9]$$

[7.2.9] together with the fact that $[M_n]^{-1} \xrightarrow{P} [M_W(\theta_0, \psi_0)]^{-1}$ will yield the result. This is easily seen as follows.

$$\sqrt{n}\, (\widehat{\theta} - \theta_0) = \sqrt{n}\ [M_n]^{-1} \Bigg\{ W_{\psi_0}(f_{\theta_0}(\lambda))[\, f_n - f_{\theta_0}(\lambda)] \odot \frac{\partial f_{\theta_0}}{\partial \theta} +$$

$$\left(W_{\psi_0}(f_{\theta_0}(\lambda))[\, f_n - f_{\theta_0}(\lambda)] \odot \frac{\partial f_{\theta_0}}{\partial \theta} - W_{\widehat{\psi}}(f_{\theta}(\lambda))[\, f_n - f_{\theta}(\lambda)] \odot \frac{\partial f_{\theta_0}}{\partial \theta} \right) \Bigg\}$$

$$= \sqrt{n}\ [M_W(\theta_0, \psi_0)]^{-1}\, W_{\psi_0}(f_{\theta_0}(\lambda))[\, f_n - f_{\theta_0}(\lambda)] \odot \frac{\partial f_{\theta_0}}{\partial \theta}$$

$$+\ \sqrt{n}\, (M_n^{-1} - [M_W(\theta_0, \psi_0)]^{-1})\, W_{\psi_0}(f_{\theta_0}(\lambda))[\, f_n - f_{\theta_0}(\lambda)] \odot \frac{\partial f_{\theta_0}}{\partial \theta}$$

$$+ \sqrt{n}\ \mathbf{M}_n^{-1}\left[\mathbf{W}_{\psi_0}(\mathbf{f}_{\theta_0}(\lambda))[\,\mathbf{f}_n - \mathbf{f}_{\theta_0}(\lambda)] \odot \frac{\partial \mathbf{f}_{\theta_0}}{\partial \theta} - \mathbf{W}_{\hat{\psi}}(\mathbf{f}_{\theta}(\lambda))[\,\mathbf{f}_n - \mathbf{f}_{\theta}(\lambda)] \odot \frac{\partial \mathbf{f}_{\theta_0}}{\partial \theta} \right]$$

The second term goes in probability to 0, since $[\mathbf{M}_n]^{-1} \xrightarrow{P} [\mathbf{M}_W(\theta_0, \psi_0)]^{-1}$ and $\mathbf{W}_{\psi_0}(\mathbf{f}_{\theta_0}(\lambda))$ $[\mathbf{f}_n - \mathbf{f}_{\theta_0}(\lambda)] \odot \frac{\partial \mathbf{f}_{\theta_0}}{\partial \theta}$ goes in distribution to a (multivariate) normal random variable. The third term goes in probability to 0 by [7.2.9] (and since $[\mathbf{M}_n]^{-1} \xrightarrow{P} [\mathbf{M}_W(\theta_0, \psi_0)]^{-1}$). We thus see that the "ϵ_n" of theorem 7.1.1 is $\frac{1}{\sqrt{n}}$ times the sum of the second and third terms.

It should be noted that corollary 7.1.1 follows trivially from the above proof (just take [7.2.1] as the starting point for the proof). Corollary 7.1.2 also immediately follows from theorem 7.1.1. To see this, create a new "Ψ" space $\Psi_1 = \Psi \times \Theta$. Then the new $\hat{\psi}$ is $\hat{\psi}_1 = (\hat{\psi}, \hat{\theta}_1)$. This asymptotically gives the same variance as if (ψ_0, θ_1) had been used. The corollary shows consistency and asymptotic optimality of IRWLS on *one iteration*, since we take $\hat{\theta}_1$ to be a consistent but not optimal variance estimate obtained by solving $[\mathbf{f}_\theta - \mathbf{y}_n] \bullet \mathbf{W}[\mathbf{f}_\theta - \mathbf{y}_n]$ for an arbitrary *fixed* operator W. This is the same approach taken to show asymptotic optimality of IRWLS in the literature, assuming, of course, that the model is correct (see, e.g. Carrol and Rupert's (1988) theorem 2.1 or Chiu's (1988) theorem 7).

7.3 Proof of Optimality Theorem

First note that by proposition 7.1.3 and the link condition, the vector

$$\sqrt{n}\ [\mathbf{M}_W^0]^{-1} \begin{bmatrix} (\mathbf{f}_n - \mathbf{f}_{\theta_0}(\lambda)) \bullet \mathbf{W}^*_{\psi_0}(\mathbf{f}_{\theta_0}(\lambda)) \left[\frac{\partial \mathbf{f}_{\theta_0}}{\partial \theta_1}\right] \\ \cdot \\ \cdot \\ \cdot \\ (\mathbf{f}_n - \mathbf{f}_{\theta_0}(\lambda)) \bullet \mathbf{W}^*_{\psi_0}(\mathbf{f}_{\theta_0}(\lambda)) \left[\frac{\partial \mathbf{f}_{\theta_0}}{\partial \theta_p}\right] \end{bmatrix} \qquad [7.3.1]$$

has variance matrix

$$[M_W^0]^{-1}\left[\frac{\partial f_{\theta_0}}{\partial\theta}\odot\left(W\ V\ W^*\ *\ \frac{\partial f_{\theta_0}}{\partial\theta'}\right)\right]\left[[M_W^0]^{-1}\right]^* \qquad [7.3.2]$$

which equals M_V^{-1} (see definition 7.1.3) if $W^*_{\psi_0}(f_{\theta_0})\left[\frac{\partial f_\theta}{\partial\theta_i}\right] = V^{-1}\left[\frac{\partial f_\theta}{\partial\theta_i}\right]$ (assuming the model contains the limiting function, or some condition such as [7.3.10]). The problem is now to show optimality.

Unless specified otherwise, the partial derivative of $f_\theta(\lambda)$ is always to be regarded as being evaluated at θ_0. We will work in the Hilbert space $\Pi_k L^2[\Lambda]$ with inner product $\langle f, g\rangle = \int_\Lambda \sum f_i(\lambda)\ g_i(\lambda)\ d\lambda$. For any QL operator W(.)[.] and random L^2 sequence (y_n, y_0, M, SL^2, V) define the linear mappings L_W and L_V: $\Pi_k L^2 \rightarrow \mathbf{R}^p$ (where θ is p dimensional) as follows. If $h \in \Pi_k L^2$,

$$L_V[h] = [M_V]^{-1}\left(\left[V^{-1} * \frac{\partial f_{\theta_0}(\lambda)}{\partial\theta}\right]\odot\ h(\lambda)\right)$$

$$L_W[h] = \left[M_W^0\right]^{-1}\left(\left[W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}(\lambda)}{\partial\theta}\right]\odot\ h(\lambda)\right). \qquad [7.3.3]$$

Let D denote the subspace of $\Pi_k L^2$ spanned by the functions $\frac{\partial f_{\theta_0}(\lambda)}{\partial\theta_i}$ i = 1..p, and let $U = S^{-1}[D]$ denote the image of this subspace under the linear transformation $S^{-1}[f]$, where S is the square root of the random L^2 sequence variance operator V. Let $P_R[f(\lambda)]$ denote the projection of $f(\lambda)$ into the subspace R. Let γ be any vector in $\mathbf{R}^p$. We will show that the asymptotic variance of $\gamma'(\hat\theta_n^V - \theta_0)$ is always less than or equal to the asymptotic variance of $\gamma'(\hat\theta_n^W - \theta_0)$, where $(\hat\theta_n^W - \theta_0) = L_W[f_n - f_{\theta_0}] + \epsilon_n^1$ and $(\hat\theta_n^V - \theta_0) = L_V[f_n - f_{\theta_0}] + \epsilon_n^2$, $\sqrt{n}\ \epsilon_n^i \xrightarrow{P} 0$, i=1,2.

Define the linear mappings K_W, K_V: $\Pi_k L^2 \rightarrow \mathbf{R}^p$ by

$$K_W(h) = \left(W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}(\lambda)}{\partial \theta} \right) \odot h(\lambda)$$

$$K_V(h) = \left(V^{-1} * \frac{\partial f_{\theta_0}(\lambda)}{\partial \theta} \right) \odot h(\lambda). \qquad [7.3.4]$$

Then $L_W[h] = [M_W^0]^{-1} \circ K_W[h]$ (or $L_V[h] = [M_V]^{-1} \circ K_V[h]$) if we regard the $p \times p$ matrix M_W^0 (or M_V) as a linear mapping $\mathbf{R}^p \rightarrow \mathbf{R}^p$. From functional analysis (see, e.g. Conway (1985) p. 31) we know that if A: X→Y is a bounded linear mapping, where X and Y are Hilbert spaces, there exists a unique mapping (called the Adjoint) A^*: Y→X satisfying $A[h] \bullet k = h \bullet A^*[k]$. What is K_W^*? I claim it is defined by

$$K_W^*[\gamma] = \gamma' \begin{pmatrix} W^*(f_{\theta_0}) \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1} \right] \\ \vdots \\ W^*(f_{\theta_0}) \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p} \right] \end{pmatrix} \qquad [7.3.5]$$

$$= \gamma_1 W^*(f_{\theta_0}) \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1} \right] + \gamma_2 W^*(f_{\theta_0}) \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_2} \right] + \cdots$$

$$+ \gamma_p W^*(f_{\theta_0}) \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p} \right]. \qquad [7.3.6]$$

To show [7.3.5] is true, all that is necessary is to show that it satisfies the defining relationship

for adjoints. Write

$$K_W[h(\lambda)] \bullet k = \begin{pmatrix} W^*(f_{\theta_0})\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1}\right] \bullet h(\lambda) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ W^*(f_{\theta_0})\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p}\right] \bullet h(\lambda) \end{pmatrix} \bullet \begin{pmatrix} k_1 \\ k_2 \\ \cdot \\ \cdot \\ \cdot \\ k_p \end{pmatrix} \qquad [7.3.7]$$

$$= k_1 W^*(f_{\theta_0})\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1}\right] \bullet h(\lambda) + \ldots + k_p W^*(f_{\theta_0})\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p}\right] \bullet h(\lambda). \qquad [7.3.8]$$

On the other hand, using [7.3.5] as the definition for K_W^*, $h(\lambda) \bullet K_W^*[k] =$

$$h(\lambda) \bullet \left[k_1 W^*(f_{\theta_0})\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1}\right] + \ldots + k_p W^*(f_{\theta_0})\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p}\right]\right]$$

$$= h(\lambda) \bullet k_1 W^*(f_{\theta_0})\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1}\right] + \ldots + h(\lambda) \bullet k_p W^*(f_{\theta_0})\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p}\right] \qquad [7.3.9]$$

which is the same as [7.3.8].

It is a property of adjoints (Conway (1985), p. 32) that if A: X→Y, B: Y→Z, then $(BA)^* = A^* B^*$. So we have $L_V^* = K_V^* (M_V^{-1})^*$ and $L_W^* = K_W^* ([M_W^0]^{-1})^*$. Also note the adjoint of a matrix is simply its transpose (conjugate transpose if complex valued, but everything's real here). The proof now consists of carrying out the following steps:

step 1

The asymptotic variance of $\gamma'(\hat{\theta}_n^W - \theta_0)$ is the same as the asymptotic variance of $L_W^*[\gamma] \bullet (f_n - f)$ (by the defining property of adjoints, $\gamma \bullet L_W[f_n - f_0] = L_W^*[\gamma] \bullet (f_n - f)$).

step 2

Write $L_W^*[\gamma] \bullet ((f_n(\lambda) - f(\lambda)) = L_W^*[\gamma] \bullet S\, S^{-1}[f_n(\lambda) - f(\lambda)]$

$$= S^*\, L_W^*[\gamma] \bullet S^{-1}[f_n(\lambda) - f(\lambda)]$$

$$= S\, L_W^*[\gamma] \bullet S^{-1}[f_n(\lambda) - f(\lambda)] .$$

step 3

Write $S\ L_W^*[\gamma]$ as the sum of two orthogonal components $P_U[S\ L_W^*[\gamma]] + P_{U^\perp}[S\ L_W^*[\gamma]]$, where $U \equiv S^{-1}[D]$ is defined earlier. Hence $S\ [L_W^*[\gamma]] \bullet S^{-1}[f_n - f] = (\psi_1 + \psi_2) \bullet S^{-1}[f_n - f]$,

(where $\psi_1 = P_U[S\ L_W^*[\gamma]]$ and $\psi_2 = P_{U^\perp}[S\ L_W^*[\gamma]]$)

$= \psi_1 \bullet S^{-1}[f_n - f] + \psi_2 \bullet S^{-1}[f_n - f]$

$= S^{-1}[\psi_1] \bullet (f_n - f) + S^{-1}[\psi_2] \bullet (f_n - f)$ (by self adjointness of S^{-1}).

step 4

Observe these two components are asymptotically uncorrelated.

$$\lim_{n\to\infty} n \operatorname{cov}\left(S^{-1}[\psi_1] \bullet (f_n - f)\ ,\ S^{-1}[\psi_2] \bullet (f_n - f)\right) =$$

$$= S^{-1}[\psi_1] \bullet V\left[S^{-1}[\psi_2]\right]$$

$$= \psi_1 \bullet \left[S^{-1} V\, S^{-1}[\psi_2]\right]$$

(But $S^{-1}V\,S^{-1}$=the identity operator)

$=\psi_1 \bullet \psi_2 = 0$ (because ψ_1 and ψ_2 are orthogonal).

step 5

Show that $P_U[S\,L_W^*[\gamma]] = S\,L_V^*[\gamma]$. To do this we need to do two things.

1) Show $S\,L_V^*[\gamma] \in U$.

2) Show $S\,L_W^*[\gamma] - S\,L_V^*[\gamma] \perp U$.

The result follows since $S\,L_W^*[\gamma]$ may be written in only one way as a sum of something in U plus something in $U^{\perp}$.

proof of (1):

$$S\,L_V^*[\gamma] = S\,K_V^*\,[M_V^{-1}]^*[\gamma] = S\,K_V^*[k] \quad (\text{where } k = [M_V^{-1}]^*[\gamma])$$

$$= S\left[k_1 V^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1}\right] + \ldots + k_p V^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p}\right]\right]$$

$$= k_1\, S^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1}\right] + \ldots + k_p\, S^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p}\right]$$

which is a linear combination of basis elements for $S^{-1}[D]$.

proof of (2):

To do this, show that the function $S\,L_W^*[\gamma] - S\,L_V^*[\gamma]$ is orthogonal to each element in a basis

for $S^{-1}[D]$. (A basis is obviously $\left\{ S^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_1}\right], \ldots S^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_p}\right]\right\}$).

$$[S\,L_W^*[\gamma] - S\,L_V^*[\gamma]] \bullet S^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i}\right] =$$

$$S^*\,L_W^*[\gamma] \bullet S^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i}\right] - S^*\,L_V^*[\gamma] \bullet S^{-1}\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i}\right] =$$

$$L^*_W[\gamma] \bullet \frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i} - L^*_V[\gamma] \bullet \frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i} = \text{(by the defining relation for adjoints)}$$

$$\gamma' L_W \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i} \right] - \gamma' L_V \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i} \right] =$$

$$\gamma' [M^0_W]^{-1} \left(\left[W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \frac{\partial f_{\theta_0}}{\partial \theta_i} \right) - \gamma' M_V^{-1} \left(\left[V^{-1} * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \frac{\partial f_{\theta_0}}{\partial \theta_i} \right) =$$

$$\gamma' \left([M^0_W]^{-1} \left(\left[W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \frac{\partial f_{\theta_0}}{\partial \theta_i} \right) - M_V^{-1} \left(\left[V^{-1} * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \frac{\partial f_{\theta_0}}{\partial \theta_i} \right) \right).$$

Now the expression $\left[W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \frac{\partial f_{\theta_0}}{\partial \theta_i}$ represents a row of the matrix

$\frac{\partial f_{\theta_0}}{\partial \theta} \odot W(f_{\theta_0}) \left[\frac{\partial f_{\theta_0}}{\partial \theta'} \right]$. Notice this matrix is M^0_W! Hence

$M_W^{-1} \left(\left[W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \frac{\partial f_{\theta_0}}{\partial \theta_i} \right)$ = vector with 1 in ith position, 0 elsewhere, and

$M_V^{-1} \left(\left[V^{-1} * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \frac{\partial f_{\theta_0}}{\partial \theta_i} \right)$ = vector with 1 in ith position, 0 elsewhere.

So for each i, the vector in parenthesis is the 0 vector and we have proven step 5.

Observe that step 5 implies that $P_{U^\perp}[S\ L^*_W[\gamma]] = 0$ for $W=V^{-1}$ (because in step 3, $S\ L^*_W[\gamma]$ is decomposed into orthogonal functions, $P_U[S\ L^*_W[\gamma]]$ and $P_{U^\perp}[S\ L^*_W[\gamma]]$. If $P_U[S\ L^*_W[\gamma]] = S\ L^*_W[\gamma]$, then $P_{U^\perp}[S\ L^*_W[\gamma]]$ must be 0). Step 4 shows that for any operator W, if $P_{U^\perp}[S\ L^*_W[\gamma]] \neq 0$, the variance is "inflated" by $S^{-1}P_{U^\perp}[S\ L^*_W[\gamma]]$ over that we would have gotten had we used $W=V^{-1}$. To see that $W^*=V^{-1}$ on span $\left\{ \frac{\partial f_{\theta_0}}{\partial \theta_i} \right\}_{i=1..p}$ implies the same optimal variance as if V^{-1} were used, look at how M_W and Q_W are constructed (i.e.

expressions [7.1.1] and [7.1.2]), and observe the condition implies $M_W = M_{V^{-1}}$ and $Q_W = Q_{V^{-1}}$. Also observe that W is invariant under multiplication by a constant by looking at the variance matrix [7.1.5] and notice that the constant will cancel out, as two inverses are taken on M_W and the constant would appear twice in Q_W. □

Notice that the model containing the limiting function means that

$$\frac{\partial f_\theta}{\partial \theta} \odot \left[\frac{\partial \Phi_\psi(f_\theta, f_\theta)}{\partial x} - \frac{\partial \Phi_\psi(f_\theta, f)}{\partial x} \right] * \left. \frac{\partial f_\theta}{\partial \theta'} \right|_{\theta_0, \psi_0} = 0$$

$$[\Phi_\psi(f_\theta, f) - \Phi_\psi(f_\theta, f_\theta)] \odot \left. \frac{\partial^2 f_\theta}{\partial \theta' \partial \theta} \right|_{\theta_0, \psi_0} = 0 . \qquad [7.3.10]$$

As we will see in chapter 9, there are ways of obtaining [7.3.10] even if the model *does not* contain the limiting function. It turns out that the conditions [7.3.10] are not as difficult to verify in the case of a misspecified model as one might think, under certain circumstances.

7.4 Conclusions

In this chapter, we have established the main representation and optimality theorems (theorems 7.1.1 and 7.1.2) regarding parametric estimates obtained by minimizing a QL distance or solving the QL equations [7.1.4]. A solid groundwork has been laid for the ideas in chapter 6, but the question of the usefulness of chapter 6's definitions still remains. We will only give a small hint of why the theory is meaningful at the end of chapter 8, showing the optimality of an IRWLS procedure (as described in the introduction to chapter 6) for the case of a non Gaussian process. The real uses of the theory will be presented in chapters 9 and 10, and hinge on the comments at the end of section 7.3. For optimality to hold, the operator

$W^*(f_{\theta_0})[\ .\ ]$ must mimic V^{-1} on span $\left\{\frac{\partial f_{\theta_0}}{\partial \theta_i}\right\}_{i=1..p}$, where V^{-1} is the true inverse variance operator. *This is extremely important, and means that it is not necessary to use the true inverse variance operator in parametric estimation in order to obtain optimal estimates.* Why it might be desirable to use "wrong" QL functions will be discussed in chapters 9 and 10. But before doing this, we must show that spectral estimation is an application of the theory presented thus far (and specifically, to verify the unproven statements made in section 6.4). This will be the main focus of chapter 8.

Chapter VIII
Time Series Applications of QL Theory

8.1 Introduction

In this chapter, we give some theorems which support the claims of the examples in chapter 6 and set the background needed for other main results of the dissertation to follow. Chapter 9 will apply chapter 7's optimality theorem to the problem of spectral estimation when the observed series is contaminated, but it is necessary to first clearly define what is meant by "contamination". It must then be shown that whatever definition is settled upon falls within the framework of the definitions in chapter 6.

There are two approaches to frequency domain spectral estimation: regarding the periodogram as a step function by defining it at a finite number of frequencies (such as the Fourier frequencies), and then extending it, or using the natural definition of the periodogram so that it is automatically defined for all frequencies in $[-\pi, \pi]$. So far as our theory is concerned, these approaches are asymptotically identical. Of course, in practice the first is usually preferred due to numerical considerations. The central problem is establishing asymptotic unbiasedness of the periodogram, i.e. showing that $\sqrt{n}\ E(\int \psi I_n d\lambda - \int \psi g\ d\lambda) \to 0$ where g is the spectrum and ψ is some function. This condition will essentially hold if *either* ψ or g is sufficiently smooth (i.e. is continuously differentiable). Actually, we really only need to verify the asymptotic unbiasedness in the link condition, i.e. $\lim_{n\to\infty} \sqrt{n}\ E\{\int W_{\psi}^{*}(f_{\theta}) \left[\frac{\partial f_{\theta}}{\partial \theta_i}\right] (I_n - f_{\theta})\ d\lambda\} = 0$. As this is not a major point of the dissertation (actually it's a somewhat bothersome side issue which if allowed could easily *obscure* the major points of the dissertation), no

attempt has been made to find the most general sufficient conditions. As *some* assumptions must be made on the process, we will for the sake of simplicity do the following. Assume

$$\sum_{j_1,j_2,j_3} | Q_{\alpha_1 ,\alpha_2 , \ldots \alpha_4} (j_1, j_2, j_3)| < \infty$$

where $Q_{\alpha_1 ,\alpha_2 , \ldots \alpha_4} (j_1, j_2, j_3)$ is the fourth cumulant of the channels α_1, α_2, α_3, α_4. This guarantees a continuous fourth cumulant spectrum, and allows the use of Hosoya and Taniguchi's (1982) lemma A2.2 if the natural definition of the periodogram is used, and Brillinger's (1981) theorem 7.6.1 if the step function definition is used. For conditions on the spectrum and model, we will assume that the model satisfies the BCC, and one of the following holds. (1) The spectrum satisfies the BCC, or (2) The spectrum is of bounded variation. We will mostly be interested in assumption 2, as chapters 9 and 10 will discuss model fitting for a "contaminated" series, where it is not unreasonable to assume that a smooth model is being fit to a discontinuous spectrum. Chapters 9 and 10, which rely on the material in the present chapter, will consider only the simplest case of Gaussian processes (the non Gaussian case is for future research!). This simplifies the supporting theorems here for those chapters, as higher order cumulants are 0.

It should also be mentioned that *under assumption 2* we will differ slightly from the literature in our definition of the "step function" periodogram extension. For our purposes, the steps will be at the frequencies $\pi t/4n$ for t between $-4n+1$ and 4n (so there are 4n steps in [0, π] rather than n/2 steps at the Fourier frequencies). The reason for this departure from usual assumptions is to easily show that the step function extension satisfies the definition of random L^2 sequence. Most of the literature assumes the more stringent assumption (1), so that the usual "Fourier frequencies" step function extension is a random L^2 sequence, which is shown

essentially by putting together published results.

Distributional results are another side issue which will not be considered here, i.e conditions guaranteeing the asymptotic normality of $\sqrt{n}(\int \psi I_n d\lambda - \int \psi g \, d\lambda)$, as we are mainly concerned with *the asymptotic variance of parametric estimates.* Sufficient conditions for normality are given by Brillinger (1981) or Hosoya and Taniguchi (1982) respectively in the step function and natural definition of the periodogram cases.

There are some important facts about Fourier series of functions of bounded variation which will be used in the sequel, and which should be mentioned now. Two of the most important of these are the following (from Zygmund (1968)).

1) Theorem 8.6 (p. 58): If f is of bounded variation, then S[f] (the sequence of partial sums of the Fourier series) converges uniformly at every point of continuity of f (definition: A sequence of functions $s_n(x)$ defined in the neighborhood of $x=x_0$ and converging for $x=x_0$ is said to converge uniformly at x_0 to a limit s, if to every $\epsilon>0$ there is a δ and a p so that $|s_n(x)-s|<\epsilon$ for $|x-x_0|<\delta$ and $n>p$).

2) Theorem 3.7 (p. 90): If f is of bounded variation, the partial sums of S[f] are uniformly bounded.

8.2 Expectation Results for the Periodogram

The main result of this section is theorem 8.2.1, which establishes the asymptotic unbiasedness of the periodogram using either the natural or step function extension. To arrive at this theorem, it is necessary to first discuss some background material. In the following, "$\bullet$" refers to the inner product on $L^2[-\pi, \pi]$ defined by $f \bullet g = 1/2\pi \int_{-\pi}^{\pi} f(\lambda)\, g(\lambda)\, d\lambda$. A "Cesaro sum" of a L^2 function f is a sum of the form $\frac{1}{n+1}(s_0+s_1+ \ldots +s_n)$, where

$s_n(x)=\sum_{k=-n}^{n} (f \bullet e^{ixk})\, e^{i\lambda k}$. "Fejer's kernel" is defined for each positive integer n, as

$$K_n(x) = \frac{1}{n}\left[\frac{1-\cos nx}{1-\cos x}\right].$$

$K_n(x)$ is also the nth Cesaro sum of the series $\sum_{k=-\infty}^{\infty} \exp(ikx)$. See Hoffman (1988), p. 16 for further information.

Proposition 8.2.1

Suppose f_n is the nth Cesaro sum of the function f, and similarly for g_n and g. If f satisfies the BCC and $g \in L^2$, then

(a) $\lim_{n\to\infty} \sqrt{n}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} (f_n - f)\, g\, d\lambda\right] = 0.$

(b) $\lim_{n\to\infty} \sqrt{n}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} (g_n - g)\, f\, d\lambda\right] = 0.$

Note that (a) and (b) will be essential ingredients in showing asymptotic unbiasedness of the periodogram in the cases of assumption (1) and assumption (2) (of the introduction) respectively.

proof of (a): By Theorem 1 of Chiu (1988), $f(\lambda)=\sum_{n=-\infty}^{\infty} a_n\, e^{in\lambda}$, where $\sum_{n=-\infty}^{\infty} n|a_n| < \infty$. We may uniformly bound the error $|f(\lambda) - f_n(\lambda)|$ where $f_n(\lambda) = \int_{-\pi}^{\pi} f(t)\, K_{n-1}(x-t)\, dt$, and $K_n(x)$ is Fejer's kernel. To do this, recall that $f_n(\lambda) = \frac{1}{n+1}(s_0+s_1+ \ldots +s_n)$, where $s_n(x)=\sum_{k=-n}^{n} (f \bullet e^{ixk})\, e^{i\lambda k}$. Let $a_k = f \bullet e^{ixk}$. Then

$$f(x) - f_n(x) = \frac{1}{n+1} \sum_{r=0}^{n} (f(x) - s_r(x))$$

$$= \frac{1}{n+1} \sum_{r=0}^{n} \left[\left(\sum_{k=-\infty}^{\infty} a_k e^{i\lambda k} \right) - \sum_{t=-r}^{r} a_t e^{i\lambda t} \right]$$

$$= \frac{1}{n+1} \left[\sum_{k=-n}^{n} |k| a_k e^{i\lambda k} + (n+1) \sum_{k \notin [-n,n]} a_k e^{i\lambda k} \right].$$

Hence $\sqrt{n}$ times this expression goes to 0.

If $\psi \in L^2$, $\int_{-\pi}^{\pi} \psi \sqrt{n} (f_n - f) d\lambda \to 0$ (by Cauchy Schwartz, for example).

proof of (b): Let $a_k = f \bullet e^{ixk}$ and $b_k = g \bullet e^{ixk}$ (where f satisfies the BCC).

$$\sqrt{n} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(g - g_n) = \sqrt{n} \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{k=-\infty}^{\infty} a_k e^{i\lambda k} \left[g(\lambda) - \frac{1}{n+1} \sum_{r=0}^{n} \sum_{t=-r}^{r} b_t e^{i\lambda t} \right] d\lambda$$

$$= \sqrt{n} \sum_{k=-n}^{n} a_k \left[b_{-k} - \left(\frac{n - |k|}{n+1} \right) b_{-k} \right] + \sqrt{n} \sum_{k \notin [-n,n]} a_k \int_{-\pi}^{\pi} (g - g_n) e^{i\lambda k} d\lambda$$

$$= \sqrt{n} \sum_{k=-n}^{n} a_k \left(\frac{1+|k|}{n+1} \right) b_{-k} + \sqrt{n} \sum_{k \notin [-n,n]} a_k \int_{-\pi}^{\pi} (g - g_n) e^{i\lambda k} d\lambda.$$

As before, using the fact b_k and $\int_{-\pi}^{\pi} (g - g_n) e^{i\lambda k} d\lambda$ are bounded, we see that both pieces go to 0. □

The next proposition is needed to establish results concerning the "step function" extension of the periodogram. It concerns the relationship between "stepfunctionized" exponential functions (defined in definition 8.2.1) and their continuous counterparts, and shows the finer the steps, the more orthogonal stepfunctionized exponentials are created.

Definition 8.2.1

For each nonnegative integer n and positive integer k between -n+1 and n we define the sequence of functions $\{\phi_k^n(\lambda)\}$ as

$$\phi_k^n(\lambda) \equiv \sum_{t=-n+1}^{n} e^{ik\lambda_t} \chi_{\Lambda_t}(\lambda)$$

where $\lambda_t=\frac{\pi t}{n}$ for t= − n+1 . . n, $\chi_{\Lambda_t}(\lambda)$ is the indicator function for the interval $\Lambda_t=[\lambda_{t-1}, \lambda_t]$, and $-\pi \leq \lambda \leq \pi$.

Proposition 8.2.2

Let $\{\phi_k^n(\lambda)\}$ be as defined in definition 8.2.1.

(i) For each n, $\{\phi_k^n(\lambda)\}_{k=-n+1}^{n}$ are a collection of 2n orthonormal functions in L^2.

(ii) $e^{i\lambda k} \bullet \phi_j^n(\lambda) = 0$ if $j \neq k$, $-n+1 \leq j, k \leq n$.

(iii) $e^{i\lambda k} \bullet \phi_k^n(\lambda) = \left(\frac{2n-1}{2n}\right) \frac{1-e^{-\pi ik/n}}{-\pi ik/n}$, $-n+1 \leq k \leq n$.

(iv) Let r be a positive integer. $\prod_{i=1}^{r} \phi_{j_i}^{rn}(\lambda) = \phi_{\Sigma j_i}^{rn}(\lambda)$ if $-n+1 \leq j_i \leq n$ for i=1, . . . ,r.

proof

Notice we always sum over $2n-1$ consecutive frequencies (as each step function $\phi_k^n(\lambda)$ has exactly $2n-1$ steps). A formula for $x^l+x^{l+1}+ \ldots +x^h$ is $\frac{x^l(1-x^{h-l+1})}{1-x}$.

proof of (i): $\phi_j^n(\lambda) \bullet \phi_k^n(\lambda)=(1/2\pi)\int_{-\pi}^{\pi}\left(\sum_t e^{ij\lambda_t} \chi_{\Lambda_t}(\lambda)\right)\left(\sum_s e^{-ik\lambda_s} \chi_{\Lambda_s}(\lambda)\right) d\lambda$

$=(1/2\pi)(2\pi/2n)\sum_t e^{i(j-k)\lambda_t} =(1/2n)\sum_t e^{i(j-k)\frac{\pi t}{n}} = (1/2n)\sum_t \exp\left[i\frac{\pi(j-k)}{n}\right] t$. For $j \neq k$ and j, k satisfying $|j-k|<2n$, the formula gives the result (note $n-(-n+1)+1=2n$, $\exp\left[i\frac{\pi(j-k)}{n}\right] 2n=1$).

proof of (ii) and (iii): Define $U(\Lambda_t)=\lambda_t$, $L(\Lambda_t)=\lambda_{t-1}$. Then

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda k} \sum_{t=-n+1}^{n} e^{-i\lambda_t j}\, \chi_{\Lambda_t}(\lambda)\, d\lambda =$$

$$\sum_t \left(\frac{1}{2\pi}\int_{\Lambda_t} e^{i\lambda k}\, d\lambda \right) e^{-i\lambda_t j} = \qquad \text{(because } U(\Lambda_t)=\lambda_t\text{)}$$

$$\sum_t \frac{1}{2\pi} \frac{e^{iU(\Lambda_t)k} - e^{iL(\Lambda_t)k}}{ik} e^{-iU(\Lambda_t)j} =$$

$$\frac{1}{2\pi ik}\left[\sum_t e^{iU(\Lambda_t)(k-j)} - \sum_t e^{i(L(\Lambda_t)k - U(\Lambda_t)j)} \right].$$

If $k \neq j$, the first term sums to 0 since we're summing over the $2n-1$ frequencies. To see that the second term also sums to 0, note that $L(\Lambda_t)k - U(\Lambda_t)j = \frac{\pi(t-1)}{n} k - \frac{\pi tj}{n} = \frac{\pi t(k-j)}{n} - \frac{\pi k}{n}$. So the second sum is $e^{-\pi ik/n}\sum_t e^{\pi it(k-j)/n} = 0$. If $k=j$ then $\sum_t e^{iU(\Lambda_t)(k-j)} = 2n-1$, and the second sum is $(2n-1)e^{-\pi ik/n}$. So the result is

$$(2n-1)\frac{1-e^{-\pi ik/n}}{2\pi ik} = \left(\frac{2n-1}{2n}\right)\frac{1-e^{-\pi ik/n}}{\pi ik/n}.$$

(iv) Note that $\sum_i j_i \leq rn$ as needed to make $\phi^{rn}_{\Sigma j_i}(\lambda)$ well defined, and do the multiplication pointwise. □

Corollary 8.2.1 (to proposition 8.2.1)

Define g_n^* to be the Cesaro sum of g formed from the basis functions $\phi_k^{r(n)}(\lambda)$, where r(n) is an integer $\geq$ n for each n, instead of $e^{ik\lambda}$ (i.e. g_n^* is the average of $s_n = \sum_{k=-n}^{n} \frac{1}{2\pi}\int_{-\pi}^{\pi} g\, e^{-i\lambda k}\, d\lambda\, \phi_k^{r(n)}(\lambda)$), and f_n^* similarly. Then proposition 8.2.1 continues to hold.

proof

For simplicity, we will give the proof assuming r(n)=n (there is no loss of generality in doing

this, as the proof using r(n) is identical).

proof of (a): Note that by the mean value theorem and the continuity of the derivative, f satisfies a Lipschitz condition of order 1. To prove (a) it suffices to show that $\lim_{n\to\infty}\sqrt{n}\left[\int (f_n^s - f)\ g\ d\lambda\right]=0$, where $f_n^s(\lambda) = \sum_i f(\lambda_i)\ \chi_{\Lambda_i}(\lambda)$ is a "stepfunctionized" version of f. This is because you can write $\sqrt{n}\ \int (f_n^* - f)\ g\ d\lambda = \sqrt{n}\ \int (f_n^* - f_n^s)\ g\ d\lambda + \sqrt{n}\ \int (f_n^s - f)\ g\ d\lambda$, and $\sqrt{n}\ \int (f_n^* - f_n^s)\ g\ d\lambda$ goes to 0 by the same argument as is in proposition 8.2.1 (a), e.g. by the uniform boundedness of $|f(x) - f_n(x)|$. Using the Lipschitz condition on f and the fact $\lambda_n - \lambda_{n-1}=\pi/n$, $f_n^s - f$ is bounded by K/n for some constant K. It follows that the second piece also goes to 0.

proof of (b): It suffices to show that $\lim_{n\to\infty}\sqrt{n}\sum_{k=-n}^{n} a_k\left[b_{-k} - \left(\frac{n-k}{n+1}\right) b_{-k}^*\right] = 0$, where $b_k^*=(1/2\pi)\int_{-\pi}^{\pi} f(\lambda)\ \phi_{-k}(\lambda)\ d\lambda$ and $\phi_k(\lambda)$ is as defined earlier (because the first few lines of the proof of proposition 8.2.1 (b) continue to hold due to the above proposition 8.2.2). This is established by the following steps:

claim 1

$$\lim_{n\to\infty}\sqrt{n}\sum_{k=-n}^{n}|a_k\ b_{-k}\ (1 - e^{i\lambda k}\bullet\phi_{-k}^n)|= 0.$$

proof

Recalling $e^{i\lambda k}\bullet\phi_{-k}^n(\lambda) = \frac{1-e^{\pi ik/n}}{\pi ik/n}$, write the Taylor expansion of e^x as $1+x+\sum_{n=2}^{\infty} x^n/n!$. So $\frac{e^x-1-x}{x} = x\ R(x)$, where $R(x)=\sum_{n=2}^{\infty} x^{n-2}/n!$ has an infinite radius of convergence. By the continuity of R(x), $\exists M_1$ so $|R(x)|<M_1$ for $|x|<1$. This implies that $\left|\frac{e^x-1-x}{x}\right| <M_1|x|$ for $|x|<1$. Hence $\left|1 - \frac{1-e^{\pi ik/n}}{\pi ik/n}\right| < M_1\ \pi k/n$. Write

$$|1 - e^{i\lambda k}\bullet\phi_{-k}^n| \le \left|1 - \frac{1-e^{\pi ik/n}}{\pi ik/n}\right| + \left|\frac{2n-1}{2n}\right| \times \left|\frac{1-e^{\pi ik/n}}{\pi ik/n}\right| \le$$

$M_1 \frac{\pi k}{n} + \frac{M_2}{n} \leq M \frac{\pi k}{n}$ for some M. Now

$$\sqrt{n} \sum_{k=-n}^{n} | a_k b_{-k} (1 - e^{i\lambda k} \bullet \phi_{-k}^{n})| < \sqrt{n} \sum_{k=-n}^{n} |a_k b_{-k}| M \pi k/n \to 0 \text{ as } n \to \infty.$$

claim 2

$$\sqrt{n} \sum_{k=-n}^{n} \left| a_k b_{-k} \left(\frac{n - |k|}{n+1} \right) \left[1 - e^{i\lambda k} \bullet \phi_{-k}^{n} \right] \right| = 0$$

proof

Since $\frac{n - |k|}{n+1} < 1$, this is true if

$\lim_{n \to \infty} \sqrt{n} \sum_{k=-n}^{n} |a_k b_{-k} (1 - e^{i\lambda k} \bullet \phi_{-k}^{n})| = 0$, which follows from claim 1. □

Theorem 8.2.1

(a) If I_n is a component of a periodogram matrix from a process which has spectrum satisfying the BCC, then $E \sqrt{n}(\int \psi I_n - \int \psi g) \to 0$ for $\psi \in L^2$. (b) If the process has a spectrum with components in $L^2[-\pi, \pi]$, then the same holds for ψ satisfying the BCC.

(a) and (b) hold regardless of whether the definition of periodogram matrix is the natural or "r(n)" step function extension, for $r(n) \geq 2n$ (see corollary 8.2.1 for definition).

proof

For the natural definition of the periodogram, this follows from proposition 8.2.1, keeping in mind that $E \int_{-\pi}^{\pi} \psi \, I_n \, d\lambda = \int_{-\pi}^{\pi} \psi \, f_n \, d\lambda$, where f_n is the Cesaro sum of the spectrum. For the

step function extension, it follows from corollary 8.2.1, keeping in mind $E(\int \psi I_n \, d\lambda) = \int \psi \sum (1-|k|/n)\gamma_{\alpha_1\alpha_2}(k) \, \phi_k^{r(n)}(\lambda) \, d\lambda$.

In the following section, we shall often mention the "4n step function extension of the periodogram". This refers to the use of functions $\{\phi_k^{4n}(\lambda)\}$ as defined in definition 8.2.1 in the step function definition of the periodogram, i.e. $I_n(\lambda) = \sum (n-|k|)\gamma(k) \, \phi_{-k}^{4n}(\lambda)$, rather than the "natural" extension $I_n(\lambda) = \sum (n-|k|)\gamma(k) \exp(-ik\lambda)$, and is so called because these functions have 4n steps in $[0, \pi]$. The reason 4n steps are needed is so that the functions $\phi_k^{4n}(\lambda)$ will "act" like their exponential counterparts when multiplied together. For example, consider the proof of the statement "$E(\int \psi I_n \, d\lambda) = \int \psi \sum (1-|k|/n)\gamma_{\alpha_1\alpha_2}(k) \exp(-i\lambda k) \, d\lambda$". We write

$$E\left[\int_{-\pi}^{\pi} \psi I_n \, d\lambda\right] = \frac{1}{2\pi} E\left[\int_{-\pi}^{\pi} \psi \sum_{t=1}^{n} X_i(t) \exp(i \, t \, \lambda) \sum_{s=1}^{n} X_j(s) \exp(-i \, s \, \lambda) \, d\lambda\right] =$$

$$\frac{1}{2\pi} E\left[\int_{-\pi}^{\pi} \psi \sum_t \sum_s X_i(t)X_j(s) \exp(i \, \lambda \, (t-s)) \, d\lambda\right] = E\left(\sum_t \sum_s X_i(t)X_j(s) \, \tilde{\psi} \, (t-s)) \right) =$$

$$\sum_{k=-n+1}^{n} (n-|k|) \, \gamma(k) \, \tilde{\psi} \, (k) = \sum_{k=-n+1}^{n} (n-|k|) \, \gamma(k) \int_{-\pi}^{\pi} \psi(\lambda) \exp(i\lambda k) \, d\lambda =$$

$$\int_{-\pi}^{\pi} \psi(\lambda) \sum_{k=-n+1}^{n} \int_{-\pi}^{\pi} (n-|k|) \, \gamma(k) \, \exp(i\lambda k) \, d\lambda.$$

For the analogous statement "$E(\int \psi I_n \, d\lambda) = \int \psi \sum (1-|k|/n)\gamma_{\alpha_1\alpha_2}(k) \, \phi_k^{r(n)}(\lambda) \, d\lambda$" made in the proof of theorem 8.2.1 to hold, replace "$\exp(i\lambda t)$ and "$\exp(-i\lambda s)$" with "$\phi_t^{r(n)}(\lambda)$" and "$\phi_{-s}^{r(n)}(\lambda)$", for some r(n). We need $\phi_t^{r(n)}(\lambda)\phi_{-s}^{r(n)}(\lambda) = \phi_{t-s}^{r(n)}(\lambda)$ in order for the second line of the proof to hold. For this to occur, it must be true that $r(n) \geq 2n$ (see proposition 8.2.2 (iv)). The same idea is used in the proofs of the propositions and theorems in section 8.3, but

now we will need four stepfunctionized exponentials to properly multiply together, requiring the 4n steps.

8.3 Variance Results for the Periodogram (Part I)

We now give some technical results needed to establish the variance conditions in the definition of random L^2 sequence. The parts of the theorems in this section to be proved are mainly for the case of the spectrum being of bounded variation rather than satisfying the BCC, and are similar to Taniguchi's (1982) lemma A2.2. This section will give results for the natural and "4n" step function extension of the periodogram, because the proofs for either case are virtually identical. Section 4 establishes the results when the spectrum satisfies the BCC and the Fourier frequencies periodogram extension is used. The methods of proof there are most similar to the proofs in Brillinger (1981), e.g. theorem 7.6.1.

Theorem 8.3.1

Suppose $\sum_{j_1,j_2,j_3=-\infty}^{\infty} |Q^z_{\alpha_1\alpha_2\ \alpha_3\alpha_4}(j_1, j_2, j_3)| < \infty.$

(a) If $f_{\alpha_1\alpha_3}$, $f_{\alpha_2\alpha_4}$ $f_{\alpha_1\alpha_4}$, $f_{\alpha_3\alpha_4}$ are of bounded variation and W_1, W_2 are in $L^2[-\pi, \pi]$, then

$$\lim_{n\to\infty} N \operatorname{cov}\left\{ \int_{-\pi}^{\pi} W_1(\lambda)\, I_{\alpha_1\alpha_2}(\lambda)\, d\lambda,\ \int_{-\pi}^{\pi} W_2(\lambda)\, I_{\alpha_3\alpha_4}(\lambda)\, d\lambda \right\} =$$

$$2\pi \int_{-\pi}^{\pi} W_1(\lambda)\, \overline{W_2(\lambda)}\, f_{\alpha_1\alpha_3}(\lambda)\, \overline{f_{\alpha_2\alpha_4}(\lambda)}\, d\lambda + 2\pi \int_{-\pi}^{\pi} W_1(\lambda)\, \overline{W_2(-\lambda)}\ f_{\alpha_1\alpha_4}(\lambda)\, \overline{f_{\alpha_2\alpha_3}(\lambda)}\, d\lambda$$

$$+2\pi \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} W_1(\lambda_1) W_2(-\lambda_2) Q^z_{\alpha_1\alpha_2\ \alpha_3\alpha_4}(\lambda_1, \lambda_2, -\lambda_2)\, d\lambda_1\, d\lambda_2 \qquad [8.3.1]$$

for the continuous extension of the periodogram.

(b) Same as (a), except the 4n step function extension of the periodogram is used.

(c) Same as (a), except the components of the spectrum satisfy the BCC and the Fourier frequencies step function extension of the periodogram is used.

Theorem 8.3.2

There exists a constant M so that if and $W_1(\lambda)$ and $W_2(\lambda)$ are L^2 functions, then the absolute value of the covariance in theorem 8.3.1 is bounded by $M\ ||W_1||_2\ ||W_2||_2$ for all n, assuming

(a) $f_{\alpha_1\alpha_1}$, $f_{\alpha_2\alpha_2}$, $f_{\alpha_3\alpha_3}$, $f_{\alpha_4\alpha_4}$ are functions of bounded variation, and the continuous extension of the periodogram is used.

(b) Same as (a), except the 4n step function extension of the periodogram is used.

(c) Same as (a), except the components of the spectrum satisfy the BCC and the Fourier frequencies step function extension of the periodogram is used.

Theorem 8.3.3

(i) If $(X_1\ ,\ X_2)$ is a bivariate series and W is a function in L^2, then

$$E\left[\int_{-\pi}^{\pi} W(\lambda)\ I_{ij}^{n}(\lambda)\ d\lambda\right] \rightarrow \int_{-\pi}^{\pi} W(\lambda)\ f_{ij}(\lambda)\ d\lambda.$$

(ii) There exists a constant K so that $\left| E\left[\int_{-\pi}^{\pi} W(\lambda)\ I_{ij}^{n}(\lambda)\ d\lambda\right] \right| \leq K\ ||W||_2$.

(i) and (ii) hold assuming conditions (a), (b) or (c) of theorem 8.3.1.

The main theorems of this section have now been stated. It may be possible to strengthen these theorems, which is a topic for future research. What follows are proofs of the

main theorems and propositions needed for these proofs.

The proof of theorems 8.3.1 and 8.3.2 will use some of the ideas in Hosoya and Taniguchi's (1982) proof. We begin by showing N cov($\int W_1 I_n \, d\lambda$, $\int W_1 I_n \, d\lambda$) can be expressed in other forms.

Proposition 8.3.1

Suppose that W_1 and W_2 are real valued functions in $L^2[-\pi, \pi]$ and $\gamma_{\alpha_i\alpha_j}(k)$ are the covariance functions for a real valued multivariate time series (i.e. $\gamma_{\alpha_i\alpha_j}(k) = \text{cov}(X_{\alpha_i}(t), X_{\alpha_j}(t+k))$). Then

$$N \text{ cov}\left\{ \int_{-\pi}^{\pi} W_1(\lambda) I_{\alpha_1\alpha_2}(\lambda) \, d\lambda, \int_{-\pi}^{\pi} W_2(\lambda) I_{\alpha_3\alpha_4}(\lambda) \, d\lambda \right\} =$$

$$\frac{1}{N} \sum_{n_1 n_2 n_3 n_4 = 1}^{N} \widetilde{W}_1(n_1 - n_2) \widetilde{W}_2(n_4 - n_3) \{\gamma_{\alpha_1\alpha_3}(n_3 - n_1) \gamma_{\alpha_2\alpha_4}(n_4 - n_2) +$$

$$\gamma_{\alpha_1\alpha_4}(n_4 - n_1) \gamma_{\alpha_2\alpha_3}(n_3 - n_2)\} \qquad [8.3.2]$$

where $\widetilde{W}_i(k) \equiv 1/2\pi \int_{-\pi}^{\pi} W_i(\lambda) e^{i\lambda k} \, d\lambda$.

proof

$$\text{cov}\left\{ \int_{-\pi}^{\pi} W_1(\lambda) I_{\alpha_1\alpha_2}(\lambda) \, d\lambda, \int_{-\pi}^{\pi} W_2(\lambda) I_{\alpha_3\alpha_4}(\lambda) \, d\lambda \right\} =$$

$$\frac{1}{N^2} \text{ cov}\left(\frac{1}{2\pi} \int \sum_{n_1 = 1}^{n} X_{\alpha_1}(n_1) e^{i\lambda n_1} \sum_{n_2 = 1}^{n} X_{\alpha_2}(n_2) e^{-i\lambda n_2} W_1(\lambda) \, d\lambda, \right.$$

$$\frac{1}{2\pi}\int \sum_{n_3=1}^{n} X_{\alpha_3}(n_3)\, e^{i\lambda n_3} \sum_{n_4=1}^{n} X_{\alpha_4}(n_4)\, e^{-i\lambda n_4}\, W_2(\lambda)\, d\lambda \Bigg)$$

$$= \frac{1}{N^2} \sum_{n_1 n_2 n_3 n_4 = 1}^{n} \operatorname{cov}\left(X_{\alpha_1}(n_1)X_{\alpha_2}(n_2),\ X_{\alpha_3}(n_3)\, X_{\alpha_4}(n_4)\right) \widetilde{W}_1(n_1 - n_2)\, \overline{\widetilde{W}_2(n_3 - n_4)}$$

$$= \frac{1}{N^2} \sum_{n_1 n_2 n_3 n_4} \widetilde{W}_1(n_1 - n_2)\, \overline{\widetilde{W}_2(n_3 - n_4)}$$

$$\{\gamma_{\alpha_1\alpha_3}(n_3 - n_1)\gamma_{\alpha_2\alpha_4}(n_4 - n_3) + \gamma_{\alpha_1\alpha_4}(n_4 - n_1)\gamma_{\alpha_2\alpha_3}(n_3 - n_1)\}.$$

As W_2 is real valued, $\overline{\widetilde{W}_2(n)} = \widetilde{W}_2(-n)$. Hence N times this last expression equals [8.3.2]. □

The next proposition shows that the covariance in [8.3.2] may be written as a Cesaro sum of the sequence [8.3.6] (below). To prove theorem 8.3.1, the sequence [8.3.6] will be shown to converge to [8.3.1]. It will follow that the Cesaro sums (and hence N times the covariance) converge to the same thing.

Proposition 8.3.2

Suppose that W_1 and W_2 are real valued functions in $L^2[-\pi, \pi]$ and $\gamma_{\alpha_i\alpha_j}(k)$ are the covariance functions for a real valued multivariate time series. Then

(a) The partial sum

$$\frac{1}{N} \sum_{l_2 l_3 l_4 = -N+1}^{N-1} \{N - S(l_2, l_3, l_4)\}\widetilde{W}_1(-l_2)\, \widetilde{W}_2(l_4 - l_3)\, \{\gamma_{\alpha_1\alpha_3}(l_3)\, \gamma_{\alpha_2\alpha_4}(l_4 - l_2) +$$

$$\gamma_{\alpha_1\alpha_4}(l_4)\, \gamma_{\alpha_2\alpha_3}(l_3 - l_2)\} \qquad [8.3.3]$$

equals [8.3.2].

$$\text{Here, } S(l_2, l_3, l_4)= \begin{cases} \max(|l_2|, |l_3|, |l_4|) \text{ if sign } l_2=\text{sign } l_3=\text{sign } l_4 \\ \max(|l_i|, |l_j|)+ |l_k| \text{ if sign } l_i=\text{sign } l_j= -\text{sign } l_k \end{cases} \qquad [8.3.4]$$

$$\text{and } \widetilde{W}_i(k) \equiv 1/2\pi \int_{-\pi}^{\pi} W_i(\lambda) e^{i\lambda k}\, d\lambda. \qquad [8.3.5]$$

(b) [8.3.3] is the same as the nth Cesaro sum of

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} W_1^n(\lambda) f_1^n(\lambda) W_2^n(\lambda)\overline{f_2^n(\lambda)}d\lambda + \frac{1}{2\pi}\int_{-\pi}^{\pi} W_1^n(\lambda) f_3^n(\lambda) W_2^n(-\lambda)\overline{f_4^n(\lambda)}d\lambda \qquad [8.3.6]$$

where

$$W_1^n(\lambda)=\sum_{k=-n}^{n} \widetilde{W}_1(k)\, e^{-ik\lambda} \qquad W_2^n(\lambda)=\sum_{k=-n}^{n} \widetilde{W}_2(k)\, e^{-ik\lambda} \qquad [8.3.7]$$

$$f_1^n(\lambda)=\sum_{k=-n}^{n} \gamma_{\alpha_1\alpha_3}(k)\, e^{-ik\lambda} \qquad f_2^n(\lambda)=\sum_{k=-n}^{n} \gamma_{\alpha_2\alpha_4}(k)\, e^{-ik\lambda}$$

$$f_3^n(\lambda)=\sum_{k=-n}^{n} \gamma_{\alpha_1\alpha_4}(k)\, e^{-ik\lambda} \qquad f_4^n(\lambda)=\sum_{k=-n}^{n} \gamma_{\alpha_2\alpha_3}(k)\, e^{-ik\lambda}\,. \qquad [8.3.8]$$

<u>proof</u>

(a) Following Hosoya and Taniguchi (1982), define $l_2=n_2-n_1$, $l_3=n_3-n_1$, $l_4=n_4-n_1$. Then $\widetilde{W}_1(n_1-n_2)\ \widetilde{W}_2(n_4-n_3)\ \{\gamma_{\alpha_1\alpha_3}(n_3-n_1)\ \gamma_{\alpha_2\alpha_4}(n_4-n_2) + \gamma_{\alpha_1\alpha_4}(n_4-n_1)\ \gamma_{\alpha_2\alpha_3}(n_3-n_2)\}= \widetilde{W}_1(-l_2)\ \widetilde{W}_2(l_4-l_3)\ \{\gamma_{\alpha_1\alpha_3}(l_3)\ \gamma_{\alpha_2\alpha_4}(l_4-l_2) + \gamma_{\alpha_1\alpha_4}(l_4)\ \gamma_{\alpha_2\alpha_3}(l_3-l_2)\}$. For l_2, l_3, l_4 fixed, how many such terms can there be? Once n_1 is chosen, all of the other n_i are completely determined. How many different ways can n_1 be chosen? If the l_i all have the same sign, then n_1 could range between 1 and $n-\max\{|l_2|, |l_3|, |l_4|\}$. If one of the l_i, say l_2, has a

different sign (say negative) from the others, then ignoring l_2's influence, n_1 could be between 1 and $n - \max\{|l_1|, |l_3|\}$. l_2 will reduce the number of ways n_1 might be chosen: n_1 must be larger than n_2 so n_1 could be between $|l_2|$ and n. Putting this together means there are $n - \{\max\{|l_1|, |l_3|\}+|l_2|\}$ ways to choose n_1.

(b) Note that the nth Cesaro sum of the sequence $\{a_n\}$ is by definition $(s_0+s_1+ \ldots +s_{n-1})/n$, where $s_i=a_0+a_1+ \ldots +a_i$ (i.e. the "average" partial sum).

As motivation for the proof, notice that if we multiply together and integrate

$$\frac{1}{2\pi}\left(\sum_{j=-n}^{n}\widetilde{W}_1(j)\,e^{-ij\lambda}\right)\times\left(\sum_{k=-n}^{n}\widetilde{W}_2(k)e^{-ik\lambda}\right)\times$$

$$\left(\sum_{l=-n}^{n}\gamma_{\alpha_1\alpha_3}(l)e^{-il\lambda}\right)\times\left(\sum_{m=-n}^{n}\gamma_{\alpha_2\alpha_4}(m)\,e^{im\lambda}\right)+$$

$$\frac{1}{2\pi}\left(\sum_{j=-n}^{n}\widetilde{W}_1(j)\,e^{-ij\lambda}\right)\times\left(\sum_{k=-n}^{n}\widetilde{W}_2(k)e^{ik\lambda}\right)\times$$

$$\left(\sum_{l=-n}^{n}\gamma_{\alpha_1\alpha_4}(l)e^{-il\lambda}\right)\times\left(\sum_{m=-n}^{n}\gamma_{\alpha_2\alpha_3}(m)\,e^{im\lambda}\right) \qquad [8.3.9]$$

$(= \frac{1}{2\pi}\int_{-\pi}^{\pi} W_1^n(\lambda)f_1^n(\lambda)W_2^n(\lambda)\overline{f_2^n(\lambda)}+W_1^n(\lambda)f_3^n(\lambda)W_2^n(-\lambda)\overline{f_4^n(\lambda)}\,d\lambda)$ then we will have terms which resemble those in [8.3.2] (excluding the integer coefficient) with the summation between $-n$ and n. This is because the only nonzero terms in each piece of the integral are those for which $-j-k-l+m=0$ in the first sum, and $-j+k-l+m=0$ in the second sum, causing the only nonzero terms in each piece to be of the form $\widetilde{W}_1(j)\widetilde{W}_2(k)\gamma_{\alpha_1\alpha_3}(l)\gamma_{\alpha_2\alpha_4}(m)$ (or $\widetilde{W}_1(j)\widetilde{W}_2(k)\gamma_{\alpha_1\alpha_4}(l)\gamma_{\alpha_2\alpha_3}(m)$ in second sum) for such j, k, l and m between $-n$ and n. But the nonzero terms in [8.3.6] after taking the integral (and each term appears only once) satisfy the same condition. The idea is to count the number of terms of the form

$$\{\widetilde{W}_1(-l_2)\ \widetilde{W}_2(l_4-l_3)\}\ \{\gamma_{\alpha_1\alpha_3}(l_3)\ \gamma_{\alpha_2\alpha_4}(l_4-l_2)\ +$$

$$\gamma_{\alpha_1\alpha_4}(l_4)\ \gamma_{\alpha_2\alpha_3}(l_3-l_2)\} \qquad [8.3.10]$$

in the Cesaro sum of $\frac{1}{2\pi}\int_{-\pi}^{\pi} W_1^n(\lambda)f_1^n(\lambda)W_2^n(\lambda)\overline{f_2^n(\lambda)}+W_1^n(\lambda)f_3^n(\lambda)W_2^n(-\lambda)\overline{f_4^n(\lambda)}\ d\lambda$, and to show there are $N-S(l_2, l_3, l_4)$ such terms. Any term of the form $\widetilde{W}_1(j)\widetilde{W}_2(k)\gamma_{\alpha_1\alpha_3}(l)\gamma_{\alpha_2\alpha_4}(m)$ with $-j-k-l+m=0$ will appear $N-\max\{|j|, |k|, |l|, |m|\}$ times in the Cesaro sum of $\frac{1}{2\pi}\int_{-\pi}^{\pi} W_1^n(\lambda)f_1^n(\lambda)W_2^n(\lambda)\overline{f_2^n(\lambda)}\ d\lambda$, and similarly for terms of the form $\widetilde{W}_1(j)\widetilde{W}_2(k)\gamma_{\alpha_1\alpha_4}(l)\gamma_{\alpha_2\alpha_3}(m)$ with $-j+k-l+m=0$. A term of the form [8.3.9] must be constructed from terms of the two forms just described. So there will be $N-\max\{|l_2|, |l_4-l_3|, |l_3|, |l_4-l_2|, |l_4|, |l_3-l_2|\}$ terms of the form [8.3.9]. If l_2, l_3, and l_4 all have the same sign, this max will equal $\max\{|l_2|, |l_3|, |l_4|\}$, because then $|l_4-l_3|\le\max\{|l_3|, |l_4|\}$ and $|l_4-l_2|,\le\max\{|l_4|, |l_2|\}$. If one l_i, say l_2, has a different sign from the others, then $|l_4-l_2|=|l_4|+|l_2|$ and $|l_3-l_2|=|l_3|+|l_2|$. Thus the maximum will be $\max\{|l_3|, |l_4|\}+|l_2|$. □

<u>proof of theorem 8.3.1</u> (a)

Note that $W_i^n(\lambda)f_i^n(\lambda)\xrightarrow{L^2}W_i(\lambda)f_i(\lambda)$, because

$$\int_{-\pi}^{\pi}|W_i^n(\lambda)f_i^n(\lambda)-W_i(\lambda)f_i(\lambda)|^2\ d\lambda\ \le\ \int_{-\pi}^{\pi}|W_i^n(\lambda)f_i^n(\lambda)-W_i^n(\lambda)f_i(\lambda)|^2\ d\lambda$$

$$+\int_{-\pi}^{\pi}|W_i^n(\lambda)f_i(\lambda)-W_i(\lambda)f_i(\lambda)|^2\ d\lambda\ \le\ \int_{-\pi}^{\pi}|W_i^n(\lambda)|^2\ |f_i^n(\lambda)-f_i(\lambda)|^2\ d\lambda\ +$$

$$\int_{-\pi}^{\pi}|W_i^n(\lambda)-W_i(\lambda)|^2\ |f_i(\lambda)|^2\ d\lambda$$

By the boundedness of f, $\int_{-\pi}^{\pi}|W_i^n(\lambda)-W_i(\lambda)|^2\ |f_i(\lambda)|^2\ d\lambda\to 0$. To show $\int_{-\pi}^{\pi}|W_i^n(\lambda)|^2\ |f_i^n(\lambda)-f_i(\lambda)|^2\ d\lambda$ also goes to 0, we need the following from real analysis (see, e.g. proposition

13, p. 85 in chapter 4 of Royden (1968)).

Proposition 8.3.3

(a) If f is a nonnegative integrable function, then for every ϵ there is a δ so that if Δ has measure less than δ, $\int_\Delta f\, d\lambda < \epsilon$.

(b) If $f_n \xrightarrow{L^2} f$, such a δ may be chosen which works for all $|f_n|^2$ simultaneously.

proof: For any ϵ one can choose a step function g so $\int |f-g|\, d\lambda < \epsilon/2$. Obviously such a δ can be chosen for g using an "ϵ" of $\epsilon/2$. Then $\int_\Delta |f|\, d\lambda \le \int_\Delta |g|\, d\lambda + \int_\Delta |f-g|\, d\lambda \le \epsilon)$. If $f_n \xrightarrow{L^2} f$, we claim such a δ may be chosen which works for all $|f_n|^2$ simultaneously. To see this is true, choose N so $n \ge N \Rightarrow ||f_n - f|| < \epsilon/2$. Then choose δ which works for $|f|^2$ and f_i, i=1..N, but with an "ϵ" of $(\epsilon/2)^2$. If $n \ge N$ and Δ has measure less than δ,

$$\sqrt{\int_\Delta |f_n|^2\, d\lambda} \le \sqrt{\int_\Delta |f_n - f|^2\, d\lambda} + \sqrt{\int_\Delta |f|\, d\lambda}$$

$$\le \epsilon/2 + \sqrt{(\epsilon/2)^2} = \epsilon. \qquad \square$$

Now consider the integral $\int |W_i^n(\lambda)|^2\ |f_i^n(\lambda) - f_i\ (\lambda)|^2\ d\lambda$. Let $\epsilon > 0$, and suppose $|f_i^n(\lambda) - f_i\ (\lambda)|^2$ is uniformly bounded by M for all n, and $\int |W_i^n(\lambda)|^2 d\lambda$ is bounded by M_2 for all n. If $S_n = \{\lambda |\ |f_i^k(\lambda) - f_i\ (\lambda)| < \epsilon$ for all $k > n\}$, then $\cap S_n^c$ consists of a countable number of points of discontinuity of f_i and hence has measure 0. This means for any δ, there exists an N so $n \ge N \Rightarrow \bigcap_{k=1}^{n} S_k^c$ has measure less than δ (proof: Define $A_1 = S_1$, $A_2 = S_2/S_1$, $A_3 = S_3/S_1 \cup S_2$, etc. Then $\cup A_i = \cup S_i$ and the A_i are disjoint. Hence $\sum m(A_i) = 2\pi$. Choose N so $\sum_{i>N} m(A_i) < \delta$. This means that everything outside of $S_1 \cup S_2 \cup \ldots S_n$ has measure less than δ, if $n \ge N$. But $(S_1 \cup S_2 \cup \ldots S_n)^c = \bigcap_{k=1}^{n} S_k^c$). Returning to our original problem, Choose δ so $m(\Delta) < \delta \Rightarrow \int_\Delta |W_i^n(\lambda)|^2\ d\lambda < \epsilon/2M$ for all n. Choose N so $n \ge N \Rightarrow |f_i^n(\lambda) - f_i\ (\lambda)|^2 < \epsilon/(2M_2)$

except on a set Δ_n where $m(\Delta_n)<\delta$. Then

$$\int |W_i^n(\lambda)|^2 \, |f_i^n(\lambda) - f_i(\lambda)|^2 \, d\lambda = \int_{\Delta_n} |W_i^n(\lambda)|^2 \, |f_i^n(\lambda) - f_i(\lambda)|^2 \, d\lambda +$$

$$\int_{\Delta_n^c} |W_i^n(\lambda)|^2 \, |f_i^n(\lambda) - f_i(\lambda)|^2 \, d\lambda$$

$$\leq M \, \epsilon/2M + \epsilon/(2M_2) \, M_2 = \epsilon.$$

To complete the proof of 8.3.1, by the continuity of the inner product we have

$$W_1^n(\lambda)f_1^n(\lambda) \bullet W_2^n(\lambda)f_2^n(\lambda) + W_1^n(\lambda)f_3^n(\lambda) \bullet W_2^n(-\lambda)f_4^n(\lambda) \rightarrow$$

$$W_1(\lambda)f_1(\lambda) \bullet W_2(\lambda)f_2(\lambda) + W_1(\lambda)f_3(\lambda) \bullet W_2(-\lambda)f_4(\lambda)$$

as $n \rightarrow \infty$. By Zygmund (1968), theorem 1.2, p. 74, the Cesaro sums must go to the same limit. Propositions 8.3.1 and 8.3.2 give the covariance in [8.3.2] equals a Cesaro sum in [8.3.6]. So the covariance in [8.3.2] has the correct limit (i.e. [8.3.1]). Hosoya and Taniguchi's (1982) proof covers the convergence of the part of the summation involving the fourth cumulant spectrum. □

proof of theorem 8.3.2 (a)

First observe we may bound $W_1^n(\lambda)f_1^n(\lambda) \bullet W_2^n(\lambda)f_2^n(\lambda)$ by $K||W_1(\lambda)||_2 \, ||W_2(\lambda)||_2$ for some constant K. This is because $||W_i^n(\lambda)f_i^n(\lambda)||_2$ is bounded, since $\int |W_i^n(\lambda)f_i^n(\lambda)|^2 \, d\lambda \leq \int M \, |W_i^n(\lambda)|^2 \, d\lambda$ (and because according to Zygmund (1968), theorem 3.7 p90, $|f_i^n(\lambda)|$ is uniformly bounded by some constant M_i). Now use the Cauchy Schwartz inequality to obtain $|W_1^n(\lambda)f_1^n(\lambda) \bullet W_2^n(\lambda)f_2^n(\lambda)| \leq ||W_1^n(\lambda)f_1^n(\lambda)||_2 ||W_2^n(\lambda)f_2^n(\lambda)||_2 \leq M_1M_2||W_1^n(\lambda)||_2||W_2^n(\lambda)||_2$ (of course, $W_i^n(\lambda)$ is the projection of $W_i(\lambda)$ onto a subspace, and hence $||W_1^n(\lambda)||_2 \leq ||W_1(\lambda)||_2$). □

proof of theorem 8.3.3 (assuming conditions (a))

$$\text{(i) } E\left[\int_{-\pi}^{\pi} W(\lambda)\, I_{ij}^{n}(\lambda)\, d\lambda\right] = \frac{1}{n} E\left[\int_{-\pi}^{\pi} W(\lambda) \sum X_1(n_1)\, e^{i\lambda n_1} \sum X_2(n_2)\, e^{-i\lambda n_2}\right] =$$

$$\frac{1}{n} E\left[\int_{-\pi}^{\pi} W(\lambda) \sum_{n_1 n_2} X_1(n_1)\, X_2(n_2) e^{i\lambda(n_1 - n_2)}\right] =$$

$$\frac{1}{n} \sum_{n_1 n_2} \gamma_{12}(n_2 - n_1)\, \widetilde{W}(n_1 - n_2).$$

Defining $W^n(\lambda) = \sum_{k=-n}^{n} \widetilde{W}(k)\, e^{-ik\lambda}$ and $f^n(\lambda) = \sum_{k=-n}^{n} \frac{n - |k|}{n}\, \gamma_{12}(k)\, e^{-ik\lambda}$, we see that this is equal to $f^n \bullet W^n$, which goes to $f \bullet W$ by the continuity of the inner product and properties of the Fourier sums of functions of bounded variation.

(ii) Choose a constant K so all Fourier sums of f are bounded by K. Then by the Cauchy Schwartz inequality, $|f^n \bullet W^n| \leq ||f^n||\, ||W^n|| \leq K\, ||W^n||$. □

The proofs of theorems 8.3.1, 8.3.2, and 8.3.3 (i and ii) under conditions "b" (i.e the "4n" step function extension case), are similar to the above proofs, just replace "$e^{ik\lambda}$" with "$\phi_k^{4n}(\lambda)$" in [8.3.5], [8.3.7] and [8.3.8] before forming the nth Cesaro sum. The discussion at the end of section 8.2 applies here, but this time four "stepfunctionized exponentials" must correctly multiply together in [8.3.9]. The proof of 8.3.2(b) will then be exactly the same as above. The proof of 8.3.1 (b) follows from the fact that $W_i^n(\lambda) \xrightarrow{L^2} W_i(\lambda)$, where $W_i^n(\lambda)$ has been formed from $\phi_k^{4n}(\lambda)$ (along with the "coefficients" $\widetilde{W}_i(k)$). This is because of proposition 8.4.2 (a) below, and the fact that the $\phi_k^{4n}(\lambda)$ form a basis for the 4n step function space.

8.4 Variance Results for the Periodogram (part II)

For the proof of theorems 8.3.1 and 8.3.2 (c) involving the step function extension at the Fourier frequencies, we will modify the proof of Brillinger's (1981) theorem 7.6.1. Specifically, define a sequence M_n of subspaces in L^2 as $M_n \equiv \text{span}\left\{\frac{N}{2\pi}\,\chi_{\Delta_j}(\lambda)\right\}$, where $\Delta_j = \left[\frac{2\pi(j-1)}{N}, \frac{2\pi j}{N}\right]$, j=1..N and χ is the indicator function. Let $\psi_i^n \equiv \frac{n}{2\pi}\,\chi_{\Delta_i}(\lambda)$ be the ith basis element of M_n. Notice that the collection $\{\psi_i^n\}$ are an orthonormal basis for M_n. Finishing the proof of theorem 8.3.1 (and 8.3.2) will consist in showing that if the term "$A_j\left(\frac{2\pi r}{n}\right)$" in Brillinger's proof is interpreted as "$A_j \bullet \psi_r^n$", then the expression

$$\left(\frac{2\pi}{n}\right)^2 \sum_{r=1}^{n} A_j\left(\frac{2\pi r}{n}\right)\overline{A_k\left(\frac{2\pi r}{n}\right)}\, f_{a_1 a_2}\left(\frac{2\pi r}{n}\right) f_{b_1 b_2}\left(-\frac{2\pi r}{n}\right)$$

$$+ A_j\left(\frac{2\pi r}{n}\right)\overline{A_k\left(2\pi - \frac{2\pi r}{n}\right) f_{a_1 b_2}\left(\frac{2\pi r}{n}\right)}\, f_{b_1 a_2}\left(-\frac{2\pi r}{n}\right)$$

$$+ \left(\frac{2\pi}{n}\right)^3 \sum_{r=1}^{n}\sum_{r=1}^{n} A_j\left(\frac{2\pi r}{n}\right)\overline{A_k\left(\frac{2\pi s}{n}\right)}\, f_{a_1 b_1 a_2 b_2}\left(\frac{2\pi r}{n}, -\frac{2\pi r}{n}, -\frac{2\pi s}{n}\right) \qquad [8.4.1]$$

(appearing in the proof of Brillinger's (1981) theorem 7.6.1 with "T − 1" in place of "n"), converges to [5.4.2] where the integrals are taken between 0 and 2π (see discussion in section 5.4 for the equivalence of [5.4.2] and [8.3.1]). For this section, define "$\bullet$" to be the unnormalized inner product on $L^2[-\pi, \pi]$, i.e. $f \bullet g = \int_{-\pi}^{\pi} f(\lambda)\, g(\lambda)\, d\lambda$.

If we use the notation P_n to denote the projection of L^2 onto M_n, then we can easily establish the following sequence of propositions (Note that the propositions are true for any sequence M_n of "uniform step function spaces" with step length approaching zero as n approaches ∞).

Proposition 8.4.1

If f is a continuous function on $[0, 2\pi]$, then $||P_n[f] - f||_\infty \to 0$.

proof

By the mean value theorem for integrals, for each interval Δ_i $\exists$ $c_i \in \Delta_i$ so that $\frac{1}{\text{length}(\Delta_i)} \int_{\Lambda_i} f(\lambda)\, d\lambda = f(c_i)$. Given $\epsilon > 0$, by uniform continuity of f (on the compact set $[0, 2\pi]$), $\exists$ δ so that $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$. Choose N so that $\frac{2\pi}{N} < \delta$. Then for $x \in \Delta_i$, $|f(x) - f(c_i)| < \epsilon$. Notice that $P_n[f] = \sum (f \bullet \psi_i^n)\psi_i^n$ (by the orthonormality of $\{\psi_i^n\}$), which equals $\sum \frac{1}{\text{length}(\Delta_i)} \int_{\Lambda_i} f(\lambda)\, d\lambda\ \chi_{\Lambda_i}$. So if $n \geq N$, on each interval Δ_i we have $|f(x) - (f \bullet \psi_i^n)\psi_i^n| < \epsilon$ and we are done. □

Proposition 8.4.2

(a) If $f \in L^2[0, 2\pi]$, then $||P_n[f] - f||_2 \to 0$

(b) If $f_n \in L^2[0, 2\pi]$ and $||f_n||_2 \to 0$, then $||P_n[f_n]||_2 \to 0$

proof

(a) Let $\epsilon > 0$. Choose g continuous so that $||g - f|| < \epsilon/3$. Choose N by proposition 8.4.1 so that $||P_n[g] - g||_2 < \epsilon/3$ for $n \geq N$ (this is obviously implied by 8.4.1). A projection operator has norm 1, so $||P_n[f] - P_n[g]||_2 \leq ||f - g||_2 \leq \epsilon/3$. We conclude $||P_n[f] - f||_2 \leq \epsilon$.

(b) $||P_n[f_n]||_2 \leq ||f_n||_2 \to 0$ □

Proposition 8.4.3

If $||A_n^i - A^i||_2 \to 0$ and $||f_n^i - f^i||_\infty \to 0$ for i=1, 2, A_n^i, $A^i \in L^2[0, 2\pi]$, f_n^i, $f^i \in L^\infty [0, 2\pi]$ (a.e. bounded functions on $[0, 2\pi]$), then $A_n^1 f_n^1 \bullet A_n^2 f_n^2 \to A^1 f^1 \bullet A^2 f^2$.

proof

Notice that $||A_n^i f_n^i - A^i f^i||_2 \to 0$, because it's less than $||A_n^i f_n^i - A_n^i f^i||_2 + ||A_n^i f^i - A^i f^i||_2 \leq ||A_n^i||_2 ||f_n^i - f^i||_2 + ||A_n^i - A^i||_2 \, || f^i||_2$. The result follows from the continuity of the inner product. □

Proposition 8.4.4

If $||A_n^i - A^i||_2 \to 0$ and $||f_n^i - f^i||_\infty \to 0$ for A_n^i, $A^i \in L^2[0, 2\pi]$, f_n^i, $f^i \in L^\infty([0, 2\pi] \times [0, 2\pi])$, then

$$A_n^1 \bullet \int f_n^1(\lambda, \mu)\, A_n^2(\lambda)\, d\lambda \to A^1 \bullet \int f^1(\lambda, \mu)\, A^2(\lambda)\, d\lambda \qquad [8.4.2]$$

proof

By the continuity of the inner product it suffices to show that

$$\left\| \int f_n^1(\lambda, \mu)\, A_n^2(\lambda)\, d\lambda - \int f^1(\lambda, \mu)\, A^2(\lambda)\, d\lambda \right\|_2 \to 0.$$

Now $\left\| \int f_n^1(\lambda, \mu)\, A_n^2(\lambda)\, d\lambda - \int f^1(\lambda, \mu)\, A^2(\lambda)\, d\lambda \right\|_2 \leq$

$$\left\| \int f_n^1(\lambda, \mu)\, A_n^2(\lambda)\, d\lambda - \int f_n^1(\lambda, \mu)\, A^2(\lambda)\, d\lambda \right\|_2 +$$

$$\left\| \int f_n^1(\lambda, \mu)\, A^2(\lambda)\, d\lambda - \int f^1(\lambda, \mu)\, A^2(\lambda)\, d\lambda \right\|_2$$

$\leq ||f_n^1||\, ||A_n^2(\lambda) - A^2(\lambda)||_2 + ||f_n^1 - f^1||\, ||A^2||_2$, where $||f_n^1||$ and $||f_n^1 - f^1||$ denote operator norms. If g is any function in $L^\infty([0, 2\pi] \times [0, 2\pi])$, then $||g|| \leq ||g||_\infty$ (this is obvious, as $\left| \int g(\lambda, \mu)\, A(\lambda) \right| \leq ||g||_\infty \int |A(\lambda)|\, d\lambda \leq ||g||_\infty \, ||A||_2$ regardless of what μ is). The result immediately follows. □

proof of theorem 8.3.1 (part (c), and the completion of the proof of theorem 8.3.2)

For any function f on $[0, 2\pi]$ use the notation f^n to denote the "stepfunctionized version"

$f^n(\lambda) \equiv \sum f\left(\frac{2\pi j}{N}\right) \chi_{\Delta_j}(\lambda)$, $\chi_{\Delta_j}(\lambda)$ being defined in the first paragraph of this section.

If f is a function on $[0, 2\pi] \times [0, 2\pi]$, let

$f^n \equiv \sum_j \sum_k f\left(\frac{2\pi j}{N}, \frac{2\pi k}{N}\right) \chi_{\Delta_j \times \Delta_k}(\lambda)$. Notice that $n/2\pi$ times [8.4.1] can be written as

$$(P_n[A_1]\, f^n_{a_1 a_2}) \bullet (P_n[A_2]\, f^n_{b_1 b_2}) + (P_n[A_1]\, f^n_{a_1 b_2}) \bullet (P_n[A_2]\, f^n_{b_1 a_2}) \; +$$

$$P_n[A_1] \bullet \int P_n[A_2](\lambda_1)\, f^n_{a_1 b_1 a_2 b_2}(\lambda_1, \lambda_2, -\lambda_2)\, d\lambda_1 \qquad [8.4.3]$$

where P_n is the projection of L^2 onto M_n (see introduction to this section). Proposition 8.4.2 gives $P_n[A_i] \to A_i$, and then proposition 8.4.3 and the fact that $||f^n - f||_\infty \to 0$ for any continuous function f gives that

$$(P_n[A_1]\, f^n_{a_1 a_2}) \bullet (P_n[A_2]\, f^n_{b_1 b_2}) + (P_n[A_1]\, f^n_{a_1 b_2}) \bullet (P_n[A_2]\, f^n_{b_1 a_2}) \to$$
$$A_1 f_{a_1 a_2} \bullet A_2\, f_{b_1 b_2} + A_1\, f_{a_1 b_2} \bullet A_2\, f_{b_1 a_2}.$$

Proposition 8.4.4 gives

$$P_n[A_1] \bullet \int P_n[A_2](\lambda_1)\, f^n_{a_1 b_1 a_2 b_2}(\lambda_1, \lambda_2, -\lambda_2)\, d\lambda_1 \to$$

$$A_1(\lambda_2) \bullet \int A_2(\lambda_1)\, f^n_{a_1 b_1 a_2 b_2}(\lambda_1, \lambda_2, -\lambda_2)\, d\lambda_1.$$

This completes the proof of theorem 8.3.1. Theorem 8.3.2 follows from the fact that $|(P_n[A_1]$

$f^n_{a_1a_2}) \bullet (P_n[A_2]\ f^n_{b_1b_2})| \leq ||(P_n[A_1]\ f^n_{a_1a_2}||_2\ ||P_n[A_2]\ f^n_{b_1b_2})||_2 \leq ||P_n[A_1]||_2\ ||f^n_{a_1a_2}||_\infty$ $||P_n[A_2]||_2\ ||f^n_{b_1b_2}||_\infty \leq ||A_1||_2\ ||f^n_{a_1a_2}||_\infty\ ||A_2||_2\ ||f^n_{b_1b_2}||_\infty \leq ||A_1||_2\ ||f_{a_1a_2}||_\infty\ ||A_2||_2\ ||f_{b_1b_2}||_\infty$, and a similar relation holds for $|(P_n[A_1]\ f^n_{a_1b_2}) \bullet (P_n[A_2]\ f^n_{b_1a_2})|$. We also have

$$|P_n[A_1] \bullet \int P_n[A_2](\lambda_1)\ f^n_{a_1b_1a_2b_2}(\lambda_1, \lambda_2, -\lambda_2)\ d\lambda_1| \leq$$

$$||P_n[A_1]||_2 \left\| \int P_n[A_2](\lambda_1)\ f^n_{a_1b_1a_2b_2}(\lambda_1, \lambda_2, -\lambda_2)\ d\lambda_1 \right\|_2 \leq$$

$$||P_n[A_1]||_2\ ||P_n[A_2]||_2\ ||f^n_{a_1b_1a_2b_2}||_\infty \leq ||A_1||_2\ ||A_2||_2\ ||f_{a_1b_1a_2b_2}||_\infty.$$

Note that an analogous relation holds for the "o(T)" part in Brillinger's (1981) expression for cov $\{J^T_{a_1b_1}(A_1),\ J^T_{a_2b_2}(A_2)\}$, in which we replace $||f^n_{a_1a_2}||_\infty$, $||f^n_{b_1b_2}||_\infty$, $||f^n_{a_1b_2}||_\infty$, $||f^n_{b_1a_2}||_\infty$ or $||f^n_{a_1b_1a_2b_2}||_\infty$ by another bounded expression (the existence of which is guaranteed by Brillinger's (1981) lemma P4.2, p. 402).

Theorem 8.3.3 (i) under assumption "c" follows from Brillinger's (1981) theorem 7.6.1. (ii) follows by an argument similar to the proof of theorem 8.3.2 above.

8.5 QL Function Results

Recall that proposition 6.2.3 showed $D(f, g) = \int \log(f) + g/f\ d\lambda$ has the correct partial derivative (with respect to f) to be a QL function. We would like to complete the proof that D is a QL function by showing that it satisfies the other defining properties, and generalize the proof to the multivariate case. The next proposition is needed to verify property (d) in the definition for "univariate" QL distances.

Proposition 8.5.1

If $L(x, \lambda)$: $\mathbf{R} \times [a, b] \rightarrow \mathbf{R}$ has a continuous partial derivative with respect to x and the mapping F: $C \rightarrow B(L^2)$ is defined by $F(f)=L(f, \lambda)$ where $L(f, \lambda)$ is the multiplication operator (i.e. $x(\lambda) \rightarrow L(f, \lambda)\ x(\lambda)$ for $x(\lambda) \in L^2$) then $DF_f[g]$ (the derivative evaluated at $f \in C$ applied to $g \in C$) is the multiplication operator $\frac{\partial L(f,\lambda)}{\partial x}\, g(\lambda)$.

proof

The proof follows from the fact that if $\phi \in C$, $||\phi||_\infty = ||M_\phi||$, where M_ϕ is the multiplication operator on L^2 and $||\ .\ ||$ is the operator norm (theorem 1.5 p 28 of Conway (1985)). So from proposition 6.2.2,

$$\frac{||L(f, \lambda) - [L(f_0, \lambda) + L'(f_0, \lambda)(f - f_0)||}{||f - f_0||_\infty} \rightarrow 0$$

as $||f - f_0||_\infty \rightarrow 0$. Here, the norm in the numerator is either the sup or operator, since they are the same.

We are now ready to obtain the general multivariate result for "independent" QL distances, which would be stated as follows:

Theorem 8.5.1

Suppose

(a) Λ is a finite collection of disjoint closed intervals contained in the interval [a, b].

(b) $W_{\psi_q}(\mathbf{x}, \lambda)$ is a positive definite matrix with domain $\Psi_q \times \mathbf{R}^k \times \Lambda$, where $\psi_q \in \mathbf{R}^q$, whose components have continuous (on $\Psi_q \times \mathbf{R}^k \times [a, b]$) partial derivatives with respect to components of $\mathbf{x}$.

(c) $l_{\psi_q}(\mathbf{x}, \lambda)$: $\Psi_q \times \mathbf{R}^k \times [a, b] \rightarrow \mathbf{R}$ and $r_{\psi_q}(\mathbf{x}, \lambda)$: $\Psi_q \times \mathbf{R}^k \times [a, b] \rightarrow \mathbf{R}^k$ are continuous functions

on their domain having the property that

$$\frac{\partial l_{\psi_q}(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = W_{\psi_q}(\mathbf{x}, \lambda)\,\mathbf{x}\,, \qquad \frac{\partial [r_{\psi_q}(\mathbf{x}, \lambda)]}{\partial \mathbf{x}'} = W'_{\psi_q}(\mathbf{x}, \lambda). \tag{8.5.1}$$

Then if

$$D^W_\psi(x(\lambda), y(\lambda)) = \int_\Lambda l_\psi(\, x(\lambda), \lambda\,)\, d\lambda + r_\psi(x(\lambda), \lambda) \bullet y(\lambda) \tag{8.5.2}$$

where $\bullet$ represents the inner product in $\Pi_k L^2$, Ψ is a compact subset of C^q (so $\psi=(\psi_1(\lambda), \ldots \psi_q(\lambda))$), and the domain of D is $\Psi \times \Pi_k C[a, b] \times \Pi_k L^2[a, b]$ (or $\Pi_k C[\Lambda] \times \Pi_k L^2[\Lambda]$), D satisfies definition 6.2.1 and is therefore a Quasi-Likelihood distance in the extended sense. Note that as in proposition 6.2.3, two results are given by this theorem. If the design space is [a, b] and Λ is a proper subset of [a, b], then estimates obtained with this QL function will not be optimal.

proof

We verify that D^W_ψ satisfies the conditions of definition 6.3.4.

condition (b): Fix $y(\lambda) \in \Pi_k L^2$, and define functions $P:\Pi_k C[a, b] \to \Pi_k C[\Lambda]$ as the truncation operator, $F_1:\Pi_k C[\Lambda] \to C[\Lambda]$ by $F_1(f)=l_\psi(f(\lambda), \lambda)$, $F_2:\Pi_k C[\Lambda] \to \Pi_k C[\Lambda]$ by $F_2(f)=r_\psi(f(\lambda), \lambda)$, F_3: $\Pi_k C[\Lambda] \to \mathbf{R}$ by $F_3(f)=y(\lambda) \bullet f(\lambda)$ where the inner product is between two functions in $\Pi_k L^2[\Lambda]$, $F_4:C[\Lambda] \to \mathbf{R}$ by $F_4(f)=\int_\Lambda f\, d\lambda$. Then for a fixed $y(\lambda)$, [8.5.2] may be expressed as $[F_4 \circ F_1 + F_3 \circ F_2] \circ P$. Split this into $F_4 \circ F_1 \circ P + F_3 \circ F_2 \circ P$ and consider the pieces separately. $[F_4 \circ F_1 \circ P]'\big|_f [h]=F'_4\big|_{F_1 \circ P[f]} \circ (F_1 \circ P)'\big|_f[h]$. Now F_4 is linear, hence is its own derivative. So this equals $F_4 \circ (F_1 \circ P)'\big|_f[h] = F_4 \circ F'_1\big|_{P[f]} \circ P'\big|_f[h] = F_4 \circ F'_1\big|_{P[f]} \circ P[h]$.

On the other hand, $[F_3 \circ F_2 \circ P]'\big|_f [h]=F'_3\big|_{F_2 \circ P[f]} \circ (F_2 \circ P)'\big|_f[h]$. F_3 is linear, so is its own derivative. This means we can rewrite this expression as

$$F_3 \circ (F_2 \circ P)'\big|_f[h] = F_3 \circ F'_2\big|_{P[f]} \circ P'\big|_f[h] = F_3 \circ F'_2\big|_{P[f]} \circ P[h].$$

Now the proof follows immediately from corollary 6.2.1. For the right hand side, note that

$$\frac{\partial[r_\psi(\mathbf{x}, \lambda) \bullet \mathbf{y}]}{\partial \mathbf{x}} = \left(\frac{\partial[r_\psi(\mathbf{x}, \lambda)]}{\partial \mathbf{x}} \right)^* [\mathbf{y}]. \qquad [8.5.3]$$

To see this, take the derivative of $r_1(x)y_1+r_2(x)y_2+ \ldots +r_k(x)\, y_k$ with respect to $\mathbf{x}$. It is

$$\begin{bmatrix} \frac{\partial r_1(\mathbf{x})}{\partial x_1} y_1 + \frac{\partial r_2(\mathbf{x})}{\partial x_1} y_2 + \ldots + \frac{\partial r_k(\mathbf{x})}{\partial x_1} y_k \\ \frac{\partial r_1(\mathbf{x})}{\partial x_2} y_1 + \frac{\partial r_2(\mathbf{x})}{\partial x_2} y_2 + \ldots + \frac{\partial r_k(\mathbf{x})}{\partial x_2} y_k \\ \cdot \\ \frac{\partial r_1(\mathbf{x})}{\partial x_k} y_1 + \frac{\partial r_2(\mathbf{x})}{\partial x_k} y_2 + \ldots + \frac{\partial r_k(\mathbf{x})}{\partial x_k} y_k \end{bmatrix}$$

Observe this is $\left(\frac{\partial[r_\psi(\mathbf{x}, \lambda)]}{\partial \mathbf{x}} \right)^* \mathbf{y}$.

Corollary 6.2.1 gives the mapping $F:\Pi_k C \rightarrow \Pi_k C$ defined by $F[x(\lambda)]=r_\psi(x(\lambda), \lambda)$ has a derivative $DF[x(\lambda)]= \frac{\partial[r_\psi(x(\lambda), \lambda)]}{\partial \mathbf{x}} [h(\lambda)]$ evaluated at $x(\lambda)$ and applied to $h(\lambda)$ (a matrix of functions applied to a vector of functions). So the mapping $\Pi_k C \rightarrow \mathbf{R}$ defined by $y(\lambda) \bullet F[x(\lambda)]$ has a derivative (evaluated at $x(\lambda)$ and applied to $h(\lambda)$)

$$y(\lambda) \bullet \frac{\partial[r_\psi(x(\lambda), \lambda)]}{\partial \mathbf{x}} [h(\lambda)] \qquad [8.5.4]$$

$$\left(= \left(\frac{\partial[r_\psi(\mathbf{x}, \lambda)]}{\partial \mathbf{x}} \right)^* [y(\lambda)] \bullet h(\lambda) \right).$$

condition (c): Obvious since the components of $W_{\psi_q}(\mathbf{x}, \lambda)$ are continuously differentiable.

condition (d): We obtain the form of $L_{\psi f}$. If M is a space of matrices, and F is the mapping $\mathbf{R}^k \rightarrow M$ defined by $x \rightarrow W(x)$, then the derivative of F may be expressed as the vector of matrices

$$\left[\frac{\partial W}{\partial x_1}, \ldots, \frac{\partial W}{\partial x_k}\right] \qquad [8.5.5]$$

in the sense that F may be approximated by a translate of the linear mapping $\mathbf{R}^k \rightarrow M$ defined by

$$(z_1, \ldots z_k)' \rightarrow \left[\frac{\partial W}{\partial x_1}, \ldots, \frac{\partial W}{\partial x_k}\right](z_1, \ldots z_k)'. \qquad [8.5.6]$$

Here, $\frac{\partial W}{\partial x_i}$ is a matrix obtained by replacing each component of W with its partial derivative with respect to x_i. It follows by the same method of proof of proposition 6.2.1 that if F is the mapping $\Pi_k C \rightarrow B(\Pi_k L^2)$ defined by $x(\lambda) \rightarrow W(x(\lambda))[\ .\ ]$ (where the matrix is applied pointwise to each function $y(\lambda)$ in $\Pi_k L^2$), then F' evaluated at $x(\lambda)$ and applied to $h(\lambda)$ is the matrix of functions

$$\left[\frac{\partial W(x(\lambda))}{\partial x_1}, \ldots, \frac{\partial W(x(\lambda))}{\partial x_k}\right][h_1(\lambda), \ldots h_k(\lambda)]' \qquad [8.5.7]$$

which is an operator on $\Pi_k L^2$. To apply the operator to an element $a(\lambda)$ of $\Pi_k L^2$, simply apply the matrix of functions to the vector of functions $a(\lambda)$. The continuity conditions on the partial derivatives of the components of W(x) yield that conditions (d) (ii) and (iii) hold. □

Corollary 8.5.1 (to theorem 8.5.1)

The bivariate distance function defined by

$$D(f_\theta, g) \equiv \int_{-\pi}^{\pi} \log \det f_\theta^m(\omega) + \text{trace}\,([f_\theta^m(\omega)]^{-1}\, g^m(\omega)\,)\, d\omega \qquad [8.5.8]$$

where for a 4 tuple x, x^m is as defined by [6.4.2], is a QL distance.

proof

As in [6.4.3], let $(1/2\pi) \times M(f_{11}, f_{22}, c, q)$ be the matrix valued function

$$\begin{bmatrix} f_{11}^2 & |f_{12}|^2 & f_{11}c_{12} & f_{11}q_{12} \\ |f_{12}|^2 & f_{22}^2 & f_{22}c_{12} & f_{22}q_{12} \\ f_{11}c_{12} & f_{22}c_{12} & \frac{1}{2}(f_{11}f_{22}+c_{12}^2-q_{12}^2) & c_{12}q_{12} \\ f_{11}q_{12} & f_{22}q_{12} & c_{12}q_{12} & \frac{1}{2}(f_{11}f_{22}+q_{12}^2-c_{12}^2) \end{bmatrix} \qquad [8.5.9]$$

Then if $v_n(\lambda)=(I_{n1}(\lambda), I_{n2}(\lambda), \hat{c}_n(\lambda), \hat{q}_n(\lambda))'$, $v(\lambda)=(f_1(\lambda), f_2(\lambda), c(\lambda), q(\lambda))'$, and $a(\lambda)$ is any k vector of L^2 functions (i.e. $a(\lambda) \in \Pi_k L^2$), the operator $V[a(\lambda)] = M(v(\lambda))\, a(\lambda)$ is the variance operator for the L^2 sequence $v_n(\lambda)$. To show that the derivative $\frac{\partial}{\partial x} D(x, y) = V^{-1}(x(\lambda))[x(\lambda) - y(\lambda)]$ it suffices to show that (1) $\mathbf{M}(\mathbf{x}) \frac{\partial}{\partial \mathbf{x}} \log \det \mathbf{x}^m = \mathbf{x}$, and (2) $\mathbf{M}(\mathbf{x}) \frac{\partial}{\partial \mathbf{x}} \text{trace}[\mathbf{x}^m]^{-1}\mathbf{y}^m = \mathbf{y}$. This is because these imply $\frac{\partial}{\partial \mathbf{x}} \log \det \mathbf{x}^m = \mathbf{M}^{-1}(\mathbf{x})\, \mathbf{x}$ and $\frac{\partial}{\partial \mathbf{x}} \text{trace}[\mathbf{x}^m]^{-1}\mathbf{y}^m = \mathbf{M}^{-1}(\mathbf{x})\, \mathbf{y}$, and then theorem 8.5.1 yields the result. Note that $\text{trace}[\mathbf{x}^m]^{-1}\mathbf{y}^m =$

$$\text{tr}\left(\begin{bmatrix} x_1 & x_3 - ix_4 \\ x_3+ix_4 & x_2 \end{bmatrix} \begin{bmatrix} y_1 & y_3 - iy_4 \\ y_3+iy_4 & y_2 \end{bmatrix} \right)$$

$$= y_1x_1+(x_3 - ix_4)(y_3 - iy_4) +(x_3+ix_4)(y_3+iy_4) +x_2y_2$$

$$=y_1x_1+2(x_3y_3+x_4y_4)+x_2y_2$$

$$=(y_1, y_2, y_3, y_4) \bullet (x_1, x_2, 2x_3, 2x_4).$$

Thus the function is of the correct form [8.5.2]. Also observe that if $D_x\, r(\mathbf{x}) \bullet \mathbf{y} = M^{-1}(\mathbf{x})[\mathbf{y}]$ for all $\mathbf{y}$, then $(D_x[r(\mathbf{x})])^* = M^{-1}(\mathbf{x})$. This is because $D_x\, [r(\mathbf{x}) \bullet \mathbf{y}] = (D_x[r(\mathbf{x})])^*[\mathbf{y}]$ (as explained earlier) and $(D_x[r(\mathbf{x})])^*[\mathbf{y}] = M^{-1}(\mathbf{x})[\mathbf{y}]$ for all $\mathbf{y}$ implies the two matrices are the same. In our case, as M^{-1} is symmetric, $(D_x[r(\mathbf{x})])^* = M^{-1}(\mathbf{x}) = (M^{-1})^*(\mathbf{x})$, showing [8.5.1].

Proof of (1): Let $\mathbf{x}=(a, b, c, d)'$. Observe that log det $\mathbf{x}^m$ = log $[ab-(c^2+d^2)]$. So $\frac{\partial}{\partial \mathbf{x}}$ log det $\mathbf{x}^m = \frac{1}{\Delta}(b, a, -2c, -2d)'$, where $\Delta=ab-(c^2+d^2)$. It is easily seen that $\mathbf{M}(\mathbf{x})\ \frac{1}{\Delta}(b, a, -2c, -2d)' = \mathbf{x}$ by multiplying the matrix and vector.

Proof of (2): Let $\mathbf{f} = (f_1, f_2, c, q)$ and $\mathbf{h} = (e, g, k_1, k_2)$. Then if $F_h(\mathbf{f})=\text{trace}([\mathbf{f}^m]^{-1}\mathbf{h}^m)$,

$$\frac{\partial F_h}{\partial f_1} = \frac{\Delta g - t f_2}{\Delta^2}, \frac{\partial F_h}{\partial f_2} = \frac{\Delta e - t f_1}{\Delta^2}, \frac{\partial F_h}{\partial c} = \frac{\Delta(-2k_1)-t(-2c)}{\Delta^2}, \frac{\partial F_h}{\partial q} = \frac{\Delta(-2k_2)-t(-2q)}{\Delta^2}$$

where $t=f_2e+f_1g-2k_1c-2k_2q$ and Δ is as above. To show the result, we take the inner product of each row of the matrix $\mathbf{M}(\mathbf{f})$ with the vector $\frac{\partial F_h}{\partial \mathbf{f}}$, and show this inner product equals minus the corresponding element of the vector $\mathbf{h}$.

First row: $[f_1^2(\Delta g-tf_2)+(c^2+q^2)(\Delta e-tf_1)+f_1c(-2k_1\Delta+2ct)+f_1q(-\Delta 2k_2+2tq)]/\Delta^2=$

$$\frac{-t[f_1^2f_2+(c^2+q^2)f_1-2c^2f_1+2q^2f_1]}{\Delta^2} + \frac{f_1^2\Delta g+(c^2+q^2)\Delta e+f_1c(-2k_1\Delta)+f_1q(-2k_2\Delta)}{\Delta^2}=$$

$$\frac{-\Delta tf_1}{\Delta^2} + \frac{\Delta[f_1^2g+(c^2+q^2)e-2k_1f_1c-2k_2f_1q]}{\Delta^2} = \frac{-e\,\Delta}{\Delta} = -e.$$

Second row: $[(c^2+q^2)(\Delta g-tf_2)+f_2^2(\Delta e-tf_1)+f_2c(-2k_1\Delta+2tc)+f_2q(-2k_2\Delta+2qt)]/\Delta^2=$

$$\frac{t(-f_2((c^2+q^2)-f_2^2f_1+2c^2f_2+2q^2f_2)}{\Delta^2} + \frac{(c^2+q^2)\Delta g+f_2^2\Delta e-2k_1f_2c\Delta-2k_2f_2q\Delta}{\Delta^2} =$$

$$\frac{-\Delta tf_2}{\Delta^2} + \frac{\Delta[(c^2+q^2)g+f_2^2e-2k_1f_2c-2k_2f_2q]}{\Delta^2} = \frac{-g\,\Delta}{\Delta} = -g.$$

Third row: $[f_1c(\Delta g-tf_2)+f_2c(\Delta e-tf_1)+\frac{1}{2}(f_1f_2+c^2-q^2)(-2k_1\Delta+2tc)+cq(-2k_2\Delta+2tq)]/\Delta^2 =$

$$\frac{t[-f_1f_2c-f_1f_2c+2c(1/2)\ (f_1f_2+c^2-q^2)+2q^2c]}{\Delta^2} + \frac{\Delta[f_1cg+f_2ce-k_1(f_1f_2+c^2-q^2)-2k_1cq]}{\Delta^2} =$$

$$\frac{-tc\Delta}{\Delta^2} + \frac{\Delta[f_1cg+f_2ce-k_1(f_1f_2+c^2-q^2)-2k_1cq]}{\Delta^2} = \frac{-k_1\Delta}{\Delta} = -k_1$$

Fourth row: $[f_1q(\Delta g-tf_2)+f_2q(\Delta e-tf_1)+cq(-2k_1\Delta+2tc)]+\frac{1}{2}(f_1f_2+q^2-c^2)(-2k_2\Delta+2tq)/\Delta^2 =$

$$\frac{t[-f_1f_2q-f_1f_2q+2c^2q+2q\ \frac{1}{2}(f_1f_2+q^2-c^2)}{\Delta^2} + \frac{\Delta[f_1qg\ +f_2qe\ -2cqk_1-\ 2k_2\ \frac{1}{2}(f_1f_2+q^2-c^2)]}{\Delta^2} =$$

$$\frac{-tq\Delta}{\Delta^2} + \frac{\Delta[f_1qg\ +f_2qe\ -2cqk_1-\ 2k_2\ \frac{1}{2}(f_1f_2+q^2-c^2)]}{\Delta^2} = \frac{-k_2\Delta}{\Delta} = -k_2 \qquad \square$$

Corollary 8.5.2 (to theorem 8.5.1)

The cospectral functions in chapter 4 defined by

$$D^A_{f_xf_y\phi}(h(\lambda), g(\lambda)) = \int_{-\pi}^{\pi} L_A(h, g, \lambda)\, d\lambda\ , \quad D^B_{f_xf_y\phi}(h(\lambda), g(\lambda)) = \int_{-\pi}^{\pi} L_B(h, g, \lambda)\, d\lambda$$

satisfy definition 6.2.1, where $\Psi = \{f_{1\theta}\} \times \{f_{2\theta}\} \times \{q_\theta\}$ (or $\{f_{1\theta}\} \times \{f_{2\theta}\} \times \phi$).

Corollary 8.5.3 (to theorem 8.5.1)

Let Ψ be the space consisting of $S \times \{f_\theta\}$, where the model space is the set of positive continuous functions, and S is compact in $\mathbf{R}^+$. Define

$$W^1_\psi[x(\lambda)] = f^2_\theta(\lambda)\ x(\lambda) + \kappa\ f_\theta(\lambda) \int_{-\pi}^{\pi} f_\theta(\lambda)\ x(\lambda)\, d\lambda \qquad [8.5.10]$$

and $W_\psi = [W^1_\psi]^{-1}$. Then if D is defined as in example 3 of section 6.4, D satisfies condition (c) of definition 6.3.4 and hence may be used in an IRWLS procedure (see remarks following

corollary 7.1.2).

proof

It's only necessary to verify the continuity of the mapping $(f, \kappa) \rightarrow W^1_{(f,\kappa)}$ where the operator space has the strong operator topology (i.e. the topology generated by the operator norm), since the inversion is continuous (and verify that W^1 is invertible). Break this operator into its two parts, $V_f[x]=f(\lambda)x(\lambda)$ and $V_{f,\kappa}[x]=\kappa\, f_\theta(\lambda) \int f_\theta(\lambda)\, x(\lambda)\, d\lambda$. Note $||V_{f_1} - V_{f_2}|| \leq$

$$\sup_{|x|=1} \sqrt{\int [(f_1 - f_2)x]^2 d\lambda} \leq \sup_{|x|=1} |f_1 - f_2| \sqrt{\int x^2 d\lambda} \leq |f_1 - f_2|, \text{ so this operator}$$

is continuous. Defining $W_f[x]=\int f\, x\, d\lambda$, the mapping F_1: $(f, g) \rightarrow g\, W_f$ $(C \times C \rightarrow B(L^2))$ is continuous since for $||x||=1$ we have

$$|| g_1 W_{f_1} - g_2 W_{f_2}[x] || \leq$$

$$\left| g_1 \int f_1 x\, d\lambda - g_2 \int f_1 x\, d\lambda \right| + \left| g_2 \int f_1 x\, d\lambda - g_2 \int f_2 x\, d\lambda \right| \leq ||g_1 - g_2||_\infty\, ||f_1|| + ||g_2||_\infty\, ||f_1 - f_2||.$$

Note the mapping F_2: $(f, \kappa) \rightarrow (\kappa f, f)$ $(C \times \mathbf{R} \rightarrow C \times C)$ is continuous, and note that $V_{f,\kappa} = F_1 \circ F_2$.

$W^1_{(f,\kappa)}$ is invertible because each piece is a nonnegative operator, and the first is positive definite due to f being strictly positive (see exercise 8, p. 249 Conway (1985)).

Remark: Theorems 10.2.2 and 9.4.5 will essentially show the stronger result that $W(f, \kappa)[\,.\,]$ is continuously differentiable as a function of f. Operator inversion is also differentiable (e.g. theorem 7.17, p. 94 Chae (1985)), so $W^{-1}(f, \kappa)$ is a QL operator under definition 6.3.4.

8.6 IRWLS for Non Gaussian Processes

We now have all of the machinery in place to define an IRWLS procedure for a filtered white noise process (which obviously includes Gaussian processes) and show that the parameter estimates obtained from such a procedure are asymptotically BLUE.

Step 1

Obtain $\phi(\lambda)$, a rough initial estimate of $f(\lambda)$ (the true spectral density) by smoothing (or fit a spline to the smoothed periodogram).

Step 2

Obtain a rough initial estimate of κ by using Chiu's (1988) theorem 6 and theorem 3, i.e. estimate

$$\int\int_{\Lambda} f_4(\lambda, -\lambda, \mu)\, d\lambda\, d\mu \quad \text{by}$$

$$\frac{2\pi}{n^2}\sum_{\Lambda}\sum_{\Lambda} I_n(\lambda)\, I_n(\mu) - \frac{2\pi}{n}\sum_{\Lambda} I_n^2 + 2\pi\left(\sum_{\Lambda} I_n\right)^2$$

and then obtain $\hat{\kappa}$ by dividing this number by $\int\int_{\Lambda} \phi(\lambda)\phi(\mu)\, d\lambda\, d\mu$.

Step 3

Obtain an estimate $\hat{\theta}$ of θ by minimizing

$$(I_n(\lambda) - f_\theta(\lambda)) \bullet W_{\hat{\psi}}\, [I_n(\lambda) - f_\theta(\lambda)]$$

with respect to θ, where $\hat{\psi} = (n, \phi(\lambda), \hat{\kappa})$ and W_ψ is as defined in corollary 8.5.3.

Step 4

Repeat the above steps, using $f_{\hat{\theta}}(\lambda)$ in place of $\phi(\lambda)$.

Remark: In practice you use matrix representations of the operator W_{ψ} in step 3 which are found as follows. Let $\lambda_1, \lambda_2, \ldots \lambda_k$ be the ordered Fourier frequencies in Λ, and let W_{ψ}^{-1} be the inverse of the matrix

$$\begin{bmatrix} f^2(\lambda_1) & 0 & . & . & . & 0 \\ 0 & f^2(\lambda_2) & 0 & . & . & 0 \\ . & 0 & . & & & \\ . & . & & . & & \\ . & . & & & . & \\ 0 & 0 & & & & f^2(\lambda_k) \end{bmatrix} +$$

$$\frac{\kappa}{n} \times \begin{bmatrix} f(\lambda_1)f(\lambda_1) & . & . & . & . & f(\lambda_1)f(\lambda_k) \\ . & & & & & . \\ . & & & & & . \\ . & & & & & . \\ . & & & & & . \\ f(\lambda_k)f(\lambda_1) & . & . & . & . & f(\lambda_k)f(\lambda_k) \end{bmatrix}$$

Replace $I_n(\lambda)$ with the vector $[I_n(\lambda_1), I_n(\lambda_2), \ldots, I_n(\lambda_k)]'$ and $f_\theta(\lambda)$ with the vector $[f_\theta(\lambda_1), f_\theta(\lambda_2), \ldots, f_\theta(\lambda_k)]'$. A (probably unnecessarily complicated) mathematical justification of this simple idea is given in proposition 8.6.1 below.

The theorems of the preceding section yield an "asymptotic BLUE" property of parameter estimates obtained by using the IRWLS procedure. This is because $\hat{\kappa}$ obtained in

step 2 and $\widehat{\theta}$ obtained in step 3 on the *first* cycle through the procedure are *consistent* (but not optimal) estimators by theorem 7.1.1. Applying theorems 7.1.1 and 7.1.2 to the *second* cycle yields that the estimate of $\widehat{\theta}$ obtained there is asymptotically BLUE. Note that the IRWLS procedure *does not require the explicit form of a non-Gaussian likelihood* (i.e. we don't have to specify the non-Gaussian "distance" D(f(λ), g(λ)) in order to obtain optimal parameter estimates, just the form of the variance operator as given in corollary 5.2.1.) The form of this non-Gaussian distance (if it exists) is not known, but is of theoretical (and possibly practical) interest and is a topic for future research. In certain cases it may also be desirable to use the incorrect variance operator in the IRWLS procedure, e.g. in the case where $f_{\widehat{\theta}}$ from the first iteration is near 0 at a frequency λ. This will force $f_{\widehat{\theta}}$ in future iterations to be (perhaps overly) influenced by the periodogram ordinate at λ. To prevent this from happening, one could use $1/\max\{c, f^2_{\widehat{\theta}}(\lambda)\}$ (for some predetermined constant c) as a weight for the next iteration rather than $1/f^2_{\widehat{\theta}}(\lambda)$. Again, a topic for future research.

Proposition 8.6.1

Let M_n be a sequence of finite dimensional "step function" subspaces of L^2, and define P_n as the projection operator onto M_n. Suppose (i) $W(x(\lambda), \kappa)$ is defined as in [8.5.10], (ii) U is a neighborhood of $x_0(\lambda)$ and K is a (positive) neighborhood of κ_0 so that for all $x \in U$, $\kappa \in K$, there exists a positive R_1, R_2 with $x(\lambda) \geq R_1$ and $||W(x(\lambda), \kappa)|| < R_2$ in operator norm (iii) $P_n[y] \xrightarrow{L^2} y$ for any $y \in L^2$. Then given $\epsilon > 0$ and $z \in L^2$, $\exists$ an integer N and a neighborhood $V \times K_1 \subset U \times K$ of $x_0 \times \kappa_0$ so that for $n \geq N$, $x \in V$, $\kappa \in K$

$$||[P_n W(x(\lambda), \kappa) P_n]^{-1} P_n[z] - W^{-1}(x_0(\lambda), \kappa_0)[z]|| < \epsilon.$$

Here, the operator $P_n W(x(\lambda), \kappa) P_n$ is restricted to M_n and the inverse is defined on this space.

proof

Step 1 Note that if λ_n is the smallest eigenvalue of $P_n W(x(\lambda), \kappa)\ P_n$ restricted to the subspace

M_n, then $\lambda_n \geq R_1$ for all n and all x, $\kappa \in U \times K$ (of course this implies the operator $P_n W_n P_n$ restricted to the subspace M_n is invertible). This is because $P_n M_\phi P_n$ may be represented with respect to the natural basis for M_n as a diagonal matrix with $1/\text{length}(\Lambda_i) \int_{\Lambda_i} \phi(\lambda)\, d\lambda$ on the diagonal, where Λ_i is the ith step. But $1/\text{length}(\Lambda_i) \int_{\Lambda_i} \phi(\lambda)\, d\lambda \geq R_1$.

Step 2 Observe the sequence $[P_n W(x(\lambda), \kappa)\, P_n]^{-1}$ (restricted to M_n) is bounded in operator norm (by $R_3 \equiv 1/R_1$) for all n and all x, $\kappa \in U \times K$. Also, $[P_n W(x(\lambda), \kappa) P_n]^{-1} P_n$ as an operator on L^2 is self adjoint ($x \bullet [P_n W(x(\lambda),\kappa) P_n]^{-1} P_n[y] = P_n[x] \bullet [P_n W(x(\lambda), \kappa) P_n]^{-1} P_n[y] = [P_n W(x(\lambda), \kappa) P_n]^{-1} P_n[x] \bullet P_n[y]$ by the self adjointness of $[P_n W(x(\lambda), \kappa) P_n]^{-1}$ on M_n)
Choose y so $z = W(x_0(\lambda), \kappa_0)[y]$. This is possible since W is invertible.

Step 3 Choose $V \times K_1 \subset U \times K$ and N_1 so

$|||[P_n W(x_0(\lambda), \kappa_0) - P_n W(x(\lambda), \kappa) P_n][y]|| < \epsilon/2R_3$ for all x, $\kappa \in V \times K_1$ and $n \geq N_1$.

Note $|||[P_n W(x_0(\lambda), \kappa_0) - P_n W(x(\lambda), \kappa) P_n][y]|| \leq$

$|||[P_n W(x_0(\lambda), \kappa_0) - P_n W(x(\lambda), \kappa)][y]|| + |||[P_n W(x(\lambda), \kappa) - P_n W(x(\lambda), \kappa) P_n][y]||$.

But $|||[P_n W(x_0(\lambda), \kappa_0) - P_n W(x(\lambda), \kappa)][y]|| = ||P_n[W(x_0(\lambda), \kappa_0) - W(x(\lambda), \kappa)][y]|| \leq$

$||P_n||\ |||[W(x_0(\lambda), \kappa_0) - W(x(\lambda), \kappa)][y]||$, and

$|||[P_n W(x(\lambda), \kappa) - P_n W(x(\lambda), \kappa) P_n][y]|| =$

$$|||[P_n W(x(\lambda), \kappa)[I - P_n][y]|| \leq ||P_n||\ ||W(x(\lambda), \kappa)||\ |||[I - P_n][y]||.$$

Therefore, $V \times K_1$ should be chosen so $|||[W(x_0(\lambda), \kappa_0) - W(x(\lambda), \kappa)][y]|| < \epsilon/2$ for all x, $\kappa \in V \times K_1$, and N_1 should be chosen so $|||[I - P_n][y]|| \leq \epsilon/2R_2$.

Step 4 Observe that for x, $\kappa \in V \times K_1$, $||[P_n W(x(\lambda), \kappa)P_n]^{-1}P_n[W(x_0(\lambda), \kappa_0) - P_n W(x(\lambda), \kappa) P_n][y]|| < \epsilon/2$. This is by the boundedness of $||[P_n W(x(\lambda), \kappa)P_n]^{-1}||$ by R_3 and the fact $||P_n[W(x_0(\lambda), \kappa_0) - P_n W(x(\lambda), \kappa) P_n][y]|| < \epsilon/2R_3$.

Step 5 Finally, write

$$[P_n W(x(\lambda), \kappa)P_n]^{-1}P_n \ [W(x_0(\lambda), \kappa_0) - P_n W(x(\lambda), \kappa)P_n][y] =$$

$$[P_n W(x(\lambda), \kappa)P_n]^{-1}P_n \ W(x_0(\lambda), \kappa_0)[y] - P_n[y].$$

Choose N_2 so $||P_n[y] - y|| < \epsilon/2$ for $n \geq N_2$. Hence if $N=\max\{N_1, N_2\}$ and $n \geq N$, then

$$|| \ [P_n W(x(\lambda), \kappa)P_n]^{-1}P_n W(x_0(\lambda), \kappa_0)[y] - y|| < \epsilon$$

for x, $\kappa \in V \times K_1$. Using the fact $z=W(x_0(\lambda), \kappa_0)[y]$,

$$|| \ [P_n W(x(\lambda), \kappa)P_n]^{-1}P_n[z] - W^{-1}(x_0(\lambda), \kappa_0)[z]|| < \epsilon$$

and we are done. □

8.7 Conclusions

In this chapter, we have given theorems supporting the application of chapter 6's theory to spectral analysis. Sections 2, 3 and 4 show that the periodogram satisfies the definition of a random L^2 sequence (definition 6.3.2). Specifically, we may conclude two main results from the theorems of this chapter. First, the periodogram calculated from the univariate "filtered white noise process" (example 5 of section 6.4) is a random L^2 sequence with limiting function $f(\lambda)$, the true spectrum, and variance operator $W(f, \kappa)[x(\lambda)] = 2\pi M_{f^2}[x(\lambda)] + 2\pi\kappa f(\lambda) \int_{-\pi}^{\pi} f(\lambda) \, x(\lambda) \, d\lambda$. The model space is the set of positive, continuous functions. Second, the components of the periodogram matrix calculated from the bivariate Gaussian process (example 4 of section 6.4) is a (multivariate) random L^2 sequence. The limiting function is $(f_{11}(\lambda), f_{22}(\lambda), c(\lambda), q(\lambda))$, and the variance operator is represented by the matrix of

functions $M(f_{11}(\lambda), f_{22}(\lambda), c(\lambda), q(\lambda))$ (defined in [6.4.3]), in the sense that if $\psi=(\psi_1, \psi_2, \psi_3, \psi_4)$ is a vector of L^2 functions, $V[\psi]$ is the vector of functions obtained by calculating $M[\psi]$ pointwise for each λ. The model space for this random L^2 sequence is $\{f=(f_1, f_2, f_3, f_4) \in \Pi_4 C[-\pi, \pi] \mid \det f^m>0\}$ (notation f^m defined in example 4 of section 6.4).

Theorems 8.2.1 and 8.3.3 establish the expectation conditions (f and g) of definition 6.3.2, while theorems 8.3.1 and 8.3.2 establish the variance conditions (d and e). For the multivariate process, one must take into consideration the comments in section 5.4 when establishing $M(f_{11}(\lambda), f_{22}(\lambda), c(\lambda), q(\lambda))$ is the variance operator, as theorems 8.3.1 and 8.3.2 are given in terms of (complex valued) cross spectra rather than the co and quad spectra.

The theorems of this chapter establish the above mentioned results for three versions of the periodogram: (1) the "natural" extension, (2) the "4n" step function extension, and (3) the "Fourier frequencies" step function extension. For (1) and (2), the results hold for the true spectrum a function of bounded variation which is greater than some positive constant c (for the multivariate case, the components of the true spectral matrix must be of bounded variation and $\det f^m$ must be greater than c). For (3), the results hold for the true spectrum a function greater than c which also satisfies the BCC.

Section 8.5 establishes the "QL function" results discussed in the examples of section 6.4. Theorem 8.5.1 actually gives more than this, giving sufficient conditions for the existence of a multivariate (on the function space $\Pi_k C$) QL function. Finally, section 8.6 gives an IRWLS procedure for a non Gaussian "filtered white noise" process, showing the procedure results in optimal parametric estimates according to the theorems in chapter 7. Section 8.6, however, is not the conclusion of the theory. The following chapter will take a closer look at "non Gaussian" type QL operators, examining why it may be desirable to use them for parametric estimation even when the observed series is Gaussian.

Chapter IX
"Almost Optimal" Estimation in Misspecified Models

All models are wrong; some models are useful - G. E. P. Box

9.1 Introduction

The obvious uses of the theory developed in the preceding two chapters are in obtaining optimal parameter estimates by an IRWLS method for (non) Gaussian univariate and multivariate series, and especially in cospectral estimation. But the most important ramifications are in *"almost optimal" parametric estimation for the case of a misspecified model.* Chiu (1988) pointed out that in the case of "contaminated" series, i.e. where the contamination only appears in certain frequency bands of the spectrum of the observed process, the contaminated bands may be excluded from analysis before fitting the model (see also Rice (1979), Cameron and Turner (1987)). He assumes, however, that the spectrum of the observed process satisfies the BCC (i.e. is smooth in some sense, such as being continuously differentiable). One might wish to assume that the spectrum of the observed series is f_X+f_N, where f_X and the model $\{f_\theta\}$ satisfy the BCC, but f_N is "rougher" in that it may not even be continuous. This is the reason behind the setup of the theorems in chapter 8: Theorems 8.2.1 and 8.3.1 show the correct asymptotic variance and bias conditions obtain for processes whose spectra are of bounded variation. However, there is an important distinction to be made between this more general approach and that of Chiu (1988) along with most of the literature. Under the assumption that the spectrum satisfies the BCC, one may restrict the periodogram to the finite collection of intervals Λ and obtain a random L^2 sequence. Alternatively, one

may consider $D(f, g)=\int_{\Lambda} \log(f)+g/f \, d\lambda$ to be a QL function for the periodogram defined on $[-\pi, \pi]$. However, if the spectrum is only of bounded variation, by the results in this dissertation one *may not consider the periodogram restricted to* Λ *to be a random* L^2 *sequence or* $D(\ .\ ,\ .)$ *to be a QL function.* This is essentially because theorem 8.2.1 will not necessarily yield asymptotic unbiasedness if the periodogram is restricted to Λ, or will not necessarily guarantee the link condition to hold if we attempt to restrict $D(\ .\ ,\ .\ )$ to Λ. In the sequel we will only consider the problem of fitting a model defined on $[-\pi, \pi]$, rather than fitting a model over frequency bands. It will be shown that the problem of not being able to eliminate certain "contaminated" bands is solved in another way.

We give a procedure which may be applied if it is *not precisely known which frequency bands are contaminated.* Under certain conditions, this procedure is equivalent to reweighted least squares applied to a smoothed version of the periodogram, and the procedure might roughly be described as reweighted least squares in which asymptotic covariance in the periodogram is incorrectly assumed. This relates to parametric estimation in non-Gaussian series, because the periodogram of a non-Gaussian series does have asymptotic covariance in its periodogram, as was seen in chapter 6.

One of the central ideas in this dissertation is a new definition of "quasi likelihood function" (Chapter 6) which assumes a *decreasing* amount of correlation in the observations (i.e. the periodogram) in a systematic fashion. The usual definition from McCullagh (1983) would assume any covariance between two observations as a fixed function of the means vector, and an intuitive discussion of how "variance operators" describe the limiting behavior of "variance matrices" is given in section 9.6. If the means model is misspecified, in some cases it may be enlightening to actually determine the QL function in the extended class to see how bias affects the estimates. Further, in the case of model misspecification, it may be desirable to

purposefully use the "wrong" QL function in order to lessen the bias. If this "incorrect" function is properly chosen, *variance of the parametric estimates will be little damaged if the model was in fact correct, and parametric variances will be "almost" optimal even if the model is incorrect.* Earlier we stated that frequency domain spectral estimation should be regarded as attempting to estimate a "one dimensional generalized model response surface". The applications to be presented will crystallize this idea, and the central concept extends beyond time series applications to the arbitrary dimensional "generalized model response surface" described in the next section.

9.2 The Generalized Model Response Surface

Consider the following situation. We are interested in modeling the means of the independent observations y_i as a function of some regressors x_i, where $x_i \in [0, 1]$ for all i. The means are described by $E(y_i) = \beta_0+\beta_1 x_i$, and the variances are $var(y_i) = v(x_i)$ for some positive function v on [0, 1]. The experiment is to be performed in stages, where at stage n observations corresponding to $x_i = i/2^n$ will have been observed. This is not a true "generalized model response surface", but a simplified example designed to illustrate a basic concept which applies in that setting.

First consider estimating the unknown parameter vector β solving the normal equations

$$\mathbf{X'W(y - X\beta)=0} \qquad [9.2.1]$$

where **W** is an arbitrary fixed matrix (i.e. not assumed to be symmetric, etc). It may be shown that Var $\hat{\beta}$, where $\hat{\beta}$ solves the above equation, is given by

$$[(\mathbf{X'WX})^{-1}\ \mathbf{X'}\ \mathbf{W}]\ \mathbf{V}\ [\mathbf{W'}\ \mathbf{X}(\mathbf{X'W'X})^{-1}]\ . \qquad [9.2.2]$$

Of course if $\mathbf{W} = \mathbf{V}^{-1}$, the true inverse variance matrix (a diagonal matrix with $v_{ii}=v(x_i)$), then var $\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ which is the "optimal" variance matrix for BLUE parametric estimates. However, if $\mathbf{W}'\mathbf{X} = \mathbf{V}^{-1}\mathbf{X}$, then var $\hat{\beta}$ will be exactly the same (use the fact that $\mathbf{X}'\mathbf{W}=(\mathbf{W}'\mathbf{X})'$ to reduce [9.2.2]). Hence if $\mathbf{W}'$ "acts" like $\mathbf{V}^{-1}$ on the columns of the $\mathbf{X}$ matrix, then the variance of parameter estimates will not be damaged. This fact will be used to construct W matrices which may be used in place of the V matrix in an IRWLS procedure (see, for example, Chiu (1988) for theorems supporting the use of IRWLS in spectral estimation). Specifically, we will use W matrices of the form $\mathbf{W}_n = \mathbf{V}_n^{-1}\,\mathbf{K}_n$ or $\mathbf{W}_n = \mathbf{K}_n'\,\mathbf{V}_n^{-1}\,\mathbf{K}_n$, where $\mathbf{V}_n^{-1}$ is the true diagonal inverse variance matrix, and $\mathbf{K}_n$ is a "fixed bandwidth kernel matrix". For example, ignoring possible problems at the boundary, $\mathbf{K}_n$ applied to the second column of the $\mathbf{X}$ matrix for our example might look like

$$\begin{bmatrix} \cdot \\ \frac{.21+.22+.23}{3} \\ \frac{.22+.23+.24}{3} \\ \frac{.23+.24+.25}{3} \\ \frac{.24+.25+.26}{3} \\ \frac{.25+.26+.27}{3} \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & \cdot & & & & \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & & & & \\ & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & & & \\ & & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & & & \end{bmatrix} \begin{bmatrix} \cdot \\ .22 \\ .23 \\ .24 \\ .25 \\ .26 \\ \cdot \end{bmatrix}$$

Notice that each element of the vector $\mathbf{K}_n\,\mathbf{x}$ is a Riemann sum of $\int_0^1 K(\lambda - x)\,x\,dx$, where K is a function called the **kernel** whose graph is a straight line centered around 0 on the x axis. The length of this line is called the "bandwidth", and by "fixed bandwidth" we mean

that the bandwidth remains the same even after more observations are taken in later stages of the experiment.

There are several points to be made about the setup just described. First, the matrices should be viewed as linear operators on the space of (continuous) functions on the unit interval. For example, $\mathbf{V}_n^{-1}\,\mathbf{x}$ "is" the function $\frac{1}{v(x)}\,x$ where v(x) is the diagonal element of $\mathbf{V}_n$, for $x \in [0, 1]$. So the variance matrix may be viewed as the operator which maps the function $f \in C[0, 1]$ to $V(f) \in C[0, 1]$, where $V[f](x) = f(x)/v(x)$. Similarly, the kernel matrix is the operator which maps the function f to

$$K[f](x) = \int_0^1 K(x-y)f(y)\,dy.$$

Second, notice that minimizing

$$(\mathbf{y} - \mathbf{X}\beta)'\,\mathbf{K}_n'\,\mathbf{V}_n^{-1}\,\mathbf{K}_n\,(\mathbf{y} - \mathbf{X}\beta) \;=\; (\mathbf{K}_n\,\mathbf{y} - \mathbf{K}_n\,\mathbf{X}\beta)'\,\mathbf{V}_n^{-1}\,(\mathbf{K}_n\,\mathbf{y} - \mathbf{K}_n\,\mathbf{X}\beta) \qquad [9.2.3]$$

is the same as minimizing

$$(\mathbf{y}_s - \mathbf{X}\beta)'\,\mathbf{V}^{-1}(\mathbf{y}_s - \mathbf{X}\beta) \qquad [9.2.4]$$

(assuming $\mathbf{K}_n\mathbf{X}=\mathbf{X}$), where $\mathbf{y}_s=\mathbf{K}_n\mathbf{y}$ (i.e. a "smoothed" version of the observations y). Furthermore, minimizing

$$(\mathbf{y} - \mathbf{X}\beta)'\,\mathbf{V}_n^{-1}\,\mathbf{K}_n\,(\mathbf{y} - \mathbf{X}\beta) \qquad [9.2.5]$$

is the same as minimizing

$$(\mathbf{y}-\mathbf{X}\beta)'\,\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta)\;+(\mathbf{y}_*-\mathbf{X}\beta)'\,\mathbf{V}^{-1}(\mathbf{y}_*-\mathbf{X}\beta)\,. \qquad [9.2.6]$$

For a proof of this, observe that [9.2.5] may be expanded as

$$\mathbf{y}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}\;-\mathbf{y}'\mathbf{V}^{-1}\mathbf{K}\mathbf{X}\beta-\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}+\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{K}\mathbf{X}\beta=$$

$$\mathbf{y}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}\;-\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}-\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}+\;\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta \qquad [9.2.7]$$

using the facts that (1) $\mathbf{KX}=\mathbf{X}$, and (2) $\mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\beta=(\mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\beta)'$ since its a scalar. [9.2.6] may be expanded as

$$(\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}-\mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\beta-\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}+\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta)+$$
$$(\mathbf{y}'\mathbf{K}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}-\mathbf{y}'\mathbf{K}'\mathbf{V}^{-1}\mathbf{K}\mathbf{X}\beta-\beta'\mathbf{X}'\mathbf{K}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}+\beta'\mathbf{X}'\mathbf{K}'\mathbf{V}^{-1}\mathbf{K}\mathbf{X}\beta)$$

$$=(\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}-2\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}+\;\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta)+$$
$$(\mathbf{y}'\mathbf{K}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}-2\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}+\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta) \qquad [9.2.8]$$

using the facts that (1) $\mathbf{KX}=\mathbf{X}$, (2) $\mathbf{X}'\mathbf{K}'=(\mathbf{KX})'$, and (3) $\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}=(\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y})'$ since its a scalar.

Now take the derivatives of [9.2.7] and [9.2.8] with respect to β. The derivative of [9.2.7] is $-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}-\mathbf{X}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}+\;2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta$, which, set to 0 and solved for β yields

$$\widehat{\beta}=1/2\;(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{K})\mathbf{y}. \qquad [9.2.9]$$

The derivative of [9.2.8] is $-2\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}+2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}-2\mathbf{X}'\mathbf{V}^{-1}\mathbf{K}\mathbf{y}+2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}$, which, set to 0 and solved for $\boldsymbol{\beta}$ also yields [9.2.9].

What this suggests is that at least so far as multiple linear regression is concerned, an IRWLS procedure may be applied to smoothed data. It will also be seen that if W is not symmetric, IRWLS using the "variance operator" W may not yield a consistent estimate of the solution to [9.2.1], and an alternative procedure is described in section 7. Assuming a symmetric kernel operator K and using the "inverse variance operator" $KV^{-1}K$, if the smoothing would "reproduce" the columns of the X matrix *and* the weighted columns of the X matrix (i.e. the columns of $V^{-1}X$), then the variance of parametric estimates will not be damaged (see corollary 9.4.4). This is assuming the model is correct. What would be the effect of smoothing if the model was misspecified? Is it really possible to smooth without damaging the model? We will attempt to provide an answer in the sequel.

9.3 A Misspecified Spectral Model

We are interested in the situation of having observed a series X(t)+N(t), where X(t) and N(t) are uncorrelated Gaussian stationary processes, but wanting to estimate the spectrum of X(t) (see, e.g. Cameron and Turner (1987)). Assume that the model $\{f_\theta\}_{\theta \in \Theta}$ contains f_X, the spectrum of X(t). Also assume that the noise series N(t) has a spectrum f_N which is very "sharp" compared to a "smooth" f_X. Such a situation might arise in practice if the observed series has not had a seasonal trend properly removed prior to the spectral analysis. The spectrum of the observed series is $f_X + f_N$, and according to Taniguchi (1979), the $\hat{\theta}$ from ML is actually estimating θ_0 which minimizes

$$D(f_\theta, f_X + f_N) = \int_{-\pi}^{\pi} \log(f_\theta(\lambda)) + \frac{f_X(\lambda) + f_N(\lambda)}{f_\theta(\lambda)}\, d\lambda. \qquad [9.3.1]$$

Unfortunately the variance of $\hat{\theta}$ is also adversely affected. Taniguchi (1979) gives an expression for the variance in the misspecified model case, the analog of which in terms of the more familiar (to most applied statisticians) multiple linear regression model would be

$$\text{Var } \hat{\theta} = [(\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}] \mathbf{V} [\mathbf{W} \mathbf{X} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}] \qquad [9.3.2]$$

where $\mathbf{V}$ is a diagonal matrix with $g^2(\lambda_j)$ on the diagonal and $g(\lambda)$ is the true spectrum (again, this corresponds to theorem 7.1.2). $\mathbf{W}$ is supposed to be the inverse variance matrix, but it has $1/f^2_{\theta_0}(\lambda_j)$ on its diagonal, where $f_{\theta_0}(\lambda)$ is *from the misspecified model*, plus some terms from the second derivative of the likelihood function and model (e.g. see [5.1.2], [7.1.1], [7.1.2]). Thus if the model is misspecified, *both* the variance and bias of the parametric estimates are affected.

In terms of our definitions in chapter 6, the "intuitive" scenario just described will be expressed as follows (which is assumed to hold in the sequel).

Assumption 9.3.1 The observed series is univariate Gaussian and has a spectrum of bounded variation.

It is not even necessary to assume the model to be fit satisfies the BCC, it just needs to be continuous. However, given a particular QL operator, it will need to be checked that the link condition holds. This will be done by verifying the following

Assumption 9.3.2 $W^*[\ .\ ]$ has range satisfying the BCC.

and then appealing to the results of chapter 8.

9.4 Applications of the New QL Theory

So far, we have not said anything regarding if and how our new definition extends the old in a meaningful way. We know from chapter 6 that the function

$$D_1(f(\lambda), g(\lambda)) = \int_{-\pi}^{\pi} \log f(\lambda) + \frac{g(\lambda)}{f(\lambda)}\, d\lambda \qquad [9.4.1]$$

is a QL function and

$$\frac{\partial}{\partial[f(\lambda)]} D_1(f(\lambda), g(\lambda)) \quad = \frac{1}{f^2(\lambda)} \quad [f(\lambda) - g(\lambda)].$$

So the inverse variance operator for D_1 applied to an arbitrary function $h(\lambda)$ would be

$$V^{-1}(f(\lambda))[h(\lambda)] = \frac{1}{f^2(\lambda)} \quad h(\lambda),$$

i.e. multiplication by the inverse variance function for independent, exponential random variables. More generally, if we take any function $L(y, \mu)$ defined on a subset of $\mathbf{R} \times \mathbf{R}$ which satisfies

$$\frac{\partial L(y, \mu)}{\partial \mu} = \frac{(y - \mu)}{V(\mu)}$$

where $V(\mu)$ is a positive function defined on $\mathbf{R}$, and define $D(f, g) = \int_{\Lambda} L(f(\lambda), g(\lambda))\, d\lambda$, the resulting "distance" will be a QL function. This corresponds to the "independent observations" case, as we essentially have replaced the sum of log likelihoods with an integral. The derivative is

$$I_{\Lambda}(\lambda) \frac{(f(\lambda) - g(\lambda))}{V(f(\lambda), \lambda)}$$

where $I_\Lambda(\lambda)$ is the indicator function, and the variance operator is multiplication by $I_\Lambda(\lambda)/V(f(\lambda), \lambda)$. Note that for the link condition to hold for this QL operator, a condition is needed such as the spectrum satisfying the BCC, which is the usual condition in most of the current literature. This is because the link condition is verified by invoking theorem 8.2.1.

A "distance" which does not simply involve integrating an "L" function is

$$D_2(f(\lambda), g(\lambda)) = \int_{-\pi}^{\pi} \log(K[f(\lambda)]) + \frac{K[g(x)]}{K[f(\lambda)]}\, d\lambda \qquad [9.4.2]$$

where $K[f]=\int k(\lambda, x)\, f(x)\, dx$, which is the same as $D_1(K[f], K[g])$. Notice that the definition *requires the concept of a "function space" for the kernel operator to make sense.* Some requirements also need to be made on K, namely (1) $K[f]>0$ for f in the model space, and (2) K^* satisfies assumption 9.3.2. Assuming these conditions, we have the following theorem concerning the derivative of D_2:

Theorem 9.4.1

$$\frac{\partial}{\partial[f(\lambda)]} D_2(f(\lambda), g(\lambda))\Big|_{f_0} = K^* M_{1/(K[f_0])^2} K[f_0 - g]$$

Remark: As mentioned in chapter 6, the function on the right is a representation of a linear functional on L^2.

proof

Use the chain rule to determine $\frac{\partial}{\partial[f(\lambda)]} D_2(f(\lambda), g(\lambda))\Big|_{f_0(\lambda)} [h(\lambda)]$ for $h(\lambda) \in C$. Because for fixed g, $D_2(f(\lambda), g(\lambda))=D_1(K[f(\lambda)], K[g(\lambda)])$, we have that

$$\frac{\partial}{\partial[f(\lambda)]} D_2(f(\lambda), g(\lambda))\Big|_{f_0(\lambda)} [h(\lambda)] =$$

$$\frac{\partial}{\partial[f(\lambda)]} D_1(f(\lambda), K[g(\lambda)])\Big|_{K[f_0(\lambda)]} \circ \left(\frac{\partial}{\partial[f(\lambda)]} \{K[f]\}\Big|_{f_0} [h(\lambda)] \right)$$

$$= \int_{-\pi}^{\pi} M_{1/(K[f_0])^2} K[f_0 - g] \cdot K[h] \; d\lambda \quad \text{(this is true by the requirements on K)}$$

$$= \int K^* M_{1/(K[f_0])^2} K[f_0 - g] \cdot h \; d\lambda$$

Hence D_2 is seen to be of the correct form with the claimed inverse variance operator. □

What type of kernel operator K should be used? The intuition from section 2 suggests that the ideal K should reproduce the columns of the X matrix, or analogously the partial derivatives of the model $\partial f_{\theta_0}/\partial\theta$ as closely as possible. However, it should be kept in mind that *section 2 was geared towards showing IRWLS on a smoothed dataset is equivalent to IRWLS with a nondiagonal inverse variance matrix for a special case.* Nothing was said regarding the general nonlinear model problems of variance and bias, which we examine now.

First, consider the bias problem. Suppose we have a kernel K so that $K[f_X]=f_X$. If $f_X \in \{f_\theta\}$ but we observe the process X(t)+N(t), then by theorem 7.1.1, $\hat{\theta}_n$ minimizing $D_1(f_\theta, I_n)$ is estimating θ_0 minimizing $D_1(f_\theta, f_X+f_N)$. By another application of theorem 7.1.1, $\hat{\theta}_n$ minimizing $D_2(f_\theta, I_n)$ is estimating θ_0 minimizing $D_2(f_\theta, f_X+f_N)$. But $D_2(f_\theta, f_X+f_N) = D_1(\int k(\lambda, x)f_\theta(x)\,dx, \int k(\lambda, x)(f_X(x)+f_N(x))\,dx$. Notice that $\int k(\lambda, x)(f_X(x)+f_N(x))\,dx = f_X+\int k(\lambda, x)f_N(x)\,dx$, that is, the spectrum we want to estimate plus a smoothed version of the noise spectrum. This smoothed noise spectrum may have greatly reduced capability for biasing our parametric estimates, assuming it is "sharp" in comparison to f_X and assuming K

is properly chosen. How to choose K is a major topic in the sequel. However, it should be mentioned at this point that one condition which we will require of K, which is not assumed symmetric, is that K^* satisfy Assumption 9.3.2 so that the variance operator will satisfy this condition. This is not restrictive in practice, as a sufficient condition is simply requiring the function $k(\lambda, x)$ to be twice continuously differentiable with respect to x. Alternatively, given an arbitrary operator K, one can always use KK_1, where K_1 is a small bandwidth symmetric smoother operator to force $(KK_1)^*=K_1^*K^*=K_1K^*$ to satisfy the condition.

What happens if the "wrong" kernel K is used, i.e if $K[f_X] \neq f_X$? If K still filters the noise, nothing so far as bias is concerned, as is shown in the following theorem. However, if K does not filter the noise, the estimate will still be biased.

Theorem 9.4.2

If (1) $K[f_N]=0$, (2) $f_X= f_{\theta_0}$, and (3) $K[f_{\theta_1}] \neq K[f_{\theta_2}]$ if $\theta_1 \neq \theta_2$, then $D_2(f_\theta, f_X+f_N)$ has a unique minimum at $\theta=\theta_0$.

proof

$D_2(f_\theta, f_X+f_N)=D_2(f_\theta, f_X)=D_1(K[f_\theta], K[f_X])$. But D_1 has a unique minimum at θ_0 by Theorem 1 of Taniguchi (1979) (since $K[f_{\theta_0}]=K[f_X]$). □

Actually, it is easily seen that any operator of the form $K_1(f)K(f)$, where $K(f_{\theta_0})[f_N]=0$, will result in unbiased estimates assuming the QL equations [7.1.4] have a *unique solution.* This is because $K_1(f)K(f)\ [f_{\theta_0} - (f_{\theta_0}+f_N)]=0$, so θ_0 will always be *a solution.*

The simple QL function of theorem 9.4.1 motivates the rationale behind why smoothing prior to IRWLS might reduce bias, but it should be remembered that we *don't* need the actual QL function in order to use its variance operator in an IRWLS procedure. As a

matter of fact, the QL function corresponding to a given variance operator *doesn't have to exist.* For example, consider the variance operator

$$W(f(\lambda))\,[h(\lambda)] = \frac{1}{f^2(\lambda)} \int K(\lambda\text{-}x)\, h(x)\, dx \qquad [9.4.3]$$

and a simplified "matrix version"

$$\begin{bmatrix} 1/x_1^2 & 0 \\ 0 & 1/x_2^2 \end{bmatrix} \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} (1/x_1^2)(k_{11}x_1+k_{12}x_2) \\ (1/x_2^2)(k_{21}x_1+k_{22}x_2) \end{bmatrix}$$

Suppose a two variable, real valued function $F(x_1, x_2)$ exists so that $\partial F/\partial x_1 = \frac{1}{x_1^2}(k_{11}x_1+k_{12}x_2)$ and $\partial F/\partial x_2 = \frac{1}{x_2^2}(k_{21}x_1+k_{22}x_2)$. Integrating $\partial F/\partial x_1$ with respect to x_1 gives $k_{11}\log(x_1) - \frac{k_{12}x_2}{x_1^2} + \phi(x_2)$ for some function ϕ of x_2. Now take the derivative of this function with respect to x_2 and set it equal to $\partial F/\partial x_2$ to solve for $\phi'(x_2)$. We have $k_{12}/x_1^2 + \phi'(x_2) = \frac{1}{x_2^2}(k_{21}x_1+k_{22}x_2)$. But this is impossible (without assuming conditions on the K matrix, such as $k_{12}=k_{21}=0$), as it implies ϕ' is also function of x_1.

Another "impossible" but useful inverse variance operator is a slight variant of D_2's operator, namely $K^*M_{1/f^2}K$. Recall this operator was discussed in section 2, and of course if K reproduces the true spectrum, D_2's inverse variance operator will reduce to this. The sequel will discuss "non self adjoint" and other inverse variance operators for which there may not exist corresponding QL functions, showing the consistency of an IRWLS method for determining the zeros of the QL equations.

How does the asymptotic variance matrix for the parametric estimates act if the model is correct and under model misspecification? Recall from chapter 7 that the asymptotic variance matrix for $\sqrt{n}(\hat{\theta} - \theta_0)$ is $[M_W(\theta_0)]^{-1}\, Q_W(\theta_0)\, [M_W^*(\theta_0)]^{-1}$ where

$$M_W(\theta) \equiv [\Phi_\psi(f_\theta, f) - \Phi_\psi(f_\theta, f_\theta)] \odot \frac{\partial^2 f_\theta}{\partial\theta' \partial\theta} +$$

$$\frac{\partial f_\theta}{\partial\theta} \odot \left[\left[W(f_\theta) + \frac{\partial\Phi(f_\theta, f_\theta)}{\partial x} - \frac{\partial\Phi(f_\theta, f)}{\partial x} \right] * \frac{\partial f_\theta}{\partial\theta'} \right]$$

and

$$Q_W(\theta) \equiv \frac{\partial f_\theta}{\partial\theta} \odot \left[W(f_\theta)\ V\ W^*(f_\theta) * \frac{\partial f_\theta}{\partial\theta'} \right]$$

(i.e. [7.1.1] and [7.1.2]). Here is a theorem formalizing the intuitive discussion in section 9.2.

Theorem 9.4.4

Assuming

$$\text{i) } W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial\theta} = V^{-1} * \frac{\partial f_{\theta_0}}{\partial\theta} \qquad [9.4.4]$$

$$\text{ii) } \frac{\partial\Phi(f_{\theta_0}, f_{\theta_0})}{\partial x} - \frac{\partial\Phi(f_{\theta_0}, f)}{\partial x} = \mathbf{0},\ \Phi_\psi(f_{\theta_0}, f) - \Phi_\psi(f_{\theta_0}, f_{\theta_0})] = \mathbf{0} \qquad [9.4.5]$$

then $[M_W(\theta_0)]^{-1}\, Q_W(\theta_0)\, [M_W^*(\theta_0)]^{-1}$ collapses to M_V^{-1} (definition 7.1.3 (c)), and the asymptotic variance matrix is optimal.

proof

To see this, notice that

$$\frac{\partial f_\theta}{\partial \theta} \odot \left[W(f_\theta) * \frac{\partial f_\theta}{\partial \theta'} \right] = \left[W^*(f_\theta) * \frac{\partial f_\theta}{\partial \theta} \right] \odot \frac{\partial f_\theta}{\partial \theta'}$$

so that (i) and (ii) together imply $M_V = M_W$, and

$$\frac{\partial f_\theta}{\partial \theta} \odot \left[W(f_\theta) V\, W^*(f_\theta) * \frac{\partial f_\theta}{\partial \theta'} \right] = \left[W^*(f_\theta) * \frac{\partial f_\theta}{\partial \theta} \right] \odot \left[V\, W^*(f_\theta) * \frac{\partial f_\theta}{\partial \theta'} \right]$$

so that (i) also implies $Q_V = Q_W$. □

Observe that for the weighted least squares QL distance, the first part of [9.4.5] is automatically satisfied as $\Phi(x(\lambda), y(\lambda))$ is constant as a function of $x(\lambda)$. For the variance operators we have discussed ($M_{1/f^2} K$ and $K^* M_{1/f^2} K$), corollary 9.4.4 below will show that if the conjugate kernel reproduces the weighted partial derivatives $(\partial f_{\theta_0}(\lambda)/\partial \theta_i)(1/f^2_{\theta_0}(\lambda))$ and the model is correct, then the variance of our estimates is not damaged. But first, it is shown in the following theorem 9.4.5 and corollaries 9.4.1 to 9.4.3 that these QL operators, which are all specific cases of the operator $K_2 M_{1/f^2} K_1$, actually *are* QL operators. They also show that the "second derivative" parts in the definition of M_W (expression [7.1.1]) are zero if the kernels are chosen to filter the noise, resulting in a simpler asymptotic variance matrix for parametric estimates just as if the model were correct. Corollary 9.4.4 looks at conditions on the kernel to achieve an optimal variance matrix if the model is misspecified.

Theorem 9.4.5

Let U be an open subset of C[a, b], and suppose $f \in U$. Let $g(x, \lambda)$: $U_1 \times [a, b] \rightarrow \mathbf{R}$ (where $U_1 \subset \mathbf{R}$) have a continuous partial derivative with respect to x, and have a domain such that $g(u(\lambda), \lambda)$ is well defined for all $u \in U$. For the mapping $U \rightarrow B(L^2[a, b])$ defined by

$F(u)=M_{g(u,\lambda)}[\ .\ ]$ (where $M_h[\ .\ ]$ is the multiplication operator on L^2 defined by $M_h[x(\lambda)] = h(\lambda)\ x(\lambda)$), we have $DF|_f[h] = M_{\frac{\partial g(f,\lambda)}{\partial x}\ h}[\ .\ ]$ (i.e. multiplication by the function $\frac{\partial g(f(\lambda),\ \lambda)}{\partial x}$ $h(\lambda)$).

proof

Recall the proof of proposition 8.5.1. As shown in theorem 1.5, p. 28 of Conway (1985), the mapping $L^\infty[a, b]\to B(L^2)$ defined by $\phi\to M_\phi$ is linear and norm preserving. For completeness, we give a short proof here.

For $f \in L^2$, $||\phi f||_2 \le ||\phi||_\infty ||f||_2$, so $||M_\phi|| \le ||\phi||_\infty$. On the other hand, given $\epsilon>0$, there exists a set A_ϵ so that $|\phi| \ge ||\phi||_\infty - \epsilon$ on A_ϵ. Then $(1/\sqrt{m(A_\epsilon)})\ I_{A_\epsilon}$ is in the unit ball of L^2 with $\left\| (1/\sqrt{m(A_\epsilon)})\ I_{A_\epsilon}\ \phi \right\| \ge ||\phi||_\infty - \epsilon$. As ϵ is arbitrary, we conclude $||M_\phi|| \ge ||\phi||_\infty$ and we are done.

Returning to the proof of theorem 9.4.5, regard the mapping F as the composition $F_1 \circ F_2$, where F_2:C[a, b]$\to$C[a, b] is defined by $u\to g(u, \lambda)$, and F_1:C[a, b]$\to B(L^2)$ is defined by $\phi\to M_\phi$. Now use proposition 6.2.1 and the chain rule. □

Corollary 9.4.1 (to theorem 9.4.5)

The mapping C[a, b]$\to B(L^2[a, b])$ defined by $F(f)=M_{g(K[f],\lambda)}[\ .\ ]$ has derivative, evaluated at f and applied to h, $DF|_f[h] = M_{\frac{\partial g(K[f],\lambda)}{\partial x}\ K[h]}[\ .\ ]$ (i.e. multiplication by $\frac{\partial g(K[f(\lambda)],\ \lambda)}{\partial x}$ $K[h(\lambda)]$).

proof

We may write $F=F_1 \circ F_2 \circ F_3$, where F_3:C[a, b]$\to$C[a, b] is defined by $F_3[f] = K[f]$, and F_1, F_2 are as defined in theorem 9.4.5. By the chain rule, $DF|_f[h] = D[F_1 \circ F_2]\big|_{F_3[f]} \circ D[F_3]\big|_f[h]$. Of

course, F_3 being linear is its own derivative. So $DF|_f[h] = D[F_1 \circ F_2]|_{K[f]} \circ K[h]$, which equals $M_{\frac{\partial g(K[f],\lambda)}{\partial x}\, K[h]}[\,\cdot\,]$. □

Corollary 9.4.2 (to theorem 9.4.5)

The mapping $C[a, b] \to B(L^2[a, b])$ defined by

$$F(u)=K_2 M_{g(K[u],\lambda)} K_1$$

where K, K_1 and K_2 are any bounded operators on $L^2[a, b]$, has a derivative (evaluated at f and applied to h)

$$DF|_f[h] = K_2 M_{\frac{\partial g(K[f],\lambda)}{\partial x}\, K[h]} K_1.$$

proof

This mapping is a composition $F_2 \circ F_1$ where $F_1: C[a, b] \to B(L^2)$ is defined by $F_1(\phi) = M_{g(K[\phi],\lambda)}$, and $F_2: B(L^2) \to B(L^2)$ is defined by $F_2[M]=K_2 M K_1$. Note that F_2 is linear and use the chain rule together with corollary 9.4.1. □

We are now able to show the important corollary 9.4.3, which says that if the noise can be filtered by K_1, a QL operator of the form $F(u)=K_2 M_{g(K[u],\lambda)} K_1$ will *yield unbiased estimates of the parameter, and the second derivative parts of the matrix [7.1.1] will vanish.*

Corollary 9.4.3 (to theorem 9.4.5)

(i) The mapping $C[a, b] \to B(L^2[a, b])$ defined by $F(u)=K_2 M_{g(K[u],\lambda)} K_1$ is a QL operator.

Suppose $\{f_\theta\}$ is a model, and $g \in L^2$ is such that $g=f_{\theta_0}+f_N$. If $K_1[f_N]=0$, then:

(ii) θ_0 is a solution of the QL equations $\frac{\partial f_\theta}{\partial \theta} \odot K_2 M_{g(K[f_\theta],\lambda)} K_1[f_\theta - g]$.

(iii) $\Phi_\psi(f_{\theta_0}, f_X+f_N) - \Phi_\psi(f_{\theta_0}, f_{\theta_0})]=0$.

(iv) $\dfrac{\partial\Phi(f_{\theta_0}, f_{\theta_0})}{\partial x} - \dfrac{\partial\Phi(f_{\theta_0}, f_X+f_N)}{\partial x} = \mathbf{0}.$

proof

(i) follows from Corollary 9.4.2. (ii) is obvious, as $K_1[f_{\theta_0} - g]=0$. (iii) is obvious, so we prove (iv). Recall $\Phi(x(\lambda), y(\lambda))$ was defined as $W(x(\lambda))[y(\lambda)]$, which, as a function of $x(\lambda)$ can be viewed as the composition $F_1 \circ F_2$, where F_2: $C \to B(L^2)$ is defined by $F_2[x(\lambda)] = W(x(\lambda))[\ .\ ]$ and F_1: $B(L^2) \to L^2$ is defined by $F_1[W] = W[y(\lambda)]$. So again by the chain rule, the partial derivative of Φ with respect to x evaluated at f and applied to y is $DF_1|_{F_2[f]} \circ DF_2|_f[h]$. Regardless of what f is, $DF_1|_{F_2[f]}$ is evaluation at y. By corollary 9.4.2,

$$DF_2|_f[h] = K_2 M_{\frac{\partial g(K[f],\lambda)}{\partial x}\, K[h]}\, K_1.$$

Because $K_1[f_N]=0$, if this operator is evaluated at $y=f_X+f_N$, it is the same as being evaluated at f_X. If $f_{\theta_0}=f_X$, (iii) follows immediately. □

The (nice) conditions (iii) and (iv) only hold for *fixed* operators K_1 and K_2. In chapter 10 it will be seen that if K_1 or K_2 is a function of the model, (iv) *most likely will not hold.*

Now that we know about bias, we can begin to examine conditions on K required for the operators $M_{1/f^2}K$ or $K^*M_{1/f^2}K$ (1) to yield optimal variance for the estimates if the model is correct, and (2) to yield optimal variance for the estimates if the model is incorrect due to contamination, while at the same time giving an unbiased estimate.

Corollary 9.4.4 (to theorem 9.4.4)

1) Suppose the inverse variance operator is $M_{1/f^2}K$, and the true spectral density $g=f_{\theta_0}+f_N$

for a univariate process with fourth cumulant spectrum identically zero satisfies:

(i) $K[f_N]=0$.

(ii) $$K^*\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial\theta_i}\frac{1}{f^2_{\theta_0}(\lambda)}\right]=\frac{\partial f_{\theta_0}(\lambda)}{\partial\theta_i}\frac{1}{f^2_{\theta_0}(\lambda)}\,,\ i=1\ldots p.$$

(iii) $$\int K^*\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial\theta_i}\frac{1}{f^2_{\theta_0}(\lambda)}\right](g^2(\lambda)-f^2_{\theta_0}(\lambda))\,K^*\left[\frac{\partial f_{\theta_0}(\lambda)}{\partial\theta_j}\frac{1}{f^2_{\theta_0}(\lambda)}\right]d\lambda=0\,,\ i=1\ldots p.$$

Then under the conditions of corollary 7.1.1 (i.e. uniqueness of θ_0), the asymptotic variance matrix for $\hat{\theta}$ solving the QL equations is optimal, *just as if* f_{θ_0} *were the true spectral density.* Note that the condition (i) guarantees θ_0 is a solution to the QL equations [7.1.4].

2) If the inverse variance operator is $K^* M_{1/f^2} K$ and the condition

(iv) $$K*\frac{\partial f_{\theta_0}}{\partial\theta}=\frac{\partial f_{\theta_0}}{\partial\theta}$$

also holds, then the conclusion is the same.

Remarks: (1) It will become apparent from the discussion in chapter 10 that if the "second derivative parts" of [7.1.1] are 0, the conclusion of corollary 9.4.4 remains the same for "model dependent" operators K(f)[.]. (2) Condition (iii) is "given" and cannot be affected by QL operator choice. As such, it will not play a role in the sequel. It is assumed that (iii) is "close" to 0, that is, the model is not misspecified *too* badly.

proof

Because of corollary 9.4.3, it only needs to be shown that

$$\frac{\partial f_{\theta_0}}{\partial\theta}\odot\left[W(f_{\theta_0})*\frac{\partial f_{\theta_0}}{\partial\theta'}\right]=\frac{\partial f_{\theta_0}}{\partial\theta}\odot\left[M_{1/f^2_{\theta_0}}*\frac{\partial f_{\theta_0}}{\partial\theta'}\right] \qquad [9.4.6]$$

and

$$\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[W(f_{\theta_0}) V\, W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] = \frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[M_{1/f^2_{\theta_0}} * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \qquad [9.4.7]$$

where $W(f)=M_{1/f^2}\, K$ or $K^*\, M_{1/f^2}\, K$. Notice that

$$\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[M_{1/f^2_{\theta_0}}\, K * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] = \left[K^*\, M_{1/f^2_{\theta_0}} * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \frac{\partial f_{\theta_0}}{\partial \theta'} \qquad [9.4.8]$$

and

$$\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[K^* M_{1/f^2_{\theta_0}}\, K * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] = \left[K * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \left[M_{1/f^2_{\theta_0}}\, K * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \qquad [9.4.9]$$

so that (i) and [9.4.8] imply [9.4.6] if $W(f)=M_{1/f^2}\, K$, and (iv) and [9.4.9] imply [9.4.6] if $W(f)=K^*\, M_{1/f^2}\, K$. Also observe

$$\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[M_{1/f^2_{\theta_0}}\, KV\, K^* M_{1/f^2_{\theta_0}} * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] =$$

$$\left[K^* M_{1/f^2_{\theta_0}} * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \left[M_{g^2}\, K^* M_{1/f^2_{\theta_0}} * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \qquad [9.4.10]$$

and

$$\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[K^*\, M_{1/f^2_{\theta_0}}\, K\, V\, K^* M_{1/f^2_{\theta_0}}\, K * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] =$$

$$\left[K^* M_{1/f^2_{\theta_0}}\, K * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \odot \left[M_{g^2}\, K^* M_{1/f^2_{\theta_0}}\, K * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \qquad [9.4.11]$$

Now [9.4.10] and (iii) imply [9.4.7] if $W(f)=M_{1/f^2}\,K$, and [9.4.11] and (iii) imply [9.4.7] if $W(f)=K^*\,M_{1/f^2}\,K$. □

Corollary 9.4.4 says that if the model is not correct, if the "remainder" term $g^2(\lambda)-f^2_{\theta_0}(\lambda)$ is "rough" relative to the weighted partial derivatives of the model, and if the conjugate kernel passes the weighted partial derivatives, then optimal parameter estimates with respect to variance are obtained even though the model is an underfit (assuming inverse variance operator $M_{1/f^2}\,K$ is used). Note that corollary 9.4.4 also says that if the model *is correct* and the conjugate kernel passes the weighted partial derivatives, then the variance is still optimal even though the "wrong" QL operator was used (e.g. take $f_{\theta_0}(\lambda)=g(\lambda)$ in the theorem). This may be compared to the analogous situation in a multiple regression problem. If $y=X\beta+\epsilon$, where $X=X_1|X_2$, $\beta=[\beta_1|\beta_2]'$ and the columns of the matrices X_1, X_2 are orthogonal, then if the model $X_1\beta_1$ is fit to the data, the resulting parameter estimates are unbiased and optimal. Either way, the optimality is given by an orthogonality condition on the model. In the spectral estimation case, one might think of the reduced model as consisting of the "principal components" of the true density, the actual density being too complicated to model (e.g. suppose it is an arbitrary positive L^2 function). Of course, in practice the conditions will be only approximately met. If, on the other hand, the model is really correct, corollary 9.4.4 says that assuming the kernel is appropriately chosen (so that its conjugate passes the weighted partial derivatives), the asymptotic variance for parametric estimates will not be damaged. Notice that the variance operator $M_{1/f^2}\,K$ has fewer requirements on the kernel in order to leave the variance undamaged, so that one might be inclined to use it for that reason. This will be discussed further in chapter 10.

In summary, for a non self-adjoint "variance operator" W, the behavior of W on the model $\{f_\theta\}_{\theta\in\Theta}$ controls the asymptotic bias of the resulting estimates, while the behavior of

W^* on the "derivative model space" $\left\{\frac{\partial f_{\theta_0}}{\partial \theta}\right\}$ controls the asymptotic variance. If W is of the form $K^* M_{1/f^2} K$ or $M_{1/f^2} K$ and K is not self adjoint, then the behavior of K on the model controls the bias and the behavior of K^* on the weighted derivatives of the model, i.e. $(\partial f_{\theta_0}(y)/\partial\theta)(1/f^2_{\theta_0}(y))$, controls the variance. In practice, it probably will not be possible to *exactly* choose W or equivalently K for our specific examples so that bias is eliminated and variance undamaged for a correct model. Hence our strategy will be to *view the bias conditions as most important, forcing W to act correctly on the model while simultaneously attempting to minimize damage to the variance assuming the model was correct.* One reason for this is because the asymptotic variance is $[(X'WX)^{-1} X' W] V [W' X(X'W'X)^{-1}]$, which decreases as n increases, regardless of the bias. However, the bias *will never change.* This means that as more observations are taken, *the bias will always overtake the variance if in fact the process is contaminated.*

9.5 An Example

As an illustration of the ideas in this chapter, the technique of "naive" IRWLS on a smoothed periodogram is illustrated on simulated data. 300 observations X(t) were observed from a MA(8) process, to which was added contamination in the form of "sine waves" of the form $N(t)=\sum_{\lambda} A_\lambda \sin(\lambda t)$ with frequencies concentrated in two bands. Figure 1 shows the true spectrum of the uncontaminated process X(t), together with the contaminated periodogram (of the process X(t)+N(t)). A regression spline was used as the model (which closely fit the true spectrum), and the resulting estimate from 5 iterations using the raw periodogram (of X(t)+N(t)) in an IRWLS procedure as described in Chiu (1988) is shown in figure 2. Suppose the analyst suspected the series was contaminated, but had no idea where the contamination might be (of course, if it was precisely known what frequency bands were contaminated, they

could be eliminated prior to the analysis as was pointed out by Chiu (1988)). We therefore smooth the periodogram (of X(t)+N(t)) with a tapered smoother involving 13 periodogram ordinates at each frequency. The result of IRWLS using the same spline model on the smoothed data is shown in figure 3. Figure 4 shows what would have happened had we used the IRWLS procedure on the smoothed periodogram where the periodogram was calculated from the *uncontaminated* series X(t). Figure 5 shows the result of doing IRWLS on the unsmoothed periodogram from the uncontaminated series X(t).

Notice that there is not much difference in figures 4 and 5, indicating we did not damage the variance much with the smoothing. However, the visually significant differences in figures 2 and 3 show how the smoothing drastically reduced the bias in our estimate.

While an example such as the above does not "prove" anything, it illustrates the ease and robustness of a naive IRWLS procedure. The kernel was rather arbitrarily chosen, and no attempt was made to verify that it satisfies the conditions described at the end of section 9.4. Yet the basic idea of "bias reduction" obviously worked (at least in this case, which is typical in my experience). How might a more refined theory be formulated? This is the main topic of chapter 10, but in the next section we digress to give an informal discussion of "non Gaussian" variance operators and their relationship to the topic at hand. The chapter will conclude by answering a question raised in section 2: how is IRWLS done if the inverse variance matrix (operator) is not symmetric?

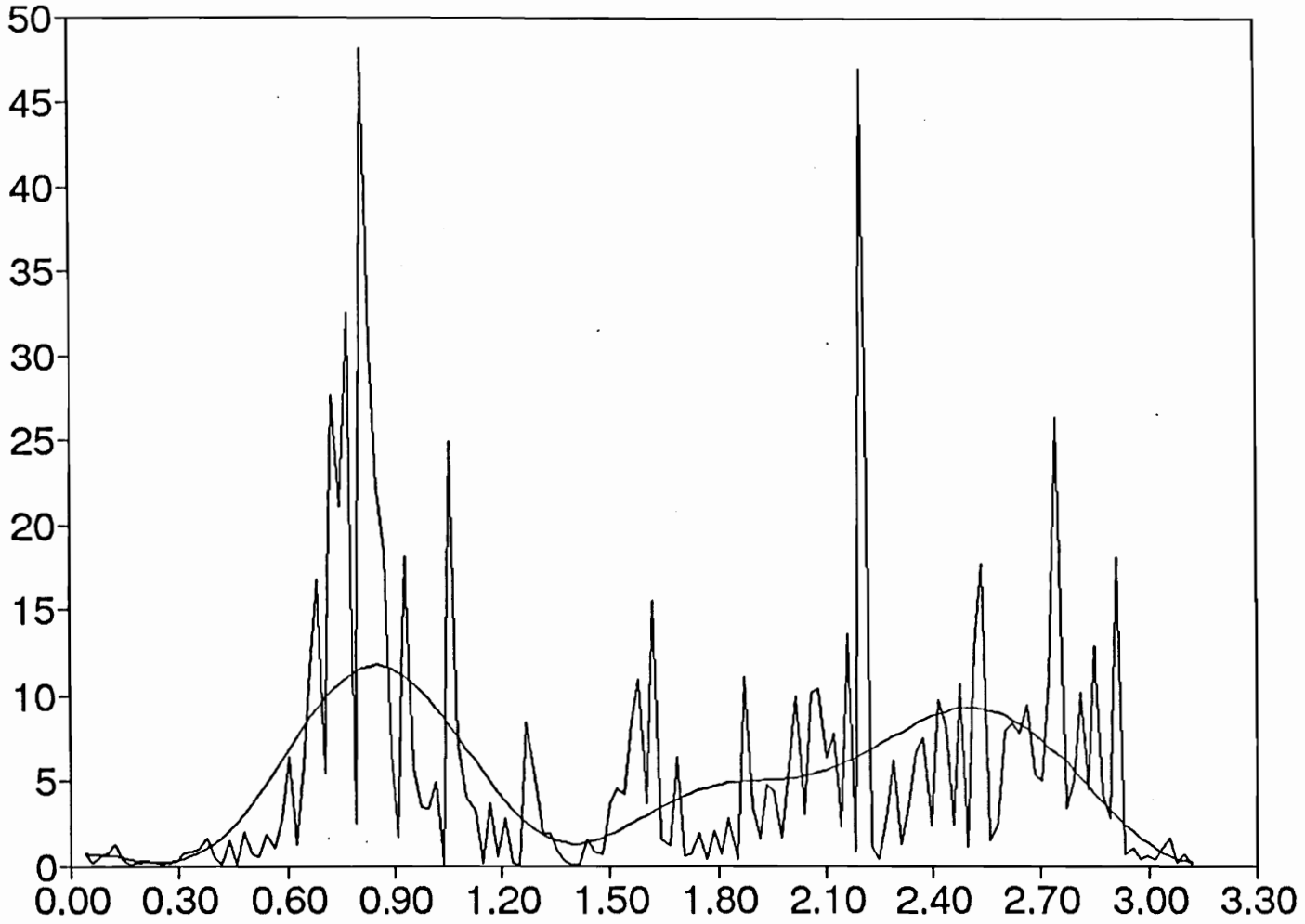

Figure 1. Contaminated Periodogram and Uncontaminated Spectrum.

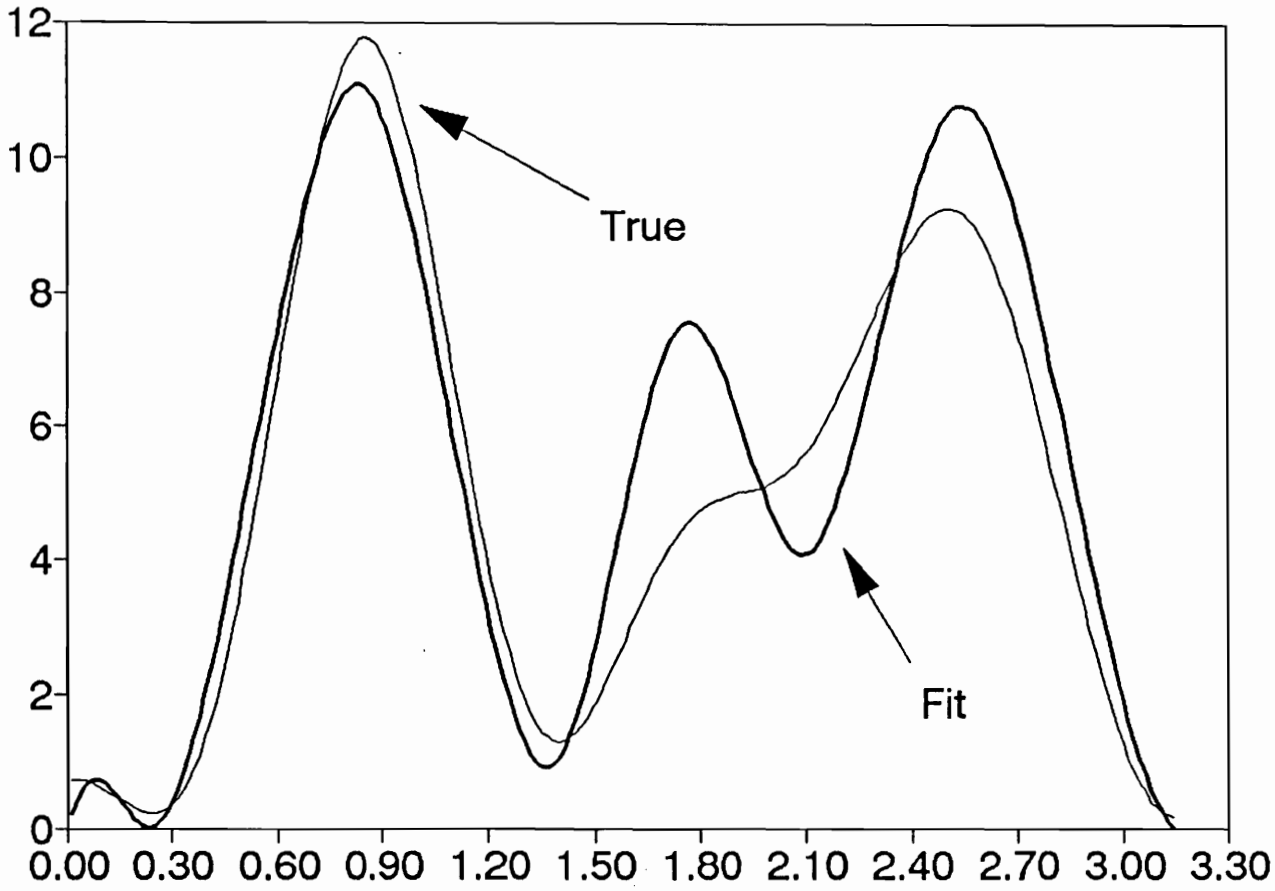

Figure 2. IRWLS on Raw, Contaminated Periodogram.

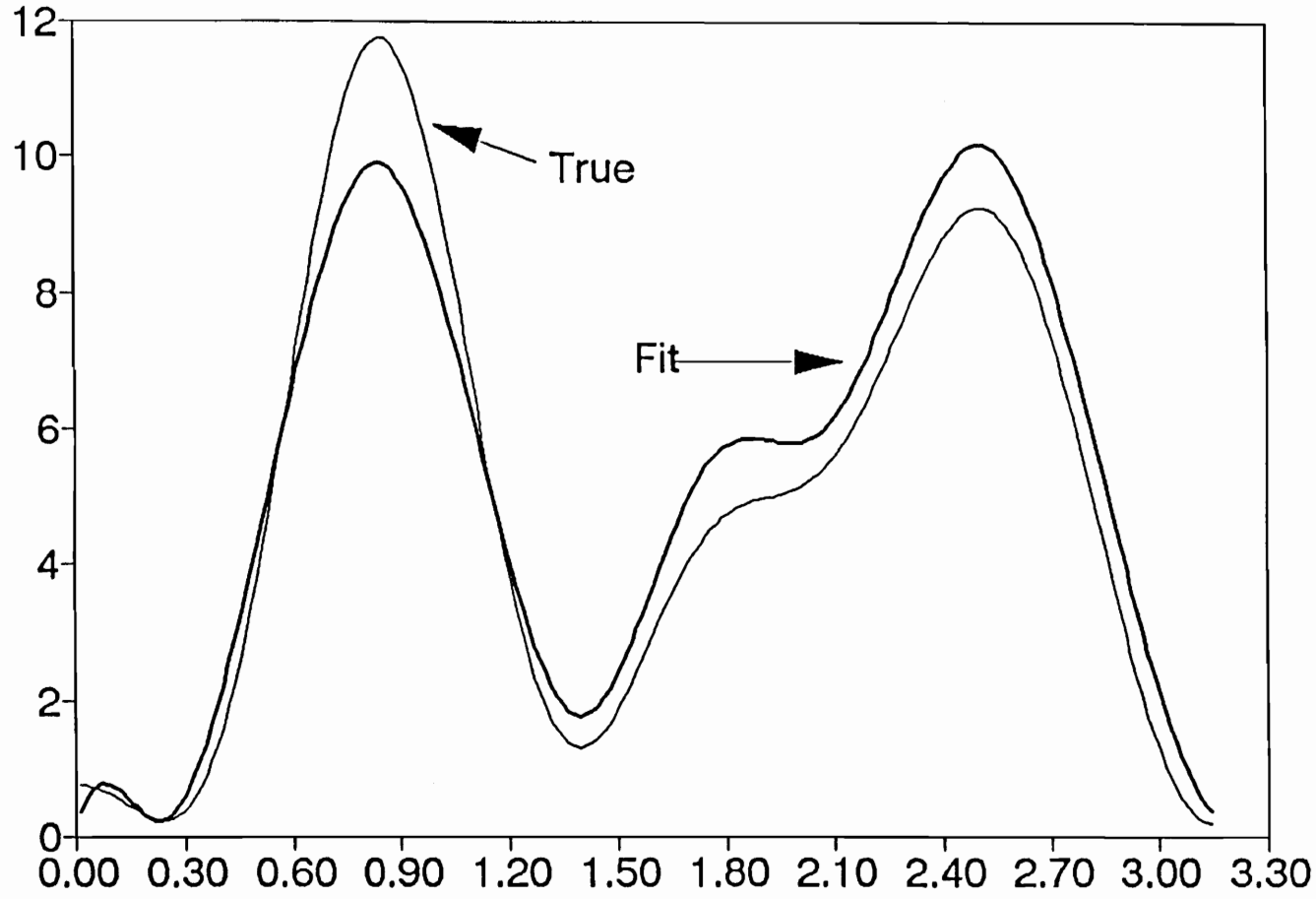

Figure 3. IRWLS on Smoothed Contaminated Periodogram.

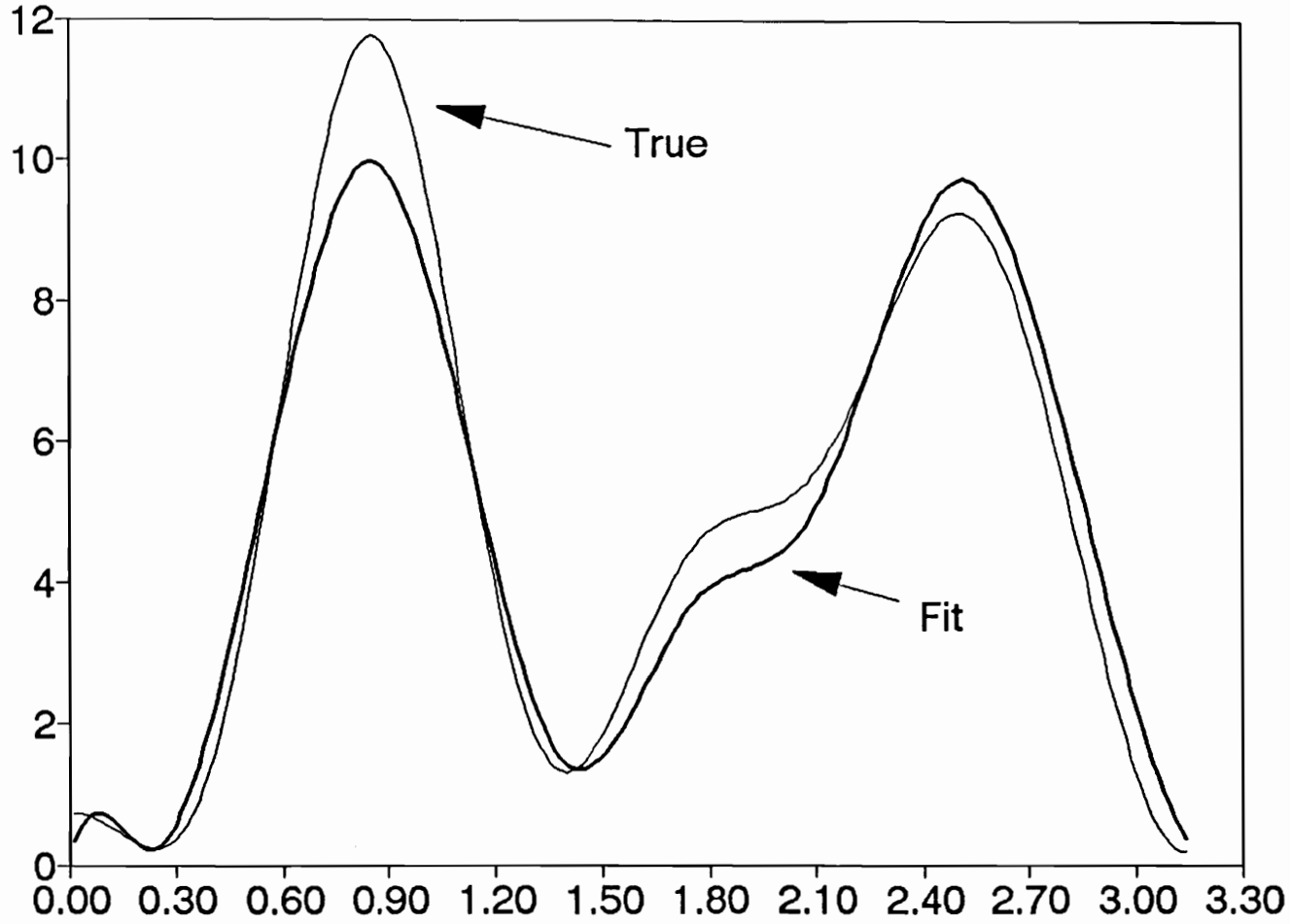

Figure 4. IRWLS on Smoothed, Uncontaminated Periodogram.

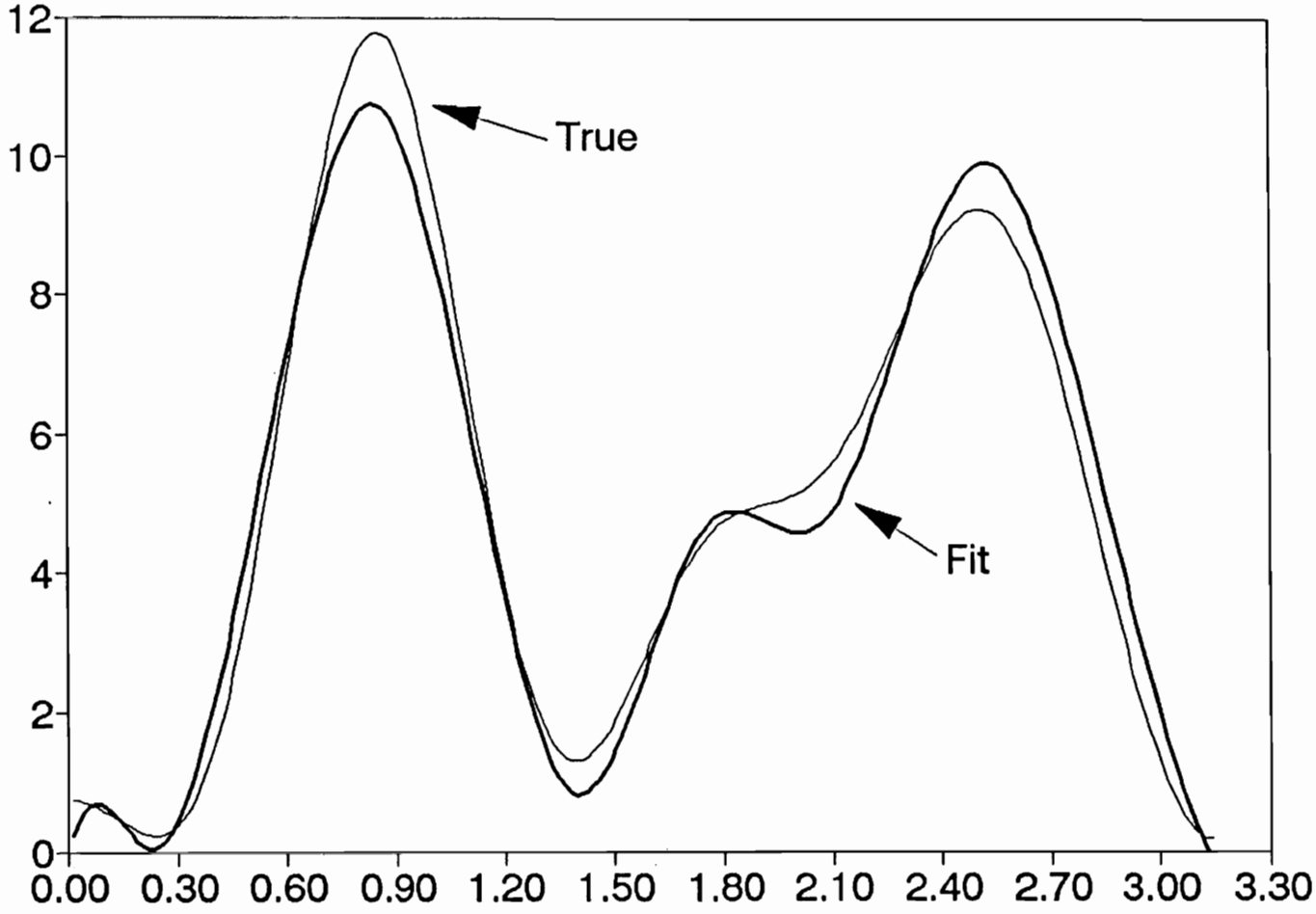

Figure 5. IRWLS on Raw, Uncontaminated Periodogram.

9.6 Non-Gaussian Variance Operators

QL functions having "variance operators" W(f)[.] which are not simply multiplication operators might be thought of as *non-Gaussian QL functions*, since a non Gaussian process has a variance operator which is not a multiplication operator (e.g. Brillinger's (1981) theorem cited earlier). As has been previously discussed, parametric estimates obtained by minimizing $\int \log(f_\theta(\lambda)) + \frac{I(\lambda)}{f_\theta(\lambda)}\, d\lambda$ *do not* have the same variance for Gaussian and non Gaussian series, the variance matrix for the latter depending upon the fourth cumulant spectrum.

We also know from Brockwell and Davis (1987), theorem 10.3.2 p. 337 that $\mathrm{cov}(I_n(\lambda_1), I_n(\lambda_2)) = O(1/n)$. So while the covariance of the periodogram evaluated at any two frequencies goes to zero, for non-Gaussian processes it doesn't go fast enough to avoid influencing the asymptotic variance of parametric estimates.

An operator acts on a function space (e.g. L^2 functions). If the "observation" in an experiment is *in a function space* rather than being a vector in $\mathbf{R}^n$, then a "variance operator for the experiment" would *characterize the behavior of linear functionals* applied to the observation function and *depends upon the design of the experiment.* For example, the variance matrix for the vector $\mathbf{y} \in \mathbf{R}^n$ has the property that if $\mathbf{a}$ and $\mathbf{b}$ are any two fixed vectors in $\mathbf{R}^n$, the covariance of the real valued random variables $\mathbf{a}'\mathbf{y}$, $\mathbf{b}'\mathbf{y}$ is $\mathbf{a}'\mathbf{V}\mathbf{b}$ where $\mathbf{V}$ is the variance of $\mathbf{y}$.

Brillinger's (1981) theorem cited above essentially says the same thing for *the periodogram viewed as a function* (a step function, for example). Here, the variance operator applied to an arbitrary function $h(\lambda)$ is

$$V[h(\lambda)] = 2\pi f^2(\lambda)h(\lambda) + 2\pi\kappa \int_{-\pi}^{\pi} f_4(\lambda,-\lambda,\mu)\, h(\mu)\, d\mu$$

and if $A_1(\lambda)$ and $A_2(\lambda)$ are any two functions (viewed as functionals, as L^2 is self dual), the asymptotic covariance of the real valued random variables $A_1 \bullet I_n$ and $A_2 \bullet I_n$ is given by $A_1 \bullet V[A_2]$. The inner product $\bullet$ is the usual one defined by $f \bullet g = \int f(\lambda)g(\lambda)d\lambda$.

Inherent in saying that a "step functionized" version of an observation vector has a non multiplication variance operator is the following. (1) The experiment is done in stages in a response surface setting where the regressors are chosen over some design space (e.g. the interval $[0, \pi]$), (2) The observations from stage n+1 have less covariance than the observations from stage n, and (3) The observations from stage n+1 are retaken from all the locations in the design space where they were taken at stage n, plus some additional locations. However for the analysis at stage n+1, only the observations taken at stage n+1 are used, *all others from previous stages are ignored.* The scenario described appears unrealistic for practical "response surface" type experiments (mainly due to the last condition), but when you calculate the periodogram at the Fourier frequencies for a time series of increasing length, it occurs automatically in a natural way.

9.7 IRWLS and Non Symmetric Variance Operators

In section 2 it was observed that minimizing $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})' \mathbf{V}_n^{-1} \mathbf{K}_n (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ ([9.2.5]) is the same as minimizing $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\mathbf{y}_s-\mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1}(\mathbf{y}_s-\mathbf{X}\boldsymbol{\beta})$, assuming $K[\mathbf{x}_i] = \mathbf{x}_i$ for each column of the X matrix. Since the $\widehat{\boldsymbol{\beta}}$ minimizing this later equation is a mixture of the $\widehat{\boldsymbol{\beta}}$'s minimizing $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ and $(\mathbf{y}_s-\mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1}(\mathbf{y}_s-\mathbf{X}\boldsymbol{\beta})$, one might suspect that the solution would *still* be biased, since the solution to $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ is biased. One also might question whether minimizing an expression such as [9.2.5] consistently estimates the appropriate θ (which is θ_0 such that $f_X = f_{\theta_0}$ assuming (1) $f_X \in \{f_\theta\}$, and (2) $K[f_N]=0$, but not necessarily assuming K reproduces the model) solving the QL equations

$$M_{1/f^2}\, K[f_\theta - (f_X + f_N)] \odot \frac{\partial f_\theta}{\partial \theta} = 0 \,. \qquad [9.7.1]$$

As will be seen, *it does not* and a modification to the IRWLS procedure will need to be made.

For motivation, again consider the linear model (i.e GLIM with the identity link). We want to solve the QL equations $\mathbf{X'M(y - X\beta) = 0}$, where first suppose **M** is a symmetric "inverse variance matrix" which depends on the unknown β vector. Because for a *symmetric* matrix **M**, $\mathbf{x'Mx}$ has derivative $2\mathbf{Mx}$ with respect to the **x** vector, it follows that the derivative of $\mathbf{(y - X\beta)'M(y - X\beta)}$ for a fixed **M** is $2\mathbf{X'M(y - X\beta)}$. But if **M** is not symmetric, *the derivative of* $\mathbf{x'Mx}$ *is not* $\mathbf{2Mx}$, *and the derivative of* $\mathbf{(y - X\beta)'M(y - X\beta)}$ *is not* $\mathbf{2X'M(y - X\beta)}$. Thus, if **M** is not symmetric, it does not make sense to minimize the quadratic form $\mathbf{(y - X\beta)'M(y - X\beta)}$ which may not even be positive! Of course, the same basic argument holds if **M** is an operator rather than a matrix.

We will make the assumptions that (1) **M** is a square matrix, not assumed symmetric, and (2) although **M** needn't be invertible, $\mathbf{X'M'X}$ is invertible. As an IRWLS procedure, we will

1) Determine a starting value $\widehat{\beta}$.

2) Calculate $\mathbf{M}_{\widehat{\beta}}$

3) Minimize $\| P_X[\mathbf{M}_{\widehat{\beta}}(\mathbf{X\beta - y})] \|^2$ with respect to β, where P_X is the projection of $\mathbf{R}^n$ onto the span of the columns of the **X** matrix.

This determines a new $\widehat{\beta}$, which is used to repeat (2) and (3). To see what is going on, let us fix a $\widehat{\beta}$ and let $\mathbf{M=M}_{\widehat{\beta}}$. Then in step 3 we are supposed to minimize $\mathbf{(y - X\beta)'}$ $\mathbf{[M'X(X'X)^{-1}X'M](y - X\beta)}$. Since the matrix in the center is now symmetric, the solution is $\mathbf{2X'M'X(X'X)^{-1}X'My = 2X'M'X(X'X)^{-1}X'MX\beta}$ or $\mathbf{[X'M'X(X'X)^{-1}]X'My = [X'M'X(X'X)^{-1}]}$

$\mathbf{X'MX}\beta$. By our assumption, the matrix in brackets on each side of the equation is invertible. Hence it may be removed to reveal the original QL equations (but with a *fixed* **M**), which are the normal equations for "weighted" regression with a nonsymmetric weight matrix.

In the general case where the model may be nonlinear, step 3 of the procedure is modified as follows:

For each θ_0, define the **model space at θ_0** to be the subspace of L^2 generated by the span of $\left\{ \frac{\partial f_{\theta_0}}{\partial \theta_i} \right\}$. Replace the "projection onto the span of the columns of the X matrix" with the "projection onto the model space at $\hat{\theta}$", and in general let P_θ denote the operator which is the projection onto the model space at θ. The following theorem shows consistency of this IRWLS procedure:

Theorem 9.7.1

If

i) W(f)[.] is a QL operator.

ii) The true limiting function is f_X+f_N, where $f_X = f_{\theta_0}$ for some θ_0.

iii) $W(f_\theta)[f_N]=0$ for all θ.

iv) $\theta_1 \neq \theta_2 \Rightarrow W(f_\theta)[f_{\theta_1}] \neq W(f_\theta)[f_{\theta_2}]$ for all θ.

v) $\frac{\partial f_\theta}{\partial \theta} \odot \left[W(f_\theta) * \frac{\partial f_\theta}{\partial \theta'} \right]$ is a nonsingular matrix for all θ.

Then

1) $W^*(f_\theta)\ P_\theta\ W(f_\theta)$ satisfies (c) in the definition of QL distance (with "ψ"=θ).

2) There exists a neighborhood U of θ_0 so that if $\theta_1, \theta_2 \in U$, $\theta_2 \neq \theta_0$, then

$$||P_{\theta_1}[W(f_{\theta_1})[f_X+f_N-f_{\theta_2}]\,]\,||_2^2 > 0 \qquad [9.7.2]$$

so that

$$||P_{\theta_1}[W(f_{\theta_1})[f_X+f_N-f_{\theta}]\,]\,||_2^2 \qquad [9.7.3]$$

is uniquely minimized (in U) at $\theta=\theta_0$.

3) If $\widehat{\theta}_1$ is a consistent estimator of θ_0, and $\widehat{\theta}$ is obtained to minimize $||P_{\widehat{\theta}_1}[W(f_{\widehat{\theta}_1})[I_n-f_{\theta}]\,]\,||_2^2$ $(=[I_n-f_{\theta}]\bullet W^*(f_{\widehat{\theta}_1})\,P_{\widehat{\theta}_1}\,W(f_{\widehat{\theta}_1})[I_n-f_{\theta}])$, then $\widehat{\theta}$ has the same asymptotic variance matrix as the $\widehat{\theta}_2$ solving the QL equations

$$W(f_{\theta})[I_n-f_{\theta}]\;\odot\frac{\partial f_{\theta}}{\partial\theta}=0.$$

proof

1) W(f) satisfies the definition, and P_{θ} is continuous in the operator norm.

2) By conditions (ii) and (iii), $||P_{\theta_1}[W(f_{\theta_1})[f_X+f_N-f_{\theta_0}]\,]\,||_2^2=0$.

By condition (v) and the continuity of the function

$$(\theta_1,\,\theta_2)\rightarrow\frac{\partial f_{\theta_1}}{\partial\theta}\odot\left[W(f_{\theta_1})*\frac{\partial f_{\theta_2}}{\partial\theta'}\right] \qquad [9.7.4]$$

there exists a neighborhood U of θ_0 so that if $(\theta_1,\,\theta_2)\in U\times U$, the matrix in [9.7.4] is invertible. $P_{\theta_1}[W(f_{\theta_1})[f_X+f_N-f_{\theta_2}]]=P_{\theta_1}[W(f_{\theta_1})[f_{\theta_0}-f_{\theta_2}]]$. Now

$$f_{\theta_2}=f_{\theta_0}+\frac{\partial f_{\theta_*}}{\partial\theta'}\,(\theta_2-\theta_0)$$

for some θ_* between θ_0 and θ_2. So this may be rewritten as

$$P_{\theta_1}\left[W(f_{\theta_1})\left[\frac{\partial f_{\theta_*}}{\partial \theta'}(\theta_2-\theta_0)\right]\right]$$

which is then nonzero (because the matrix in [9.7.4] being invertible means that the inner product of $W(f_{\theta_1})\left[\frac{\partial f_{\theta_*}}{\partial \theta'}(\theta_2-\theta_0)\right]$ with the partials can't all be 0, hence the projection can't be 0). So we conclude $||P_{\theta_1}[W(f_{\theta_1})[f_X+f_N-f_{\theta_2}]||>0$.

3) The self adjoint variance operator for the IRWLS procedure is $W^*(f_\theta)\ P_\theta\ W(f_\theta)$. Because of (2), the IRWLS corollary from chapter 7 (corollary 7.1.2) may be applied. We need to show that the asymptotic variance of the parameter estimates using this operator (at θ_0) is $M_W^{-1}Q_W\left[M_W^{-1}\right]'$, where

$$Q_W=\frac{\partial f_{\theta_0}}{\partial \theta}\odot\left[W(f_{\theta_0})VW^*(f_{\theta_0})*\frac{\partial f_{\theta_0}}{\partial \theta'}\right],\ \text{and}\ M_W=\frac{\partial f_{\theta_0}}{\partial \theta}\odot\left[W(f_{\theta_0})*\frac{\partial f_{\theta_0}}{\partial \theta'}\right]. \qquad [9.7.5]$$

From the IRWLS corollary, the asymptotic variance for the IRWLS procedure is $M^{-1}_{W^*PW}Q_{W^*PW}M^{-1}_{W^*PW}$, where

$$Q_{W^*PW}=\frac{\partial f_{\theta_0}}{\partial \theta}\odot\left[W^*(f_{\theta_0})P_{\theta_0}W(f_{\theta_0})VW^*(f_{\theta_0})P_{\theta_0}W(f_{\theta_0})*\frac{\partial f_{\theta_0}}{\partial \theta'}\right],$$

$$M_{W^*PW}=\frac{\partial f_{\theta_0}}{\partial \theta}\odot\left[W^*(f_{\theta_0})P_{\theta_0}W(f_{\theta_0})*\frac{\partial f_{\theta_0}}{\partial \theta'}\right]. \qquad [9.7.6]$$

So it suffices to show that $M_W^{-1}Q_WM_W^{-1}$ and $M^{-1}_{W^*PW}Q_{W^*PW}M^{-1}_{W^*PW}$ are really the same matrix. As motivation, consider the "matrix" case. $M^{-1}_{W^*PW}Q_{W^*PW}M^{-1}_{W^*PW}$ would be written as

$$[X'W^*X(X'X)^{-1}X'WX]^{-1}$$

$$\{X'W^*X(X'X)^{-1}X'WVW^*X(X'X)^{-1}X'WX\}[X'W^*X(X'X)^{-1}X'W\ X]^{-1} \qquad [9.7.7]$$

$$=[X'WX]^{-1}(X'X)[X'W^*X]^{-1}$$

$$\{[X'W^*X](X'X)^{-1}X'WVW^*X(X'X)^{-1}[X'WX]\}\ [X'WX]^{-1}(X'X)[X'W^*X]^{-1} \qquad [9.7.8]$$

$$=[X'WX]^{-1}[X'WVW^*X][X'W^*X]^{-1} \qquad [9.7.9]$$

which is $M_W^{-1}Q_W\left[M_W^{-1}\right]'$. The steps involved in the matrix case may essentially be copied to obtain the result.

First, observe that for $y \in L^2$, $P_\theta[y]$ may be represented as

$$P_\theta[y]= \frac{\partial f_{\theta_0}}{\partial \theta'}\left(\frac{\partial f_{\theta_0}}{\partial \theta'} \odot \frac{\partial f_{\theta_0}}{\partial \theta} \right)^{-1}\left(\frac{\partial f_{\theta_0}}{\partial \theta} \odot y \right). \qquad [9.7.10]$$

Interpret this as forming the column vector $\frac{\partial f_{\theta_0}}{\partial \theta} \odot y$, applying the matrix

$$\left(\frac{\partial f_{\theta_0}}{\partial \theta'} \odot \frac{\partial f_{\theta_0}}{\partial \theta} \right)^{-1}$$

to $\frac{\partial f_{\theta_0}}{\partial \theta} \odot y$, and then applying the row vector of functions to the column vector so that the result is a linear combination of functions (see, e.g., theorem 2.5.1 p. 59 of Brockwell and Davis for an analogous proof of why this is true). Thus, Q_{W^*PW} may be rewritten as

$$\left(\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \right)\left(\frac{\partial f_{\theta_0}}{\partial \theta'} \odot \frac{\partial f_{\theta_0}}{\partial \theta} \right)^{-1}\left(\frac{\partial f_{\theta_0}}{\partial \theta} \odot \right.$$

$$\left. \left[W(f_{\theta_0})VW^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \right)\left(\frac{\partial f_{\theta_0}}{\partial \theta'} \odot \frac{\partial f_{\theta_0}}{\partial \theta} \right)^{-1}\left(\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[W(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \right)$$

and M_{W^*PW} may be rewritten as

$$\left(\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[W^*(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \right) \left(\frac{\partial f_{\theta_0}}{\partial \theta'} \odot \frac{\partial f_{\theta_0}}{\partial \theta} \right)^{-1} \left(\frac{\partial f_{\theta_0}}{\partial \theta} \odot \left[W(f_{\theta_0}) * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \right).$$

As in the "matrix" case, pieces of the above will cancel to yield that

$$M^{-1}_{W^*PW} Q_{W^*PW} M^{-1}_{W^*PW} = M^{-1}_{W} Q_{W} \left[M^{-1}_{W} \right]'. \qquad \square$$

9.8 Conclusions

In summary, this chapter has introduced some examples of "non Gaussian" (and possibly non self adjoint) variance operators. We have concentrated on operators involving a kernel function, and reduced the problem of QL function choice to kernel choice. There are different ways to construct QL operators using kernels, and the chapter has examined some of the problems involved in these different operators. In general, for a non self adjoint inverse variance operator W, the behavior of W determines the bias properties and the behavior of W^* (on the partial derivatives) determines the variance properties of estimates obtained by solving the QL equations. This basic concept carries over to the *non self adjoint kernels* which may compose the kernel type operators presented here. The operator $K^* M_{1/f^2} K$ is a natural extension of the operator in a smoothed Taniguchi (1979) distance D(K[f], K[g]). If the kernel is chosen to filter the noise and pass the model, its use is equivalent to an IRWLS procedure on a smoothed periodogram with a diagonal variance matrix (multiplication variance operator), but this does not address the problem of non optimal variance if the model was in fact correct. In such a situation, corollary 9.4.4 suggests there are fewer restrictions on the kernel for a non self adjoint operator such as $M_{1/f^2} K$ to not damage variance. As IRWLS is a practical

approach to solving the QL equations, modifications had to be made in the IRWLS procedure so that non symmetric variance matrices could be dealt with. Theorem 9.7.1 gives a specific method for carrying out IRWLS in such cases.

There is a fundamental problem with the kernel type operators in this chapter. They are all of "fixed kernel" type; once the kernels are initially determined, they do not change as a function of the model. The theorems of section 4 (especially corollary 9.4.4) suggest that in order to obtain good variances with the "wrong" QL operator, certain conditions should be satisfied with *regard to the partial derivatives or weighted partial derivatives of the model*, rather than the model itself. But according to definition 6.3.4, a QL operator can depend only upon the means function (i.e. the model), not the partial derivatives. This motivates a redefinition of what is meant by "QL operator" and "model". In the next chapter, we will explore how the definitions in chapter 6 might have been made in order to allow "model driven" QL operators which could adjust themselves to the model derivatives in the iterative process. Chapter 10 will be the culmination of the theory as presented in this dissertation, giving a strong argument for the usefulness of the function space approach to generalized nonlinear models.

Chapter X
Unified Theory of Quasi Likelihood

10.1 Kernel Choice/Model Driven Kernels

In the preceding chapter, we have given some examples of new QL operators involving "kernel" operators, but little indication as to how one should choose the kernel. The purpose of this section is to examine some of the problems involved in kernel selection, and the discussion will be limited to inverse variance operators of the forms $K^*M_{1/f^2}K$, $K_1M_{1/f^2}K_2$, or $M_{1/f^2}K$ to be used on univariate Gaussian series. Of course, the basic ideas easily extend to GLIM response surfaces.

We begin by considering the inverse variance operator $M_{1/f^2}K$, because corollary 9.4.4 has only one "optimal variance" requirement, that the conjugate kernel must pass the weighted partial derivatives. Given a specific model $\{f_\theta\}_{\theta \in \Theta}$, how might K be chosen to minimize damage to the variance?

One method of achieving minimal damage to the weighted partials upon application of K might be to use a symmetric kernel operator with the same bandwidth for all λ (so $K=K^*$). For example, we take $k(\lambda)$ to be a function ($[-1, 1] \rightarrow \mathbf{R}$) symmetric about 0. For a given bandwidth b, define $k_b(x, \lambda)=k((\lambda - x)/b)$ if $|\lambda - x| \leq b$ and 0 otherwise. Then define the operator K_b by

$$K_b[h] = \frac{\int_{-\pi}^{\pi} k_b(x, \lambda)h(\lambda)\, d\lambda}{b\int_{-\pi}^{\pi} k(x)\, dx}$$

where $h(\lambda) \in L^2[-\pi, \pi]$ is extended to **R** as a periodic function when taking the integral. The problem of choosing the kernel would then be reduced to choosing the bandwidth b. One method of doing this might be by deciding upon some "upper bound", say ϵ, so that

$$\left\| K_b\left[\frac{1}{f_\theta^2}\frac{\partial f_\theta}{\partial \theta_i}\right] - \left[\frac{1}{f_\theta^2}\frac{\partial f_\theta}{\partial \theta_i}\right] \right\|_\infty \leq \epsilon \qquad [10.1.1]$$

for i=1..p and K_b the kernel with bandwidth b, and then making b as large as possible so that [10.1.1] still holds. This would be done *at each stage of the IRWLS procedure*, so that different iterations might have different bandwidths. The procedure would be as follows:

Step 1 Obtain $\hat{\theta}_{old}$, a robust estimate of θ_0 (say, by arbitrarily deciding a bandwidth), and set $\epsilon>0$.

Step 2 Choose $\hat{b}$ as large as possible so that

$$\left\| K_b\left[\frac{1}{f_{\hat{\theta}}^2}\frac{\partial f_{\hat{\theta}}}{\partial \theta_i}\right] - \left[\frac{1}{f_{\hat{\theta}}^2}\frac{\partial f_{\hat{\theta}}}{\partial \theta_i}\right] \right\|_\infty \leq \epsilon.$$

Step 3 Minimize $||P_{\hat{\theta}}\,[M_{1/f_{\hat{\theta}}^2}\,K_{\hat{b}}\,[I_n - f_\theta]]||^2$, where P_θ is the projection onto the span of the partials at θ to obtain a new estimate $\hat{\theta}_{new}$.

Step 4 Set $\hat{\theta}_{old} = \hat{\theta}_{new}$ and go to step 2.

Implicitly assumed in such an approach is that one does not have any preconceived notions about where the noise is, and wants to smooth as much as possible without damaging variance. This is similar to the example in section 5, and would probably act in a similar

fashion to reduce bias in practice. However, there are two basic problems with the procedure.

First, it is apparent that while a kernel such as described might *reduce* bias, *a positive kernel cannot eliminate a positive noise spectrum.* For example, if $K(\lambda, x)>0$ and $f_N>0$, then $\int K(\lambda, x)\, f_N(x)\, dx>0$. Hence, such a kernel cannot theoretically satisfy the conditions needed for unbiased estimation, and this problem will be addressed shortly. Of course, if one does not view f_N as being a true spectrum, but just the difference between g, the true spectrum, and f_{θ_0}, the "best" estimate of g from the model, then f_N may well be negative. Second, the QL operator as described *does not depend solely on* f_θ as required by definition 6.3.4, because the bandwidth (and hence the kernel operator) depends also upon the partial derivatives. In other words, the QL operator depends upon *the specific parametrization of the model*, a condition we will define as being **model driven**. Other examples of model driven kernels will appear in this section, but their theoretical justification will not be discussed until later. Let us begin by first considering the bias problem.

In the example from section 5, the kernel was chosen to have a bandwidth not too extreme, so that the model would pass through "relatively undamaged". This was before we were aware of corollary 9.4.4, so variance did not play a role in kernel choice. However, if the kernel for the operator $K^*M_{1/f^2}K$ passes the columns of the X matrix (or for each θ we create a new kernel which passes the partial derivatives of the model in the nonlinear model case, another example of a model driven kernel), then the procedure simplifies to IRWLS (or in the nonlinear case, another similar procedure called Iterated, Reweighted, Resmoothed Least Squares to be described in section 10.2). Although not mentioned in section 9.5, a *variable bandwidth* kernel might just as well have been chosen to achieve a balance between the dual goals of noise suppression and model passage. If one knew frequency bands where contamination *might* be concentrated, one would apply a (hopefully model reproducing) kernel

only to those bands. The bandwidth might depend on the *extent of the contamination we expect*; if it is suspected that the noise spectrum is "light" (i.e. has little power), then a smaller bandwidth might be chosen than if it is suspected that the noise spectrum is "heavy", since then more smoothing would be needed to reduce its influence. What this suggests is that the kernel choice *carries information about the contamination we are trying to protect against.* The fact that an arbitrarily chosen kernel will most likely not exactly reproduce the model is not a serious problem, *if the kernel eliminates the noise, i.e* $K[f_N]=0$, since the estimate will still be unbiased. You just can't do an easy reweighted least squares procedure. The real problem is how to use the intuitive information contained in the choice of a kernel supplied by the user to create a *new* kernel which does truly filter the noise.

Let us call the "naive" kernel given by the analyst K_1 and the "constructed" kernel utilizing its information K_2. One possible solution to the construction of K_2 is to define a **noise space** N_1 by finding the singular value decomposition (SVD) of the (compact) operator K_1, which can be viewed as finding an orthogonal basis $\{V_i\}$ for L^2 and an orthogonal basis $\{U_i\}$ for the closure of the range of K_1. The action of K_1 on f can be determined as follows: Write $f=\sum \alpha_i V_i$. Then $Kf=\sum \sqrt{\lambda_i}\, \alpha_i U_i$, where λ_i are the eigenvalues of the (compact) self adjoint operator K^*K and V_i are the corresponding eigenvectors (See Conway (1985), theorem 2.7.6 p. 56). Although as we have stated before, K_1 cannot satisfy $K_1[f_N]=0$ if both K_1 and f_N are positive, we would expect K_1 to have been chosen so that $||K_1[f_N]||_2$ is much smaller than $||f_N||_2$. Put into other words, this is saying that *while f_N is not strictly in the null space of K_1, it should lie primarily in the spaces V_i which correspond to small eigenvalues.* So the "working" noise space N_1 is the kernel of K_1 plus eigenvectors of $K_1^*K_1$ corresponding to "small" (definition to be determined by the analyst) eigenvalues.

What use is made of the noise space depends upon one's objectives. If the inverse

variance operator is of the form $M_{1/f^2}K$, it might be desirable to choose K as the orthogonal projection onto the (orthogonal) complement of the noise space. As projections are self adjoint, $K^*=K$, and the extent to which the variance will be damaged depends upon the extent to which the noise spectrum is a component of the weighted partials. For example, if the noise space is perpendicular to the weighted partial derivative space, then there will be *no* loss in asymptotic variance. This situation also illustrates the usual "tradeoff" between variance and bias: the bigger the noise space, the more likely it is to impinge upon the weighted partial derivative space to damage the variance.

Another use of the noise space might be in the construction of $K^*M_{1/f^2}K$ type operators. If the objective is to pass the partial derivatives of the model (say, in order to apply an IRWLS type procedure on smoothed data, see theorem 10.2.3), then one might consider using a *non orthogonal projection* operator K. We attempt to divide L^2 into two not necessarily orthogonal (under the usual inner product $f \bullet g = \int fg$) subspaces M and N, corresponding to "model" and "noise" spaces, respectively, so that $L^2 = M \oplus N$. M here is not the same as the "model space" from chapter 6; we hope it is as large as possible, making the noise space as small as possible and resulting in less the damage to variance. The implication of L^2's decomposition is that each function f in L^2 can be written in a unique way as $f_M + f_N$ where $f_M \in M$ and $f_N \in N$, but it is not necessarily true that $f_M \bullet f_N = 0$. The desired "kernel operator" K_2 is simply the (non) orthogonal projection onto M, i.e. to find $K_2[f]$, first find the unique f_M and f_N so that $f = f_M + f_N$. Then $K_2[f] = f_M$. To define the model space M, first define M_1 to be the space spanned by the partial derivatives. Then $M \equiv M_1 \cap (M_1 \cap N)^{\perp} + (M_1 + N)^{\perp}$, the idea being to make M as large as possible and take out noise if $M_1 \cap N$ is not empty. Thus defined, $L^2 = M \oplus N$. In practice, the IRWLS procedure would look like the following.

Step 1 Obtain $\hat{\theta}_{old}$, a robust estimate of θ_0, and N, a subspace of L^2 where it is suspected the noise spectrum lies.

Step 2 Defining $M_{1\theta}=\text{span}\left\{ f_\theta, \frac{\partial f_\theta}{\partial \theta_i} \right\}$, $M_\theta = M_{1\theta} \cap (M_{1\theta} \cap N)^\perp + (M_{1\theta}+N)^\perp$, and K_θ to be the nonorthogonal projection onto M_θ, calculate $K_{\hat{\theta}}$. Note: N should be chosen so that $N \cap M_{1\theta}$ is empty for all θ if possible, or else the kernel will damage the partials.

Step 3 Minimize $[I_n - f_\theta] \bullet K^*_{\hat{\theta}} M_{1/f^2_{\hat{\theta}}} K_{\hat{\theta}} [I_n - f_\theta]$ to obtain a new estimate $\hat{\theta}_{new}$.

Step 4 Set $\hat{\theta}_{old} = \hat{\theta}_{new}$ and go to step 2.

Notice that since K_2 depends upon the partials at each θ, we have again described a *model driven kernel.*

As previously mentioned, there is a problem with variance which the above procedure has not addressed. By corollary 9.4.4, if the model was correct, a sufficient condition for the asymptotic variance for parametric estimates to be undamaged is that

$$K^* \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i} \frac{1}{f^2_{\theta_0}(\lambda)} \right] = \frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i} \frac{1}{f^2_{\theta_0}(\lambda)} \text{ and } K \left[\frac{\partial f_{\theta_0}(\lambda)}{\partial \theta_i} \right] = \frac{\partial f_\theta(\lambda)}{\partial \theta_i} \qquad [10.1.2]$$

but the procedure has only taken into account the second condition (note, however, that the second condition is sufficient for M_W to be correct). While the choice of K in this situation remains a topic for future research, we will now consider a second QL operator which may reduce damage to the asymptotic variance, assuming the model is correct. This operator will be called the **filter-reconstructor** (FR) operator, and is of the form $K_2 M_{1/f^2} K_1$. It is more difficult to work with in practice because it is not self adjoint, meaning (among other things)

we can't reduce its use to a simpler problem of a reweighted least squares procedure on a smoothed periodogram with a diagonal variance matrix.

As motivation for the FR operator, consider the two problems of variance and bias inherent in the contaminated series problem. Suppose K_1 is some operator chosen to solve the bias problem (without regard to the variance problem), called the **filter**. Recall from the form of the QL equations $W(f_\theta)[f_\theta - g] \odot \frac{\partial f_\theta}{\partial \theta} = 0$ that unbiasedness will be obtained if $W(f_\theta)[f_N]=0$, a condition which holds if $K_1[f_N]=0$. K_2 will be chosen, hopefully to help the variance problem, *after K_1 has been chosen, and assuming the model was in fact correct.* To do this, we need

$$\frac{\partial f_\theta}{\partial \theta} \odot \left[K_2 \, M_{1/f_\theta^2} \, K_1 * \frac{\partial f_\theta}{\partial \theta'} \right] = \frac{\partial f_\theta}{\partial \theta'} \odot \left[M_{1/f_\theta^2} * \frac{\partial f_{\theta_0}}{\partial \theta'} \right] \qquad [10.1.3]$$

and

$$\frac{\partial f_\theta}{\partial \theta} \odot \left[[K_2 \, M_{1/f_\theta^2} \, K_1] M_{f^2} \, [K_1^* \, M_{1/f_\theta^2} \, K_2^*] * \frac{\partial f_\theta}{\partial \theta'} \right] =$$

$$\frac{\partial f_\theta}{\partial \theta'} \odot \left[M_{1/f_\theta^2} * \frac{\partial f_{\theta_0}}{\partial \theta} \right] \qquad [10.1.4]$$

where f is the true spectrum. For motivational purposes, assume $f=f_{\theta_0}$. Take the viewpoint of minimizing damage to variance assuming the model is correct. If the model is incorrect, we will need a condition such as that in corollary 9.4.4 (iii).

To achieve [10.1.3], it suffices to have

$$K_1^* \left[\frac{1}{f_\theta^2} \, K_2^* \left[\frac{\partial f_\theta}{\partial \theta_i} \right] \right] = \frac{1}{f_\theta^2} \, \frac{\partial f_\theta}{\partial \theta_i} \qquad [10.1.5]$$

so *if possible*, K_2 should be chosen to achieve this result. Note that this motivates calling K_2 the **reconstructor**, as it attempts to reconstruct the weighted partial derivatives. But [10.1.5] is also sufficient to achieve [10.1.4], because the left side of [10.1.4] equals

$$\left[K_1^* M_{1/f_\theta^2} K_2^* * \frac{\partial f_\theta}{\partial \theta}\right] \odot \left[M_{f^2} K_1^* M_{1/f_\theta^2} K_2^* * \frac{\partial f_\theta}{\partial \theta}\right] \qquad [10.1.6]$$

which equals the right side of [10.1.4] due to [10.1.5].

It may not be possible to achieve [10.1.5], because for any operator K, cl(Range K^*) = $\ker(K)^{\perp}$ (Conway (1985), p. 36). Here, ker(K) is the "null space" of K, that is, $\{x \in L^2 | K[x]=0\}$, and for any set A, cl(A) is the topological closure of A. Also, for any subset N of L^2, $N^{\perp} = \{x \in L^2 | x \bullet n = 0 \ \forall \ n \in N\}$. However, the noise may not be exactly perpendicular to the weighted partial derivatives, and so the weighted partial derivatives may not be exactly in the range of K^*. As an alternative, K_2 might be chosen to minimize

$$\left\| K_1^* \left[\frac{1}{f_\theta^2} K_2^* \left[\frac{\partial f_\theta}{\partial \theta_i} \right] \right] - \frac{1}{f_\theta^2} \frac{\partial f_\theta}{\partial \theta_i} \right\|_2 . \qquad [10.1.7]$$

This would be easy to do in practice, as it is equivalent to finding functions (vectors) $x_1(\lambda)$, . . . , $x_p(\lambda)$ so that

$$\left\| K_1^* [x_i] - \frac{1}{f_\theta^2} \frac{\partial f_\theta}{\partial \theta_i} \right\|_2 \qquad [10.1.8]$$

is minimized, and then defining K_2^* so that $K_2^* \left[\frac{\partial f_\theta}{\partial \theta_i} \right] = f_\theta^2 x_i$. In practice, where K_1 and K_2 are matrices viewed as linear operators on a finite dimensional Hilbert space, the appropriate ranges are always closed and it is possible to achieve the minimum. If the noise space is too big, it will not be possible to recover the weighted partials well. Components of these weighted

partials which lie in the noise space as determined by K_1 represent loss in terms of variance. The IRWLS procedure corresponding to this discussion is:

Step 1 Obtain $\widehat{\theta}_{old}$, a robust estimate of θ_0, and K_1, a filter for the noise spectrum.

Step 2 Choose $K_{2\widehat{\theta}}$ so that for the functions $\frac{\partial f_{\widehat{\theta}}}{\partial \theta_i}$,

$$\left\| K_1^* \left[\frac{1}{f_{\widehat{\theta}}^2} K_{2\theta}^* \left[\frac{\partial f_{\widehat{\theta}}}{\partial \theta_i} \right] \right] - \frac{1}{f_{\widehat{\theta}}^2} \frac{\partial f_{\widehat{\theta}}}{\partial \theta_i} \right\|_2$$

is minimized. For functions in the space perpendicular to the space of partial derivatives, define $K_{2\widehat{\theta}}$ as the identity mapping.

Step 3 Minimize $|| P_{\widehat{\theta}} [K_{2\widehat{\theta}} M_{1/f_{\widehat{\theta}}^2} K_1 [I_n - f_\theta]] ||^2$, where P_θ is the projection onto the span of the partials at θ to obtain a new estimate $\widehat{\theta}_{new}$.

Step 4 Set $\widehat{\theta}_{old} = \widehat{\theta}_{new}$ and go to step 2.

Again, the process which has been described in choosing K_2 is iterative; it must be repeated on each IRWLS iteration. So K_2 is again another example of a model driven kernel operator.

Our final attempt at creating new QL functions will be the **nonparametric noise constructor** (nnc). It's basic form is $M_{1/f^2} K_{f,\psi}$, where Ψ is the space of bounded functions with the supremum norm. $\widehat{\psi}_n$ will be a consistent, nonparametric estimate of the contaminated spectrum ψ_0. The procedure is the following.

Step 1 Obtain $\widehat{\theta}_{old}$, a robust estimate of θ_0.

Step 2 Construct an estimate $\hat{\phi}$ of the noise spectrum using $f_{\hat{\theta}}$ and $\hat{\psi}_n$ (e.g. $\hat{\phi}= \hat{\psi}_n - f_{\hat{\theta}}$). $K_{f_{\hat{\theta}}, \hat{\psi}}$ will then be the orthogonal projection onto the space perpendicular to $\hat{\phi}$.

Step 3 Minimize $||P_{\hat{\theta}} [M_{1/f^2_{\hat{\theta}}} K_{f_{\hat{\theta}}, \hat{\psi}} [I_n - f_\theta]]||^2$, where P_θ is the projection onto the span of the partials at θ to obtain a new estimate $\hat{\theta}_{new}$.

Step 4 Set $\hat{\theta}_{old}= \hat{\theta}_{new}$ and go to step 2.

The nnc is an attempt to minimize the damage done to the weighted partials, as it has the smallest possible dimension noise space, defined by a consistent estimator of the noise. Note the (perhaps unusual) definition of Ψ as a nonparametric function space. This definition takes full advantage of the lenient requirements on Ψ as defined in definition 6.3.4. Observe also that the projection changes on each iteration since it is constructed from $f_{\hat{\theta}}$, but it is *not* a model driven QL operator as we do not need to consider the partials of the model in its construction. The "noise space" does not have to remain fixed, but may change with θ in an attempt to find a "good" model and in a sense model the noise along with the means. This concept of "nonlinear" noise removal is undoubtedly very important in future development of the theory as here outlined, because the ability of the user to exactly and completely specify the noise space is extremely doubtful in practice.

10.2 The Model Driven QL Operator: A Redefinition

Definition 6.3.4 does not cover "model driven QL operators" as described in section 10.1. For concreteness, let us consider the "bandwidth" problem in the model driven QL operator $M_{1/f^2}K_b$ discussed in the opening paragraphs of section 10.1. Recall this operator cannot possibly satisfy definition 6.3.4 because the bandwidth depends upon the specific partial

derivatives of the model, rather than the function f_θ. A different parametrization of f_θ might result in a different bandwidth. The apparent practical usefulness of model driven "QL operators" motivates a rethinking of the concepts of "model" and "QL operator".

The dependence of the model driven QL operator upon the partials is only apparent, and disappears with a redefinition of both "model" and "QL operator". In the following definition, let U be an open subset of C[a, b].

Definition 10.2.1 A **k dimensional QL operator** is a mapping between $\Psi \times U \times C[a, b] \times \ldots \times C[a, b]$ and $B(L^2)$ defined by $\psi \times (x(\lambda), f_1(\lambda), \ldots f_k(\lambda)) \rightarrow W(x(\lambda), f_1(\lambda), \ldots, f_k(\lambda))[\ .\ ]$. This mapping must satisfy (a), (c), and (d) of definition 6.3.4, with all occurrences of "$x(\lambda)$" in definition 6.3.4 (c) and (d) replaced by the k+1 dimensional vector of functions "$(x(\lambda), f_1(\lambda), \ldots f_k(\lambda))$".

Definition 10.2.2 A **k dimensional QL function** $D_\psi(f, g)$ is a real valued function with domain $\Psi \times \Omega \times L^2[a, b]$, where $\Omega = U \times C[a, b] \times \ldots \times C[a, b]$. Here, there are k factors C[a, b], and $U \subset C[a, b]$ (our previous definition of model would be contained in U). If $f=(f_0, f_1, \ldots f_k) \in \Omega$ and $g \in L^2$, D(f, g) must satisfy

$$\frac{\partial D(f, g)}{\partial f_0} = W(f)[f_0 - g] \qquad [10.2.1]$$

where W(f)[.] is a QL operator according to definition 10.2.1.

Definition 10.2.3 A **k dimensional model** for a random L^2 sequence is the k+1 dimensional vector of functions

$$\left(f_\theta, \frac{\partial f_\theta}{\partial \theta_1}, \ldots \frac{\partial f_\theta}{\partial \theta_k}\right)$$

where $\theta \in \mathbf{R}^p$ and $f_\theta \in M$, the model space of the random L^2 sequence.

For simplicity in notation, these definitions are made in the univariate case. As in chapter 6, the multivariate case is completely analogous, e.g. Ω becomes $U \times \Pi_k C[a, b] \times \ldots \times \Pi_k C[a, b]$, where $U \subseteq \Pi_k C[a, b]$.

Intuitively, what definition 10.2.2 means is that the operator $W(x(\lambda), f_1(\lambda), \ldots, f_k(\lambda))[\,.\,]$ depends on the argument vector of functions in a "differentiable" way. For example, it can be shown that $W(x(\lambda), f_1(\lambda), f_2(\lambda))=M_{1/x^2} P_{f_1,f_2}[\,.\,]$ is a QL operator, where $P_{f_1,f_2}[\,.\,]$ is the projection onto the span of f_1 and f_2. This is because the projection is *differentiable as a function of f_1 and f_2* (see theorem 10.2.2).

With these definitions, all of the theorems and proofs of chapters 6 and 7 for QL operators, e.g. proposition 6.5.3, theorem 6.5.1 (b), etc., follow *in a similar fashion.* We will now indicate what the differences are.

The main changes in the theory from chapter 7 involve derivatives of $\Phi(x(\lambda), y(\lambda))=W(x(\lambda))[y(\lambda)]$, where $x(\lambda) \in \Omega$ and $y(\lambda) \in L^2$. For example, proposition 7.2.3 should be changed to read

Proposition 10.2.1

$$D_z\Phi(z(\lambda), z_0(\lambda))\Big|_{a(\lambda)}[h] = \frac{\partial\Phi(a(\lambda), a_0(\lambda))}{\partial x}[h] + \frac{\partial\Phi(a(\lambda), a_0(\lambda))}{\partial y}[h_0] \qquad [10.2.2]$$

for $a(\lambda)=(a_0(\lambda), a_1(\lambda), \ldots a_k(\lambda))$, $h(\lambda)=(h_0(\lambda), h_1(\lambda), \ldots h_k(\lambda))$ both in Ω.

Remark: we are taking the derivative of the mapping $z(\lambda)\rightarrow\Phi(z(\lambda), z_0(\lambda))$, for $z(\lambda)=(z_0(\lambda), z_1(\lambda), \ldots z_k(\lambda)) \in \Omega$, evaluated at $a(\lambda) \in \Omega$ and applied to $h(\lambda) \in \Omega$. $\frac{\partial\Phi(a(\lambda), a_0(\lambda))}{\partial x}[\,.\,]$ is a linear mapping from $\Omega\rightarrow L^2$, while $\frac{\partial\Phi(a(\lambda), a_0(\lambda))}{\partial y}[\,.\,]$ is a linear mapping from $C\rightarrow L^2$.

proof

View $z(\lambda) \rightarrow \Phi(z(\lambda), z_0(\lambda))$ as the composition of the linear map $F[z(\lambda)]=(z(\lambda), z_0(\lambda))$ (a mapping from Ω to $\Omega \times L^2$) with $\Phi(x(\lambda), y(\lambda))$, where $x(\lambda) \in \Omega$ and $y(\lambda) \in L^2$. By the chain rule and proposition 7.2.2, the derivative evaluated at $a(\lambda)$ and applied to $h(\lambda)$ is

$$D\Phi(x(\lambda), y(\lambda))\Big|_{(a(\lambda), a_0(\lambda))} [h(\lambda), h_0(\lambda)]' =$$

$$\left[\frac{\partial \Phi(x(\lambda), y(\lambda))}{\partial x} \Big|_{(a(\lambda), a_0(\lambda))}, \frac{\partial \Phi(x(\lambda), y(\lambda))}{\partial y} \Big|_{(a(\lambda), a_0(\lambda))} \right] \begin{bmatrix} h(\lambda) \\ h_0(\lambda) \end{bmatrix}.$$

But this equals [10.2.2].

Now by the chain rule, we have the following analog to corollary 7.2.1.

Corollary 10.2.1 (to proposition 10.2.1)

$$\frac{\partial}{\partial \theta} \Phi(f_\theta, f_{\theta 0}) = \frac{\partial \Phi(f_\theta(\lambda), f_{\theta 0}(\lambda))}{\partial x} * \frac{\partial f_\theta}{\partial \theta} + \frac{\partial \Phi(f_\theta(\lambda), f_{\theta 0}(\lambda))}{\partial y} * \frac{\partial f_{\theta 0}}{\partial \theta} \qquad [10.2.3]$$

where $f_\theta=(f_{\theta 0}, f_{\theta 1}, \ldots f_{\theta k}) \in \Omega$ is a model according to definition 10.2.3. Note that the partial evaluated at θ is a linear mapping $\mathbf{R}^p \rightarrow L^2$.

proof

The proof essentially follows from the chain rule. The derivative of $\theta \rightarrow f_\theta$, evaluated at θ and applied to θ_1 is $\frac{\partial f_\theta}{\partial \theta'} \theta_1$, a function which is a linear combination of the partial derivatives $\left\{ \frac{\partial f_\theta}{\partial \theta_i} \right\}_{i = 1..p}$. The derivative of $\Phi(x, y)$, evaluated at f_θ and applied to $\frac{\partial f_\theta}{\partial \theta'} \theta_1$ is the transpose of the column vector [10.2.3] (matrix) multiplied with the column vector θ_1. Hence the representation [10.2.3].

Corollary 7.2.2 generalizes as follows:

Corollary 10.2.2 (to proposition 10.2.1)

$$\frac{\partial}{\partial\theta'}\,\Phi(f_\theta,\, f_{\theta 0})\odot\frac{\partial f_{\theta 0}}{\partial\theta} \;=\; \Phi(f_\theta,\, f_{\theta 0})\odot\frac{\partial^2 f_{\theta 0}}{\partial\theta'\partial\theta} +$$

$$\frac{\partial f_{\theta 0}}{\partial\theta}\odot\left(\frac{\partial\Phi(f_\theta(\lambda),\, f_{\theta 0}(\lambda))}{\partial x} * \frac{\partial f_\theta}{\partial\theta} + \frac{\partial\Phi(f_\theta(\lambda),\, f_{\theta 0}(\lambda))}{\partial y} * \frac{\partial f_{\theta 0}}{\partial\theta}\right) \qquad [10.2.4]$$

proof

The proof follows from the preceding corollary and the product rule proposition (7.1.1).

The analog of corollary 7.2.3 is easily obtained in a similar way.

Corollary 10.2.3 (to proposition 10.2.1)

$$\frac{\partial}{\partial\theta'}\,\Phi(f_\theta,\, f_n)\odot\frac{\partial f_{\theta 0}}{\partial\theta} \;=$$

$$\Phi(f_\theta,\, f_n)\odot\frac{\partial^2 f_{\theta 0}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta 0}}{\partial\theta}\odot\left(\frac{\partial\Phi(f_\theta(\lambda),\, f_n(\lambda))}{\partial x} * \frac{\partial f_\theta}{\partial\theta'}\right) \qquad [10.2.5]$$

We now have all of the pieces to use in the proof of the representation theorem. Replacing the parts from the corollaries in chapter 7 with the analogous parts from the above corollaries results in a proof of the representation theorem, with $Q_W(\psi,\, \theta)$ as defined before (in [7.1.2]), and $M_W(\psi,\, \theta)$ is defined as

$$[\Phi_\psi(f_\theta,\, f) - \Phi_\psi(f_\theta,\, f_{\theta 0})]\odot\frac{\partial^2 f_{\theta 0}}{\partial\theta'\partial\theta} + \frac{\partial f_{\theta 0}}{\partial\theta}\odot\left(\left\{W_\psi(f_\theta) * \frac{\partial f_{\theta 0}}{\partial\theta'}\right\}\right.$$

$$+\left\{ \frac{\partial\Phi(f_\theta(\lambda),\, f_{\theta 0}(\lambda))}{\partial x} - \frac{\partial\Phi(f_\theta(\lambda),\, f(\lambda))}{\partial x} * \frac{\partial f_\theta}{\partial \theta'} \right\} \Bigg). \qquad [10.2.6]$$

The QL equations [7.1.4] are replaced by their model driven version

$$W_\psi(f_\theta)[f_{\theta 0} - g] \odot \frac{\partial f_{\theta 0}}{\partial \theta} = 0 \qquad [10.2.7]$$

keeping to the basic principle of replacing any "old definition" model function which has an operator applied to it by $f_{\theta 0}$, but leaving alone any "old definition" model function which is the argument of an operator. Observe that there are a few differences in the opening lines of the proof of theorem 7.1.1. Specifically, $\frac{\partial}{\partial \theta} D_\psi(f_\theta, f)$ can be written as *two equations set to 0,* one of which is the QL equations [10.2.7]. As it is not necessary to solve the other set, we only solve the QL equations [10.2.7].

Write $f=(f_0, f_1, \ldots f_k)$, and split f into the parts f_0 and $f_r=(f_1, \ldots f_k)$. By proposition 7.2.2 we may represent $\frac{\partial}{\partial f} D(f, g)$ as $\left[\frac{\partial}{\partial f_0} D(f, g), \frac{\partial}{\partial f_r} D(f, g) \right]$. Similarly, $\frac{\partial f_\theta}{\partial \theta}$ can be represented as $\left[\frac{\partial f_{\theta 0}}{\partial \theta}, \frac{\partial f_{r\theta}}{\partial \theta} \right]'$, where $f_{r\theta} = (f_{1\theta}, \ldots f_{k\theta})$. By the chain rule, the derivative $\frac{\partial}{\partial \theta} D(f_\theta, g)$ is

$$\left[\frac{\partial}{\partial f_0} D(f, g), \frac{\partial}{\partial f_r} D(f, g) \right] \begin{bmatrix} \frac{\partial f_{\theta 0}}{\partial \theta} \\ \frac{\partial f_{r\theta}}{\partial \theta} \end{bmatrix}.$$

Set to 0, the QL equations are now $\frac{\partial}{\partial f_0} D(f, g) \odot \frac{\partial f_{\theta 0}}{\partial \theta} = 0$, and $\frac{\partial}{\partial f_r} D(f, g) \odot \frac{\partial f_{\theta r}}{\partial \theta} = 0$. Fortunately, there is no reason to solve the second set, as the proof of theorem 7.1.1 is taken to begin at [10.2.7].

There are several important conclusions that may be drawn from this generalization. First, let us make the following definition.

Definition 10.2.4 A **k dimensional model driven kernel** $K(x_0, x_1, \ldots x_k)$ is a bounded linear operator satisfying definition 10.2.1.

We consider "model driven kernel analogs" of the QL operators discussed in chapter 9, attempting to study the variance and bias properties. Regarding bias conditions, we have the following:

Theorem 10.2.1

Any model driven QL operator of the form $K_2(x_0, \ldots x_k)\ K_1(x_0, \ldots x_k)$, where $K_1(f_{\theta_0})[f_N] = 0$ and $K_2(f_{\theta_0})[(\{D_x K_1(f_{\theta_0})\}[\partial f_{\theta_0}/\partial\theta])\ [f_N]] = 0$ satisfies

$$M_W(\psi, \theta_0) = \frac{\partial f_{\theta_0 0}}{\partial\theta} \odot \left\{ W_\psi(f_{\theta_0}) * \frac{\partial f_{\theta_0 0}}{\partial\theta'} \right\}. \qquad [10.2.8]$$

Remarks: (1) In the following, f_{θ_0} refers to the (k+1 dimensional) model vector evaluated at θ_0, while $f_{\theta_0 0}$ refers to the (1 dimensional) first component of the model vector evaluated at θ_0. (2) Any QL operator from chapter 9 has the form K_1K_2 (e.g. $K_1M_{1/f^2}K_2 = (K_1M_{1/f^2})\ K_2$). (3) If the model is correct, $K_1(f_{\theta_0})[f_N]=0$ automatically, and we are no worse off using model driven kernels than their fixed counterparts.

proof

Observe

$$\Phi_\psi(f_{\theta_0}, f_{\theta_0 0}+f_N) = \Phi_\psi(f_{\theta_0}, f_{\theta_0 0}) \qquad [10.2.9]$$

(obvious by the definition of Φ, since $K_1[f_N]=0$), so we will show

$$\frac{\partial\Phi(f_{\theta_0}(\lambda),\, f_{\theta_0 0}(\lambda))}{\partial x}\left[\frac{\partial f_{\theta_0}}{\partial\theta_i}\right]=\frac{\partial\Phi(f_{\theta_0}(\lambda),\, f_{\theta_0 0}+f_N)}{\partial x}\left[\frac{\partial f_{\theta_0}}{\partial\theta_i}\right]. \qquad [10.2.10]$$

To see [10.2.10], first note

$$\frac{\partial\Phi(f_{\theta_0}(\lambda),\, y(\lambda))}{\partial x}[h(\lambda)]=\left\{\frac{\partial W(f_{\theta_0}(\lambda))}{\partial x}[h(\lambda)]\right\}[y(\lambda)] \qquad [10.2.11]$$

for $h(\lambda)\in\Omega$, $y(\lambda)$ in L^2, and $W(x)=K_2(x)K_1(x)$. This is by the chain rule on $x\to W(x)[\ .\ ]\to W(x)[y]$. Evaluate the right side of [10.2.11] for $h(\lambda)= f_{\theta_0 0}(\lambda)+f_N(\lambda)$ or $f_{\theta_0 0}(\lambda)$, and $y(\lambda)$ equal to the partials of f_θ evaluated at θ_0. To do this requires taking the derivative of $W(x)[\ .\ ]$ with respect to $x\in\Omega$.

First observe K_2K_1 may be written as the composition $F\circ G$, where G: $\Omega\to B(L^2)\times B(L^2)$ is defined by $G[x]=[K_2(x),\, K_1(x)]'$ and F: $B(L^2)\times B(L^2)\to B(L^2)$ is defined by $F[A_2,\, A_1]'=A_2A_1$. By proposition 7.2.2, DF (the derivative of F evaluated at $(A_2,\, A_1)$) may be represented as

$$\left[\frac{\partial F}{\partial A_2},\frac{\partial F}{\partial A_1}\right] \qquad [10.2.12]$$

where $\frac{\partial F}{\partial A_2}[W]=WA_1$, and $\frac{\partial F}{\partial A_1}[W]=A_2W$. So $\left.\frac{\partial F}{\partial A_2}\right|_{K_1(x)}[W]=WK_1(x)$, and

$\left.\frac{\partial F}{\partial A_1}\right|_{K_2(x)}[W]=K_2(x)W$ for $W\in B(L^2)$. By the same proposition, DG evaluated at x may be

represented as

$$\begin{bmatrix} \dfrac{\partial K_2(x)}{\partial x} \\ \dfrac{\partial K_1(x)}{\partial x} \end{bmatrix}. \qquad [10.2.13]$$

So $$F|_{G(x)} \circ DG|_x [h] = \left[\frac{\partial F}{\partial A_2}, \frac{\partial F}{\partial A_1} \right] \begin{bmatrix} \dfrac{\partial K_2(x)}{\partial x} \\ \dfrac{\partial K_1(x)}{\partial x} \end{bmatrix} [h] =$$

$$\frac{\partial F}{\partial A_2} \frac{\partial K_2(x)}{\partial x} [h] + \frac{\partial F}{\partial A_1} \frac{\partial K_1(x)}{\partial x} [h] =$$

$$\left\{ \frac{\partial K_2(x)}{\partial x} [h] \right\} K_1(x) + K_2(x) \left\{ \frac{\partial K_1(x)}{\partial x} [h] \right\}. \qquad [10.2.14]$$

[10.2.10] follows from [10.2.14] and [10.2.11]. In [10.2.14], letting h be the partials of f_θ evaluated at θ_0 and evaluating the derivatives of the K's at f_{θ_0}, the resultant operator will map $f_{\theta_0 0}+f_N$ and $f_{\theta_0 0}$ to the same function in L^2. This means that [10.2.6] reduces to [10.2.8]. □

Of course, the hypotheses in theorem 10.2.1 will always hold if K_1 is a fixed operator satisfying $K_1[f_N]=0$, and $K_1[f_N]$ always equals 0 if $f_N=0$, indicating no "second derivative damage" for a correct model (see, e.g., the discussion at the end of chapter 7). The theorem suggests construction of "filter reconstructor" type operators; for example, choose $K_1(f_\theta)[\ .\ ]$ in some manner so that it filters the noise f_N, assumed to be in some noise space N. Then choose $K_2(f_\theta)\ [\ .\ ]$ so that it filters $(\{D_x K_1(f_\theta)\}[\partial f_\theta/\partial\theta])\ [f_N]$ at each stage of the IRWLS procedure while simultaneously attempting to straighten out the variance. It also would seem to make sense that to filter the noise, the partial derivatives of $f_{\theta 0}$ would not be needed. $K_1(x)$ being a

function of x_0 only would simplify taking $DK_1(x)$ (see corollary 10.2.4). The details involved in the practical implementation of this discussion will not be given here and are for future research.

What is an example of a model driven QL function? The "Taniguchi distance"!

Proposition 10.2.2

$D_1(K[x_1, \ldots x_k]f, K[x_1, \ldots x_k]g)$ is a k dimensional QL operator for any model driven kernel $K[x_1, \ldots x_k]$, *which does not depend on* x_0 and has derivative $K^*(x)M_{1/[K(x)[f]]^2}K(x)[f-g]$.

proof

Follows immediately from theorem 9.4.1.

We now give a simple example of a model driven QL kernel. By itself, the following probably is not useful in practice, although it may be a component of a more useful QL kernel. It does, however, show some of the difficulties involved in taking operator derivatives.

Theorem 10.2.2

If $f=\{f_1, \ldots f_k\}$ are any functions in C[a, b], define $P_f[\ .\ ]$ to be the L^2 orthogonal projection onto the space spanned by the components of f. Then $P_f[\ .\ ]$ is a model driven QL kernel.

proof It suffices to show this for $f=\{f_1\}$, since $P_f[\ .\ ] = P_{f_1}[\ .\ ]+ \ldots +P_{f_k}[\ .\ ]$. Now $P_f[x]=\frac{x \bullet f}{||f||^2} f$, so we can break up the mapping $f \rightarrow P_f[\ .\ ]$ as

$$f \xrightarrow{F} \begin{bmatrix} \frac{1}{||f||^2} \\ f^* \\ f \end{bmatrix} \xrightarrow{G} \frac{1}{||f||^2} f^* f$$

where F: $C \to \mathbf{R} \times (L^2)^* \times L^2$, and f^* is the functional defined by integrating against f. To find DF, it is necessary to take the derivative of the nonlinear functional $f \to 1/||f||^2$. This is a composition $f \to f^2 \to \int f^2 \to 1/(\int f^2)$. By proposition 6.2.1 and the chain rule, the derivative evaluated at f and applied to h is $-\frac{1}{||f||^4}\int 2fh\, d\lambda$. So the derivative of F, evaluated at f, must be

$$\begin{bmatrix} \frac{-2f^*}{||f||^4} \\ I_{fnal} \\ I \end{bmatrix}$$

where $I_{fnal}[f]=f^*$ for $f \in L^2$ and I is the identity from $C \to L^2$. The mapping G: $\mathbf{R} \times (L^2)^* \times L^2 \to B(L^2)$ defined by $G(a, b, c) = a \times b \times c$, has derivative, evaluated at (a, b, c) and applied to (d, e, f)

$$[\, b \times c\ ,\ a \times c,\ a \times b] \begin{bmatrix} d \\ e \\ f \end{bmatrix} = b \times c \times d + a \times c \times e + a \times b \times f$$

$$\text{So } DG|_{F(f)} \circ DF|_f[h] = \left[f^*f,\ \frac{1}{||f||^2} f,\ \frac{1}{||f||^2} f^*\right] \begin{bmatrix} \frac{-2f^*}{||f||^4} \\ I_{fnal} \\ I \end{bmatrix} [h]$$

$$= f^*f \frac{-2f^*[h]}{||f||^4} + \frac{1}{||f||^2} f\, h^* + \frac{1}{||f||^2} f^*h$$

$$= f\left[-\frac{2}{||f||^4}\left(\int fh\, d\lambda\right) f^* + \frac{1}{||f||^2}\, h^*\right] + \frac{1}{||f||^2}\, h\, f^*. \qquad [10.2.15]$$

[10.2.15] satisfies (d) of definition 6.3.4 because it is continuous as a function of f. □

In general, proofs involving derivatives of operators are not easy, and it is beyond the scope of this dissertation to prove that all of the examples of section 10.1 are QL operators. It is clear that at least *intuitively* those examples fit the framework described here, and we can show an interesting application of theorem 10.2.2 to the nonparametric noise constructor of section 10.1.

Corollary 10.2.4 (to theorem 10.2.2)

Let Ψ be the space of positive functions in L^2, and let ψ represent an element of Ψ. Suppose $\psi_0=f_{\theta_0 0}+f_N$. Define the operator $K(f, \psi)$ by $K(f, \psi)[h]= I[h] - P_{\{f-\psi\}}[h]$ for h, $f \in L^2$ and I the identity operator on L^2. Define the operator $L(f, h)[\ .\ ]$ to be

$$M_{-(1/f^3)h}\, K(f, \psi) \; -$$

$$M_{1/f^2}\left(f\left[\frac{1}{||f||^2}\, h^* - \frac{1}{||f||^4}\left(\int 2fh\, d\lambda\right) f^*\right] + h\,\frac{f^*}{||f||^2}\right). \qquad [10.2.16]$$

For the operator $M_{1/f^2}K(f, \psi)[\ .\ ]$ (i.e. the nonparametric noise constructor),

[10.2.10] holds if $L\left(f_{\theta_0 0}, \frac{\partial f_{\theta_0 0}}{\partial \theta_i}\right)[f_N] = 0$ for i=1..p. A sufficient condition for this to occur is if $K(f_{\theta_0 0}, \psi_0)[f_N]=0$, $f_{\theta_0 0} \perp f_N$ and $\frac{\partial f_{\theta_0}}{\partial \theta_i} \perp f_N$ for all i.

proof

For $f \in \Omega$, the derivative $D_f K(f_0)[h][\ .\ ] = D_{f_0} K(f_0)[h_0][\ .\ ]$ by the argument following [10.2.7] concerning $\frac{\partial}{\partial f} D(f, g)$. The mapping $f \rightarrow M_{1/f^2}K(f, \psi)[\ .\ ]$ can be broken down as

$$f \xrightarrow{G} \begin{bmatrix} M_{1/f^2} \\ K(f, \psi) \end{bmatrix} \xrightarrow{F} M_{1/f^2} K(f, \psi)$$

Regard this as the composition F o G, where G: $C \to B(L^2) \times B(L^2)$ and F: $B(L^2) \times B(L^2) \to B(L^2)$ are as indicated above (F is defined the same as in the proof of theorem 10.2.1). Using the notation in the proof of 10.2.1, the derivative is the following (evaluated at f and applied to $h \in C$).

$$\left[\frac{\partial F}{\partial A_2}, \frac{\partial F}{\partial A_1}\right]\Big|_{F \circ G(f)} \begin{bmatrix} D_f M_{1/f^2} \\ D_f K(f, \psi) \end{bmatrix} =$$

$$\{(D_f M_{1/f^2})[h]\}\ K(f, \psi) + M_{1/f^2}\ \{(D_f K(f, \psi))[h]\} \qquad [10.2.17]$$

Theorem 10.2.2 gives that $D_f K(f, \psi)[h] = -L_1(f)[h]$, where $L_1(f)$ is the operator defined by [10.2.15]. By theorem 9.4.5 and theorem 10.2.2, [10.2.17] can be evaluated as the operator

$$M_{-(1/f^3)h}\ K(f, \psi) +$$

$$M_{1/f^2}\left(-f\left[\frac{1}{||f||^2} h^* - \frac{1}{||f||^4}\left(\int 2fh\ d\lambda\right) f^*\right] - h\ \frac{f^*}{||f||^2}\right). \qquad [10.2.18]$$

To determine how close [10.2.10] is to being satisfied, [10.2.18] must be evaluated for $f=f_{\theta_0 0}$, $h=\frac{\partial f_{\theta_0 0}}{\partial \theta_i}$, i=1..p, and then applied to $f_{\theta_0 0}$ and $f_{\theta_0 0}+f_N$. The difference will be [10.2.18] applied to f_N. Note that the first piece of [10.2.18] applied to f_N is 0, and the second piece will also be 0 if $f_{\theta_0 0} \perp f_N$ and $\frac{\partial f_{\theta_0 0}}{\partial \theta_i} \perp f_N$ for all i. □

Recalling corollary 9.4.4 (which continues to hold for the nnc assuming [10.2.10]), we see that allowing a variable kernel most likely will damage the variance more than a fixed one due to [10.2.10] not being satisfied.

We can also use theorems 10.2.1 and 10.2.2 to show that for any QL operator W(.)[.], and any model $\{f_\theta\}_{\theta \in \Theta}$, $W^*P_\theta W$ is a QL operator. Recall this was the operator used in the IRWLS procedure for nonsymmetric W (see theorem 9.7.1). Assuming a correct model, there is no difference in the asymptotic variance obtained by solving the QL equations formed from either operator.

Proposition 10.2.3

Let W(.)[.] be a QL operator, and $\{f_\theta\}$ be a model. Then

(a) $W^*P_\theta W$ is a QL operator

(b) Suppose the QL equations

$$\frac{\partial f_\theta}{\partial \theta} \odot W_\psi(f_{\theta_0})[f_\theta - f_{\theta_0}] = 0 \text{ and } \frac{\partial f_\theta}{\partial \theta} \odot W^*_\psi(f_\theta)P_\theta W_\psi(f_\theta)[f_\theta - f_{\theta_0}] = 0$$

both have θ_0 as a unique solution, and suppose the model is correct. Then the asymptotic variance of $\hat{\theta}_1$ solving

$$\frac{\partial f_\theta}{\partial \theta} \odot W_\psi(f_{\theta_0})[f_\theta - I_n] = 0$$

is the same as the asymptotic variance of $\hat{\theta}_2$ solving

$$\frac{\partial f_\theta}{\partial \theta} \odot W^*_\psi(f_\theta)P_\theta W_\psi(f_\theta)[f_\theta - I_n] = 0.$$

proof

(a) By theorem 10.2.2, $W^*P_\theta W$ is a QL operator.

(b) As in the proof of theorem 9.7.1, it may be seen that $M^{-1}_{W^*PW}Q_{W^*PW}M^{-1}_{W^*PW} = M^{-1}_W Q_W\left[M^{-1}_W\right]^{-1}$ by direct calculation of these two matrices. □

The revised QL definitions, together with corollary 10.2.4 also motivate the following conjecture:

Theorem 10.2.3

The following procedure is equivalent to minimizing the QL equations with any QL operator of the form $K^*M_{1/f^2}K$, where $K(x(\lambda), f_1(\lambda), \ldots, f_k(\lambda))$ is as defined in the second IRWLS procedure of section 10.1 That is, $K(x(\lambda), f_1(\lambda), \ldots, f_k(\lambda))$ [h] is the nonorthogonal projection of h onto $M_v = M_{1v} \cap (M_{1v} \cap N)^{\perp} + (M_{1v}+N)^{\perp}$, where $v=(x(\lambda), f_1(\lambda), \ldots, f_k(\lambda))$, M_{1v} is the subspace spanned by the components of v, and N is the "noise space". Furthermore, assuming $f_{\theta_0 0} \perp f_N$ and $\frac{\partial f_{\theta_0 0}}{\partial \theta_i} \perp f_N$ for all i (or some similar conditions), [10.2.9] and [10.2.10] are satisfied.

step 1 Begin with a robust starting value $\hat{\theta}$ for the unknown parameter vector θ.

step 2 Find $K_{\hat{\theta}}$. The inverse variance operator is now $K^*_{\hat{\theta}} V^{-1}_{\hat{\theta}} K_{\hat{\theta}}$, where $V^{-1}_{\hat{\theta}}$ is the diagonal variance matrix with $1/f^2_{\hat{\theta}}(\lambda_i)$ on the diagonal.

step 3 Resmooth the data (periodogram) using the new kernel, finding $y_s = K_{\hat{\theta}}\, y$.

step 4 Find the new $\hat{\theta}$ minimizing $(y_s - f_\theta)' V^{-1}_{\hat{\theta}} (y_s - f_\theta)$.

This general procedure might be described as iterated, reweighted, *resmoothed* least squares, (IRWRSLS) since *the kernel in step 3 changes with the iteration.* In the case of identity link, the model space remains the same, so that the kernel does not change in this step and the procedure is the same as IRWLS. The proof of theorem 10.2.3 is for future research.

10.3 Conclusions

The "near optimality" theorems of chapter 9 all involve the partial derivatives of the model, but the QL theorems there all involve "fixed" kernels in that the kernels are initially set, and do not change with θ. If the kernels are supposed to act a certain way on the partials, it makes sense that the kernels should be *dependent on the partials* in order to act in the appropriate manner. This violates the basic definition 6.3.4, which says that the QL operator is a function of the means function only, not the parametrization. Section 10.1 gives some practical examples of *how* one might go about constructing model dependent kernels, such as the filter reconstructor or M_{1/f^2} K_b type operators with bandwidths to be determined in the iterative process. The nonparametric noise constructor is an example of a QL operator utilizing a *model dependent kernel*, but not a *model driven kernel*, since the kernel depends only on the means function, not the partial derivatives.

The focus then shifts to *how* the definitions of chapter 6 should be altered so that the theorems of chapters 6 and 7 would still apply to the types of operators discussed in section 10.1. Section 10.2 gives the revised definitions and shows specific changes that would need to be made in some of the theory of chapter 7.

Section 10.2 begins to examine the "second derivative" conditions [10.2.9] and [10.2.10], needed so that the matrix M_W will not contain terms with a nonzero second derivative in the case of a misspecified model. The "smoothed Taniguchi distance" [9.4.2]

using model driven kernels not dependent on the means function is an example of a model driven QL distance (proposition 10.2.2), and the projection operator is seen be a model driven kernel in theorem 10.2.2. Theorem 10.2.1 gives conditions under which model driven QL operators of the basic forms under discussion will satisfy the "second derivative conditions" [10.2.9] and [10.2.10]. Corollary 10.2.4 shows specific conditions needed for the nonparametric noise constructor to satisfy the hypotheses of theorem 10.2.1. Proofs such as that of theorem 10.2.2 are not as easy for bandwidth type operators, and it is left for future research to prove the other operators (or variants thereof) defined in section 10.1 are QL operators under the new definition. The chapter concludes with an unproven conjecture relating to the motivating example involving IRWLS of section 9.2.

Chapter XI
Conclusions and Areas for Further Research

In this dissertation, we have given a "functional analysis" extension of McCullagh's (1983) QL theory and showed how it may be applied to multivariate "generalized nonlinear model response surfaces" in the context of spectral estimation. Chapter 5 argues that multivariate spectral estimation is essentially a problem in multivariate generalized nonlinear model response surfaces, and suggests a way in which "separately parametrized" models may be easily fit to the multivariate spectrum. Using this as a starting point, chapter 6 begins to examine the problem of parametric spectral estimation in non Gaussian univariate series. This is seen to deviate from the usual GLIM type theory, in that the published asymptotic results for parametric estimates involve the fourth cumulant spectrum, indicating the periodogram cannot be viewed as being a collection of "asymptotically independent" random variables. It is established that one way of viewing this situation is in terms of functionals and operators on L^2, where the "variance matrix" becomes a variance operator on the function space L^2 which contains the "observation vector" (i.e. the periodogram). A new definition of QL functions is given (chapter 6), together with an optimality theorem (chapter 7) for parametric estimates obtained by using these functions. We further argue that non-Gaussian series must have a QL operator in this extended class as their "real" QL operator. As in the literature, it is shown that IRWLS and minimizing a "QL distance" (or solving the "QL equations") result in parametric estimates with the same asymptotic variance.

The focus of the dissertation then is to examine the ramifications of this new definition, which is begun in chapter 9 , i.e. what are some examples of "new" QL functions,

and what are they used for? But first, chapter 8 sets the background and shows the problems involved in spectral estimation fit into the framework of chapters 6 and 7. The theorems in chapter 8 assume only that the true spectrum is of bounded variation, whereas most of the literature assumes the spectrum satisfies the BCC.

Chapter 9 indicates that the theory of chapters 6 and 7 is not a trivial extension, but rather a useful tool in dealing with the problem of contamination and model misspecification. Some simple examples of "new" QL functions are given, and it is shown why using the "wrong" (i.e. a non Gaussian) QL function on a contaminated Gaussian series actually may remove bias from the parametric estimates. Recalling the equivalence of IRWLS and minimizing a QL function, we show that in some cases the use of a non Gaussian QL function is equivalent to applying an IRWLS procedure to the smoothed periodogram. The question of what damage would be done to the variance of the parametric estimates, assuming we incorrectly believed the series to be contaminated and used a non Gaussian QL estimate, is then considered. It is again seen that the damage will be insignificant, provided the QL function (i.e kernel) is properly chosen.

It should be remembered that the actual QL function is not needed (and may not even exist) in order to implement IRWLS in practice, just the QL operator. If the variance operator is not self adjoint, the usual method of IRWLS (i.e. minimizing $(y_n - f_\theta) \bullet W[y_n - f_\theta]$) is not equivalent to solving the QL equations, but section 9.7 gives a new method by which IRWLS may be carried out in practice.

Motivated by the theorems in chapter 9, which indicate that the kernel operators need to act in certain ways on the model partial derivatives (or functions thereof), in order to obtain good variances for parametric estimates, chapter 10 explores what happens if we have model driven operators; that is, operators which *do* depend upon the parametrization of the model.

This in turn raises new problems, for model dependent kernels may have nonzero "second derivative" parts of the QL function in the asymptotic variance matrix. In short, the beginnings of a new QL theory are established which includes the old as a special case.

Besides the questions raised in chapter 10, there are many directions for future research. For example, no serious simulation study of the different QL operators suggested in chapters 9 or 10 has yet been attempted. There are undoubtedly many other unknown QL operators, the ones given are only a first attempt at finding useful applications of QL theory. What are the situations in which the various operators work best? What are good ways to achieve a balance between variance and bias when choosing the QL operator kernel? Furthermore, the ideas in chapters 9 and 10 were geared mainly to univariate Gaussian series, virtually ignoring the non Gaussian and multivariate cases. What would the similarities and differences for these cases be?

Another possible extension of the dissertation would be establishing distributional results. Recall that all of the theorems were stated in terms of second order moment assumptions, no "central limit theorem" dependent theorems were established. There are different setups under which CLT type results may be obtained (e.g. Taniguchi (1982), Brillinger (1981)). This would be very important in any future hypothesis testing type theory.

Other topics not included in the dissertation might involve diagnostics specifically geared for spectral estimation, or "goodness of fit" tests for a parametric hypothesis against a nonparametric alternative. How do you do hypothesis testing in "contaminated" series? Presumably, using a non Gaussian QL function for hypothesis testing would result in "slightly less" power if the QL operator is correctly chosen. What are the implications of chapter 10's theory for "semiparametric" models, i.e fitting the parametric part of a model which includes both parametric and nonparametric pieces?

Finally, how does all of this apply to the true "generalized model response surface", leaving the context of spectral estimation? The dissertation strongly relies upon a "functional" approach of regarding the observation vector as a function. If the design points aren't asymptotically uniformly distributed throughout the design space, how do things change?

Bibliography

Azzalini, A., Bowman, A.W., and Hardle, W. (1989), On the use of nonparametric regression for model checking. Biometrika, 76, 1, 1-11.

Azencott, Robert and Dacunha-Castelle, Didier (1986), *Series of Irregular Observations: Forecasting and Model Building*, Springer Verlag, New York.

Bartle, Robert G. (1976), *The Elements of Real Analysis*, John Wiley and Sons, New York.

Bloomfield, P. (1973), An Exponential Model for the Spectrum of a Scalar Time Series, Biometrika 60, 217-226.

Box, George E. P., and Jenkins, Gwilym M. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day, Inc., Oakland, California.

Brillinger, D.R. (1981), *Time Series Analysis: Data Analysis and Theory*, Holt, Rinehart & Winston, New York.

Brockwell, P.J. and Davis, R.A (1987), *Time Series: Theory and Application.* Springer Verlag, New York.

Cameron, M. and Turner, R. (1987), Fitting Models to Spectra Using Regression Packages. Appl. Stat. 36, 1, 47-57.

Carroll, R.J. and Rupert, D. (1988), *Transforming and Weighing in Regression*, Chapman and Hall, New York.

Chae, S.B. (1985) *Holomorphy and Calculus in Normed Spaces*, Marcel Dekker, New York.

Chiu, S. (1988), Weighted Least Squares Estimators on the Frequency Domain for the Parameters of a Time Series, Annals, 16, 3, 1315-1326.

Chiu, S. (1990), Peak-Insensitive Parametric Spectrum Estimation, Stochastic Processes and their Applications, 35, 121-140.

Conway, John B. (1985) *A Course in Functional Analysis*, Springer Verlag, New York.

Dahlhaus, R. (1988), Small Sample Effects in Time Series Analysis: A New Asymptotic Theory and a New Estimate, Annals of Statistics, 16, 2, 808-841.

Davies, R. (1973), Asymptotic Inference in Stationary Gaussian Time Series, Adv. Appl. Prob. 5, 469-497.

Dunsmuir, W. (1979), A Central Limit Theorem for Parameter Estimation in Stationary Vector Time Series and its Application to Models for a Signal observed with Noise, Annals of Statistics, 7, 3, 490-506.

Dunsmuir, M. and Hannan, E. (1976), Vector Linear Time Series Models, Adv. Appl. Prob. 8, 339-364.

Dzhaparidze, K. (1974), A New Method for Estimating Spectral Parameters of a Stationary Regular Time Series, Th. Prob. Appl. 14, 1, 122-132.

Good, I.J. and Gaskins, R.A (1971). Nonparametric Roughness penalties for probability densities. Biometrika 58 255-277.

Green, P. (1984), Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives, J. R. Statist. Soc. B, 46, 2, 149-192.

Hannan, E. (1973), The Asymptotic Theory of Linear Time Series Models, J. Appl. Prob. 10, 130-145.

Hoffman, K (1988), *Banach Spaces of Analytic Functions*, Dover, New York.

Hosoya, Y. and Taniguchi, M. (1982), A Central Limit Theorem for Stationary Processes and the Parameter Estimation of Linear Processes, Annals of Statistics 10, 1, 132-153.

Ibragimov, I. (1967), On Maximum Likelihood Estimation of Parameters of the Spectral Density of a Stationary Time Series, Theory Prob. Appl. 12, 115-119.

Jorgensen, Bent (1983), Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models, Biometrika 70, 19-21.

Kolmogorov, A.N. (1941), Local structure of turbulence in fluid for very large Reynolds numbers, Transl. in Turbulence (S. K. Friedlander and L. Topper, eds.), 1961. New York, Interscience Publishers.

Koopmans, L.H. (1974), *The Spectral Analysis of Time Series*, Academic Press, New York.

Kulperger, R. (1985), On an Optimality Property of Whittle's Gaussian Estimate of the Parameter of the Spectrum of a Time series, J. Time Ser. Anal. 6,4, 253-259.

LeCam, L. M. (1969), *Theorie Asymptotique de la Decision Statistique*, Montreal University Press.

Marx, Brian D. (1988), Ill conditioned Information Matrices and the Generalized Linear Model: An Asymptotically Biased Estimation Approach, Ph.D. dissertation: VPI and SU.

McCullagh, P. (1983), Quasi-Likelihood Functions, Annals of Statistics, 11, 1, 59-67.

McCullagh, P. and Nelder, J.A. (1983), *Generalized Linear Models*, Chapman and Hall, New York.

Myers, Raymond H. (1988). Class Notes.

Nelder, J. and Wedderburn, R. (1972), Generalized Linear Models. J. R. Statist. Soc. A, 135, 3, 370-384.

O'Sullivan, Finbar, Yandell, Brian, and Raynor, William (1986), Automatic Smoothing of Regression Functions in Generalized Linear Models, Journal of the American Statistical Society, 81, 393, 96-103.

Rice, J. (1979), On the Estimation of the Parameters of a Power Spectrum, Journal of Multivariate Analysis 9, 378-392.

Robinson, P. (1978), Alternative Models for Stationary Stochastic Processes, Stoch. Process. Appl. 8, 141-152.

Rosenblatt, Murray (1985), *Stationary Sequences and Random Fields*, Birkhauser Boston, Inc., Boston.

Royden, H.L. (1968), *Real Analysis*, MacMillian Publishing Co., N.Y.

Rozanov, Y.A. (1967), *Stationary Random Processes*, Holden-Day, San Francisco, California.

Smith, Patricia L., (1979), Splines as a Useful and Convenient Statistical Tool, The American Statistician, 33, 2 57-62.

Staniswalis, Joan (1989), The Kernel Estimate of a Regression Function in Likelihood-Based Models, Journal of the American Statistical Society 84, 405, 276-283.

Staniswalis, Joan and Severini, Thomas (1991), Diagnostics for Assessing Regression Models, Journal of the American Statistical Society 86, 415, 684-692.

Taniguchi, M. (1979), On Estimation of Parameters of Gaussian Stationary Processes, J. Appl. Prob. 16, 575-591.

Tibshirani, Robert and Hastie, Trevor (1987), Local Likelihood Estimation, Journal of the American Statistical Society 82, 398 559-567.

Walker, A. (1964), Asymptotic properties of Least Squares estimates of Parameters of the Spectrum of a Stationary Non-Deterministic Time Series, J. Australian Math Soc. IV, 3, 363-384.

Wedderburn, R. (1974), Quasi-likelihood functions, Generalized linear models, and the Gauss-Newton Method, Biometrika, 61, 3, 439-447.

_____ (1976), Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, Biometrika, 61,3, 439-447.

Whittle, P. (1951), *Hypothesis Testing in Time Series Analysis*, Almquist and Wicksell, Uppsala.

Whittle, P. (1953), The analysis of multiple stationary time series J. R. Statist. Soc. B 15, 125-129.

Zygmund, A. (1968), *Trigonometric Series*, Cambridge University Press, N.Y.

Vita

Keith N. Selander was born November 20, 1960 in Raleigh, North Carolina to Edwin V. and Mary V. Selander. He received his B.S. degree in Mathematics from Virginia Tech in 1982, and a M.S. degree, also in mathematics, in 1984. He continued his studies in the mathematics department, working as a graduate teaching assistant and completing all requirements for the Ph. D. degree except the dissertation. In 1987, the author entered the Ph. D program in statistics at Virginia Tech, where he was also employed as a graduate assistant. He received the Boyd Harshbarger award for scholarship during the first year of graduate studies in statistics, and the Best Paper Award in 1991 from the Virginia Academy of Sciences. The author is married to Sindee Sutherland, whom he met during his stay in Blacksburg.