

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Essays on the optimal selection of series functions

Permalink

<https://escholarship.org/uc/item/3w63v3s1>

Author

Pascual, Francisco L.

Publication Date

2007

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Essays on the Optimal Selection of Series Functions

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Economics

by

Francisco L. Pascual

Committee in charge:

Professor Halbert White, Chair
Professor Ian Abramson
Professor Clive W. J. Granger
Professor Allan Timmermann
Professor Ruth Williams

2007

Copyright
Francisco L. Pascual, 2007
All rights reserved.

The dissertation of Francisco L. Pascual is approved, and
it is acceptable in quality and form for publication on
microfilm:

Chair

University of California, San Diego

2007

To my family

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Table of Contents	v
List of Tables	vii
Acknowledgments	viii
Vita, Publications, and Fields of Study	ix
Abstract	x
I Asymptotic Optimality in Linear Model Selection under Ridge Estimation	1
1. Introduction	1
2. Notation and definitions	4
3. Asymptotic optimality	9
A. Mallows's C_L	10
B. Leave-1-out cross-validation	13
C. Generalized cross-validation	16
D. GIC_{λ_n}	18
4. Simulation	19
5. Conclusion and further research	23
6. Proofs of theorems	25
7. References	41
II Optimal Selection of Series Functions under Dependent and Heterogeneous Processes	44
1. Introduction	44
2. Notation and definitions	48
3. Asymptotic loss-efficiency	51
A. GC_L	52
B. Leave-1-out cross-validation	56
4. Conclusion	58
5. Proofs of theorems	60
6. References	86
III Optimal Selection of Series Functions by h -Block Cross-validation	89
1. Introduction	89
2. Notation and definitions	91
3. Asymptotic loss-efficiency of h -block cross-validation	94
4. Simulation	97

5. Conclusion	100
6. Proofs	101
7. References	119

LIST OF TABLES

Table I.1:	Means of Ratios of Average Square Losses for DGP1	21
Table I.2:	Standard Deviations of Ratios of Average Square Losses for DGP1	21
Table I.3:	Means of Ratios of Average Square Losses for DGP2	22
Table I.4:	Standard Deviations of Ratios of Average Square Losses for DGP2	22
Table III.1:	Means of Ratios of Average Square Losses for DGP1	99
Table III.2:	Standard Deviations of Ratios of Average Square Error Losses for DGP1	99
Table III.3:	Means of Ratios of Average Square Losses for DGP2	100
Table III.4:	Standard Deviations of Ratios of Average Square Error Losses for DGP2	100

ACKNOWLEDGEMENTS

I am indebted to my advisor Hal White for his help and guidance throughout the whole research process. I also would like to thank the other members of my committee: Clive Granger, Allan Timmermann, Ruth Williams and Ian Abramson for their helpful comments and suggestions.

While at UCSD I benefited from the exceptional teaching of Hal White in econometric theory and causal modeling. I was also able to learn time series from Clive Granger and benefit from his discussions on my research as well as other econometrics topics. I also thank James Hamilton, Graham Elliott, Yixiao Sun, Mark Machima, Ted Groves, Ross Starr, Vincent Crawford and Julian Betts .

At the same time I must express my gratitude to the UCSD economics staff members and in particular to Mary Jane Hubbard, Devaney Kerr, Rafael Acebedo and Mike Bacci. I can not forget the help and support from students and colleagues and in particular from Max, Philip, Hilary, Zack, Alex, Kosit, Andrei, Alberto, Hiroaki, Yong-Gook, Hee-Seung, Danielken, Lucas, Augusto, Christian, Carlos, Munir, George and Marius.

Finally and very importantly I have to thank my parents, sisters, nephews and other members of my family for their constant strong support and encouragement.

VITA

1998	Bachelor of Arts in Economics Universidad Autonoma de Madrid
1999	Master of Science in Economics and Econometrics University of Southampton
2000-2001	Teaching Assistant, Linguistics Department University of California, San Diego
2001-2006	Teaching Assistant, Economics Department University of California, San Diego
2007	Doctor of Philosophy in Economics University of California San Diego

ABSTRACT OF THE DISSERTATION

Essays on the Optimal Selection of Series Functions

by

Francisco L. Pascual

Doctor of Philosophy in Economics

University of California, San Diego, 2007

Professor Halbert White, Chair

The object of study of the present dissertation is the asymptotic optimality of model selection procedures for choosing an optimal model whose parameters have been estimated using ridge regression. Thus, we have to select simultaneously both the optimal model and the optimal ridge parameter.

Our problem focuses on obtaining a forecast of a target variable given a fixed vector of predictors. It is well known that given a quadratic loss function (that we assumed throughout) the optimal forecast is the conditional expectation of the target variable given the predictors, which is an unknown function of the predictors. Therefore our task reduces to finding an appropriate model capable of approximating this unknown function.

In particular and following White (2006) we focus on series estimators which are linear in the coefficients. They are as flexible as non-linear models and avoid at the same time the computational difficulties that arise in the implementation of the latter.

How well series estimators perform (in terms of approximating an unknown function) depends crucially on how optimally the basis or series functions

are chosen. In addition, the use of ridge estimation improves the predictive ability of the model (and therefore its approximation capabilities) as long as the shrinkage parameter is optimally chosen.

Ideally we would select the series functions and ridge parameter that minimize prediction mean squared error (PMSE) but since this is unknown (depends on the joint distribution of the data) we have to use an estimator. Most model selection procedures can be regarded as either direct or indirect estimates of PMSE and that is why we want to study their optimality.

The asymptotic optimality property we analyze is known in the literature as asymptotic loss-efficiency. It implies that the accuracy of the estimator based on the selected model is asymptotically the same as that based on the best model in the list (which is unknown).

The first chapter studies the conditions delivering the asymptotic loss-efficiency of Mallows C_L , leave-1-out cross-validation, generalized cross-validation and GIC_{λ_n} , the latter embedding a number of other model selection criteria. We assume stochastic regressors and *iid* observations. Therefore the chapter constitutes an extension of the work by Li (1986, 1987) and Shao (1997) to cover model selection under ridge estimation, which generalizes OLS.

The second chapter studies asymptotic loss-efficiency in an environment with dependent and heterogeneous observations and therefore it allows us to deal with financial and macroeconomic data. We consider a generalized version of Mallows C_L that we called GC_L (following Andrews (1991)) and which is appropriate under this more general data structure. We also cover leave-1-out cross-validation. As we show there, when the errors of the data generating process (that we define at the outset of chapter one) are correlated, the optimality of these criteria breaks

down.

The last chapter analyzes a general version of leave-1-out cross-validation which is known as h -block cross-validation and that is robust to error correlation. It concludes with a small simulation showing the relative performance of the different model selection procedures under dependence and heterogeneity.

I

Asymptotic Optimality in Linear Model Selection under Ridge Estimation

I.1 Introduction

In many instances in economics and finance we need to obtain a point forecast of a target variable Y_i given a $d \times 1$ vector of predictors Z_i (with d a finite integer). One common way to proceed, under quadratic loss, is to construct point forecasts as approximations to the conditional expectation of Y_i given Z_i

$$\mu(Z_i) = E[Y_i|Z_i]$$

which yields the best possible prediction of Y_i given Z_i under prediction mean squared error (PMSE), provided Y_i has finite variance. That is, μ solves the problem

$$\min_{m \in \mathfrak{M}} E [(Y_i - m(Z_i))^2]$$

where $\mathfrak{M}$ is the set of functions m of Z_i having finite variance and the expectation is taken with respect to the joint distribution of Y_i and Z_i .

Throughout this chapter we will focus on quadratic loss functions, which yields the conditional expectation as optimal predictor. Different optimal predictors would arise under other losses (e.g. the conditional quantile under asymmetric loss) but for simplicity we don't undertake that analysis in this paper, which is subject to further research.

As the function μ is unknown (it depends on the unknown conditional distribution of Y_i given Z_i) the researcher has to specify a model for μ , that is, a collection $\mathfrak{M}$ of functions of Z_i . If μ belongs to $\mathfrak{M}$ we say that the model is correctly specified, otherwise the model is misspecified. In economics and finance, as well as other sciences, models are commonly misspecified and are just approximations to the true data generating process (DGP) about which we have at best little information.

Given the above we are forced to use models to get as an accurate as possible approximation on the underlying conditional expectation. These can be classified as parametric or non-parametric and those falling in the parametric class can be linear or non-linear. Here, as in White (2006), we advocate the use of non-linear parametric models that can be regarded as “series functions” of the following form

$$f(z, \theta) = z'\gamma + \sum_{j \in \alpha} \psi_j(z)\beta_j$$

where, as will be clarified in section 2, α is a subset of $\mathbb{N}$ containing $q \leq p_n$ elements belonging to A_n , which is a set of subsets of $\mathbb{N}$ with at most p_n elements; and the basis functions $\psi_j(\cdot)$ are non-linear functions of z .

Following White (2006), one of the advantages of this specification is that the model is non-linear in z (thus achieving the degree of flexibility featured by non-linear models in general) and linear in the parameters $\theta \equiv (\gamma', \beta')'$, $\beta \equiv (\beta_1, \dots, \beta_q)'$ (overcoming the computational difficulties typical of any model non-

linear in the parameters). Thus under this setting OLS estimation becomes readily available.

When choosing a forecasting model under the above parameterization, one crucial question is that of selecting an appropriate model $\alpha \in A_n$ (this is equivalent to selecting optimally the number of basis functions within those proposed for a given n). Another crucial point is that our choice of basis functions should yield a better approximation the larger their number becomes (this is the requirement that the span (the set of all linear combinations) of the basis functions should be dense in the function space where μ lies).

Focusing on the first question in the paragraph above, we would ideally select the regressors or basis functions (among those available for a given size n) so as to minimize PMSE with respect to $\alpha \in A_n$. Nevertheless, this depends on the joint distribution of Y_i and Z_i , which is unknown so we have to use model selection criteria as direct or indirect estimates of PMSE.

The second point above implies that larger models (containing an increasing number of basis functions) should deliver better approximations. It is well known that as we increase the dimension or complexity of a model we reduce the bias in the PMSE but as more parameters need to be estimated, the component in PMSE accounting for parameter uncertainty becomes larger. If we estimate the model using shrinkage techniques, we achieve a substantial reduction in variance in exchange for little additional bias. This would allow model selection criteria to pick larger and more accurate models, which justifies our use of ridge regression to estimate models throughout the paper.

This chapter examines the asymptotic loss-efficiency (an asymptotic optimality property defined in the next section) of a number of the most widely known

model selection procedures for choosing models under ridge estimation. There are several results in the literature about asymptotic loss-efficiency for choosing models under OLS estimation (a good review can be found in the paper by Shao (1997)) and for selecting the ridge parameter in Mallows C_L and generalized cross-validation (Li (1986)). Here we focus on selecting models under ridge estimation, which implies choosing simultaneously the optimal number of basis functions as well as the optimal amount of shrinkage and thus, delivering better accuracy as discussed above.

The chapter is organized as follows. In section 2 we introduce the main notation and definitions which will be used in later sections. Section 3 examines the asymptotic loss-efficiency for linear model selection under ridge estimation of several of the most commonly used statistics in applied research, which include Mallows C_L , leave-1-out cross-validation, generalized cross-validation and following Shao (1997) the GIC_{λ_n} , which embeds a number of model selection criteria and is related to leave-d out cross-validation. Section 4 contains a simulation illustrating the main results in the paper and including a larger number of statistics such as AIC and BIC. We end with a conclusion(including several directions for further research) and an appendix containing the proofs of the main results.

I.2 Notation and definitions

As stated in the introduction, our focus is on predicting a target variable Y_i given a $d \times 1$ vector of predictors Z_i . We regard our forecasting model as an approximation to the conditional expectation of Y_i given Z_i (the optimal predictor in terms of PMSE) so we are implicitly working with a quadratic loss. We assume that Y_i is generated as

$$Y_i = \mu(Z_i) + \varepsilon_i$$

where $\mu(Z_i) \equiv E[Y_i|Z_i]$ and ε_i is a zero mean error term with constant variance representing the deviations between Y_i and $\mu(Z_i)$.

Given a sample of n observations on the Y_i 's and Z_i 's, we write $Y \equiv [Y_1, \dots, Y_n]'$, $\varepsilon \equiv [\varepsilon_1, \dots, \varepsilon_n]'$ and $Z^n \equiv [Z_1, \dots, Z_n]'$ to denote the vectors containing the n observations on Y_i and ε_i respectively and the matrix whose i^{th} row contains the $1 \times d$ vector of predictors Z_i' . We also write $Y = \eta + \varepsilon$ where $\eta \equiv [\mu(Z_1), \dots, \mu(Z_n)]'$ denotes the $n \times 1$ vector of conditional expectations corresponding to each i .

We now introduce more formally assumptions A.1 and A.2, which refer to the data generating process (DGP) and the model respectively and will be used in the main results of the paper.

A.1 (Data generating process)

- (a) Let $\{(Y_i, Z_i)\}$ be a sequence of *iid* random vectors such that Y_i is real-valued and Z_i is a $\mathbf{Z}$ -valued vector, $\mathbf{Z} \subset \mathbb{R}^d$, $d \in \mathbb{N}$.
- (b) $\varepsilon_i \equiv Y_i - E[Y_i|Z_i] = Y_i - \mu(Z_i)$ is such that $Var[\varepsilon_i|Z_i] = Var[\varepsilon_i] \equiv \sigma^2 < \infty$.

A.2 (Model) Let $\{\psi_j : \mathbf{Z} \rightarrow \mathbb{R}\}$ be a sequence of measurable functions. Let $\{p_n\}$ be a given non-decreasing sequence of integers and $\{A_n\}$ a given sequence of sets where each A_n is a set of subsets of $\mathbb{N}$ and each subset contains at most p_n elements, so that $\alpha \in A_n$ is a subset of $\mathbb{N}$ containing $q_n \equiv \#\alpha \leq p_n$ elements.

The model is given by

$$\mathfrak{M}_n = \{m : \mathbf{Z} \rightarrow \mathbb{R} \mid m(z) = z'\gamma + \sum_{j \in \alpha} \beta_j \psi_j(z) \quad \gamma \in \mathbb{R}^d, \beta_j \in \mathbb{R}, \alpha \in A_n\}$$

We next define the ridge estimator; for that purpose let $X_n(\alpha)$ be the $n \times (d + q)$ matrix whose i^{th} row has elements Z_i and $\psi_j(Z_i)$, $j \in A_n$, $\alpha \in A_n$, $i = 1, \dots, n$, $n = 1, 2, \dots$. We consider the estimator given by

$$\hat{\eta}(\alpha, h) \equiv [\hat{\eta}_1(\alpha, h), \dots, \hat{\eta}_n(\alpha, h)]' \equiv M_n(\alpha, h)Y$$

where $M_n(\alpha, h) \equiv X_n(\alpha)[X_n(\alpha)'X_n(\alpha) + hI_q]^{-1}X_n(\alpha)'$ for given $h \geq 0$ and $\alpha \in A_n$.

Note that h (the shrinkage or ridge parameter) takes values ranging from 0 (which corresponds to the OLS estimator) to ∞ . As we increase h we decrease the variance of our estimator in exchange for additional bias. The optimal h involves a trade-off between bias and variance.

The conditional PMSE given Y, Z^n and an out-of-sample Z_i is

$$\sigma^2 + L_n(\alpha, h) \tag{I.2.1}$$

where $L_n(\alpha, h) \equiv n^{-1}\|\eta - \hat{\eta}(\alpha, h)\|^2 = n^{-1}\sum_{i=1}^n(\mu_i - \hat{\mu}_i(\alpha, h))^2$ is the squared error loss.

Note that (2.1) describes the prediction performance of a model whereas $L_n(\alpha, h)$ measures the efficiency of model α when estimation of the conditional expectation is the target of the analysis.

The loss can be decomposed as

$$L_n(\alpha, h) = \Delta_n(\alpha, h) + n^{-1}\varepsilon'M_n^2(\alpha, h)\varepsilon - n^{-1}2\eta'[I_n - M_n(\alpha, h)]M_n(\alpha, h)\varepsilon$$

where $\Delta_n(\alpha, h) \equiv n^{-1}\|\eta - M_n(\alpha, h)\eta\|^2$, so the risk conditional on Z^n , $R_n(\alpha, h; Z^n) \equiv E(L_n(\alpha, h)|Z^n)$ can be expressed as

$$R_n(\alpha, h; Z^n) = \Delta_n(\alpha, h) + \sigma^2 n^{-1} \text{tr} M_n^2(\alpha, h)$$

As mentioned in the introduction we want to select models under ridge estimation. This implies selecting simultaneously the optimal terms or basis functions ($\alpha \in A_n$) and the ridge parameter ($h \geq 0$). Ideally we would select the pair (α, h) that minimizes the unconditional PMSE or its conditional version (2.1), which is equivalent to minimizing $L_n(\alpha, h)$ with respect to α and h . As these quantities contain the unknown vector of conditional expectations η , we can only minimize an estimated version of them.

Some model selection criteria are direct estimates of the unconditional PMSE (e.g. Mallows's C_L , AIC, ...). As pointed in Friedman et al. (2001) it doesn't make much difference whether we focus on the unconditional or conditional PMSE for *iid* data when it comes to selecting the best model.

We next introduce the modern notions of consistency and asymptotic loss-efficiency for model selection. The latter is highly desirable for a model selection procedure to behave well and constitutes the main focus of the paper, as shown in the next section.

Let $(\hat{\alpha}_n, \hat{h}_n)$ denote the model and ridge parameter chosen by a given selection procedure and let (α_n^*, h_n^*) be the pair minimizing $L_n(\alpha, h)$ over $\alpha \in A_n$ and $h \geq 0$.

Following Shao (1997), we say that the selection procedure is **consistent** if

$$P\{(\hat{\alpha}_n, \hat{h}_n) = (\alpha_n^*, h_n^*) | Z^n\} \rightarrow 1 \quad a.s.$$

As Shao (1997) makes explicit, this consistency is in terms of model selection and is not related to the consistency of $\hat{\eta}(\hat{\alpha}_n, \hat{h}_n)$ as an estimator of η , i.e. we do not require

$$L_n(\hat{\alpha}_n, \hat{h}_n) = o_p(1)$$

This last consistency is not very useful since, as Shao (1997) points out, sometimes there no consistent estimator of η (e.g. $\inf_{\alpha \in A_n} \inf_{h \geq 0} L_n(\alpha, h) \neq o_p(1)$) because the model may be misspecified and sometimes there are too many consistent estimators (e.g. $\sup_{\alpha \in A_n} \sup_{h \geq 0} L_n(\alpha, h) = o_p(1)$), as when one has a correctly specified model with irrelevant predictors.

In some cases a selection procedure is not consistent but $(\hat{\alpha}_n, \hat{h}_n)$ is still “close” to (α_n^*, h_n^*) in the sense that

$$\frac{L_n(\hat{\alpha}_n, \hat{h}_n)}{L_n(\alpha_n^*, h_n^*)} \longrightarrow 1 \quad \text{in prob.}$$

A selection procedure satisfying this condition is said by Li (1986) among others to be **asymptotic loss-efficient** and this is the property we analyze in the present paper.

We conclude this section by making some remarks about assumptions A.1 and A.2 which will be useful for the rest of the paper:

1. First note that in A.1 we assume our observations are *iid*. We could relax the independence assumption and our results would hold under uncorrelated observations. An extension to dependent data is also possible though that would require an analysis of more general versions of the model selection procedures we treat in section 4.
2. We also assume our errors are conditionally homoskedastic. Our results could be easily generalized to accommodate conditional heteroskedasticity under additional condition on the errors to those we impose in the next section (see Andrews (1991) for the case of linear model selection under OLS).
3. In A.2 our model $\mathfrak{M}_n$ could potentially include a non-linear component containing any linear combination of p_n basis functions for each n . This generalizes the case where the basis functions follow a strict order (e.g. zero order

polynomials first, follow by first order polynomials, followed by second order polynomials, and so on) and any linear combination of a given number of them is possible.

In the approximation theory literature this is known as “non-linear approximations” (see De Vore (1998) for a survey and White (2005) for a summary). This is because the functions $g_\alpha(z, \beta) = \sum_{j \in \alpha} \psi_j(z) \beta_j$ define a non-linear space of functions, in the sense that linear combinations of the form $ag_\alpha + bg_K$, where K has also q elements, generally have up to $2q$ terms. This contrasts with the so-called “linear approximations” where the basis functions follow a strict order as shown above, so any linear combination of them would be of the form $g_q(z, \beta) = \sum_{j=1}^q \psi_j(z) \beta_j$. They define a space of functions that is linear in that any linear combination of two elements of this space is again a linear combination of the first q of the ψ_j 's.

Non-linear approximations are more flexible than linear ones and we might obtain a better approximation to the conditional expectation for a given number of basis functions q . Thus we would ideally consider for each n , all subsets of p_n basis functions. Nevertheless, for the statistics in the next section to select asymptotically an optimal model as illustrated above, we will need to impose some constraints on the rate at which we can increase the number of terms with n .

I.3 Asymptotic optimality

In this section we show the conditions under which Mallows's C_L , leave-1-out cross-validation and generalized cross-validation (GCV) are asymptotic loss-efficient. The GIC_{λ_n} statistic, under this framework, is asymptotic-loss efficient for an extended loss that incorporates a penalty on model complexity. Neverthe-

less Shao (1997) conjectures that GIC_{λ_n} might not be asymptotic loss-efficient for $L_n(\alpha, h)$.

Li (1987) showed that Mallows's C_L , leave-1-out cross-validation and GCV are asymptotic loss-efficient for linear model selection under OLS. These results were shown for the case of fixed regressors and models that if of fixed dimension, must be misspecified. They hold under some restrictions on the number of models we can compare for each sample size n and moment restrictions on the errors. Shao (1997) extends these results, for a larger number of model selection procedures, to the case where there are also correctly specified models of fixed dimension. Finally Li (1986) analyzes asymptotic loss efficiency of Mallows's C_L and GCV for choosing the ridge parameter though his results hold under normally distributed errors.

Here we extend these results to linear model selection under ridge estimation, considering the case of stochastic errors without restrictions on the error distribution. Our theorems only hold for the case of misspecified models. Nevertheless, as we argued before, this is the common situation in economics and finance and this is why we use series approximations.

I.3.A Mallows's C_L

Mallows's C_L selects $\alpha \in A_n$ and $h \geq 0$ by minimizing

$$\Gamma_{n,2}(\alpha, h) \equiv n^{-1} \|Y - \hat{\eta}(\alpha, h)\|^2 + 2\hat{\sigma}_n^2 \text{tr} M_n(\alpha, h)$$

where $\hat{\sigma}_n^2$ is an estimator of the error variance σ^2 and $\Gamma_{n,2}$ is the notation for Mallows's C_L used by Shao (1997).

Let (α^{C_L}, h^{C_L}) minimize $\Gamma_{n,2}(\alpha, h)$ over $A_n \times [0, \infty)$; our goal is to show

asymptotic loss-efficiency, that is

$$\frac{L_n(\alpha^{C_L}, h^{C_L})}{\inf_{\alpha \in A_n} \inf_{h \geq 0} L_n(\alpha, h)} \longrightarrow 1 \quad \text{in prob.}$$

where $L_n(\alpha, h) \equiv n^{-1} \|\eta - \hat{\eta}(\alpha, h)\|^2$.

Now assume the following regularity conditions:

$$\text{A.3 } \sum_{\alpha \in A_n} [\inf_{h \geq 0} n R_n(\alpha, h; Z^n)]^{-m} \longrightarrow 0$$

where m is some fixed positive integer such that $E[|\varepsilon_i|^{2m}] < \infty$.

$$\text{A.4 } \limsup_{n \rightarrow \infty} \sum_{\alpha \in A_n} \frac{\iota_n(\alpha) - k}{[\inf_{h \geq 0} n R_n(\alpha, h; Z^n)]^{m^*}} < \infty$$

where k is a positive integer not depending on n such that $\lambda_k(\alpha) > 0$ for all n ($\lambda_k(\alpha)$ denoting the k^{th} largest eigenvalue of $X_n(\alpha)' X_n(\alpha)$), $\iota_n(\alpha)$ is the largest i such that $\lambda_i(\alpha) > 0$, $i \leq q_n = \#\alpha \leq p_n$, and m^* is some fixed positive integer such that $E[|\varepsilon_i|^{2m^*}] < \infty$.

$$\text{A.5 } |\hat{\sigma}_n^2 - \sigma^2| = o_p(1)$$

and where A.3-A.5 hold for all sequences $\{Z^n\}$ a.s.; that is, in a set of probability 1.

Theorem I.3.1. *Under A.1-A.5 the pair (α^{C_L}, h^{C_L}) selected by Mallows's C_L is asymptotically loss-efficient.*

Some remarks on conditions A.3-A.5:

1. Condition A.3 is a blend of the conditions imposed by Li (1987) for model selection under OLS and Li (1986) for optimally choosing the ridge parameter. Since $R_n(\alpha, h; Z^n) = \Delta_n(\alpha, h) + \sigma^2 n^{-1} \text{tr} M_n(\alpha, h)$, condition A.3 doesn't hold for correctly specified models of fixed dimension (since in that case $\Delta_n(\alpha, h) = 0$). This condition also generalizes Li's (1986) results in that we

don't impose normality in the errors; instead, we only require finite moments of order $\max(2m, 2m^*)$.

In the case of series function approximations to η , we might expect the optimal conditional risk $R_n(\alpha, h; Z^n)$ to be of order $O(n^{-1+\delta})$ for some $\delta > 0$. So as long as the cardinality of A_n is of polynomial order $n^{\delta'}$ for some $\delta' > 0$, there always exists an m such that A.3 is satisfied.

Nevertheless, under non-linear approximations (as illustrated in the earlier section), we might want A_n to contain all possible subsets of dimension $\leq p_n$. But this violates A.3, as A_n would increase at an exponential rate and there wouldn't exist any finite m such that A.3 holds. Yang (1999) shows this is not the case if we add a model complexity penalty term to Mallows's C_L , though his result holds under the restriction of normally distributed errors.

2. Condition A.4 arises specifically for model selection under ridge estimation. As mentioned above, for series approximations $R_n(\alpha, h; Z^n)$ is typically $O(n^{-1+\delta})$ with $\delta > 0$, so $nR_n(\alpha, h; Z^n) \simeq O(n^\delta)$. On the other hand $\iota_n(\alpha)$ is $O(n)$ so for any $m^* \geq 3/\delta$ the limit above is bounded and A.4 would hold.
3. Finally, condition A.5 requires $\hat{\sigma}_n^2$ to be consistent for σ^2 . As Shao (1997) shows, a popular choice for $\hat{\sigma}_n^2$ is $(n - p_n)^{-1} \|Y - \hat{\eta}(\bar{\alpha}_n)\|^2$ where $\bar{\alpha}_n$ is the largest model in A_n (assuming that A_n always includes the largest model of dimension p_n) and where $\hat{\eta}(\bar{\alpha}_n)$ is the OLS estimator ($h = 0$) of η . Because $\hat{\sigma}_n^2 = (n - p_n)^{-1} \|Y - \hat{\eta}(\bar{\alpha}_n)\|^2 = (n - p_n)^{-1} \varepsilon' [I_n - M_n(\bar{\alpha}_n)] \varepsilon + (n - p_n)^{-1} n \Delta(\bar{\alpha}_n) + 2(n - p_n)^{-1} \varepsilon' [I_n - M_n(\bar{\alpha}_n)] \eta$, as long as $\Delta(\bar{\alpha}_n) \rightarrow 0$ and $p_n/n \nrightarrow 1$, $\hat{\sigma}_n^2$ is consistent and A.5 will be satisfied.

I.3.B Leave-1-out cross-validation

The leave-1-out cross-validation (CV_{-1}) method minimizes

$$CV_{n,-1}(\alpha, h) \equiv n^{-1} \sum_{i=1}^n \left(Y_i - x_i(\alpha)' [X_{n,(-i)}(\alpha)' X_{n,(-i)}(\alpha) + hI_{q+d}]^{-1} \right. \\ \left. \times X_{n,(-i)}(\alpha) Y_{(-i)} \right)^2 \quad (\text{I.3.1})$$

with respect to $\alpha \in A_n$ and $h \geq 0$, where $x_i(\alpha)$ denotes the i^{th} row of the $X_n(\alpha)$ matrix, $X_{n,(-i)}$ denotes the $X_n(\alpha)$ matrix with all the rows except the i^{th} and $Y_{(-i)}$ is the vector Y with all its components except the i^{th} .

It is easy to show that (3.1) can be expressed as

$$CV_{n,-1}(\alpha, h) = n^{-1} \left\| [I_n - \ddot{M}_n(\alpha, h)]^{-1} [Y - \hat{\eta}(\alpha, h)] \right\|^2, \quad (\text{I.3.2})$$

where $\ddot{M}_n(\alpha, h)$ is a diagonal $n \times n$ matrix whose i^{th} diagonal element is $m_{ii}(\alpha, h)$, which is the i^{th} diagonal element of $M_n(\alpha, h)$.

Following Li (1987) we can use the results for Mallows' C_L to establish the asymptotic loss-efficiency of CV_{-1} by noticing that the CV_{-1} statistic is just Mallows' C_L applied to the class of delete-1 estimates (defined below) and then showing under some conditions that asymptotically the ridge estimator is equivalent to the delete-one estimator.

To verify this, notice that (3.1) implies that CV_{-1} selects the pair (α, h) that minimizes the sum of squared prediction errors for Y_i with Y_i being itself excluded from the data set. Given $Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n$, we may write the predictor of Y_i as

$$\hat{Y}_{-i} = \sum_{j=1}^n m_{ij}^d(\alpha, h) Y_j$$

where m_{ii}^d is zero, m_{ij}^d is the $(i, j)^{th}$ element of $M_n^d(\alpha, h) \equiv D_n(\alpha, h)[M_n(\alpha, h) - I_n] + I_n$ where $D_n(\alpha, h)$ is an $n \times n$ matrix with i^{th} diagonal element $(1 - m_{ii}(\alpha, h))^{-1}$,

and as before $m_{ii}(\alpha, h)$ is the i^{th} diagonal element of $M_n(\alpha, h)$.

Define the delete-one estimate of η as

$$\eta^d(\alpha, h) \equiv M_n^d(\alpha, h)Y$$

where $M_n^d(\alpha, h)$ has zeros as diagonal elements as specified above.

It is easy to see from this that CV_{-1} equivalently minimizes

$$n^{-1} \|Y - m_n^d(\alpha, h)Y\|^2$$

with respect to $\alpha \in A_n$ and $h \geq 0$. So CV_{-1} is just the C_L procedure applied to the class of delete-one estimates $\{\eta^d(\alpha, h)\}$.

To show asymptotic loss-efficiency of CV_{-1} , we will first show that Mallows's C_L shares this property when applied to the delete-one estimate under some conditions. Then we will address the asymptotic equivalence between the ridge and the delete-one estimate (in fact, as pointed by Li (1987), the replacement of $\hat{\eta}(\alpha, h)$ by $\eta_n^d(\alpha, h)$ can be justified by arguing that for large n the delete-one estimate which is based on a sample of size n , should be nearly the same as the ridge estimator).

Now we define

$$L_n^d(\alpha, h) \equiv n^{-1} \|\eta - \hat{\eta}^d(\alpha, h)\|^2$$

$$R_n^d(\alpha, h; Z^n) \equiv E(L_n^d(\alpha, h) | Z^n)$$

and introduce the following conditions:

$$\text{A.6 } \inf_{\alpha \in A_n} \inf_{h \geq 0} L_n(\alpha, h) = o_p(1)$$

$$\text{A.7 } \limsup_{n \rightarrow \infty} \sup_{\alpha \in A_n} \sup_{h \geq 0} \bar{\lambda}(M_n(\alpha, h)) < 1$$

where $\bar{\lambda}(\cdot)$ denotes the maximal diagonal element of $M_n(\alpha, h)$

A.8 For any sequence (α_n, h_n) , with $\alpha \in A_n$ and $h \geq 0$, we have

$$R_n^d(\alpha, h; Z^n)/R_n(\alpha, h; Z^n) \rightarrow 1$$

if either $R_n^d(\alpha, h; Z^n) \rightarrow 0$ or $R_n(\alpha, h; Z^n) \rightarrow 0$

A.9 $\exists$ a positive constant K such that for any n, α and h

$$\bar{\lambda}(M_n(\alpha, h)) \leq Kn^{-1} \text{tr} M_n^2(\alpha, h)$$

By convention A.6-A.9 hold for all sequences $\{Z^n\}$ a.s.

Theorem I.3.2. *Under A.1-A.8 the pair (α^{CV}, h^{CV}) selected by CV_{-1} is asymptotically loss-efficient.*

Some remarks on conditions A.6-A.9:

1. Condition A.6 will hold as long as our series function estimator is consistent for the unknown conditional expectation η . This will be the case for a wide range of choices of basis functions.
2. Condition A.7 is crucial for the leave-1-out method to work and requires that the diagonal elements of $M_n(\alpha, h)$ be uniformly bounded away from 1. As $M_n(\alpha, h)$ is a pseudo-projection matrix with $\sup_{h \geq 0} \bar{\lambda}(M_n(\alpha, h)) \leq 1$, this seems a weak assumption (for any $h \geq 0$ this condition is clearly satisfied). It is also important to notice that this condition seems to prevent the dimension of the largest model in A_n from being $O(n)$. Nevertheless, as we are using ridge estimation, models with a large number of basis functions (for each n) will certainly require some shrinkage and this condition will be satisfied.
3. Though A.8 is hard to verify, condition A.9 will imply it. The latter is the analog of condition (5.2) in Li (1987) and excludes unbalanced designs. We omit the proof of this implication as it is just a simple adaptation of that of Li (1987) to the present context.

I.3.C Generalized cross-validation

In this section we closely follow the development by Li (1987) of generalized cross-validation (GCV) for linear model selection under OLS. Our results are an adaptation of his to the present setting of ridge estimation.

Generalized cross-validation (Craven and Wahba (1979)) chooses $\alpha \in A_n$ and $h \geq 0$ so as to minimize

$$GCV_n(\alpha, h) \equiv \frac{n^{-1} \|Y - \hat{\eta}(\alpha, h)\|^2}{(1 - n^{-1} \text{tr} M_n(\alpha, h))^2}$$

Following Li (1987) and similar to the strategy we used to show the asymptotic optimality of leave-1-out cross-validation, we relate GCV to Mallows's C_L through the so-called nil-trace estimator of η .

The nil-trace estimator was introduced by Li (1987) and is defined as

$$\bar{\eta}(\alpha, h) \equiv -\rho Y + (1 + \rho)\hat{\eta}(\alpha, h)$$

where $\rho \equiv n^{-1} \text{tr} M_n(\alpha, h) / (1 - n^{-1} \text{tr} M_n(\alpha, h))$.

Note that this estimator can also be written as

$$\bar{\eta}(\alpha, h) = \bar{M}_n(\alpha, h)$$

where $\bar{M}_n(\alpha, h) \equiv -\rho I_n + (1 + \rho)M_n(\alpha, h)$, whose trace is zero as can be easily seen.

The nil-trace estimator shrinks our ridge estimator ($\hat{\eta}(\alpha, h)$) toward the vector Y . It is not very useful by itself and as Li (1987) affirms, it is just a device in order to get some understanding about GCV. In fact, if we apply Mallows's C_L to select the pair (α, h) using the nil-trace class $\{\bar{\eta}(\alpha, h)\}$ as estimators of η , the resulting minimization procedure is the same as GCV. Thus, if we find the conditions under which Mallows's C_L is asymptotic loss-efficient under nil-trace estimation and those under which the nil-trace and ridge estimators are asymptotically equivalent,

the asymptotic optimality of GCV will follow.

We introduce the following conditions:

A.10 For any sequence $\{\alpha_n \in A_n, h_n \geq 0\}$ such that

$$n^{-1} \text{tr} M_n^2(\alpha_n, h_n) \rightarrow 0$$

we have that

$$\left(n^{-1} \text{tr} M_n(\alpha_n, h_n) \right)^2 / n^{-1} \text{tr} M_n^2(\alpha_n, h_n) \rightarrow 0$$

A.11 $\lim_{n \rightarrow \infty} \sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} \text{tr} M_n(\alpha, h) \leq \gamma_1$ for some $0 < \gamma_1 < 1$.

A.12 $\lim_{n \rightarrow \infty} \sup_{\alpha \in A_n} \sup_{h \geq 0} \left(n^{-1} \text{tr} M_n(\alpha_n, h_n) \right)^2 / n^{-1} \text{tr} M_n^2(\alpha_n, h_n) \leq \gamma_2$ for some $0 < \gamma_2 < 1$.

By convention, we take A.10-A.12 to hold for all sequences $\{Z^n\}$ a.s.

We next state a theorem specifying the conditions under which GCV is asymptotically optimal. The proof is omitted as it is a simple adaptation of that in Li (1987) to the present context.

Theorem I.3.3. *Under A.1-A.6 and A.10-A.12 the pair (α^{GCV}, h^{GCV}) selected by GCV is asymptotic loss-efficient.*

Some remarks on conditions A.11-A.12:

1. Condition A.11 is milder than A.7 in that it requires that the diagonal elements of the $M_n(\alpha, h)$ matrix be asymptotically strictly smaller than one on average. This condition seems necessary for the GCV method to work and lies behind its motivation (see Craven and Wahba (1979)). Similar comments to those for A.7 apply here with respect to the number of basis functions we can include for each model.

2. Condition A.12 is related to A.11. Note that the average variance of $\hat{\eta}(\alpha, h)$, which equals $\sigma^2 n^{-1} \text{tr} M_n(\alpha, h)$, comes in part from the diagonal elements of $M_n(\alpha, h)$. As $(n^{-1} \text{tr} M_n(\alpha, h))^2 \leq n^{-1} \sum_{i=1}^n d_{ii}^2(\alpha, h)$, A.12 requires that the asymptotic contribution of this elements to the variance of $M_n(\alpha, h)$ not be dominant.

I.3.D GIC $_{\lambda_n}$

The GIC $_{\lambda_n}$ method (Shao (1997)) minimizes

$$\Gamma_{n, \lambda_n} \equiv n^{-1} \|Y - \hat{\eta}(\alpha, h)\|^2 + n^{-1} \lambda_n \hat{\sigma}_n^2 \text{tr} M_n(\alpha, h)$$

with respect to $\alpha \in A_n$ and $h \geq 0$ and where $\{\lambda_n\}$ is a sequence of non-random numbers greater than or equal to 2 such that $\lambda_n/n \rightarrow 0$.

GIC $_{\lambda_n}$ with $\lambda_n \rightarrow \infty$ is the GIC method of Rao and Wu (1989); with $\lambda_n \equiv 2$ this is Mallows's C_L ; and with $\lambda_n \equiv \lambda > 2$ this is the FPE $_{\lambda}$ method of Shibata (1984).

Note that the GIC $_{\lambda_n}$ statistic can be decomposed as

$$\Gamma_{n, \lambda_n} = n^{-1} \|\varepsilon\|^2 + L_n(\alpha, h) + 2n^{-1} \varepsilon' P_n(\alpha, h) \eta$$

$$+ 2n^{-1} (\sigma^2 \text{tr} M_n(\alpha, h) - \varepsilon' M_n(\alpha, h) \varepsilon) + n^{-1} (\lambda_n \hat{\sigma}_n^2 - 2\sigma^2) \text{tr} M_n(\alpha, h)$$

Letting $\ddot{L}_n(\alpha, h) \equiv L_n(\alpha, h) + n^{-1} (\lambda_n \hat{\sigma}_n^2 - 2\sigma^2) \text{tr} M_n(\alpha, h)$, GIC $_{\lambda_n}$ can be expressed as

$$n^{-1} \|\varepsilon\|^2 + \ddot{L}_n(\alpha, h) + 2n^{-1} \varepsilon' P_n(\alpha, h) \eta + 2n^{-1} (\sigma^2 \text{tr} M_n(\alpha, h) - \varepsilon' M_n(\alpha, h) \varepsilon)$$

As remarked by Shao (1997), $\ddot{L}_n(\alpha, h)$ is a loss function (it is the squared error loss plus a penalty on the model complexity, $\text{tr} M_n(\alpha, h)$).

It is easy to see that under conditions A.1-A.2 and A.4-A.5 the GIC_{λ_n} with $\lambda_n > 2$ is asymptotic loss-efficient under $\ddot{L}_n(\alpha, h)$ as loss. Shao (1997) conjectures that it might not have this optimality property for $L_n(\alpha, h)$.

It is also worth noting that Shao shows that the delete- d cross-validation model selection procedure has the same asymptotic behavior as the GIC_{λ_n} method with $\lambda_n = \frac{n}{n-d} + 1$.

In the same paper Shao proves that whenever there are correctly specified models with fixed dimension, the GIC_{λ_n} method is asymptotic loss-efficient and since under some regularity conditions it has the same asymptotic behavior as delete- d cross-validation, this also holds for the latter. He nevertheless conjectures that the GIC_{λ_n} statistic might not have that optimality property under the setting of this paper (series approximations to an unknown conditional expectation). We shall not resolve this question here. Nevertheless, it appears to be possible to establish optimality of delete- d cross-validation when the loss function is based on $n - d$ rather than n observations. We leave this for future research.

I.4 Simulation

We use neural network models with the ridgelet activation function to approximate two data generating processes (referred to as DGP1 and DGP2 respectively). The ridgelet activation function was proposed by Candes (1998,1999,2003) and has the ability to approximate nonlinear multivariate functions which contain singularities along lines and are otherwise continuous.

DGP1 is given by $Y_i = 1 + \ln X_{1i}^2 + \ln X_{2i}^2 + \varepsilon_i$ where the regressors were generated from a multivariate normal distribution with mean 0 and variance-covariance matrix equal to the identity matrix. The error term is normal with mean 0 and variance 1. DGP2 is given by $Y_i = 10\sin(\pi X_{1i} + 20(X_{2i} - 5)^2) + \varepsilon_i$

where the regressors and error term were generated in the same way as for DGP1.

We applied a wide range of model selection criteria which includes C_L , AIC, corrected AIC (AICc), Leave-1-out cross-validation (CV-1), generalized cross-validation (GCV) and BIC in order to compare their relative performances for choosing the number of ridgelet basis functions and the ridge parameter.

As we are advocating from the outset the use of models which are linear in the parameters and nonlinear in the predictors, we follow the methodology proposed in White (2006) to obtain such parameterizations. This is based on earlier results by Bierens (1990) and Stinchcombe and White (1998). It works by starting with a linear model in the predictors, generating a large number of basis functions by setting their internal parameters randomly and selecting the one with the highest correlation with the residual of the linear model (in the sense that is the one capturing more nonlinearities among those generated in this way). In a second step, if we want to add a second basis, we generate again a large number of them and pick the one having the highest correlation with the residual of the model generated in the first step (which contains the linear term plus one basis function). We can keep adding more basis functions following the same procedure and the more we include the better our approximation becomes, though we have to estimate a larger number of parameters. For those looking for a justification of this method, we refer them to White (2006) and the other papers mentioned above.

We thus generate 33 models as described above. The first containing the two predictors plus one basis and we keep incorporating basis functions one by one until we arrive at the last model which includes the predictors (linear term) plus 33 basis functions (at each step we generate 500 new basis functions and pick the one with the highest correlation as described above). Notice that all these models are linear in the parameters and we estimate them using the ridge method.

For each DGP we compute the ratio of the average squared error loss evaluated at the model and ridge parameter (the latter from a grid of 20 values) chosen by the corresponding model selection procedure and average squared error loss evaluated at the pair (α, h) that minimize it. We did that for sample sizes of 100, 250, 500 and 1000 observations and always kept the number of models fixed at 33.

We compute 100 replications of the above ratio for each sample size and model selection criteria under consideration, reporting the mean and standard deviation of the ratios across replications as shown in tables 1 and 2 for DGP1 and 3 and 4 for DGP2. We would expect the mean of these ratios to converge to 1 with a decreasing variance for those criteria which are asymptotic loss-efficient.

Table I.1: Means of Ratios of Average Square Losses for DGP1

	n=100	n=250	n=500	n=1000
CV-1	1.2415	1.1213	1.0330	1.0049
AIC	1.1618	1.0843	1.0241	1.0040
Mallow's	1.2330	1.0933	1.0265	1.0040
GCV	1.2176	1.1022	1.0287	1.0041
BIC	1.2990	1.1745	1.1418	1.1219
AICc	1.1747	1.0918	1.0268	1.0040

Table I.2: Standard Deviations of Ratios of Average Square Losses for DGP1

	n=100	n=250	n=500	n=1000
CV-1	0.1596	0.0648	0.0388	0.0100
AIC	0.1339	0.0611	0.0323	0.0083
Mallow's	0.1518	0.0616	0.0329	0.0083
GCV	0.1520	0.0627	0.0340	0.0083
BIC	0.1708	0.0672	0.0420	0.0289
AICc	0.1380	0.0595	0.0341	0.0083

These properties clearly holds for all except BIC whose mean lies well above the others and shows a very slow rate of convergence under DGP1 and no

tendency to converge under DGP2. This is expected as BIC is not asymptotic loss-efficient and has the same behavior as leave- d -out cross-validation with the ratio d/n tending to one.

Table I.3: Means of Ratios of Average Square Losses for DGP2

	n=100	n=250	n=500	n=1000
CV-1	1.1192	1.0723	1.0392	1.0145
AIC	1.1938	1.0851	1.0364	1.0166
Mallow's	1.1378	1.0837	1.0331	1.0167
GCV	1.1316	1.0733	1.0327	1.0166
BIC	1.1492	1.1201	1.1153	1.1256
AICc	1.1655	1.0829	1.0351	1.0167

Table I.4: Standard Deviations of Ratios of Average Square Losses for DGP2

	n=100	n=250	n=500	n=1000
CV-1	0.1499	0.0598	0.0429	0.0137
AIC	0.2376	0.0812	0.0413	0.0162
Mallow's	0.1705	0.0764	0.0384	0.0163
GCV	0.1729	0.0650	0.0392	0.0164
BIC	0.1438	0.0993	0.0489	0.0632
AICc	0.2187	0.0768	0.0401	0.0163

These property clearly holds for all except BIC whose mean lies well above the others and shows a very slow rate of convergence under DGP1 and no tendency to converge under DGP2. This is expected as BIC is not asymptotic loss-efficient and has the same behavior as leave- d -out cross-validation with the ratio d/n tending to one.

The others behave very similarly for larger samples. Notice that AIC and Mallow's C_L differ by how the error variance is estimated but are otherwise almost identical (many authors place them in the same family). Mallow's C_L uses a unique consistent estimator of the variance for every model while AIC implicitly

estimates that variance under each model. A requirement for its asymptotic optimality would be that all the models under consideration approximate arbitrarily well the unknown DGP and p_n/n tends to 0. In practice, as the smaller models never fit well and the bigger models achieve a good degree of approximation, AIC behavior is very similar to Mallows's C_L .

The optimal ridge parameter tends to zero as we increase the sample size. This is because, for computational simplicity, we keep the number of models fixed from the outset. We could have alternatively increased then with the sample size n and/or consider a much larger number of models for each n . This might keep constant or even increase the optimal ridge parameter with respect to n and would achieve even better approximations.

I.5 Conclusion and further research

We have examined the conditions under which a number of widely used model selection procedures are asymptotic loss-efficient in the sense defined above when selecting linear models under ridge estimation. This represents a crucial property that our criteria should have if we expect to use them to obtain models with good predictive ability and thus avoid in-sample over-fitting. The use of shrinkage estimation enhances the stability of the procedures (which are otherwise criticized of being instable in the sense of choosing very different models under small changes in the data). It also allows them to optimally choose larger models and thus better approximations to an unknown DGP. We also showed in the simulations how criteria with a bigger penalty (such as BIC) don't share this property as they favor smaller models. This also holds for the family of leave- d -out cross-validation statistics with d/n tending to a constant strictly greater than 0. Shao (1997) shows for the OLS case that these last two statistics are asymptotically optimal when there are several correctly specified models within those considered

and they tend to select the one with the smallest dimension. Nevertheless, in economics, finance and many other fields our models are typically misspecified and we can only hope to achieve good approximations. It is for this reason that we focus on models which are series function approximations.

Though leave- d -out with d specified as above is not asymptotically optimal under misspecification, it seems intuitive that it would be for a loss function based on $n - d$ observations instead of the whole sample. The same would hold for BIC and related criteria. It would be interesting to provide a proof for that case.

We implicitly assume a quadratic loss function, which yields the conditional expectation as optimal predictor. An asymmetric loss would yield the conditional quantile but this would require nonlinear estimation methods which are not covered in the present study. Other shrinkage methods such as the Lasso (Tibshirani 1996) or more generally Bridge regression (Frank and Friedman 1993, Fu 1998) are also non-linear. An extension to cover these cases is also subject to further research and would also require the generalization of some procedures to accommodate nonlinear estimation methods.

Finally we assume that our data was *iid* and excluded the time series case. If the errors in the DGP are martingale differences, we could analyze the conditions under which some model selection procedures are asymptotic loss-efficient. Under error correlation the criteria get a bias term that can not be consistently estimated and methods that are robust to correlation in the errors are needed. We undertake this study in the following chapters.

I.6 Proofs of theorems

We will use the following lemma in the proof of Mallow's C_L to get appropriate bounds:

Lemma I.6.1. *Let $\{w_i\}$ be a sequence of random variables, let Z^n be a $\mathbf{Z}^n$ -valued random matrix, let $\{\varphi_i : \mathbf{Z}^n \rightarrow \mathbb{R}\}$ be a sequence of measurable nonnegative functions and let $\varrho = \{(c_1, \dots, c_n) \in \mathbb{R}^n : 0 \leq c_1 \leq \dots \leq c_n \leq \xi \leq 1\}$. Also assume that $E[|\sum_{i=1}^k \varphi_i(Z^n)w_i|^q | Z^n] \leq K(\sum_{i=1}^k \varphi_i(Z^n))^p$ a.s., for all $1 \leq i \leq n$ and with $p > 1$ and K a constant. Then for all $n \geq 1$*

$$P\left\{\sup_{\varrho}\left|\sum_{i=1}^n c_i \varphi_i(Z^n)w_i\right| \geq \delta \mid Z^n\right\} \leq \delta^{-q} C \left(\sum_{i=1}^n \varphi_i(Z^n)\right)^p$$

where C is a constant depending on q and p .

Proof:

Following lemma 4.4 of Speckman (1985) we get that

$$\sup_{\varrho}\left|\sum_{i=1}^n c_i \varphi_i(Z^n)w_i\right| = \max_{1 \leq k \leq n}\left|\sum_{i=1}^k \varphi_i(Z^n)w_i\right|$$

so

$$\begin{aligned} P\left\{\sup_{\varrho}\left|\sum_{i=1}^n c_i \varphi_i(Z^n)w_i\right| \geq \delta \mid Z^n\right\} &= P\left\{\max_{1 \leq k \leq n}\left|\sum_{i=1}^k \varphi_i(Z^n)w_i\right| \geq \delta \mid Z^n\right\} \\ &\leq \delta^{-q} E\left[\left(\max_{1 \leq k \leq n}\left|\sum_{i=1}^k \varphi_i(Z^n)w_i\right|\right)^q \mid Z^n\right] \end{aligned}$$

by applying the generalized Chebychev inequality.

The lemma then holds by applying a moment inequality from Moricz (1976) to the expectation above.

Proof of theorem 3.1:

Following Li (1986) and to help in the proof of the theorem, rather than using $X_n(\alpha)$ we will work with the matrix $\Lambda_n(\alpha) = [D_n(\alpha)', \mathbf{0}']'$, where $D_n(\alpha)$ is the $(d+q) \times (d+q)$ diagonal matrix with $d_i(\alpha) = \lambda_i^{1/2}(\alpha)$ as diagonal elements, the $\lambda_i(\alpha)$'s denote the eigenvalues of $X_n(\alpha)'X_n(\alpha)$ and $\mathbf{0}$ is an $n - (d+q) \times (d+q)$ matrix of zeros. This is achieved by applying the singular value decomposition to the $X_n(\alpha)$ matrix; that is, $X_n(\alpha) = U_n(\alpha)\Lambda_n(\alpha)V_n(\alpha)$ where $U_n(\alpha)$ and $V_n(\alpha)$ are $n \times n$ and $(d+q) \times (d+q)$ orthogonal matrices respectively. $\Lambda_n(\alpha)$ has a very simple structure. Our statistics don't change as long as we pre-multiply the vectors Y, η and ε by $U_n(\alpha)'$, yielding $\tilde{Y} \equiv [\tilde{Y}_1, \dots, \tilde{Y}_n]'$, $\tilde{\eta} \equiv [\tilde{\mu}_1, \dots, \tilde{\mu}_n]'$ and $\tilde{\varepsilon} \equiv [\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_n]'$. By transforming ε in this way, the components of $\tilde{\varepsilon}$ are uncorrelated but are no longer independent, and we will have to account for that in this proof.

Write $\Gamma_{n,2}$ as

$$\Gamma_{n,2}(\alpha, h) = n^{-1}\|\varepsilon\|^2 + L_n(\alpha, h) + 2n^{-1}\varepsilon'P_n(\alpha, h)\eta \quad (\text{I.6.1})$$

$$+ 2n^{-1}(\sigma^2 \text{tr} M_n(\alpha, h) - \varepsilon' M_n(\alpha, h)\varepsilon) \quad (\text{I.6.2})$$

$$+ 2n^{-1}(\hat{\sigma}_n^2 - \sigma^2) \text{tr} M_n(\alpha, h) \quad (\text{I.6.3})$$

where $P_n(\alpha, h) \equiv I - M_n(\alpha, h)$.

Asymptotic loss-efficiency for $\Gamma_{n,2}$ will hold if

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} |L_n(\alpha, h)/R_n(\alpha, h; Z^n) - 1| = o_p(1), \quad (\text{I.6.4})$$

and the last term in (6.1) as well as in (6.2) and (6.3) are of smaller order of magnitude in probability than the conditional risk $R_n(\alpha, h; Z^n)$ uniformly with respect to α and h , that is

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\varepsilon' P_n(\alpha, h)\eta| / R_n(\alpha, h; Z^n) = o_p(1) \quad (\text{I.6.5})$$

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} |\hat{\sigma}_n^2 - \sigma^2| n^{-1} \text{tr} M_n(\alpha, h) / R_n(\alpha, h; Z^n) = o_p(1) \quad (\text{I.6.6})$$

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\sigma^2 \text{tr} M_n(\alpha, h) - \varepsilon' M_n(\alpha, h)\varepsilon| / R_n(\alpha, h; Z^n) = o_p(1) \quad (\text{I.6.7})$$

where all the limits above hold for all sequences $\{Z^n\}$ a.s.

This will imply that

$$\frac{L_n(\alpha^{C_L}, h^{C_L})}{\inf_{\alpha \in A_n} \inf_{h \geq 0} L_n(\alpha, h)} \longrightarrow 1 \quad \text{in prob.}$$

and

$$\frac{R_n(\alpha^{C_L}, h^{C_L}; Z^n)}{\inf_{\alpha \in A_n} \inf_{h \geq 0} R_n(\alpha, h; Z^n)} \longrightarrow 1 \quad \text{in prob.,}$$

again for all sequences $\{Z^n\}$ in a set of probability 1. Before proving (6.4)-(6.7) we define

$$\tilde{R}_n(\alpha, h; Z^n) \equiv \tilde{\Delta}_n(\alpha, h) + \sigma^2 n^{-1} \text{tr} \tilde{M}_n(\alpha, h)$$

and

$$\tilde{P}_n(\alpha, h) \equiv I_n - \tilde{M}_n(\alpha, h),$$

where $\tilde{\Delta}_n(\alpha, h) = n^{-1} \|\tilde{\eta} - \tilde{M}_n(\alpha, h)\|^2$ and

$$\tilde{M}_n(\alpha, h) \equiv \Lambda_n(\alpha) [\Lambda_n(\alpha)' \Lambda_n(\alpha) + h I_q]^{-1} \Lambda_n(\alpha)'.$$

We first verify condition (6.5).

$$\text{Define } \tilde{B}_n(\alpha, h; Z^n) \equiv \sum_{i=1}^n \tilde{\mu}_i^2(\lambda_i(\alpha) + h)^2,$$

as $n\tilde{R}_n(\alpha, h; Z^n) \geq h^2 \tilde{B}_n(\alpha, h; Z^n)$, we get that

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\varepsilon' P_n(\alpha, h) \eta|}{n R_n(\alpha, h; Z^n)} = \sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\tilde{\varepsilon}' \tilde{P}_n(\alpha, h) \tilde{\eta}|}{n \tilde{R}_n(\alpha, h; Z^n)} \quad (\text{I.6.8})$$

$$\leq \sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\sum_{i=1}^n \tilde{\varepsilon}_i \tilde{\mu}_i(\lambda_i(\alpha) + h)^{-1}|}{\tilde{B}_n^{1/2}(\alpha, h; Z^n) (n \tilde{R}_n(\alpha, h; Z^n))^{1/2}} \quad (\text{I.6.9})$$

so for (6.5) to hold it is enough to show that (6.9) tends to zero in probability.

For each n , let $I_1(j) = \{1, 2, \dots, j\}$ and $I_2(j) = \{j+1, \dots, n\}$. Recall that $\iota_n(\alpha)$ is the largest i such that $\lambda_i(\alpha) > 0$.

Let

$$\tilde{Q}_n(\alpha; Z^n) \equiv \inf_{h \geq 0} n\tilde{R}_n(\alpha, h; Z^n)$$

and

$$\tilde{V}_n(\alpha, h; Z^n) \equiv \sum_{i=1}^n \lambda_i^2(\alpha) (\lambda_i(\alpha) + h)^{-2}$$

(observe that $n\tilde{R}_n(\alpha, h; Z^n) \geq \sigma^2 \tilde{V}_n(\alpha, h; Z^n)$).

(6.9) will be $o_p(1)$ if for any natural k and for $l = 1, 2$

$$\sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \leq h} \frac{|\sum_{i \in I_l(k)} \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha) + h)^{-1}|}{\tilde{B}_n^{1/2}(\alpha, h; Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} = o_p(1) \quad (\text{I.6.10})$$

and $\forall \epsilon > 0 \exists c_1(\epsilon)$ and $c_2(\epsilon)$ such that $c_1(\epsilon) \rightarrow 0, c_2(\epsilon) \rightarrow 0, \epsilon \rightarrow \infty$,

and

$$P \left\{ \sup_{\alpha \in A_n} \sup_{j=k, \dots, \iota_n(\alpha)} \sup_{\lambda_{j+1}(\alpha) \leq h \leq \lambda_j(\alpha)} \frac{|\sum_{i \in I_l(j)} \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha) + h)^{-1}|}{\tilde{B}_n^{1/2}(\alpha, h; Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} > \varepsilon | Z^n \right\} \leq c_l(\epsilon) \quad (\text{I.6.11})$$

for $l = 1, 2$

Proof of (6.10)

When $l = 1$ the left side of (6.10) doesn't exceed

$$\sup_{\alpha \in A_n} Q_n^{-1/2}(\alpha; Z^n) \max_{1 \leq i \leq k} |\tilde{\varepsilon}_i| \sum_{i=1}^k \sup_{h \geq \lambda_k(\alpha)} |\tilde{\mu}_i| (\lambda_i(\alpha) + h)^{-1} / \tilde{B}_n^{1/2}(\alpha, h; Z^n) \quad (\text{I.6.12})$$

$$\leq \sup_{\alpha \in A_n} Q_n^{-1/2}(\alpha; Z^n) \max_{1 \leq i \leq k} |\tilde{\varepsilon}_i| k \quad (\text{I.6.13})$$

which tends to 0 in probability if $\inf_{\alpha \in A_n} \inf_{h \geq 0} n\tilde{R}_n(\alpha, h; Z^n)$

$= \inf_{\alpha \in A_n} \inf_{h \geq 0} nR_n(\alpha, h; Z) \rightarrow \infty$ a.s, since $\max_{1 \leq i \leq k} |\tilde{\varepsilon}_i| = O_p(1)$.

When $l = 2$, it suffices to show that $\forall \varepsilon > 0$

$$P \left\{ \sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \leq h} \frac{|\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i h (\lambda_i(\alpha) + h)^{-1}|}{h \tilde{B}_n^{1/2}(\alpha, h; Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \varepsilon | Z^n \right\} \rightarrow 0 \quad a.s. \quad (\text{I.6.14})$$

Since $h^2 \tilde{B}_n(\alpha, h; Z^n)$ is non-decreasing in h , the left-side of (6.14) doesn't exceed

$$P \left\{ \sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \leq h} \frac{|\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i h (\lambda_i(\alpha) + h)^{-1}|}{\lambda_k(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \varepsilon | Z^n \right\} \leq \quad (\text{I.6.15})$$

$$P \left\{ \sup_{\alpha \in A_n} \frac{|\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i \lambda_k(\alpha) (\lambda_i(\alpha) + \lambda_k(\alpha))^{-1}|}{\lambda_k(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \varepsilon/2 | Z^n \right\} + \quad (\text{I.6.16})$$

$$P \left\{ \sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \leq h} \frac{|\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i [h (\lambda_i(\alpha) + h)^{-1} - \lambda_k(\alpha) (\lambda_i(\alpha) + \lambda_k(\alpha))^{-1}]|}{\lambda_k(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \varepsilon/2 | Z^n \right\}. \quad (\text{I.6.17})$$

By applying the generalized Chebychev inequality, (6.16) is no greater than

$$\sum_{\alpha \in A_n} \frac{E \left[\left(\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha) + \lambda_k(\alpha))^{-1} \right)^{2m} | Z^n \right]}{\left(\frac{1}{2} \varepsilon \right)^{2m} \tilde{B}_n^m(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^m(\alpha; Z^n)}. \quad (\text{I.6.18})$$

Now we have to bound the expectation in (6.18). For that purpose let $v_i(\alpha) \equiv \tilde{\mu}_i (\lambda_i(\alpha) + \lambda_k(\alpha))^{-1}$ and note that $\tilde{\varepsilon}_i = U_i(\alpha)' \varepsilon = u_{1i}(\alpha) \varepsilon_1 + \dots + u_{ni}(\alpha) \varepsilon_n$ where $U_i(\alpha) \equiv [u_{1i}(\alpha), \dots, u_{ni}(\alpha)]'$ denotes the i^{th} column of the orthogonal matrix $U(\alpha)$ and ε is our original vector of *iid* errors.

So we get that

$$\begin{aligned} & E \left[\left(\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha) + \lambda_k(\alpha))^{-1} \right)^{2m} | Z^n \right] \\ &= E \left[\left(\sum_{i=k+1}^n (u_{1i}(\alpha) \varepsilon_1 + \dots + u_{ni}(\alpha) \varepsilon_n) v_i(\alpha) \right)^{2m} | Z^n \right] \end{aligned}$$

$$\begin{aligned}
&= E \left[\left(\varepsilon_1 \sum_{i=k+1}^n u_{1i} v_i(\alpha) + \dots + \varepsilon_n \sum_{i=k+1}^n u_{ni} v_i(\alpha) \right)^{2m} | Z^n \right] \\
&= E \left[\left(\sum_{j=1}^n \varepsilon_j \phi_j(\alpha) \right)^{2m} | Z^n \right] \leq C_1 \left(\sum_{j=1}^n \phi_j^2(\alpha) \right)^m,
\end{aligned}$$

where $\phi_j(\alpha) \equiv \sum_{i=k+1}^n u_{ji}(\alpha) v_i(\alpha)$, C_1 is a constant depending on $2m$ and the last inequality holds by applying theorem 2 of Whittle (1960) (recall the moment bound for the ε_i 's). Now,

$$\begin{aligned}
\sum_{j=1}^n \phi_j^2 &= \sum_{j=1}^n \left(\sum_{i=k+1}^n u_{ji}(\alpha) v_i(\alpha) \right)^2 \\
&= \sum_{j=1}^n \sum_{i=k+1}^n u_{ji}^2(\alpha) v_i^2(\alpha) + \sum_{j=1}^n \sum_{i \neq i'}^n u_{ji}(\alpha) u_{ji'}(\alpha) v_i(\alpha) v_{i'}(\alpha) \\
&= \sum_{i=k+1}^n v_i^2(\alpha) \sum_{j=1}^n u_{ji}^2(\alpha) + 2 \sum_{i \neq i'}^n v_i(\alpha) v_{i'}(\alpha) \sum_{j=1}^n u_{ji}(\alpha) u_{ji'}(\alpha) \\
&= \sum_{i=k+1}^n v_i^2(\alpha) \\
&= \sum_{i=k+1}^n \tilde{\mu}_i^2(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2},
\end{aligned}$$

which holds by the orthogonality of $U(\alpha)$.

We therefore get that (6.18) is no greater than

$$C_1 \left(\frac{1}{2} \epsilon \right)^{-2m} \sum_{\alpha \in A_n} \frac{[(\sum_{i=k+1}^n \tilde{\mu}_i^2(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2})^m]}{\tilde{B}_n^m(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^m(\alpha; Z^n)}. \quad (\text{I.6.19})$$

As $\sum_{i=k+1}^n \tilde{\mu}_i^2(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2} / (\sum_{i=1}^n \tilde{\mu}_i^2(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2}) \leq 1$, where $\sum_{i=1}^n \tilde{\mu}_i^2(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2} = \tilde{B}_n^m(\alpha, \lambda_k(\alpha); Z^n)$ we get that (6.19) is no greater than

$$C_1 \left(\frac{1}{2} \epsilon \right)^{-2m} \sum_{\alpha \in A_n} \tilde{Q}_n^{-m}(\alpha; Z^n) = C_1 \left(\frac{1}{2} \epsilon \right)^{-2m} \sum_{\alpha \in A_n} [\inf_{h \geq 0} n \tilde{R}_n(\alpha, h; Z^n)]^{-m}, \quad (\text{I.6.20})$$

which tends to zero if $\sum_{\alpha \in A_n} [\inf_{h \geq 0} n \tilde{R}_n(\alpha, h; Z^n)]^{-m} \rightarrow 0$ a.s.

To bound (6.17) we fix j and write

$$h(\lambda_i(\alpha) + h)^{-1} - \lambda_k(\alpha)(\lambda_i(\alpha) + \lambda_k(\alpha))^{-1} = (\lambda_i(\alpha) + \lambda_k(\alpha))^{-1} (h - \lambda_k(\alpha))(\lambda_i(\alpha) + h)^{-1}$$

Note that for $i \geq k+1$ and $h \geq \lambda_k(\alpha)$ we have that $(h - \lambda_k(\alpha))\lambda_i(\alpha)(\lambda_i(\alpha) + h)^{-1}$ is non-increasing in i and no greater than $\lambda_k(\alpha)$.

We apply lemma 6.1 with $\varphi_i(Z^n) = \tilde{\mu}_i^2(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2}$, $c_i = (h - \lambda_k(\alpha))\lambda_i(\alpha)(\lambda_i(\alpha) + h)^{-1}$, $w_i = \tilde{\varepsilon}_i$, $\xi = \lambda_k(\alpha)$ and $\delta = \frac{\varepsilon}{2}\lambda_k(\alpha)\tilde{B}_n^{1/2}(\alpha, \lambda_k(\alpha); Z^n) \times \tilde{Q}_n^{1/2}(\alpha; Z^n)$. Note that the c_i 's here depend on h .

Lemma 6.1 gives that (6.17) is bounded by

$$B_1\left(\frac{1}{2}\varepsilon\right)^{-2m} \sum_{\alpha \in A_n} \frac{\left[\left(\sum_{i=k+1}^n \tilde{\mu}_i^2(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2}\right)^m\right]}{\tilde{B}_n^m(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^m(\alpha; Z^n)}, \quad (\text{I.6.21})$$

thus achieving a similar bound as for (6.16) and the same reasoning applies.

This concludes the proof of (6.10).

Proof of (6.11)

For $l = 1$ $\tilde{B}_n(\alpha, h)$ is non-increasing in h , so the left-side of (6.11) doesn't exceed

$$\sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} P \left\{ \sup_{\lambda_{j+1}(\alpha) \leq h \leq \lambda_j(\alpha)} \frac{\left| \sum_{i=1}^j \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha) + h)^{-1} \right|}{\tilde{B}_n^{1/2}(\alpha, \lambda_j(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \varepsilon |Z^n| \right\} \quad (\text{I.6.22})$$

$$\leq \sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} P \left\{ \frac{\left| \sum_{i=1}^j \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha) + \lambda_j(\alpha))^{-1} \right|}{\tilde{B}_n^{1/2}(\alpha, \lambda_j(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \frac{\varepsilon}{2} |Z^n| \right\} + \quad (\text{I.6.23})$$

$$\begin{aligned} \sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} P \left\{ \sup_{\lambda_{j+1}(\alpha) \leq h \leq \lambda_j(\alpha)} \frac{\left| \sum_{i=1}^j \tilde{\varepsilon}_i \tilde{\mu}_i [(\lambda_i(\alpha) + h)^{-1} - (\lambda_i(\alpha) + \lambda_j(\alpha))^{-1}] \right|}{\tilde{B}_n^{1/2}(\alpha, \lambda_j(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \right. \\ \left. \geq \frac{\varepsilon}{2} |Z^n| \right\} \end{aligned} \quad (\text{I.6.24})$$

By the generalized Chebychev inequality, (6.23) doesn't exceed

$$\begin{aligned} & \sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} \left(\frac{1}{2} \epsilon \tilde{B}_n^{1/2}(\alpha, \lambda_j(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n) \right)^{-2m^*} \\ & \times E \left(\sum_{i=1}^j \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha) + \lambda_j(\alpha))^{-1} | Z^n \right)^{2m^*} \end{aligned} \quad (\text{I.6.25})$$

By using the same argument as for bounding the expectation in (6.18) and applying Whittle's theorem we get that

$$E \left[\left(\sum_{i=1}^j \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha) + \lambda_j(\alpha))^{-1} \right)^{2m^*} | Z^n \right] \leq C_2 \left(\sum_{i=1}^j \tilde{\mu}_i^2 (\lambda_i(\alpha) + \lambda_j(\alpha))^{-2} | Z^n \right)^{m^*}$$

As $\left(\sum_{i=1}^j \tilde{\mu}_i^2 (\lambda_i(\alpha) + \lambda_j(\alpha))^{-2} | Z^n \right)^{m^*} / \tilde{B}_n^{-m^*}(\alpha, \lambda_j(\alpha); Z^n) \leq 1$, (6.25) is no greater than

$$\begin{aligned} (16)^{-m^*} \epsilon^{-2m^*} C_2 \sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} \tilde{Q}_n^{-m^*}(\alpha; Z^n) &= K_1(\epsilon) \sum_{j=k}^{\iota_n(\alpha)} \tilde{Q}_n^{-m^*}(\alpha; Z^n) \\ &= K_1(\epsilon) \sum_{\alpha \in A_n} \frac{[\iota_n(\alpha) - k]}{\tilde{Q}_n^{m^*}(\alpha; Z^n)} \end{aligned}$$

Observe that $K_1(\epsilon) = (16)^{-m^*} \epsilon^{-2m^*} C_2$ is decreasing in ϵ , so (6.23) will be $o_p(1)$ if condition A.4 holds.

To bound (6.24) we fix j and write

$$(\lambda_i(\alpha) + h)^{-1} - (\lambda_i(\alpha) + \lambda_j(\alpha))^{-1} = (\lambda_i(\alpha) + \lambda_j(\alpha))^{-1} (\lambda_j(\alpha) - h) (\lambda_i(\alpha) + h)^{-1}$$

Note that for $\lambda_{j+1}(\alpha) \leq h \leq \lambda_j$ and $i \leq j$ we have that $(\lambda_j(\alpha) - h)(\lambda_i(\alpha) + h)^{-1}$ is non-decreasing in i and is no greater than one.

We apply lemma 6.1 with $\varphi_i(Z^n) = \mu_i^2(\lambda_i(\alpha) + \lambda_j(\alpha))^{-2}$, $c_i = (\lambda_j(\alpha) - h)(\lambda_i(\alpha) + h)^{-1}$, $w_i = \tilde{\varepsilon}_i \xi = 1$ and $\delta = \frac{\epsilon}{2} \tilde{B}_n^{1/2}(\alpha, \lambda_j(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)$. Note that

the c_i 's here depend again on h .

Lemma 6.1 then gives that (6.24) is bounded by

$$\begin{aligned} \sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} \left(\frac{\varepsilon}{2}\right)^{-2m^*} \tilde{B}_n^{-m^*}(\alpha, \lambda_j(\alpha); Z^n) \tilde{Q}_n^{-m^*}(\alpha; Z^n) B_2 \left(\sum_{i=1}^j \tilde{\mu}_i^2(\lambda_i(\alpha) + \lambda_j(\alpha))^{-2} \right)^{m^*} \\ \leq K_2(\epsilon) \sum_{\alpha \in A_n} \frac{[\iota_n(\alpha) - k]}{\tilde{Q}_n^{m^*}(\alpha; Z^n)} \end{aligned}$$

where the last inequality holds by using the same reasoning as for bounding (6.21).

The proof of (6.11) for $l = 1$ is now complete.

Now when $l = 2$, since $h^2 \tilde{B}_n(\alpha, h; Z)$ is non-decreasing in h , we get that (6.11) is no greater than

$$\sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} P \left\{ \sup_{\lambda_{j+1}(\alpha) \leq h \leq \lambda_j(\alpha)} \frac{|\sum_{i=j+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i h(\lambda_i(\alpha) + h)^{-1}|}{\lambda_{j+1}(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_{j+1}(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \epsilon |Z^n| \right\} \quad (\text{I.6.26})$$

which doesn't exceed

$$\sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} P \left\{ \frac{|\sum_{i=1}^j \tilde{\varepsilon}_i \tilde{\mu}_i \lambda_{j+1}(\alpha) (\lambda_i(\alpha) + \lambda_{j+1}(\alpha))^{-1}|}{\lambda_{j+1}(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_{j+1}(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \frac{\epsilon}{2} |Z^n| \right\} + \quad (\text{I.6.27})$$

$$\begin{aligned} \sum_{\alpha \in A_n} \sum_{j=k}^{\iota_n(\alpha)} P \left\{ \sup_{\lambda_{j+1}(\alpha) \leq h \leq \lambda_j(\alpha)} \frac{|\sum_{i=1}^j \tilde{\varepsilon}_i \tilde{\mu}_i [h(\lambda_i(\alpha) + h)^{-1} \right. \\ \left. - \frac{\lambda_{j+1}(\alpha) (\lambda_i(\alpha) + \lambda_{j+1}(\alpha))^{-1}}{\lambda_{j+1}(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_{j+1}(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)}]|}{\lambda_{j+1}(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_{j+1}(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} \geq \frac{\epsilon}{2} |Z^n| \right\} \quad (\text{I.6.28}) \end{aligned}$$

Using the generalized Chebychev inequality for the first term and then Whittle's theorem as before we obtain a similar bound than for (6.23); that is, (6.28) is bounded by

$$K_3(\epsilon) \sum_{\alpha \in A_n} \frac{[\iota_n(\alpha) - k]}{\tilde{Q}_n^{m^*}(\alpha; Z^n)} \quad (\text{I.6.29})$$

where $K_3(\epsilon)$ depends on ϵ and m^* and $K_3(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow \infty$

By applying lemma 4.1 and Whittle's theorem to (6.29) we can also bound it by

$$K_4(\epsilon) \sum_{\alpha \in A_n} \frac{[\iota_n(\alpha) - k]}{\tilde{Q}_n^{m^*}(\alpha; Z^n)} \quad (\text{I.6.30})$$

where $K_4(\epsilon)$ depends on ϵ and m^* and $K_4(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow \infty$

For that purpose observe that $h(\lambda_i(\alpha) + h)^{-1} - \lambda_{j+1}(\alpha)(\lambda_i(\alpha) + \lambda_{j+1}(\alpha))^{-1} = (\lambda_i(\alpha) + \lambda_{j+1}(\alpha))^{-1} \lambda_i(\alpha)(h - \lambda_{j+1}(\alpha)(\lambda_i(\alpha) + h)^{-1}$

Note that for $\lambda_{j+1}(\alpha) \leq h \leq \lambda_j(\alpha)$ and $i \geq j + 1$ we have that $\lambda_i(\alpha)(h - \lambda_{j+1}(\alpha)(\lambda_i(\alpha) + h)^{-1}$ is non-increasing in i and is no greater than $\lambda_{j+1}(\alpha)$.

We apply lemma 6.1 with $\varphi_i(Z^n) = \mu_i^2(\lambda_i(\alpha) + \lambda_{j+1}(\alpha))^{-2}$, $c_i = \lambda_i(\alpha)(h - \lambda_{j+1}(\alpha)(\lambda_i(\alpha) + h)^{-1}$, $w_i = \tilde{\varepsilon}_i$, $\xi = \lambda_{j+1}(\alpha)$ and $\delta = \frac{1}{2}\epsilon\lambda_{j+1}(\alpha)\tilde{B}_n^{1/2}(\alpha, \lambda_{j+1}(\alpha); Z^n) \times \tilde{Q}_n^{1/2}(\alpha; Z^n)$.

We have now established (6.11) for $l = 2$. The proof of condition (6.5) in now complete.

To verify condition (6.6) notice that

$$\sup_{\alpha \in n} \sup_{h \geq 0} \frac{|\hat{\sigma}_n^2 - \sigma^2| \text{tr} M_n(\alpha, h)}{n R_n(\alpha, h; Z^n)} = \sup_{\alpha \in n} \sup_{h \geq 0} \frac{|\hat{\sigma}_n^2 - \sigma^2| \text{tr} \tilde{M}_n(\alpha, h)}{n \tilde{R}_n(\alpha, h; Z^n)} \quad (\text{I.6.31})$$

$$= \sup_{\alpha \in n} \sup_{h \geq 0} \frac{|\hat{\sigma}_n^2 - \sigma^2| \sum_{i=1}^n \lambda_i(\alpha) (\lambda_i(\alpha) + h)}{n \tilde{R}_n(\alpha, h; Z^n)} \quad (\text{I.6.32})$$

As $\sum_{i=1}^n \lambda_i(\alpha) (\lambda_i(\alpha) + h) / n \tilde{R}_n(\alpha, h; Z^n) \leq 1$, (6.32) will tend to zero in probability as long as $|\hat{\sigma}_n^2 - \sigma^2| = o_p(1)$.

Now we verify condition (6.7), begin observing that

$$\sup_{\alpha \in n} \sup_{h \geq 0} \frac{|\sigma^2 \text{tr} M_n(\alpha, h) - \varepsilon' M_n(\alpha, h) \varepsilon|}{n R_n(\alpha, h; Z^n)} = \sup_{\alpha \in n} \sup_{h \geq 0} \frac{|\sigma^2 \text{tr} \tilde{M}_n(\alpha, h) - \tilde{\varepsilon}' \tilde{M}_n(\alpha, h) \tilde{\varepsilon}|}{n \tilde{R}_n(\alpha, h; Z^n)} \quad (\text{I.6.33})$$

$$\leq \sup_{\alpha \in n} \sup_{h \geq 0} \frac{|\sum_{i=1}^n (\sigma^2 - \tilde{\varepsilon}_i^2) \lambda_i(\alpha) (\lambda_i(\alpha) + h)^{-1}|}{\tilde{V}_n^{1/2}(\alpha, h; Z^n) (n \tilde{R}_n(\alpha, h; Z^n))^{1/2}} \quad (\text{I.6.34})$$

Compare (6.35) with the left side of (6.10). Making the correspondences $\tilde{\varepsilon}_i^2 - \sigma^2$ to $\tilde{\varepsilon}_i$, $\lambda_i(\alpha)$ to μ_i and $\tilde{V}_n^{1/2}(\alpha, h; Z^n)$ to $\tilde{B}_n^{1/2}(\alpha, h; Z^n)$ and arguing as in the proof of (6.10), we get that (6.35) is $o_p(1)$.

The proof of (6.7) is complete.

To check condition (6.4) notice that

$$\begin{aligned} L_n(\alpha, h) - R_n(\alpha, h; Z^n) &= n^{-1} \varepsilon' M_n(\alpha, h) \varepsilon - n^{-1} \sigma^2 \text{tr} M_n^2(\alpha, h) \\ &\quad - 2n^{-1} \eta' [I_n - M_n(\alpha, h)] M_n(\alpha, h) \varepsilon \\ &= n^{-1} \tilde{\varepsilon}' \tilde{M}_n(\alpha, h) \tilde{\varepsilon} - n^{-1} \sigma^2 \text{tr} \tilde{M}_n^2(\alpha, h) \\ &\quad - 2n^{-1} \eta' [I_n - \tilde{M}_n(\alpha, h)] \tilde{M}_n(\alpha, h) \tilde{\varepsilon} \\ &= n^{-1} \sum_{i=1}^n \tilde{\varepsilon}_i^2 \lambda_i^2(\alpha) (\lambda_i(\alpha) + h)^{-2} - n^{-1} \sigma^2 \sum_{i=1}^n \lambda_i^2(\alpha) (\lambda_i(\alpha) + h)^{-2} \end{aligned}$$

$$-2n^{-1}\sum_{i=1}^n\tilde{\varepsilon}_i\tilde{\mu}_i h\lambda_i(\alpha)(\lambda_i(\alpha)+h)^{-2}$$

So it is enough to show that

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\sum_{i=1}^n \tilde{\varepsilon}_i \tilde{\mu}_i \lambda_i(\alpha)(\lambda_i(\alpha)+h)^{-2}|}{\tilde{B}_n^{1/2}(\alpha, h; Z)(n\tilde{R}_n(\alpha, h; Z^n))^{1/2}} = o_p(1) \quad (\text{I.6.35})$$

and

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\sum_{i=1}^n (\tilde{\varepsilon}_i^2 - \sigma^2) \lambda_i^2(\alpha)(\lambda_i(\alpha)+h)^{-2}|}{\tilde{V}_n^{1/2}(\alpha, h; Z^n)(n\tilde{R}_n(\alpha, h; Z^n))^{1/2}} = o_p(1) \quad (\text{I.6.36})$$

To verify (6.35) we have to show that for all fixed natural k and $l = 1, 2$

$$\sup_{\alpha \in A_n} \sup_{h \geq \lambda_k(\alpha)} \frac{|\sum_{i \in I_l(k)} \tilde{\varepsilon}_i \tilde{\mu}_i \lambda_i(\alpha)(\lambda_i(\alpha)+h)^{-2}|}{\tilde{B}_n^{1/2}(\alpha, h; Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} = o_p(1) \quad (\text{I.6.37})$$

and $\forall \epsilon > 0 \exists c_1^*(\epsilon)$ and $c_2^*(\epsilon)$ such that $c_1^*(\epsilon) \rightarrow 0$, $c_2^*(\epsilon) \rightarrow 0$ and $\epsilon \rightarrow \infty$

and

$$P \left\{ \sup_{\alpha \in A_n} \sup_{j=k, \dots, \iota_n(\alpha)} \sup_{\lambda_{j+1}(\alpha) \leq h \leq \lambda_j(\alpha)} \frac{|\sum_{i \in I_l(j)} \tilde{\varepsilon}_i \tilde{\mu}_i \lambda_i(\alpha)(\lambda_i(\alpha)+h)^{-2}|}{\tilde{B}_n^{1/2}(\alpha, h; Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} > \epsilon \mid Z^n \right\} \quad (\text{I.6.38})$$

$\leq c_l^*(\epsilon)$ for $l=1,2$.

We first proof (6.37)

For $l = 1$, (6.37) doesn't exceed

$$\sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \geq 0} \frac{|\sum_{i \in I_1(k)} \tilde{\varepsilon}_i \tilde{\mu}_i (\lambda_i(\alpha)+h)^{-1}|}{\tilde{B}_n^{1/2}(\alpha, h; Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} = o_p(1) \quad (\text{I.6.39})$$

since $\lambda_i(\alpha)(\lambda_i(\alpha) + h)^{-1} < 1$ so the proof is exactly the same as in (6.10)

For $l = 2$ we get

$$P\left\{\sup_{\alpha \in A_n} \sup_{h \geq \lambda_k(\alpha)} \frac{|\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i h \lambda_i(\alpha)(\lambda_i(\alpha) + h)^{-2}|}{h \tilde{B}_n^{1/2}(\alpha, h; Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} > \epsilon | Z^n\right\} \quad (\text{I.6.40})$$

which is no greater than

$$P\left\{\sup_{\alpha \in A_n} \sup_{h \geq \lambda_k(\alpha)} \frac{|\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i h \lambda_i(\alpha)(\lambda_i(\alpha) + h)^{-2}|}{\lambda_k(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} > \epsilon | Z^n\right\} \quad (\text{I.6.41})$$

$$\leq P\left\{\sup_{\alpha \in A_n} \frac{|\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i \lambda_k(\alpha) \lambda_i(\alpha)(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2}|}{\lambda_k(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} > \frac{\epsilon}{2} | Z^n\right\} + \quad (\text{I.6.42})$$

$$P\left\{\sup_{\alpha \in A_n} \sup_{h \geq \lambda_k(\alpha)} \frac{|\sum_{i=k+1}^n \tilde{\varepsilon}_i \tilde{\mu}_i h \lambda_i(\alpha)[h(\lambda_i(\alpha) + h)^{-2} - \lambda_k(\alpha)(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2}]|}{\lambda_k(\alpha) \tilde{B}_n^{1/2}(\alpha, \lambda_k(\alpha); Z^n) \tilde{Q}_n^{1/2}(\alpha; Z^n)} > \frac{\epsilon}{2} | Z^n\right\} \quad (\text{I.6.43})$$

By the generalized Chebychev inequality and because $\lambda_i(\alpha)(\lambda_i(\alpha) + \lambda_k(\alpha)) < 1$ (6.42) doesn't exceed

$$C_3 \left(\frac{1}{4}\epsilon\right)^{-2m} \sum_{\alpha \in A_n} \tilde{Q}_n^{-m}(\alpha; Z^n) = C_3 \left(\frac{1}{4}\epsilon\right)^{-2m} \sum_{\alpha \in A_n} \left[\inf_{h \geq 0} n R_n(\alpha, h; Z^n)\right]^{-m} \quad (\text{I.6.44})$$

which tends to zero if $\sum_{\alpha \in A_n} \left[\inf_{h \geq 0} n R_n(\alpha, h; Z^n)\right]^{-m} \rightarrow 0 \quad a.s.$

The second term has a similar bound by lemma 6.1.

To apply lemma 6.1 observe that

$$\begin{aligned} h(\lambda_i(\alpha) + h)^{-2} - \lambda_k(\alpha)(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2} = \\ (\lambda_i(\alpha) + \lambda_k(\alpha))^{-2} (h - \lambda_k(\alpha))(\lambda_i^2(\alpha) - h\lambda_k(\alpha))(\lambda_i(\alpha) + h)^{-2} \end{aligned}$$

Note that for $i \geq k+1$ and $h \geq \lambda_k(\alpha)$ we have that $(h - \lambda_k(\alpha))(\lambda_i^2(\alpha) - h\lambda_k(\alpha))(\lambda_i(\alpha) + h)^{-2}$ is non-decreasing in i and no greater than $\lambda_k(\alpha)$ so we can apply lemma 1 with $\varphi_i(Z^n) = \tilde{\mu}\lambda_i(\alpha)(\lambda_i(\alpha) + \lambda_k(\alpha))^{-2}$, $c_i = (h - \lambda_k(\alpha))(\lambda_i^2(\alpha) - h\lambda_k(\alpha))(\lambda_i(\alpha) + h)^{-2}$, $w_i = \tilde{\varepsilon}_i$ and $\xi = \lambda_k(\alpha)$.

The proof of (6.35) is now complete.

The proof of (6.36) follows a similar argument than that of (6.7).

Proof of theorem 3.2:

As stated in section 3.2 leave-1-out cross-validation is equivalent to applying Mallows's C_L to the delete-one estimate, so we show that under A.1-A.5 and A.7 CV_{-1} is asymptotic loss-efficient for the loss $L_n^d(\alpha, h)$. The asymptotic equivalence between the ridge estimator and the delete-one holds under A.1-A.6 and A.8 and follows the same argument as in Li (1987), where the proof can be found.

Mallows's C_L applied to $\eta^d(\alpha, h)$ equals

$$\begin{aligned} & n^{-1} \|Y - M_n^d(\alpha, h)Y\|^2 + 2n^{-1}\sigma^2 \text{tr} M_n^d(\alpha, h) \\ &= n^{-1} \|\varepsilon\|^2 + L_n^d(\alpha, h) + 2n^{-1}\varepsilon' P_n^d(\alpha, h)\eta + 2n^{-1}(\hat{\sigma}_n^2 - \sigma^2) \text{tr} M_n^d(\alpha, h) + \\ & \quad 2n^{-1}(\sigma^2 \text{tr} M_n^d(\alpha, h) - \varepsilon' M_n^d(\alpha, h)\varepsilon) \end{aligned}$$

where $P_n^d(\alpha, h) \equiv I_n - M_n^d(\alpha, h)$ and $\text{tr} M_n^d(\alpha, h) = 0$.

So asymptotic loss-efficiency will follow if we show that

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} |L_n^d(\alpha, h)/R_n^d(\alpha, h; Z^n) - 1| = o_p(1) \quad (\text{I.6.45})$$

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\varepsilon' P_n^d(\alpha, h) \eta| / R_n^d(\alpha, h; Z^n) = o_p(1) \quad (\text{I.6.46})$$

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} |\hat{\sigma}_n^2 - \sigma^2| n^{-1} \text{tr} M_n^d(\alpha, h) / R_n^d(\alpha, h; Z^n) = o_p(1) \quad (\text{I.6.47})$$

and

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\sigma^2 \text{tr} M_n^d(\alpha, h) - \varepsilon' M_n^d(\alpha, h) \varepsilon| / R_n^d(\alpha, h; Z^n) = o_p(1) \quad (\text{I.6.48})$$

where all the limits above hold for all sequences $\{Z^n\}$ a.s.

We first show (6.46)

For that purpose notice that

$$\begin{aligned} nR_n^d(\alpha, h; Z^n) &= \eta' [I_n - M_n^d(\alpha, h)]' [I_n - M_n^d(\alpha, h)] \eta + \sigma^2 \text{tr} M_n^d(\alpha, h) M_n^d(\alpha, h)' \\ &= \eta' (I_n - D_n(\alpha, h) [M_n(\alpha, h) - I_n] - I_n)' (I_n - D_n(\alpha, h) [M_n(\alpha, h) - I_n] - I_n) \eta \\ &\quad + \sigma^2 \text{tr} (D_n(\alpha, h) [M_n(\alpha, h) - I_n] + I_n)' (D_n(\alpha, h) [M_n(\alpha, h) - I_n] + I_n) \\ &= \eta' (M_n(\alpha, h) - I_n) D_n^2(\alpha, h) (M_n(\alpha, h) - I_n) \eta \\ &\quad + \sigma^2 \text{tr} ([M_n(\alpha, h) - I_n] D_n^2(\alpha, h) [M_n(\alpha, h) - I_n] \\ &\quad + 2D_n(\alpha, h) [M_n(\alpha, h) - I_n] + I_n) \\ &\geq \eta' [M_n(\alpha, h) - I_n]^2 \eta + \sigma^2 \text{tr} M_n^2(\alpha, h) \end{aligned}$$

as the elements along the diagonal of $D_n(\alpha, h)$ are all greater or equal than one.

Also observe that

$$\begin{aligned} \varepsilon' P_n^d(\alpha, h) \eta &= \varepsilon' [M_n^d(\alpha, h) - I_n] \eta \\ &= \varepsilon' (I_n - D_n(\alpha, h) [M_n(\alpha, h) - I_n] - I_n) \eta \\ &\leq \sup_{\alpha \in A_n} \sup_{h \geq 0} \max_{i \leq n} d_{ii}(\alpha, h) \varepsilon' P_n(\alpha, h) \eta \end{aligned}$$

where $d_{ii}(\alpha, h)$ is the $(ii)^{th}$ element of $D_n(\alpha, h)$.

So we get that

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\varepsilon' P_n^d(\alpha, h) \eta|}{n R_n^d(\alpha, h; Z^n)} \leq h_n \sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\varepsilon' P_n(\alpha, h) \eta|}{n R_n(\alpha, h; Z^n)} \quad (\text{I.6.49})$$

where $h_n = \sup_{\alpha \in A_n} \sup_{h \geq 0} \max_{i \leq n} d_{ii}(\alpha, h)$. The right side term of (6.50) will converge to 0 in probability under A.1-A.5 if $h_n = O_p(1)$, which is implied by B.2.

As for Mallows's C_L , (6.48) will hold if condition A.5 is satisfied (as $\text{tr} M_n^d(\alpha, h)/n R_n^d(\alpha, h; Z) \leq 1$).

To show (6.48) notice that

$$\begin{aligned} & \left| \varepsilon' M_n^d(\alpha, h) \varepsilon - \sigma^2 \text{tr} M_n^d(\alpha, h) \right| = \\ & \left| \varepsilon' (D_n(\alpha, h)[M_n(\alpha, h) - I_n] + I_n) \varepsilon - \sigma^2 \text{tr} (D_n(\alpha, h)[M_n(\alpha, h) - I_n] + I_n) \right| \\ & \leq \sup_{\alpha \in A_n} \sup_{h \geq 0} \max_{i \leq n} d_{ii} \left| \varepsilon' M_n(\alpha, h) \varepsilon - \sigma^2 \text{tr} M_n(\alpha, h) \right| \\ & = h_n \left| \varepsilon' M_n(\alpha, h) \varepsilon - \sigma^2 \text{tr} M_n(\alpha, h) \right| \end{aligned}$$

So we get that

$$\begin{aligned} & \sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\sigma^2 \text{tr} M_n^d(\alpha, h) - \varepsilon' M_n^d(\alpha, h) \varepsilon|}{n R_n^d(\alpha, h; Z^n)} \\ & \leq h_n \sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\varepsilon' M_n(\alpha, h) \varepsilon - \sigma^2 \text{tr} M_n(\alpha, h)|}{n R_n(\alpha, h; Z^n)} \end{aligned}$$

which will converge to zero in probability under A.1-A.5 and B.2.

Finally, convergence to zero in probability of (6.45) and (6.47) also follows under A.1-A.5 and B.2 by similar arguments.

I.7 References

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, 2nd International Symposium in Information Theory, Petrov, B.N. and Csaki, F. eds., Akademiai Kiado, Budapest, 267-281.

Andrews, D.W.K. (1991), Asymptotic Optimality of Generalized C_L , Cross-validation and Generalized Cross-validation in Regression with Heteroskedastic Errors, *Journal of Econometrics*, 47, 359-377.

Andrews, D.W.K. (1991), Asymptotic Normality of Series Estimators for Various Nonparametric and Semi-parametric Regression Models, *Econometrica*, 59, 307-345.

Bierens, H. (1990), A Consistent Conditional Moment Test of Functional Form, *Econometrica*, 58, 1443-1458.

Burnham, K. and Anderson, D. (2003), *Model Selection and Multi-Model Inference*, Springer.

Candes, E. (1998), *Ridgelets: Theory and Applications*. Ph.D. Dissertation, Department of Statistics, Stanford University.

Candes, E. (1999), Harmonic Analysis of Neural Networks, *Applied and Computational Harmonic Analysis*, 6, 197-218.

Candes, E. (2003), Ridgelets: Estimating with Ridge Functions, *Annals of Statistics*, 33, 1561-1599.

Craven, P. and Wahba, G. (1979), Smoothing Noisy Data with Spline Functions, *Numerical Mathematics*, 31, 377-403.

Frank, I.E. and Friedman, J. (1993), A Statistical View of Some Chemometrics Regression Tools, *Technometrics*, 35, 109-148.

Fu, W. (1998), The Bridge versus the Lasso, *Journal of Computational and Graphical Statistics*, 7, 397-416.

Li, K.-C. (1985), From Stein's Unbiased Risks Estimates to the Method of Generalized Cross-validation, *Annals of Statistics*, 33, 1352-1377.

Li, K.-C. (1986), Asymptotic Optimality of C_L and Generalized Cross-validation in Ridge regression with Applications to Spline Smoothing, *Annals of Statistics*, 14, 1101-1112.

- Li, K.-C. (1987), Asymptotic Optimality for C_p , C_L , Cross-validation and Generalized Cross-validation: Discrete Index Set, *Annals of Statistics*, 15, 958-975.
- Mallows, C. (1973), Some Comments on C_p , *Technometrics*, 15, 661-675.
- Moricz, F. (1976), Moment Inequalities and the Strong Laws of Large Numbers, *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, 35, 299-314.
- Opsomer, J., Wang, Y. and Yang, Y. (2001), Nonparametric Regression with Correlated Errors, *Statistical Science*, 16, 134-153.
- Rao, C.R. and Wu, Y. (1989), A Strongly Consistent Procedure for Model Selection in a regression Problem, *Biometrika*, 76, 369-374.
- Speckman, P. (1985), Spline Smoothing and Optimal Rates of Convergence in Nonparametric Regression Models, *Annals of Statistics*, 13, 970-983.
- Shao, J. (1993), Linear Model Selection by Cross-validation, *Journal of the American Statistical Association*, 88, 486-494.
- Shao, J. (1997), An Asymptotic Theory for Linear Model Selection, *Statistica Sinica*, 7, 221-264.
- Stinchcombe, M. and White, H. (1998), Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative, *Econometric Theory*, 14, 295-325.
- Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, B*, 58, 267-288.
- White, H. (2006), Approximate Nonlinear Forecasting Methods, *Handbook of Economic Forecasting*, Edited by Graham Elliott, Clive Granger and Allan Timmermann, Elsevier.
- White, H. (2001), *Asymptotic Theory for Econometricians*, Academic Press.
- Whittle, P. (1960), Bounds for the Moments of Linear and Quadratic Forms in Independent Variables, *Theory of Probability and its Applications*, 5, 302-305.
- Yang, Y. (1999), Model Selection for Nonparametric Regression, *Statistica Sinica*, 9, 475-499.

Zhang, P. (1993), Model Selection via Multifold Cross-validation, *Annals of Statistics*, 21, 299-313.

II

Optimal Selection of Series Functions under Dependent and Heterogeneous Processes

II.1 Introduction

This chapter extends the results in the first one to accommodate data dependence and heterogeneity and thus achieves as much generality as possible. Recall that one of the goals of forecasting is to obtain a point forecast of a target variable Y_t given a $d \times 1$ vector of predictors Z_t (which might include lagged values of Y_t) and with d a finite integer. The conditional expectation of Y_t given Z_t

$$\mu(Z_t) = E[Y_t|Z_t]$$

yields the best possible prediction of Y_t given Z_t under prediction mean squared error (PMSE), provided Y_t has finite variance. That is, μ solves the problem

$$\min_{m \in \mathfrak{M}} E[(Y_t - m(Z_t))^2]$$

where $\mathfrak{M}$ is the set of functions m of Z_t having finite variance and the expectation is taken with respect to the joint distribution of Y_t and Z_t .

There are several classes of non-parametric models capable of approximating the unknown conditional expectation above with more or less degree of accuracy. Here, following White (2006), we focus on series estimators which are linear in the parameters. That is, they have the following form

$$f(z, \theta) = z'\gamma + \sum_{j \in \alpha} \psi_j(z)\beta_j$$

where α is a subset of $\mathbb{N}$ containing $q \leq p_n$ elements belonging to A_n , which is a set of subsets of $\mathbb{N}$ with at most p_n elements; and the basis functions $\psi_j(\cdot)$ are non-linear functions of z .

As stated in the previous chapter and following White (2006), they offer the advantage of being non-linear in z (achieving the degree of flexibility of non-linear models) and linear in the parameters (avoiding the computational problems of models lacking this property).

The performance of series estimators will depend on the properties of the family of basis functions chosen (polynomials, trigonometric functions, neural networks with sigmoidal activation functions, wavelets, etc...) but most crucially on the optimal selection of the number of basis functions for a given sample size. We thus trade bias (decreasing as we incorporate more basis functions) for variance (parameter estimation uncertainty).

Ideally we would select the number of basis functions minimizing PMSE but since it is unknown (it depends in the process generating the data) we can only work with an estimate. There is a wide range of model selection procedures which can be regarded as either direct or indirect estimates of PMSE. Among the first lie the family of cross-validation methods and among the indirect ones we find information selection criteria such as AIC, BIC, Mallow's C_L , etc...

Here we analyze the optimality (asymptotic loss-efficiency) of a more general version of the Mallows's C_L statistic that we call "generalized" Mallows's C_L (GC_L) following Andrews (1991). It is asymptotically equivalent to the family of generalized information criteria studied by Konishi and Kitagawa (1996) particularized to ridge estimation. GC_L accounts for model misspecification and in that sense is more general than the TIC information criterion (Takeuchi, 1976), which reduces to AIC (Akaike, 1973) when the models are correctly specified. We also study the optimality of leave-1-out cross-validation, though any other cross-validation method with the validation set size of smaller order than the sample size will have the same asymptotic behavior.

We will see that the model selection procedures we analyze are asymptotically optimal for simultaneously choosing the number of basis functions and the ridge parameter under some regularity conditions that include some mild restrictions in the process generating the data and misspecification of the approximating models. This is the common case in economics and finance. Other procedures such as BIC (Schwarz, 1978) and H-Q (Hannan and Quinn, 1979) and leave- d -out cross-validation with d/n tending to a constant greater than zero are only optimal when the models are correctly specified.

Our framework is very general as it applies to any series estimator linear in the parameters. It also accommodates dependence and heterogeneity in the data generating process (DGP). We allow our data to be near-epoch dependent as many time series processes satisfy this memory requirement (as long as they possess some mild dynamic stability conditions) and is therefore more general than mixing.

We also allow for heterogeneity, which in economics and finance occurs very often in the form of breaks. Some basis functions (such as the family of wavelets or the more powerful ridgelets and curvelets (Candes, 1998, 1999, 2003))

are robust to this type of phenomena and they constitute our main motivation for relaxing the stationarity assumption commonly used in model selection.

Li (1986, 1987) studied the asymptotic loss-efficiency of Mallows's C_L (Mallows, 1973), leave-1-out cross-validation (Stone, 1974) and generalized cross-validation (GCV) (Craven and Wahba, 1979) for selecting the number of regressors or basis functions under the *iid* assumption and fixed regressors. Also that of C_L and GCV for choosing the ridge parameter under *iid* and normally distributed errors. Andrews (1991) studied the optimality of C_L , GCV and leave-1-out cross-validation under independence and fixed regressors but allowing for heteroskedastic errors. He finds that GCV is not optimal for series estimators under those conditions. Shao (1997) analyzes this property for a large number of model selection criteria finding that those mentioned in the earlier papers are optimal under misspecification while others such as BIC or leave- d -out cross-validation are optimal when there are two or more correctly specified models. His paper represents a good summary of the optimality properties of model selection procedures. Finally, Burman and Nolan (1992) studied the asymptotic loss-efficiency of leave-1-out cross-validation under time series stationarity though their results are tailored at regression splines and their conditions are much stronger than ours.

In the next section we introduce the notation and concepts that will be used throughout the paper. Section 3 deals with the asymptotic loss-efficiency of GC_L and leave-1-out cross-validation. We then conclude and offer the proof of the results in the last section.

II.2 Notation and definitions

Our focus is on predicting a target variable Y_t given a $d \times 1$ vector of predictors Z_t . We regard our forecasting model as an approximation to the conditional expectation of Y_t given Z_t (the optimal predictor in terms of PMSE) so we are implicitly working with a quadratic loss. We assume that Y_t is generated as

$$Y_t = \mu(Z_t) + \varepsilon_t$$

where $\mu(Z_t) \equiv E[Y_t|Z_t]$ and ε_t is a zero mean error term representing the deviations between Y_t and $\mu(Z_t)$.

Given a sample of n observations on the Y_t 's and Z_t 's, we write $Y \equiv [Y_1, \dots, Y_n]'$, $\varepsilon \equiv [\varepsilon_1, \dots, \varepsilon_n]'$ and $Z^n \equiv [Z_1, \dots, Z_n]'$ to denote the vectors containing the n observations on Y_t and ε_t respectively and the matrix whose t^{th} row contains the $d \times 1$ vector of predictors Z_t' . We also write $Y = \eta + \varepsilon$ where $\eta \equiv [\mu(Z_1), \dots, \mu(Z_n)]'$ denotes the $n \times 1$ vector of conditional expectations corresponding to each t .

We now introduce assumptions B.1 and B.2, which refer to the data generating process (DGP) and the model respectively and will be used in the main results of the paper.

B.1 (Data generating process)

- (a) Let $\{(Y_t, Z_t)\}$ be a sequence of dependent and possibly heterogeneous random vectors such that Y_t is real-valued and Z_t is a $\mathbf{Z}$ -valued vector, $\mathbf{Z} \subset \mathbb{R}^d$, $d \in \mathbb{N}$.
- (b) $\varepsilon_t \equiv Y_t - E[Y_t|Z_t] = Y_t - \mu(Z_t)$.

B.2 (Model) Let $\{\psi_j : \mathbf{Z} \rightarrow \mathbb{R}\}$ be a sequence of measurable functions. Let $\{p_n\}$ be a given non-decreasing sequence of integers and $\{A_n\}$ a given sequence

of sets where each A_n is a set of subsets of $\mathbb{N}$ and each subset contains at most p_n elements, so that $\alpha \in A_n$ is a subset of $\mathbb{N}$ containing $q_n \equiv \#\alpha \leq p_n$ elements.

The model is given by

$$\mathfrak{M}_n = \{m : \mathbf{Z} \rightarrow \mathbb{R} \mid m(z) = z'\gamma + \sum_{j \in \alpha} \beta_j \psi_j(z) \quad \gamma \in \mathbb{R}^d, \beta_j \in \mathbb{R}, \alpha \in A_n\}$$

We next define the ridge estimator; for that purpose let $X_n(\alpha)$ be the $n \times (d + q)$ matrix whose t^{th} row has elements Z_t and $\psi_j(Z_t)$, $j \in A_n$, $\alpha \in A_n$, $t = 1, \dots, n$, $n = 1, 2, \dots$. We consider the estimator given by

$$\hat{\eta}(\alpha, h) \equiv [\hat{\eta}_1(\alpha, h), \dots, \hat{\eta}_n(\alpha, h)]' \equiv M_n(\alpha, h)Y$$

where $M_n(\alpha, h) \equiv X_n(\alpha)[X_n(\alpha)'X_n(\alpha) + hI_q]^{-1}X_n(\alpha)'$ for given $h \geq 0$ and $\alpha \in A_n$.

We also define $X_{p_n}(\cdot) \equiv [\psi_1(\cdot), \dots, \psi_{p_n}(\cdot)]'$ to denote a $p_n \times 1$ vector containing the maximum number of basis functions given a sample of size n and $X_{np_n} \equiv [X_{p_n}(Z_1), \dots, X_{p_n}(Z_n)]'$ to denote the corresponding $n \times p_n$ matrix which includes all the observations.

We next define the concept of near-epoch dependence, a memory requirement on the sequences of observations that will be used throughout the paper:

Given a probability space (Ω, F, P) , let V_t be a F -measurable vector-valued sequence, let $F_{t-m}^{t+m} \equiv \sigma(V_{t-m}, \dots, V_{t+m})$ such that $\{F_{t-m}^{t+m}\}_{m=0}^{\infty}$ is an increasing sequence of σ -fields contained in F . Also let $\{S_{it}\}$ be a sequence of F -measurable random variables with $E|S_{it}| < \infty$ for all $i = 1, \dots, d$ and $t = 1, \dots, n$. Then $\{S_{it}\}$ is L_2 -NED on $\{V_t\}$ of size $-a$ iff

$$\nu_{mi} \equiv \sup_t \|S_{it} - E_{t-m}^{t+m} S_{it}\|_2 = O(m^{-a^*})$$

for $a^* > a$.

We define $L_n(\alpha, h) \equiv n^{-1}\|\eta - \hat{\eta}(\alpha, h)\|^2 = n^{-1}\sum_{t=1}^n(\mu_t - \hat{\mu}_t(\alpha, h))^2$ as the average squared error loss, which can be decomposed as

$$L_n(\alpha, h) = \Delta_n(\alpha, h) + n^{-1}\varepsilon' M_n^2(\alpha, h)\varepsilon - n^{-1}2\eta'[I_n - M_n(\alpha, h)]M_n(\alpha, h)\varepsilon$$

where $\Delta_n(\alpha, h) \equiv n^{-1}\|\eta - M_n(\alpha, h)\eta\|^2$, and the unconditional risk is defined as its expectation; that is, $R_n(\alpha, h; Z^n) \equiv E(L_n(\alpha, h))$.

We also defined $L_n^*(\alpha, h) \equiv \Delta_n(\alpha, h) + n^{-1}\varepsilon' M_n^2(\alpha, h)\varepsilon$ which is the sum of the first two components of the average squared error loss. As we will show in the proofs, the difference between $L_n(\alpha, h)$ and $L_n^*(\alpha, h)$ is asymptotically negligible relative to average squared error loss.

We next introduce the modern notions of consistency and asymptotic loss-efficiency for model selection . The latter is highly desirable for a model selection procedure to behave well and constitutes the main focus of the paper, as shown in the next section.

Let $(\hat{\alpha}_n, \hat{h}_n)$ denote the model and ridge parameter chosen by a given selection procedure and let (α_n^*, h_n^*) be the pair minimizing $L_n(\alpha, h)$ over $\alpha \in A_n$ and $h \geq 0$.

Following Shao (1997), we say that the selection procedure is **consistent** if

$$P\{(\hat{\alpha}_n, \hat{h}_n) = (\alpha_n^*, h_n^*)\} \rightarrow 1.$$

As Shao (1997) makes explicit, this consistency is in terms of model selection and is not related to the consistency of $\hat{\eta}(\hat{\alpha}_n, \hat{h}_n)$ as an estimator of η , i.e. we do not

require

$$L_n(\hat{\alpha}_n, \hat{h}_n) = o_p(1)$$

This last consistency is not very useful since, as Shao (1997) points out, sometimes there no consistent estimator of η (e.g. $\inf_{\alpha \in A_n} \inf_{h \geq 0} L_n(\alpha, h) \neq o_p(1)$) because the model may be misspecified and sometimes there are too many consistent estimators (e.g. $\sup_{\alpha \in A_n} \sup_{h \geq 0} L_n(\alpha, h) = o_p(1)$), as when one has a correctly specified model with irrelevant predictors.

In some cases a selection procedure is not consistent but $(\hat{\alpha}_n, \hat{h}_n)$ is still “close” to (α_n^*, h_n^*) in the sense that

$$\frac{L_n(\hat{\alpha}_n, \hat{h}_n)}{L_n(\alpha_n^*, h_n^*)} \longrightarrow 1 \quad \text{in prob.}$$

A selection procedure satisfying this condition is said by Li (1986) among others to be **asymptotic loss-efficient** and this is the property we analyze in the present paper.

II.3 Asymptotic loss-efficiency

In this section we show the conditions under which GC_L and Leave-1-out cross-validation are asymptotic loss-efficient.

We will extend previous results to accommodate dependent and possibly heterogeneous observations. We will impose near-epoch dependence as memory requirement on the observations in order to achieve as much generality as possible. As stated in the introduction, our theorems only hold for the case of misspecified models; nevertheless, as we argued there, this is the common situation in economics and finance and this is why we use series approximations.

Leave-1-out cross-validation is asymptotically equivalent to GC_L but unlike this one it has the additional advantage of not requiring estimation of the variance-covariance matrix term that appears in the penalty of GC_L . When the errors in the DGP are correlated both GC_L and leave-1-out cross-validation lose their optimality and methods capable of dealing with it (such as h -block cross-validation (Burman et al., 1994; Racice, 1997)) are required.

II.3.A GC_L

GC_L selects $\alpha \in A_n$ and $h \geq 0$ by minimizing

$$GC_{Ln}(\alpha, h) \equiv n^{-1} \|Y - \hat{\eta}(\alpha, h)\|^2 + n^{-1} \text{tr} W_n^{-1}(\alpha, h) V_n(\alpha)$$

where $W_n(\alpha, h) \equiv E[X_n(\alpha)' X_n(\alpha) + h I_{q_n}]$,

$V_n(\alpha) \equiv E[X_n(\alpha)' \varepsilon \varepsilon' X_n(\alpha)]$ and $\hat{\eta}(\alpha, h) \equiv M_n(\alpha, h) Y$.

As $W_n(\alpha, h)$ and $V_n(\alpha)$ are unknown, we have to estimate them consistently. The corresponding estimators will be given by

$$\hat{W}_n(\alpha, h) \equiv n^{-1} [X_n(\alpha)' X_n(\alpha) + h I_{q_n}]$$

and

$$\hat{V}_n(\alpha) \equiv n^{-1} \sum_{t=1}^n \hat{\varepsilon}_t^2 x_t(\alpha) x_t(\alpha)'$$

where $x_t(\alpha)$ denotes the t^{th} row of the $X_n(\alpha)$ matrix.

We minimize in practice $GC_{Ln}^*(\alpha, h) \equiv n^{-1} \|Y - \hat{\eta}(\alpha, h)\|^2 + n^{-1} \text{tr} \hat{W}_n^{-1} \hat{V}_n$. Let $(\alpha^{GC_L^*}, h^{GC_L^*})$ be the corresponding minimizers over $A_n \times [0, \infty)$; our goal is to show asymptotic loss-efficiency, that is

$$\frac{L_n(\alpha^{GC_L^*}, h^{GC_L^*})}{\inf_{\alpha \in A_n} \inf_{h \geq 0} L_n(\alpha, h)} \longrightarrow 1 \quad \text{in prob.}$$

where $L_n(\alpha, h) \equiv n^{-1} \|\eta - \hat{\eta}(\alpha, h)\|^2$.

Now assume the following regularity conditions:

$$\text{B.3 } \sum_{\alpha \in A_n} (E[\inf_{h \geq 0} nL_n(\alpha, h)])^{-m} \longrightarrow 0$$

where m is some fixed positive integer such that the memory and moment restrictions below are satisfied.

$$\text{B.4 } \limsup_{n \rightarrow \infty} \sum_{\alpha \in A_n} [\iota_n(\alpha) - k] (E[\inf_{h \geq 0} nL_n(\alpha, h)])^{-m^*} < \infty$$

where k is a positive integer not depending on n such that $\lambda_k(\alpha) > 0$ for all n (where $\lambda_k(\alpha)$ denotes the k^{th} eigenvalue of the $X_n(\alpha)'X_n(\alpha)$ matrix), $\iota_n(\alpha)$ is the largest t such that $\lambda_t(\alpha) > 0$, $t \leq q = \#\alpha \leq p_n$, and m^* is some fixed positive integer such that the memory and moment restrictions above are satisfied.

B.5 (Memory requirements)

- (a) The sequence of errors $\{\varepsilon_t\}$ is L_2 - NED on $\{V_t\}$ of size $-\max\{2^{2m-1}(\frac{r-1}{r-2})^{2m-1} + (m-1), 2^{2m^*-1}(\frac{r-1}{r-2})^{2m^*-1} + (m^*-1), 4(\frac{r-1}{r-2})^2\}$ for some $r > 2$ and where $\{V_t\}$ is an α -mixing sequence of size $-(\frac{2r}{r-2} + m - 1)$.
- (b) $\{Z_{it}\}$ is a L_2 - NED sequence on $\{V_t\}$ of size $-8(\frac{r-1}{r-2})^3$ for some $r > 2$ and $i = 1, \dots, d$ and where $\{V_t\}$ is the same α -mixing sequence as before.

B.6 (Moment restrictions)

- (a) $\|\varepsilon_t\|_{\max\{(2m-1)r, (2m^*-1)r, 4r, 4p\}} \leq \Delta < \infty$, for some $r > 2$ and $p \geq 1$.
- (b) $\|\psi_j(Z_t)\|_{\max\{4r, 4p\}} \leq \Delta < \infty$, for all j and t .

B.7 The sequence of basis functions $\{\psi_j(\cdot)\}$ is such that $|\psi_j(Z^1) - \psi_j(Z^2)| \leq B(Z^1, Z^2)\rho(Z^1, Z^2)$ for all j , where $B(z^1, z^2) : \mathbf{Z} \times \mathbf{Z} \longrightarrow \mathbb{R}^+$ is a non-negative $\mathbb{B}^{2d}$ -measurable function and $\rho(z^1, z^2) \equiv \sum_{i=1}^d |z_i^1 - z_i^2|$, where z_i is the i^{th} component of the $d \times 1$ vector z .

We further assume that $\|B(Z_t, E_{t-m}^{t+m} Z_t)\|_{q/(q-1)} < \infty$, $\|\rho(Z_t, E_{t-m}^{t+m} Z_t)\|_q < \infty$ and $\|B(Z_t, E_{t-m}^{t+m} Z_t)\rho(Z_t, E_{t-m}^{t+m} Z_t)\|_r < \infty$, where $1 \leq q \leq 2$ and $r > 2$.

$$\text{B.8 } E[\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_1, Z_t, Z_{t-1}, Z_{t-2}, \dots, Z_1] = 0.$$

B.9 For some $0 < \delta < 1$ and same p as in B.6,

$$(p_n^2 n)^{1/p} \left(\sum_{\tau=1}^{n-1} (\tau + 2) \right)^{1/p} \left(\inf_{\alpha \in A_n} \inf_{h \geq 0} n L_n(\alpha, h) \right)^{-\delta} = O_p(1).$$

$$\text{B.10 } \lambda_{\min} \left\{ n^{-1} \sum_{t=1}^n E X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\} = C_n$$

where $C_n (\inf_{\alpha \in A_n} \inf_{h \geq 0} n L_n(\alpha, h))^{-(1-\delta)} = o_p(1)$.

$$\text{B.11 } \left[\inf_{\alpha \in A_n} \inf_{h \geq 0} n L_n(\alpha, h) \right]^{-\delta} p_n \sup_{t \leq n} |\hat{\mu}(Z_t) - \mu(Z_t)| = O_p(1)$$

Theorem II.3.1. *Under B.1-B.11 the pair $(\alpha^{GC_L^*}, h^{GC_L^*})$ selected by GC_L^* is asymptotically loss-efficient.*

Remarks on conditions B.3-B.11:

1. Condition B.3 is similar to A.3 in chapter 1 and similar comments apply. For the case of series function approximations to η , we might expect $\inf_{\alpha \in A_n} E[\inf_{h \geq 0} L_n(\alpha, h)]$ to be of order $O(n^{-1+\delta})$ for some $\delta > 0$. So as long as the cardinality of A_n is of polynomial order $n^{\delta'}$ for some $\delta' > 0$, there always exists an m such that B.3 is satisfied. Also notice that since $\inf_{\alpha \in A_n} E[\inf_{h \geq 0} L_n(\alpha, h)] \leq \inf_{\alpha \in A_n} \inf_{h \geq 0} R_n(\alpha, h)$ we might need slightly stronger moment restrictions on the errors than those needed under independence for this condition to hold. Finally remember from the first chapter that this assumption requires that we expand the size the model set at a polynomial rate and breaks down when one or more models are correctly specified.

2. Condition B.4 is the analog of A.4 in chapter 1 and arises specifically for model selection under ridge estimation. As mentioned above, for series approximations $\inf_{\alpha \in A_n} E[\inf_{h \geq 0} L_n(\alpha, h)]$ is typically $O(n^{-1+\delta})$ with $\delta > 0$, so $\inf_{\alpha \in A_n} E[\inf_{h \geq 0} nL_n(\alpha, h)] \simeq O(n^\delta)$. On the other hand $\iota_n(\alpha)$ is $O(n)$ at most so for any $m^* \geq 3/\delta$ the limit above is bounded and B.4 would hold.

3. Conditions B.5 and B.6 impose the memory and moment restrictions on the errors and predictors required for the results to hold. We allow the observations to be near epoch dependence (*NED*), which is a weaker memory requirement than mixing and it is satisfied by a wide range of dynamically stable time series processes. The size requirements on the *NED* coefficients allow for transformations or functions of the observations to be equally *NED*. Condition B.7 is also a condition imposed on the basis functions to inherit the *NED* property from the predictors.

4. Condition B.8 is sufficient for $\{\varepsilon_t \psi_j(Z_t)\}$ to be a martingale difference sequence for all j (see Hayashi, 2000). Notice that B.8 implies that the error term is serially uncorrelated and is also uncorrelated with the current and past predictors. When the errors are correlated this requirement breaks down and GC_L gets a bias that can not be consistently estimated. The same problem occurs with leave-1-out cross-validation and methods that circumvent this problem such as h -block cross-validation are required.

5. Condition B.9 requires the existence of a large enough p so that it holds and by B.6 this will imply that at least we will need the first p moments of the errors and regressors to be uniformly bounded. There is a trade-off between how big p should be and how slow is the convergence of the series estimator to the unknown function it is approximating. The slower the rate

of convergence, the bigger p (and corresponding moment restrictions) that we will need.

6. Condition B.10 sets the maximum rate at which we may allow the minimum eigenvalue of the information matrix to decrease as we add more basis functions into the models to improve the level of approximation. This rate is unknown as it depends on how fast the optimal $nL_n(\alpha, h)$ goes to infinity, which at the same time depends on the rate at which the series estimator converges to the true conditional expectation.
7. Finally condition B.11 imposes a restriction on the rate at which we can increase the number of basis functions with the sample size to achieve a better degree of approximation. Notice that it might imply a very slow rate if the rate of convergence of the series estimator is also slow. Nevertheless there exists a trade-off between this condition and B.8 in the sense that we can tolerate a faster degree of increase in the number of basis functions as long as the multicollinearity problem doesn't get substantially worse as the model gets bigger. In any case this all depends on the way we implement our approximation and the data generating process, which remains unknown.

II.3.B Leave-1-out cross-validation

Leave-1-out cross-validation can be regarded as a direct estimate of prediction mean squared error (PMSE). It works by splitting the data into two parts. The first part contains $n-1$ observations used for fitting the model (model construction) and the observation left is used for assessing the predictive ability of the model (model validation). This procedure is repeated for each single observation and finally the model minimizing the average of all these predictive ability measures is the chosen one.

A more general version of cross-validation is leave- d -out cross-validation. It differs from leave-1-out in that the validation set contains d observations instead of just one and the construction (or test) set contains the $n - d$ remaining ones. It is computationally less attractive than leave-1-out since unlike the latter it doesn't have closed form though some simulations have shown some evidence of better performance than leave-1-out in small samples. Asymptotically it has the same asymptotic behavior than leave-1-out as long as d/n tends to zero.

Dependence and heterogeneity in the data doesn't affect the asymptotic optimality of cross-validation as long as the errors are uncorrelated (to be more precise as long as the sequence of products of the error with the regressors is a martingale difference, which is ensured by condition B.8). When the errors are correlated we will need more general cross-validation procedures such as h -block cross-validation, whose asymptotic optimality is studied in the next chapter.

The leave-1-out cross-validation (CV_{-1}) method minimizes

$$CV_{n,-1}(\alpha, h) \equiv n^{-1} \sum_{t=1}^n \left(Y_t - x_t(\alpha)' [X_{n,(-t)}(\alpha)' X_{n,(-t)}(\alpha) + hI_{q+d}]^{-1} \right. \\ \left. \times X_{n,(-t)}(\alpha) Y_{(-t)} \right)^2 \quad (\text{II.3.1})$$

with respect to $\alpha \in A_n$ and $h \geq 0$, where $x_t(\alpha)$ denotes the t^{th} row of the $X_n(\alpha)$ matrix, $X_{n,(-t)}$ denotes the $X_n(\alpha)$ matrix with all the rows except the t^{th} and $Y_{(-t)}$ is the vector Y with all its components except the t^{th} .

It is easy to show that (2.1) can be expressed as

$$CV_{n,-1}(\alpha, h) = n^{-1} \left\| [I_n - \ddot{M}_n(\alpha, h)]^{-1} [Y - \hat{\eta}(\alpha, h)] \right\|^2, \quad (\text{II.3.2})$$

where $\ddot{M}_n(\alpha, h)$ is a diagonal $n \times n$ matrix whose t^{th} diagonal element is $m_{tt}(\alpha, h)$, the t^{th} diagonal element of $M_n(\alpha, h)$.

Now we introduce the following condition required for the asymptotic loss-efficiency of (CV_{-1}) :

$$\text{B.12 } h_n = o_p(1)$$

where $h_n = \sup_{\alpha \in A_n} \sup_{h \geq 0} \bar{\lambda}(M_n(\alpha, h))$ with $\bar{\lambda}(\cdot)$ denoting the maximal diagonal element of a matrix.

Theorem II.3.2. *Under B.1-B.10 and B.12 the pair $(\alpha^{CV-1}, h^{CV-1})$ selected by leave-1-out cross-validation is asymptotically loss-efficient.*

Remark: Condition B.12 is much milder than B.11 and thus allows much faster rates of increase in the number of basis functions than those required for the optimality of GC_L . The difference lies in that the GC_L model selection procedure requires for its implementation the estimation of the V_n variance-covariance term that appears on its penalty and the uncertainty arising from its estimation negatively affects the performance of the statistic. This does not occur for leave-1-out cross-validation, which increases its flexibility.

II.4 Conclusion

We have analyzed the conditions under which GC_L and leave-1-out model selection procedures are asymptotic loss-efficient under model misspecification as well as under time series and potentially heterogeneous data. This enhances the applicability of these model selection criteria when choosing the optimal number of basis functions in series estimation as well as the optimal shrinkage parameter whenever ridge regression is used to estimate the basis functions coefficients.

It is worth noticing that the naive AIC wouldn't be optimal since we need to account for model misspecification. On the other hand, GC_L requires the

estimation of its penalty term and that will affect its performance. When the misspecification problem is not too serious AIC might work better for selecting the optimal number of regressors or basis functions. Leave-1-out cross-validation doesn't incorporate any term requiring estimation too and in this sense it will be more efficient than GC_L , delivering better approximations to the unknown conditional expectation (optimal prediction under quadratic loss).

Other model selection procedures such as BIC, H-Q, FPE and so forth, which penalize more than AIC or GC_L will not be asymptotically optimal under model misspecification. Therefore, they should never be employed for implementing series estimators. The same conclusion holds for any leave- d -out cross-validation method with d/n tending to a positive constant since they will be asymptotically equivalent to some of the statistics above.

Accounting for heterogeneity will make it possible to implement series estimators that are robust to the presence of structural breaks (such as wavelets, rigelets or curvelets basis functions) as usually encountered in macroeconomic and financial data. Accounting for near epoch dependence will make the procedures work under very general time series processes.

It is also important to remark that error correlation will break down the asymptotic optimality of both selection procedures and methods that are robust to it are required. h -block cross-validation is a generalization of cross-validation that circumvents that problem and will be studied in depth in the next chapter.

II.5 Proofs of theorems

In the proof of theorem 3.1 we will call lemma 6.1 from chapter 1 and will use the following theorem which generalizes Whittle's result and is proved at the end of this section.

Theorem II.5.1. *Let $\{\varepsilon_j\}$ be a L_2 -NED on $\{V_j\}$ of size $-2^{2m-1}(\frac{r-1}{r-2})^{2m-1}$ where $\{V_j\}$ is an α -mixing sequence of size $-\frac{2r}{r-2}$ for some $r > 2$. Also let $E\varepsilon_j = 0$, $\|\varepsilon_j\|_{(2m-1)r} \leq \Delta < \infty$, where $m \geq 1$ and let*

$$\sum_{j=1}^{\infty} j^{m-1} \nu_{[j/4]}^{(r-2)^{2m-1}/2^{m-1}(r-1)^{2m-1}} < \infty \text{ and } \sum_{j=1}^{\infty} j^{m-1} \alpha_{[j/4]}^{1/2-1/r} < \infty;$$

where ν and α denote the L_2 -NED and α -mixing coefficients respectively and $[.]$ denotes the integer part of $.$ Then, for any sequence $\{b_j\}$ whose elements are uniformly $O_{a.s}(1)$ and any integer n , we have

$$E\left(\sum_{j=1}^n b_j \varepsilon_j\right)^{2m} \leq C \left(\sum_{j=1}^n \tilde{b}_j^2\right)^m$$

where C is a constant depending on m and $\tilde{b}_j \equiv E|b_j|$.

Proof of theorem 3.1:

As we did in chapter 1 and following Li (1986) rather than using $X_n(\alpha)$ we will work with the matrix $\tilde{X}_n(\alpha) = [\Lambda_n(\alpha)', \mathbf{0}']'$, where $\Lambda_n(\alpha)$ is the $(d+q) \times (d+q)$ diagonal matrix with $d_t(\alpha) = \lambda_t^{1/2}(\alpha)$ as diagonal elements, the $\lambda_t(\alpha)$'s denote the eigenvalues of $X_n(\alpha)'X_n(\alpha)$ and $\mathbf{0}$ is an $n - (d+q) \times (d+q)$ matrix of zeros. This is achieved by applying the singular value decomposition to the $X_n(\alpha)$ matrix; that is, $X_n(\alpha) = U_n(\alpha)\Lambda_n(\alpha)V_n(\alpha)$ where $U_n(\alpha)$ and $V_n(\alpha)$ are $n \times n$ and $(d+q) \times (d+q)$ orthogonal matrices respectively. $\Lambda_n(\alpha)$ has a very simple structure. Our statistics don't change as long as we pre-multiply the vectors Y, η and ε by $U_n(\alpha)'$, yielding $\tilde{Y} \equiv [\tilde{Y}_1, \dots, \tilde{Y}_n]'$, $\tilde{\eta} \equiv [\tilde{\mu}_1, \dots, \tilde{\mu}_n]'$ and $\tilde{\varepsilon} \equiv [\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_n]'$.

Write GC_{Ln}^* as

$$GC_{Ln}^*(\alpha, h) = n^{-1}\|\varepsilon\|^2 + L_n(\alpha, h) + 2n^{-1}\varepsilon'P_n(\alpha, h)\eta \quad (\text{II.5.1})$$

$$+ 2n^{-1}(trW_n^{-1}(\alpha, h)V_n(\alpha) - \varepsilon'M_n(\alpha, h)\varepsilon) \quad (\text{II.5.2})$$

$$+ 2n^{-1}(tr\hat{W}_n^{-1}(\alpha, h)\hat{V}_n(\alpha) - trW_n^{-1}(\alpha, h)V_n(\alpha)) \quad (\text{II.5.3})$$

where $P_n(\alpha, h) \equiv I - M_n(\alpha, h)$.

Asymptotic loss-efficiency for $GC_{Ln}^*(\alpha, h)$ will follow if

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} |L_n(\alpha, h)/L_n^*(\alpha, h) - 1| = o_p(1), \quad (\text{II.5.4})$$

and the last term in (5.1) as well as (5.2) and (5.3) are of smaller order of magnitude in probability than $L_n^*(\alpha, h)$ and $L_n(\alpha, h)$ respectively, uniformly with respect to α and h ; that is,

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1}|\varepsilon'P_n(\alpha, h)\eta|/L_n^*(\alpha, h) = o_p(1) \quad (\text{II.5.5})$$

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1}|trW_n^{-1}(\alpha, h)V_n(\alpha) - \varepsilon'M_n(\alpha, h)\varepsilon|/L_n(\alpha, h) = o_p(1) \quad (\text{II.5.6})$$

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1}|tr\hat{W}_n^{-1}(\alpha, h)\hat{V}_n(\alpha) - trW_n^{-1}(\alpha, h)V_n(\alpha)|/L_n(\alpha, h) = o_p(1) \quad (\text{II.5.7})$$

This will imply that

$$\frac{L_n(\alpha^{GC_L^*}, h^{GC_L^*})}{\inf_{\alpha \in A_n} \inf_{h \geq 0} L_n(\alpha, h)} \longrightarrow 1 \quad \text{in prob.}$$

.

Before proving (5.4)-(5.7) we define as in chapter 1

$$\tilde{L}_n^*(\alpha, h) \equiv \tilde{\Delta}_n(\alpha, h) + n^{-1}\tilde{\varepsilon}'\tilde{M}_n^2(\alpha, h)\tilde{\varepsilon}$$

and

$$\tilde{P}_n(\alpha, h) \equiv I_n - \tilde{M}_n(\alpha, h),$$

where $\tilde{\Delta}_n(\alpha, h) = n^{-1}\|\tilde{\eta} - \tilde{M}_n(\alpha, h)\|^2$ and

$$\tilde{M}_n(\alpha, h) \equiv \Lambda_n(\alpha)[\Lambda_n(\alpha)'\Lambda_n(\alpha) + hI_q]^{-1}\Lambda_n(\alpha)'.$$

We first verify condition (5.5).

Define $\tilde{B}_n^*(\alpha, h) \equiv \sum_{t=1}^n \tilde{\mu}_t^2(\lambda_t(\alpha) + h)^2$, as $n\tilde{L}_n^*(\alpha, h) \geq h^2\tilde{B}_n^*(\alpha, h)$, we get that

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\varepsilon' P_n(\alpha, h) \eta|}{n\tilde{L}_n^*(\alpha, h)} = \sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\tilde{\varepsilon}' \tilde{P}_n(\alpha, h) \tilde{\eta}|}{n\tilde{L}_n^*(\alpha, h)} \quad (\text{II.5.8})$$

$$\leq \sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\sum_{t=1}^n \tilde{\varepsilon}_t \tilde{\mu}_t(\lambda_t(\alpha) + h)^{-1}|}{(\tilde{B}_n^*(\alpha, h))^{1/2} (n\tilde{L}_n^*(\alpha, h))^{1/2}} \quad (\text{II.5.9})$$

so for (5.5) to hold it is enough to show that (5.9) tends to zero in probability.

For each n , let $I_1(j) = \{1, 2, \dots, j\}$ and $I_2(j) = \{j+1, \dots, n\}$. Recall that $\iota_n(\alpha)$ is the largest t such that $\lambda_t(\alpha) > 0$.

(5.9) will be $o_p(1)$ if for any natural k and for $l = 1, 2$

$$\sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \leq h} \frac{|\sum_{t \in I_l(k)} \tilde{\varepsilon}_t \tilde{\mu}_t(\lambda_t(\alpha) + h)^{-1}|}{(\tilde{B}_n^*(\alpha, h))^{1/2} (n\tilde{L}_n^*(\alpha, h))^{1/2}} = o_p(1) \quad (\text{II.5.10})$$

and $\forall \epsilon > 0 \exists c_1(\epsilon)$ and $c_2(\epsilon)$ such that $c_1(\epsilon) \rightarrow 0$, $c_2(\epsilon) \rightarrow 0$, $\epsilon \rightarrow \infty$,
and

$$P \left\{ \sup_{\alpha \in A_n} \sup_{j=k, \dots, \iota_n(\alpha)} \sup_{\lambda_{j+1}(\alpha) \leq h \leq \lambda_j(\alpha)} \frac{|\sum_{t \in I_l(j)} \tilde{\varepsilon}_t \tilde{\mu}_t(\lambda_t(\alpha) + h)^{-1}|}{(\tilde{B}_n^*(\alpha, h))^{1/2} (n\tilde{L}_n^*(\alpha, h))^{1/2}} > \varepsilon \right\} \leq c_l(\epsilon) \quad (\text{II.5.11})$$

for $l = 1, 2$

Proof of (5.10)

When $l = 1$ the proof is identical to the analog expression under independence shown in the first chapter.

When $l = 2$, it suffices to show that $\forall \varepsilon > 0$

$$P \left\{ \sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \leq h} \frac{|\sum_{t=k+1}^n \tilde{\varepsilon}_t \tilde{\mu}_t h (\lambda_t(\alpha) + h)^{-1}|}{(h^2 \tilde{B}_n^*(\alpha, h))^{1/2} (n \tilde{L}_n^*(\alpha, h))^{1/2}} \geq \epsilon \right\} \rightarrow 0 \quad (\text{II.5.12})$$

Since $h^2 \tilde{B}_n^*(\alpha, h)$ is non-decreasing in h , the left-side of (5.12) doesn't exceed

$$P \left\{ \sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \leq h} \frac{|\sum_{t=k+1}^n \tilde{\varepsilon}_t \tilde{\mu}_t h (\lambda_t(\alpha) + h)^{-1}|}{\lambda_k(\alpha) (\tilde{B}_n^*(\alpha, \lambda_k(\alpha)))^{1/2} (n \tilde{L}_n^*(\alpha, h))^{1/2}} \geq \epsilon \right\} \leq \quad (\text{II.5.13})$$

$$P \left\{ \sup_{\alpha \in A_n} \frac{|\sum_{t=k+1}^n \tilde{\varepsilon}_t \tilde{\mu}_t \lambda_k(\alpha) (\lambda_t(\alpha) + \lambda_k(\alpha))^{-1} \lambda_t(\alpha) + \lambda_k(\alpha))^{-1}|}{\lambda_k(\alpha) (\tilde{B}_n^*(\alpha, \lambda_k(\alpha)))^{1/2} (n \tilde{L}_n^*(\alpha, h))^{1/2}} \geq \epsilon/2 \right\} + \quad (\text{II.5.14})$$

$$P \left\{ \sup_{\alpha \in A_n} \sup_{\lambda_k(\alpha) \leq h} \frac{|\sum_{t=k+1}^n \tilde{\varepsilon}_t \tilde{\mu}_t [h (\lambda_t(\alpha) + h)^{-1} - \lambda_k(\alpha) (\lambda_t(\alpha) + \lambda_k(\alpha))^{-1}]|}{\lambda_k(\alpha) (\tilde{B}_n^*(\alpha, \lambda_k(\alpha)))^{1/2} (n \tilde{L}_n^*(\alpha, h))^{1/2}} \geq \epsilon/2 \right\}. \quad (\text{II.5.15})$$

By applying the generalized Chebychev inequality, (5.14) is no greater than

$$\sum_{\alpha \in A_n} \left(\frac{1}{2} \epsilon \right)^{-2m} E \left[\left(\sum_{t=k+1}^n \frac{\tilde{\varepsilon}_t \tilde{\mu}_t \lambda_k(\alpha) (\lambda_t(\alpha) + \lambda_k(\alpha))^{-1}}{\lambda_k(\alpha) (\tilde{B}_n^*(\alpha, \lambda_k(\alpha)))^{1/2} (n \tilde{L}_n^*(\alpha, h))^{1/2}} \right)^{2m} \right]. \quad (\text{II.5.16})$$

Now we have to bound the expectation in (5.16). For that purpose let $\gamma_t(\alpha) \equiv \tilde{\mu}_t \lambda_k(\alpha) (\lambda_t(\alpha) + \lambda_k(\alpha))^{-1}$, $G_n(\alpha, h) \equiv \lambda_k(\alpha) (\tilde{B}_n^*(\alpha, \lambda_k(\alpha)))^{1/2} (n \tilde{L}_n^*(\alpha, h))^{1/2}$ and note that $\tilde{\varepsilon}_t = U_t(\alpha)' \varepsilon = u_{1t}(\alpha) \varepsilon_1 + \dots + u_{nt}(\alpha) \varepsilon_n$ where $U_t(\alpha) \equiv [u_{1t}(\alpha), \dots, u_{nt}(\alpha)]'$ denotes the t^{th} column of the orthogonal matrix $U(\alpha)$ and ε is the vector of errors.

So we get that the expectation term in (5.16) equals

$$E \left[G_n^{-2m}(\alpha, h) \left(\sum_{t=k+1}^n (u_{1t}(\alpha) \varepsilon_1 + \dots + u_{nt}(\alpha) \varepsilon_n) \gamma_t(\alpha) \right)^{2m} \right] =$$

$$\begin{aligned}
& E \left[G_n^{-2m}(\alpha, h) \left(\varepsilon_1 \sum_{t=k+1}^n u_{1t} \gamma_t(\alpha) + \dots + \varepsilon_n \sum_{t=k+1}^n u_{nt} \gamma_t(\alpha) \right)^{2m} \right] = \\
& E \left[G_n^{-2m}(\alpha, h) \left(\sum_{j=1}^n \varepsilon_j \phi_j(\alpha) \right)^{2m} \right] = E \left[\left(\sum_{j=1}^n \varepsilon_j \tilde{\phi}_j(\alpha, h) \right)^{2m} \right] \quad (\text{II.5.17})
\end{aligned}$$

where $\phi_j(\alpha) \equiv \sum_{t=k+1}^n u_{jt}(\alpha) v \gamma_t(\alpha)$ and $\tilde{\phi}_j(\alpha, h) \equiv \phi_j(\alpha) / G_n(\alpha, h)$.

Now, by noticing that

$$\frac{\phi_j(\alpha)}{G_n(\alpha, h)} = \frac{\sum_{t=k+1}^n u_{jt}(\alpha) \tilde{\mu}_t \lambda_k(\alpha) (\lambda_t(\alpha) + \lambda_k(\alpha))^{-1}}{(\sum_{t=1}^n \tilde{\mu}_t^2 \lambda_k^2(\alpha) (\lambda_t(\alpha) + \lambda_k(\alpha))^{-2})^{1/2} (n L_n^*(\alpha, h))^{1/2}} \quad (\text{II.5.18})$$

$$\leq \frac{\sum_{t=k+1}^n |u_{jt}(\alpha)| |\tilde{\mu}_t| \lambda_k(\alpha) (\lambda_t(\alpha) + \lambda_k(\alpha))^{-1}}{\sum_{t=1}^n \tilde{\mu}_t^2 \lambda_k^2(\alpha) (\lambda_t(\alpha) + \lambda_k(\alpha))^{-2}} \quad (\text{II.5.19})$$

and taking into account that both $u_{jt}(\alpha)$ and $\lambda_k(\alpha)(\lambda_t(\alpha) + \lambda_k(\alpha))^{-1}$ are $O(1)$, it is easy to see that (5.18) is uniformly bounded in j and then we can applied theorem (5.1) with $b_j = \tilde{\phi}_j(\alpha, h)$, yielding that

$$E \left[\left(\sum_{j=1}^n \varepsilon_j \tilde{\phi}_j(\alpha, h) \right)^{2m} \right] \leq C \left(\sum_{j=1}^n (E |\tilde{\phi}_j(\alpha, h)|)^2 \right)^m$$

where C is a constant depending on m .

Now,

$$\begin{aligned}
\sum_{j=1}^n (E |\tilde{\phi}_j(\alpha, h)|)^2 &= \sum_{j=1}^n (E |G_n^{-1}(\alpha, h) \sum_{t=k+1}^n u_{jt}(\alpha) \gamma_t(\alpha)|)^2 \\
&\leq \sum_{j=1}^n E \left[G_n^{-2}(\alpha, h) \left(\sum_{t=k+1}^n u_{jt}(\alpha) \gamma_t(\alpha) \right)^2 \right] \quad (\text{II.5.20})
\end{aligned}$$

$$= E \left[G_n^{-2}(\alpha, h) \sum_{j=1}^n \left(\sum_{t=k+1}^n u_{jt}(\alpha) \gamma_t(\alpha) \right)^2 \right] \quad (\text{II.5.21})$$

$$= E \left[G_n^{-2}(\alpha, h) \left(\sum_{j=1}^n \sum_{t=k+1}^n u_{jt}^2(\alpha) \gamma_t^2(\alpha) + \sum_{j=1}^n \sum_{t \neq t'}^n u_{jt}(\alpha) u_{jt'}(\alpha) \gamma_t(\alpha) \gamma_{t'}(\alpha) \right) \right] \quad (\text{II.5.22})$$

$$= E \left[G_n^{-2}(\alpha, h) \left(\sum_{t=k+1}^n \gamma_t^2(\alpha) \sum_{j=1}^n u_{jt}^2(\alpha) + 2 \sum_{t \neq t'} \gamma_t(\alpha) \gamma_{t'}(\alpha) \sum_{j=1}^n u_{jt}(\alpha) u_{jt'}(\alpha) \right) \right] = \quad (\text{II.5.23})$$

$$E \left[G_n^{-2}(\alpha, h) \sum_{t=k+1}^n \gamma_t^2(\alpha) \right] = E \left[G_n^{-2}(\alpha, h) \sum_{t=k+1}^n \tilde{\mu}_t^2 \lambda_k(\alpha))^2 (\lambda_t(\alpha) + \lambda_k(\alpha))^{-2} \right] \quad (\text{II.5.24})$$

where (5.20) follows from Holder's inequality and (5.24) from the orthogonality of $U(\alpha)$.

Now, (5.24) doesn't exceed

$$E \left[\frac{\sum_{t=k+1}^n \tilde{\mu}_t^2 \lambda_k(\alpha))^2 (\lambda_t(\alpha) + \lambda_k(\alpha))^{-2}}{(\sum_{t=1}^n \tilde{\mu}_t^2 \lambda_k(\alpha))^2 (\lambda_t(\alpha) + \lambda_k(\alpha))^{-2}} \tilde{Q}_n^*(\alpha) \right] \quad (\text{II.5.25})$$

$$= E[O(1)(\tilde{Q}_n^*(\alpha))^{-1}] \leq C_1 E[(\tilde{Q}_n^*(\alpha))^{-1}]$$

where C_1 is a constant and $\tilde{Q}_n^*(\alpha) \equiv \inf_{h \geq 0} n \tilde{L}_n^*(\alpha, h)$.

So we get that (5.16) is no greater than

$$C_2 \left(\frac{1}{2} \epsilon \right)^{-2m} \sum_{\alpha \in A_n} [E(\tilde{Q}_n^*(\alpha))^{-1}]^m \leq C_2 \left(\frac{1}{2} \epsilon \right)^{-2m} \sum_{\alpha \in A_n} (E[\tilde{Q}_n^*(\alpha)])^{-m} \quad (\text{II.5.26})$$

where the last inequality follows from Jensen's inequality, and the result will follow from assumption B.3.

A similar bound follows for (5.15) which can be checked by following the same steps as its analog in chapter 1 (6.17) involving the application of lemma 6.1

and theorem 5.1 as we did above.

This concludes the proof of (5.10).

The proof of (5.11) also follows from its analog in the first chapter (6.11). The only difference is that here we use theorem 5.1 to replace Whittle's result and apply the same reasoning as for (5.10). The bound we will achieve is of the type

$$K(\epsilon) \sum_{\alpha \in A_n} [l_n(\alpha) - k] (E[\inf_{h \geq 0} nL_n^*(\alpha, h)])^{-m^*}$$

where $K(\epsilon)$ is a constant depending on ϵ and the result will follow from assumption B.4.

To verify condition (5.4) notice that

$$\begin{aligned} |L_n(\alpha, h) - L_n^*(\alpha, h)| &= 2n^{-1} |\eta'[I_n - M_n(\alpha, h)]M_n(\alpha, h)\varepsilon| \\ &= 2n^{-1} |\tilde{\eta}'[I_n - \tilde{M}_n(\alpha, h)]\tilde{M}_n(\alpha, h)\tilde{\varepsilon}| = 2n^{-1} \left| \sum_{t=1}^n \tilde{\varepsilon}_t \tilde{\mu}_t h \lambda_t(\alpha) (\lambda_t(\alpha) + h)^{-2} \right| \end{aligned}$$

and it therefore suffices to show that

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\sum_{t=1}^n \tilde{\varepsilon}_t \tilde{\mu}_t h \lambda_t(\alpha) (\lambda_t(\alpha) + h)^{-2}|}{(\tilde{B}_n^*(\alpha, h))^{1/2} (n\tilde{L}_n^*(\alpha, h))^{1/2}} = o_p(1) \quad (\text{II.5.27})$$

The proof of (5.27) is similar to that of (6.35) in chapter 1, with the exception that here we will need to apply theorem 5.1 instead of Whittle result as we did when verifying condition (5.5).

Now we verify condition (5.6).

First, observe that the left-hand side of (5.6) doesn't exceed

$$\left[\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h) \right]^{-1} \sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\varepsilon' M_n(\alpha, h) \varepsilon - \text{tr} W_n^{-1}(\alpha, h) V_n(\alpha)| \quad (\text{II.5.28})$$

$$\begin{aligned} &= \left[\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h) \right]^{-1} \sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\text{tr}[X_n(\alpha)' X_n(\alpha) + hI_{q_n}]^{-1} \\ &\quad \times X_n(\alpha)' \varepsilon \varepsilon' X_n(\alpha) - \text{tr} W_n^{-1}(\alpha, h) V_n(\alpha)| \end{aligned} \quad (\text{II.5.29})$$

so it will be enough to show that for some $0 < \delta < 1$

$$\begin{aligned} &\left[\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h) \right]^{-(1-\delta)} \sup_{\alpha \in A_n} \sup_{h \geq 0} \left\{ \left(n^{-1} [X_n(\alpha)' X_n(\alpha) + hI_{q_n}] \right)^{-1} \right. \\ &\quad \left. - \left(n^{-1} E[X_n(\alpha)' X_n(\alpha) + hI_{q_n}] \right)^{-1} \right\} = o_p(1) \end{aligned} \quad (\text{II.5.30})$$

and

$$\begin{aligned} &\left[\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h) \right]^{-\delta} \sup_{\alpha \in A_n} \left\{ n^{-1} X_n(\alpha)' \varepsilon \varepsilon' X_n(\alpha) \right. \\ &\quad \left. - E[n^{-1} X_n(\alpha)' \varepsilon \varepsilon' X_n(\alpha)] \right\} = O_p(1) \end{aligned} \quad (\text{II.5.31})$$

In the following we will denote by $\|\cdot\|$ the euclidian norm of a matrix or vector.

Now, (5.30) will hold if

$$\left\| n^{-1} [X'_{np_n} X_{np_n} - E[X'_{np_n} X_{np_n}]] \right\| = O_p(1) \quad (\text{II.5.32})$$

and

$$\lambda_{\min} \left\{ n^{-1} \sum_{t=1}^n E X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\} = C_n \quad (\text{II.5.33})$$

where $C_n (\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h))^{-(1-\delta)} = o_p(1)$.

We verify condition (5.32)

$$E\|n^{-1}[X'_{np_n}X_{np_n} - E[X'_{np_n}X_{np_n}]]^2 = \quad (\text{II.5.34})$$

$$E\left\{\sum_{j=1}^{p_n}\sum_{h=1}^{p_n}\left[n^{-1}\sum_{t=1}^n\psi_j(Z_t)\psi_h(Z_t) - n^{-1}E\sum_{t=1}^n\psi_j(Z_t)\psi_h(Z_t)\right]^2\right\} = \quad (\text{II.5.35})$$

$$\sum_{j=1}^{p_n}\sum_{h=1}^{p_n}n^{-2}\left\{\sum_{t=1}^nE\left[\psi_j(Z_t)\psi_h(Z_t) - E\psi_j(Z_t)\psi_h(Z_t)\right]^2 + \quad (\text{II.5.36})\right.$$

$$\left.2\sum_{\tau=1}^{n-1}\sum_{t=\tau+1}^nE\left[\psi_j(Z_t)\psi_h(Z_t) - E\psi_j(Z_t)\psi_h(Z_t)\right]\left[\psi_j(Z_{t-\tau})\psi_h(Z_{t-\tau}) - E\psi_j(Z_{t-\tau})\psi_h(Z_{t-\tau})\right]\right\} \quad (\text{II.5.37})$$

To bound (5.36) we follow the same argument as in chapter 6 of Galant and White (1988). Letting $\eta_t \equiv \psi_j(Z_t)\psi_h(Z_t) - E\psi_j(Z_t)\psi_h(Z_t)$ and $\hat{\eta}_{t\tau} \equiv E_{t-\tau-[\tau/2]}^{t-\tau+[\tau/2]}\eta_{t-\tau}$ where $E_{t-\tau-[\tau/2]}^{t-\tau+[\tau/2]}(\cdot) \equiv E(\cdot|F_{t-\tau-[\tau/2]}^{t-\tau+[\tau/2]})$ and get that

$$|E(\eta_t\eta_{t-\tau})| = |E[\eta_t\hat{\eta}_{t\tau} + \eta_t(\eta_{t-\tau} - \eta_t\hat{\eta}_{t\tau})]| \quad (\text{II.5.38})$$

$$\leq |E[\eta_t\hat{\eta}_{t\tau}]| + |E[\eta_t(\eta_{t-\tau} - \eta_t\hat{\eta}_{t\tau})]| \quad (\text{II.5.39})$$

Since $\hat{\eta}_{t\tau}$ is $F^{t-\tau+[\tau/2]}$ -measurable we get that

$$|E[\eta_t\hat{\eta}_{t\tau}]| = |E[E(\eta_t\hat{\eta}_{t\tau}|F^{t-\tau+[\tau/2]})]| \quad (\text{II.5.40})$$

$$= |E[E(\eta_t|F^{t-\tau+[\tau/2]})\hat{\eta}_{t\tau}]| \quad (\text{II.5.41})$$

$$\leq \|E^{t-\tau+[\tau/2]}\eta_t\|_2\|\hat{\eta}_{t\tau}\|_2 \quad (\text{II.5.42})$$

By the law of iterated expectations and the conditional Jensen's and Holder's inequalities we get that $\|\hat{\eta}_{t\tau}\|_2 \leq \|\eta_{t-\tau}\|_2 \leq \|\eta_{t-\tau}\|_r$ where $r > 2$.

Now, by applying Minkowski and Holder inequalities

$$\|\eta_{t-\tau}\|_r = \|\psi_j(Z_{t-\tau})\psi_h(Z_{t-\tau}) - E\psi_j(Z_{t-\tau})\psi_h(Z_{t-\tau})\|_r \quad (\text{II.5.43})$$

$$\leq \|\psi_j(Z_{t-\tau})\psi_h(Z_{t-\tau})\|_r + |E\psi_j(Z_{t-\tau})\psi_h(Z_{t-\tau})| \quad (\text{II.5.44})$$

$$\leq \|\psi_j(Z_{t-\tau})\|_{2r} \|\psi_h(Z_{t-\tau})\|_{2r} + \|\psi_j(Z_{t-\tau})\|_2 \|\psi_h(Z_{t-\tau})\|_2 \quad (\text{II.5.45})$$

$$\leq 2\|\psi_j(Z_{t-\tau})\|_{2r} \|\psi_h(Z_{t-\tau})\|_{2r} \leq 2\Delta^2 \quad (\text{II.5.46})$$

where the second inequality in (5.45) follows from the moment restrictions in B.6 (b).

Now,

$$\|E^{t-\tau+[\tau/2]}\eta_t\|_2 \leq 5\alpha_{[\tau/4]}^{1/2-1/r} \|\eta_t\|_r + \|\eta_t - E_{t-[\tau/4]}^{t+[\tau/4]}\eta_t\|_2 \quad (\text{II.5.47})$$

where the inequality holds by following the same argument as in the proof of lemma 3.14 of Gallant and White (1988).

It can also be shown that $\|\eta_t - E_{t-[\tau/4]}^{t+[\tau/4]}\eta_t\|_2 \leq K_1 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}}$, where $\nu \equiv \sum_{i=1}^d \nu_{im}$ and this follows from example 17.17 of Davidson (1994) under our assumptions.

Thus we get that

$$|E[\eta_t \hat{\eta}_{t\tau}]| \leq 2\Delta^2 (10\Delta^2 \alpha_{[\tau/4]}^{1/2-1/r} + K_1 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) \quad (\text{II.5.48})$$

which together with

$$|E[\eta_t(\eta_{t-\tau} - \eta_t \hat{\eta}_{t\tau})]| \leq \|\eta_t\|_2 \|\eta_{t-\tau} - \eta_t \hat{\eta}_{t\tau}\|_2 \quad (\text{II.5.49})$$

$$\leq \|\eta_t\|_2 K_2 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}} \quad (\text{II.5.50})$$

$$\leq 2\Delta^2 K_2 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}} \quad (\text{II.5.51})$$

yields that

$$|E(\eta_t \eta_{t-\tau})| \leq 2\Delta^2 (10\Delta^2 \alpha_{[\tau/4]}^{1/2-1/r} + K_1 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}} + K_2 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) \quad (\text{II.5.52})$$

$$= 2\Delta^2 (10\Delta^2 \alpha_{[\tau/4]}^{1/2-1/r} + K_3 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) \quad (\text{II.5.53})$$

We thus get that (5.36) doesn't exceed

$$2n \sum_{\tau=1}^{n-1} 2\Delta^2 (10\Delta^2 \alpha_{[\tau/4]}^{1/2-1/r} + K_3 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) \leq 4\Delta^2 n C_1 \quad (\text{II.5.54})$$

where $C_1 = \sum_{\tau=1}^{\infty} 2\Delta^2 (10\Delta^2 \alpha_{[\tau/4]}^{1/2-1/r} + K_3 \nu_{[\tau/4]}^{\frac{(r-2)^2}{4(r-1)^2}})$ which is finite under the mixing and near epoch dependence sizes in B.5.

The expectation term in (5.35) is no greater than

$$2\{E\psi_j^2(Z_t)\psi_h^2(Z_t) + (E\psi_j(Z_t)\psi_h(Z_t))^2\} \leq \quad (\text{II.5.55})$$

$$2\{\|\psi_j(Z_t)\|_4^2 \|\psi_h(Z_t)\|_4^2 + \|\psi_j(Z_t)\|_2^2 \|\psi_h(Z_t)\|_2^2\} \quad (\text{II.5.56})$$

$$\leq 4\|\psi_j(Z_t)\|_4^2 \|\psi_h(Z_t)\|_4^2 \leq 4\Delta^4 \quad (\text{II.5.57})$$

So finally we find that (5.34) is bounded by

$$\sum_{j=1}^{p_n} \sum_{h=1}^{p_n} (4\Delta^4 n^{-1} + 4\Delta^2 n^{-1} C_1) = C_2 p_n^2 n^{-1}$$

and (5.32) will hold if $p_n n^{-1/2} = O(1)$ as result of the generalized Chebyshev's inequality.

This concludes the proof of (5.30).

Now we verify (5.31)

For that purpose it suffices to show that

$$(\inf_{\alpha \in A_n} \inf_{h \geq 0} n L_n(\alpha, h))^{-\delta} \|n^{-1} X'_{np_n} \varepsilon \varepsilon' X_{np_n} - E[n^{-1} X'_{np_n} \varepsilon \varepsilon' X_{np_n}]\| = O_p(1) \quad (\text{II.5.58})$$

First, notice that

$$\begin{aligned}
E[n^{-1}X'_{np_n}\varepsilon\varepsilon'X_{np_n}] &= n^{-1}\sum_{t=1}^n E[\varepsilon_t^2 X_{p_n}(Z_t)X_{p_n}(Z_t)'] \\
&+ n^{-1}\sum_{\tau=1}^{n-1}\sum_{t=\tau+1}^n \{E[\varepsilon_t\varepsilon_{t-\tau}X_{p_n}(Z_t)X_{p_n}(Z_{t-\tau})'] \\
&+ E[\varepsilon_t\varepsilon_{t-\tau}X_{p_n}(Z_{t-\tau})X_{p_n}(Z_t)']\} \\
&= n^{-1}\sum_{t=1}^n E[\varepsilon_t^2 X_{p_n}(Z_t)X_{p_n}(Z_t)']
\end{aligned}$$

This is because

$$\begin{aligned}
E[\varepsilon_t\varepsilon_{t-\tau}X_{p_n}(Z_t)X_{p_n}(Z_{t-\tau})'] &= E[E[\varepsilon_t\varepsilon_{t-\tau}X_{p_n}(Z_t)X_{p_n}(Z_{t-\tau})'|\mathfrak{T}_{t-\tau}]] \\
&= E[E[\varepsilon_tX_{p_n}(Z_t)|\mathfrak{T}_{t-\tau}]\varepsilon_{t-\tau}X_{p_n}(Z_{t-\tau})'] = 0
\end{aligned}$$

where $\mathfrak{T}_t = \sigma(\varepsilon_t Z_t, \varepsilon_{t-1} Z_{t-1}, \dots)$.

Now,

$$\begin{aligned}
&\|n^{-1}X'_{np_n}\varepsilon\varepsilon'X_{np_n} - E[n^{-1}X'_{np_n}\varepsilon\varepsilon'X_{np_n}]\| \\
&= \left\| n^{-1}\sum_{t=1}^n \varepsilon_t^2 X_{p_n}(Z_t)X_{p_n}(Z_t)' - n^{-1}\sum_{t=1}^n E[\varepsilon_t^2 X_{p_n}(Z_t)X_{p_n}(Z_t)'] \right. \\
&+ n^{-1}\sum_{\tau=1}^{n-1}\sum_{t=\tau+1}^n \{\varepsilon_t\varepsilon_{t-\tau}X_{p_n}(Z_t)X_{p_n}(Z_{t-\tau})' + \varepsilon_t\varepsilon_{t-\tau}X_{p_n}(Z_{t-\tau})X_{p_n}(Z_t)'\} \left. \right\| \\
&\leq \left\| n^{-1}\sum_{t=1}^n \varepsilon_t^2 X_{p_n}(Z_t)X_{p_n}(Z_t)' - n^{-1}\sum_{t=1}^n E[\varepsilon_t^2 X_{p_n}(Z_t)X_{p_n}(Z_t)'] \right\| \quad (\text{II.5.59})
\end{aligned}$$

$$+ \left\| n^{-1}\sum_{\tau=1}^{n-1}\sum_{t=\tau+1}^n \{\varepsilon_t\varepsilon_{t-\tau}X_{p_n}(Z_t)X_{p_n}(Z_{t-\tau})' + \varepsilon_t\varepsilon_{t-\tau}X_{p_n}(Z_{t-\tau})X_{p_n}(Z_t)'\} \right\| \quad (\text{II.5.60})$$

and we will handle these two terms by first showing that (5.59) is $O_p(1)$.

Begin noticing that (5.59) equals

$$\begin{aligned}
& E \left\{ \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} \left[n^{-1} \sum_{t=1}^n \varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t) - n^{-1} E \sum_{t=1}^n \varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t) \right]^2 \right\} \\
&= \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} n^{-2} \left\{ \sum_{t=1}^n E \left[\varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t) - E \varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t) \right]^2 \right. \quad (\text{II.5.61}) \\
&\quad \left. + 2 \sum_{\tau=1}^{n-1} \sum_{t=\tau+1}^n E \left[\varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t) - E \varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t) \right] \right. \\
&\quad \left. \times \left[\varepsilon_{t-\tau}^2 \psi_j(Z_{t-\tau}) \psi_h(Z_{t-\tau}) - E \varepsilon_{t-\tau}^2 \psi_j(Z_{t-\tau}) \psi_h(Z_{t-\tau}) \right] \right\} \quad (\text{II.5.62})
\end{aligned}$$

The expectation term in (5.61) is no greater than

$$2 \{ E \varepsilon_t^4 \psi_j^2(Z_t) \psi_h^2(Z_t) + (E \varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t))^2 \} \leq \quad (\text{II.5.63})$$

$$2 \{ \|\varepsilon\|_8^4 \|\psi_j(Z_t)\|_8^2 \|\psi_h(Z_t)\|_8^2 + \|\varepsilon\|_4^4 \|\psi_j(Z_t)\|_4^2 \|\psi_h(Z_t)\|_4^2 \} \quad (\text{II.5.64})$$

$$\leq 4 \|\varepsilon\|_8^4 \|\psi_j(Z_t)\|_8^2 \|\psi_h(Z_t)\|_8^2 \leq 4\Delta^8 \quad (\text{II.5.65})$$

To bound (5.62) we let $\eta_t \equiv \varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t) - E \varepsilon_t^2 \psi_j(Z_t) \psi_h(Z_t)$ and follow the same steps taken to bound (5.36) above, arriving at the conclusion that (5.59) doesn't exceed

$$2n \sum_{\tau=1}^{n-1} 2\Delta^4 (10\Delta^4 \alpha_{[\tau/4]}^{1/2-1/r} + K \nu_{[\tau/4]}^{\frac{(r-2)^3}{8(r-1)^3}}) \leq 4\Delta^4 n C_3 \quad (\text{II.5.66})$$

where $C_3 = \sum_{\tau=1}^{\infty} 2\Delta^4 (10\Delta^4 \alpha_{[\tau/4]}^{1/2-1/r} + K \nu_{[\tau/4]}^{\frac{(r-2)^3}{8(r-1)^3}})$ which is finite under the mixing and near epoch dependence sizes in B.5.

So we finally get that (5.59) is bounded by

$$\sum_{j=1}^{p_n} \sum_{h=1}^{p_n} (4\Delta^8 n^{-1} + 4\Delta^4 n^{-1} C_3) = C_4 p_n^2 n^{-1}$$

and it will be bounded in probability if $p_n n^{-1/2} = O(1)$ as result of Chebychev's inequality.

We will handle now (5.60).

Now,

$$\begin{aligned}
& \left\| n^{-1} \sum_{\tau=1}^{n-1} \sum_{t=\tau+1}^n \left\{ \varepsilon_t \varepsilon_{t-\tau} X_{p_n}(Z_t) X_{p_n}(Z_{t-\tau})' + \varepsilon_t \varepsilon_{t-\tau} X_{p_n}(Z_{t-\tau}) X_{p_n}(Z_t)' \right\} \right\|_p \\
& \leq n^{-1} \sum_{\tau=1}^{n-1} \left\| \sum_{t=\tau+1}^n \left\{ \varepsilon_t \varepsilon_{t-\tau} X_{p_n}(Z_t) X_{p_n}(Z_{t-\tau})' + \varepsilon_t \varepsilon_{t-\tau} X_{p_n}(Z_{t-\tau}) X_{p_n}(Z_t)' \right\} \right\|_p \\
& \leq n^{-1} \sum_{\tau=1}^{n-1} \left\{ \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} E \left| \sum_{t=\tau+1}^n \left\{ \varepsilon_t \varepsilon_{t-\tau} \psi_j(Z_t) \psi_h(Z_{t-\tau}) + \varepsilon_t \varepsilon_{t-\tau} \psi_j(Z_{t-\tau}) \psi_h(Z_t) \right\} \right|^p \right\}^{1/p} \\
& \leq n^{-1} \sum_{\tau=1}^{n-1} \left\{ \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} 2^{p-1} \left(E \left| \sum_{t=\tau+1}^n \varepsilon_t \varepsilon_{t-\tau} \psi_j(Z_t) \psi_h(Z_{t-\tau}) \right|^p \right. \right. \\
& \quad \left. \left. + E \left| \sum_{t=\tau+1}^n \varepsilon_t \varepsilon_{t-\tau} \psi_j(Z_{t-\tau}) \psi_h(Z_t) \right|^p \right) \right\}^{1/p}
\end{aligned}$$

where the last inequality follows from applying the c_r inequality.

Now, because

$$E \left| \varepsilon_t \varepsilon_{t-\tau} \psi_j(Z_t) \psi_h(Z_{t-\tau}) \right|^p \leq \|\varepsilon_t\|_{4p}^p \|\varepsilon_{t-\tau}\|_{4p}^p \|\psi_j(Z_t)\|_{4p}^p \|\psi_h(Z_{t-\tau})\|_{4p}^p \leq \Delta^{4p}$$

and by a straightforward generalization of lemma 6.7(b) of Gallant and White (1988) to L^p norms we get that the last two summands above don't exceed $C p_n^{2/p} n^{1/p} (\sum_{\tau=1}^{n-1} (\tau+2))^{1/p}$, where C is a constant that depends on p . Therefore (5.31) will hold if for some $p > 0$ and δ as specified above

$$(p_n^2 n)^{1/p} \left(\sum_{\tau=1}^{n-1} (\tau+2) \right)^{1/p} \left(\inf_{\alpha \in A_n} \inf_{h \geq 0} n L_n(\alpha, h) \right)^{-\delta} = O_p(1)$$

This concludes the proof of (5.6)

Now we verify condition (5.7)

Notice that the left-hand side of (5.7) doesn't exceed

$$\left[\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h) \right]^{-1} \sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |tr \hat{W}_n^{-1}(\alpha, h) \hat{V}_n(\alpha) - tr W_n^{-1}(\alpha, h) V_n(\alpha)|$$

so it will be enough to show that for $0 < \delta < 1$

$$\begin{aligned} & \left[\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h) \right]^{-(1-\delta)} \sup_{\alpha \in A_n} \sup_{h \geq 0} \left\{ \left(n^{-1} [X_n(\alpha)' X_n(\alpha) + h I_{q_n}] \right)^{-1} \right. \\ & \quad \left. - \left(n^{-1} E[X_n(\alpha)' X_n(\alpha) + h I_{q_n}] \right)^{-1} \right\} = o_p(1) \end{aligned} \quad (\text{II.5.67})$$

and

$$\begin{aligned} & \left[\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h) \right]^{-\delta} \sup_{\alpha \in A_n} \left\{ n^{-1} \sum_{t=1}^n \hat{\varepsilon}_t^2 x_t(\alpha) x_t(\alpha)' - E[n^{-1} X_n(\alpha)' \varepsilon \varepsilon' X_n(\alpha)] \right\} \\ & \quad = O_p(1) \end{aligned} \quad (\text{II.5.68})$$

Condition (5.67) follows from (5.30) so we only have to show (5.68), which will hold if

$$\left[\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h) \right]^{-\delta} \left\| n^{-1} \sum_{t=1}^n \hat{\varepsilon}_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)' - E[n^{-1} X'_{np_n} \varepsilon \varepsilon' X_{np_n}] \right\| = O_p(1) \quad (\text{II.5.69})$$

Let $\hat{V}_n^* \equiv n^{-1} X'_{np_n} \hat{\varepsilon} \hat{\varepsilon}' X_{np_n}$ and $\tilde{V}_n^* \equiv E[n^{-1} X'_{np_n} \varepsilon \varepsilon' X_{np_n}]$ and notice that $\hat{\varepsilon}_t = \varepsilon_t - (\hat{\mu}(Z_t) - \mu(Z_t))$, where $\hat{\mu}(Z_t)$ denotes a consistent estimator of $\mu(Z_t)$.

So we get that

$$\hat{V}_n^* - \tilde{V}_n^* = n^{-1} \sum_{t=1}^n (\varepsilon_t - (\hat{\mu}(Z_t) - \mu(Z_t)))^2 X_{p_n}(Z_t) X_{p_n}(Z_t)'$$

$$-n^{-1} \sum_{t=1}^n E[\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)'] \quad (\text{II.5.70})$$

$$= n^{-1} \sum_{t=1}^n (\varepsilon_t^2 + (\hat{\mu}(Z_t) - \mu(Z_t))^2 - 2\varepsilon_t(\hat{\mu}(Z_t) - \mu(Z_t))) X_{p_n}(Z_t) X_{p_n}(Z_t)'$$

$$-n^{-1} \sum_{t=1}^n E[\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)'] \quad (\text{II.5.71})$$

$$= n^{-1} \sum_{t=1}^n [\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)' - E\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)']$$

$$-2n^{-1} \sum_{t=1}^n \varepsilon_t(\hat{\mu}(Z_t) - \mu(Z_t)) X_{p_n}(Z_t) X_{p_n}(Z_t)'$$

$$+n^{-1} \sum_{t=1}^n (\hat{\mu}(Z_t) - \mu(Z_t))^2 X_{p_n}(Z_t) X_{p_n}(Z_t)' \quad (\text{II.5.72})$$

$$= V_{1n}^* + V_{2n}^* + V_{3n}^*$$

So $\|\hat{V}_n^* - \tilde{V}_n^*\| \leq \|V_{1n}^*\| + \|V_{2n}^*\| + \|V_{3n}^*\|$ and it suffices to show that $\|V_{in}^*\| = O_p(1)$ for $i = 1, 2, 3$.

The bound for $\|V_{1n}^*\|$ follows from (5.58), which was shown above.

Now we bound $\|V_{2n}^*\|$.

Now,

$$\|V_{2n}^*\| = 4 \left\| n^{-1} \sum_{t=1}^n \varepsilon_t(\hat{\mu}(Z_t) - \mu(Z_t)) X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\|$$

$$\leq 4 \left\| n^{-1} \sum_{t=1}^n |\varepsilon_t| |\hat{\mu}(Z_t) - \mu(Z_t)| X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\|$$

$$\leq 4 \sup_{t \leq n} |\hat{\mu}(Z_t) - \mu(Z_t)| \left\| n^{-1} \sum_{t=1}^n |\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\|$$

$$= 4 \sup_{t \leq n} |\hat{\mu}(Z_t) - \mu(Z_t)| \left\| n^{-1} \sum_{t=1}^n (|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)' - E|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)') \right.$$

$$\left. - E|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)' + E|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\|$$

$$\leq 4 \sup_{t \leq n} |\hat{\mu}(Z_t) - \mu(Z_t)| \left\| n^{-1} \sum_{t=1}^n (|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)' - E|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)') \right\| \quad (\text{II.5.73})$$

$$+ 4 \sup_{t \leq n} |\hat{\mu}(Z_t) - \mu(Z_t)| \left\| n^{-1} \sum_{t=1}^n E|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\| \quad (\text{II.5.74})$$

and we need to show that both (5.73) and (5.74) are $O_p(1)$

(5.73) will be $O_p(1)$ if $p_n n^{-1/2} = O(1)$ and $\sup_{t \leq n} |\hat{\mu}(Z_t) - \mu(Z_t)| = O_p(1)$ by following the same argument as in (5.58).

Now we bound (5.74).

Notice for that purpose that

$$\begin{aligned} \left\| n^{-1} \sum_{t=1}^n E|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\| &\leq n^{-1} \sum_{t=1}^n \left\| E|\varepsilon_t| X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\| \\ &= n^{-1} \sum_{t=1}^n \left\{ \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} (E[|\varepsilon_t| \psi_j(Z_t) \psi_h(Z_t)])^2 \right\}^{1/2} \\ &\leq n^{-1} \sum_{t=1}^n \left\{ \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} \|\varepsilon_t\|_2^2 \|\psi_j(Z_t)\|_4^2 \|\psi_h(Z_t)\|_4^2 \right\}^{1/2} \leq \Delta^3 p_n \end{aligned}$$

and therefore (5.74) will be bounded in probability if

$$\left[\inf_{\alpha \in A_n} \inf_{h \geq 0} n L_n(\alpha, h) \right]^{-\delta} p_n \sup_{t \leq n} |\hat{\mu}(Z_t) - \mu(Z_t)| = O_p(1).$$

Finally, we find a bound for $\|V_{3n}\|$.

Now,

$$\begin{aligned}
\|V_{3n}^*\| &= \left\| n^{-1} \sum_{t=1}^n (\hat{\mu}(Z_t) - \mu(Z_t))^2 X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\| \\
&\leq \sup_{t \leq n} (\hat{\mu}(Z_t) - \mu(Z_t))^2 \left\| n^{-1} \sum_{t=1}^n X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\| \\
&= \sup_{t \leq n} (\hat{\mu}(Z_t) - \mu(Z_t))^2 \left\| n^{-1} \sum_{t=1}^n (X_{p_n}(Z_t) X_{p_n}(Z_t)' - E[X_{p_n}(Z_t) X_{p_n}(Z_t)']) \right. \\
&\quad \left. + E[X_{p_n}(Z_t) X_{p_n}(Z_t)'] \right\| \\
&\leq \sup_{t \leq n} (\hat{\mu}(Z_t) - \mu(Z_t))^2 \left\| n^{-1} \sum_{t=1}^n (X_{p_n}(Z_t) X_{p_n}(Z_t)' \right. \\
&\quad \left. - E[X_{p_n}(Z_t) X_{p_n}(Z_t)']) \right\| \\
&\quad + \sup_{t \leq n} (\hat{\mu}(Z_t) - \mu(Z_t))^2 \left\| n^{-1} \sum_{t=1}^n [X_{p_n}(Z_t) X_{p_n}(Z_t)'] \right\|
\end{aligned}$$

and from here it is obvious that $\|V_{3n}\|$ will be $O_p(1)$ under the same conditions that guarantee that $\|V_{2n}\|$ is also bounded.

Proof of theorem 3.2:

First notice that the squared norm in (3.2) can be expressed in the following way

$$\begin{aligned}
\| [I_n - \ddot{M}_n(\alpha, h)]^{-1} [Y - \hat{\eta}(\alpha, h)] \|^2 &= (Y - \hat{\eta}(\alpha, h))' (I_n - \ddot{M}_n(\alpha, h))^{-2} (Y - \hat{\eta}(\alpha, h)) \\
&= (Y - \hat{\eta}(\alpha, h))' (I_n + 2\ddot{M}_n(\alpha, h) + 3\ddot{M}_n^2(\alpha, h) + \dots) (Y - \hat{\eta}(\alpha, h)) \\
&= (Y - \hat{\eta}(\alpha, h))' (Y - \hat{\eta}(\alpha, h)) + 2(Y - \hat{\eta}(\alpha, h))' \ddot{M}_n(\alpha, h) (Y - \hat{\eta}(\alpha, h)) \\
&\quad + 3(Y - \hat{\eta}(\alpha, h))' \ddot{M}_n^2(\alpha, h) (Y - \hat{\eta}(\alpha, h)) + \dots
\end{aligned}$$

and therefore

$$CV_{n,-1}(\alpha, h) = n^{-1} (Y - \hat{\eta}(\alpha, h))' (Y - \hat{\eta}(\alpha, h))$$

$$+2n^{-1}(Y - \hat{\eta}(\alpha, h))' \ddot{M}_n(\alpha, h)(Y - \hat{\eta}(\alpha, h)) + O_p\left(\frac{h_n T_n(\alpha, h)}{n}\right)$$

where $T_n(\alpha, h) \equiv (Y - \hat{\eta}(\alpha, h))' \ddot{M}_n(\alpha, h)(Y - \hat{\eta}(\alpha, h))$ and

$h_n = \sup_{\alpha \in A_n} \sup_{h \geq 0} \bar{\lambda}(M_n(\alpha, h))$ with $\bar{\lambda}(\cdot)$ denoting the maximal diagonal element of a matrix.

It can be easily shown that

$$CV_{n,-1}(\alpha, h) = n^{-1} \|\varepsilon\|^2 + L_n(\alpha, h) + 2n^{-1} \varepsilon'(I_n - M_n(\alpha, h))\eta \\ + 4n^{-1} T_{1n}(\alpha, h) + 2n^{-1} T_{2n}(\alpha, h) + 2n^{-1} T_{3n}(\alpha, h) + O_p\left(\frac{h_n T_n(\alpha, h)}{n}\right)$$

where

$$T_{1n}(\alpha, h) \equiv \varepsilon'(I_n - M_n(\alpha, h)) \ddot{M}_n(\alpha, h)(I_n - M_n(\alpha, h))\eta$$

$$T_{2n}(\alpha, h) \equiv \eta'(I_n - M_n(\alpha, h)) \ddot{M}_n(\alpha, h)(I_n - M_n(\alpha, h))\eta$$

and

$$T_{3n}(\alpha, h) \equiv \varepsilon'[(I_n - M_n(\alpha, h)) \ddot{M}_n(\alpha, h)(I_n - M_n(\alpha, h)) - M_n(\alpha, h)]\eta$$

and asymptotic loss-efficiency of $CV_{n,-1}$ will follow if we show that

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} |L_n(\alpha, h)/L_n^*(\alpha, h) - 1| = o_p(1), \quad (\text{II.5.75})$$

and

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\varepsilon' P_n(\alpha, h)\eta|/L_n^*(\alpha, h) = o_p(1) \quad (\text{II.5.76})$$

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |T_{in}(\alpha, h)|/L_n^*(\alpha, h) = o_p(1) \quad (\text{II.5.77})$$

for $i = 1, 2, 3$.

We already showed for the generalized AIC model selection procedure that (5.75) and (5.76) are $o_p(1)$ under the assumptions.

Now we show that (5.77) is $o_p(1)$ for $i = 1, 2, 3$.

For $i = 1$ the results follows as

$$\begin{aligned} \varepsilon'(I_n - M_n(\alpha, h))\ddot{M}_n(\alpha, h)(I_n - M_n(\alpha, h))\eta \\ \leq h_n \varepsilon'(I_n - M_n(\alpha, h))^2 \eta \end{aligned}$$

so it is enough to show that

$$\sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\varepsilon'(I_n - M_n(\alpha, h))^2 \eta| / L_n^*(\alpha, h) = o_p(1) \quad (\text{II.5.78})$$

but (5.78) equals

$$\begin{aligned} \sup_{\alpha \in A_n} \sup_{h \geq 0} n^{-1} |\varepsilon'(I_n - \tilde{M}_n(\alpha, h))^2 \tilde{\eta}| / \tilde{L}_n^*(\alpha, h) \\ \leq \sup_{\alpha \in A_n} \sup_{h \geq 0} \frac{|\sum_{t=1}^n \tilde{\varepsilon}_t \tilde{\mu}_t h^2 (\lambda_t(\alpha) + h)^{-2}|}{h(\tilde{B}_n^*(\alpha, h))^{1/2} (n\tilde{L}_n^*(\alpha, h))^{1/2}} \end{aligned} \quad (\text{II.5.79})$$

and it can be readily seen that (5.79) is $o_p(1)$ by following the same lines as in (5.5) and taking into account that $h^2((\lambda_t(\alpha) + h)^{-2}) \leq 1$.

We now verify (5.77) for $i = 2$.

For this purpose notice that

$$\begin{aligned} T_{2n}(\alpha, h) &= \eta'(I_n - M_n(\alpha, h))\ddot{M}_n(\alpha, h)(I_n - M_n(\alpha, h))\eta \\ &\leq h_n \eta'(I_n - M_n(\alpha, h))^2 \eta \leq h_n n L_n^*(\alpha, h) \end{aligned}$$

so $|T_{2n}(\alpha, h)|/nL_n^*(\alpha, h) \leq h_n$ and the result follows as $h_n = o_p(1)$ by assumption.

Finally we verify (5.77) for $i = 3$.

Now,

$$\begin{aligned}
|T_{3n}(\alpha, h)| &= |\varepsilon'[(I_n - M_n(\alpha, h))\ddot{M}_n(\alpha, h)(I_n - M_n(\alpha, h)) - M_n(\alpha, h)]\varepsilon| \\
&= |\varepsilon'[\ddot{M}_n(\alpha, h) - M_n(\alpha, h)\ddot{M}_n(\alpha, h) - \ddot{M}_n(\alpha, h)M_n(\alpha, h) \\
&\quad + M_n(\alpha, h)\ddot{M}_n(\alpha, h)M_n(\alpha, h) - M_n(\alpha, h)]\varepsilon| \\
&\leq C|\varepsilon'\ddot{M}_n(\alpha, h)\varepsilon - \varepsilon'M_n(\alpha, h)\varepsilon|
\end{aligned}$$

where C denotes a finite constant.

By following the same steps than for showing condition (5.6) it is straightforward to see that (5.77) for $i = 3$ will hold if

$$\|n^{-1}[X'_{np_n}X_{np_n} - E[X'_{np_n}X_{np_n}]]\| = O_p(1),$$

$$\lambda_{\min}\left\{n^{-1}\sum_{t=1}^n EX_{p_n}(Z_t)X_{p_n}(Z_t)'\right\} = C_n$$

where $C_n(\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h))^{-(1-\delta)} = o_p(1)$

and

for some $p > 0$ and δ as above

$$(p_n^2 n)^{1/p}(\inf_{\alpha \in A_n} \inf_{h \geq 0} nL_n(\alpha, h))^{-\delta} = O_p(1).$$

Proof of theorem 5.1:

For simplicity we will only consider the case $m = 2$. Analog results follow for $m > 2$ though the algebra becomes more complicated. We let C_1 and C_2 be generic constants that will be used throughout.

Now,

$$E\left[\left(\sum_{j=1}^n \varepsilon_j b_j\right)^4\right] = \sum_{j=1}^n E[\varepsilon_j^4 b_j^4] + \sum_{j \neq k} E[\varepsilon_j^2 b_j^2 \varepsilon_k^2 b_k^2] + \sum_{j \neq k} E[\varepsilon_j^3 b_j^3 \varepsilon_k b_k] \quad (\text{II.5.80})$$

$$+ \sum_{j \neq k \neq l} E[\varepsilon_j^2 b_j^2 \varepsilon_k b_k \varepsilon_l b_l] + \sum_{j \neq k \neq l \neq m} E[\varepsilon_j b_j \varepsilon_k b_k \varepsilon_l b_l \varepsilon_m b_m] \quad (\text{II.5.81})$$

We will proceed to bound each of the summands above.

Now the first summand in (5.80) equals

$$\sum_{j=1}^n E[\varepsilon_j^4 (E|b_j|)^4 \frac{b_j^4}{(E|b_j|)^4}] = \sum_{j=1}^n E[\varepsilon_j^4 \tilde{b}_j^4 d_j^4] \quad (\text{II.5.82})$$

where $\tilde{b}_j \equiv E|b_j|$ and $d_j \equiv b_j/E|b_j| = b_j/\tilde{b}_j$.

By Holder's inequality the right hand-side of (5.82) doesn't exceed

$$\begin{aligned} \sum_{j=1}^n E|\varepsilon_j^4 \tilde{b}_j^4| \|d_j^4\|_\infty &= \sum_{j=1}^n E|\varepsilon_j^4| \|\tilde{b}_j^4\| \|d_j^4\|_\infty \\ &\leq C_1 \sum_{j=1}^n \tilde{b}_j^4 \leq C_2 \left(\sum_{j=1}^n (E|b_j|)^2\right)^2 \end{aligned}$$

where the first inequality follows since $\sup_j E|\varepsilon_j^4| < \infty$ and $\sup_j \|d_j^4\|_\infty < \infty$. under the theorem assumptions.

Similarly the second summand in (5.80) equals

$$\sum_{j \neq k} E[\varepsilon_j^2 \tilde{b}_j^2 d_j^2 \varepsilon_k^2 \tilde{b}_k^2 d_k^2] \leq \sum_{j \neq k} \|\varepsilon_j\|_4^2 \|\varepsilon_k\|_4^2 \|d_j^2 d_k^2\|_\infty \tilde{b}_j^2 \tilde{b}_k^2 \leq C_2 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2$$

where the second inequality holds as $\sup_{j,k} \|d_j^2 d_k^2\|_\infty < \infty$.

To bound the third summand in (5.80) we begin with

$$\begin{aligned} & \left| \sum_{j \neq k} E[\varepsilon_j^3 b_j^3 \varepsilon_k b_k] \right| \leq \sum_{j \neq k} |E[\varepsilon_j^3 b_j^3 \varepsilon_k b_k]| \\ &= \sum_{j \neq k} |E[\varepsilon_j^3 \tilde{b}_j^3 d_j^3 \varepsilon_k \tilde{b}_k d_k]| \leq \sum_{j \neq k} \tilde{b}_j^3 \tilde{b}_k \|d_j^3 d_k\|_\infty |E \varepsilon_j^3 \varepsilon_k| \\ &\leq \sum_{j \neq k} \tilde{b}_j^3 \tilde{b}_k \|d_j^3 d_k\|_\infty \left(\|\varepsilon_j^3\|_r (5\alpha_{[\frac{k-j}{4}]}^{1/2-1/r} \|\varepsilon_k\|_r + B_1 \nu_{[\frac{k-j}{4}]}) + \|\varepsilon_k\|_r B_2 \nu_{[\frac{k-j}{4}]}^{\frac{(r-2)^3}{8(r-1)^3}} \right) \quad (\text{II.5.83}) \end{aligned}$$

where $r > 2$.

Let

$$\vartheta_{[\frac{k-j}{4}]}^1 \equiv 5\alpha_{[\frac{k-j}{4}]}^{1/2-1/r} \|\varepsilon_k\|_r + B_1 \nu_{[\frac{k-j}{4}]} + B_2 \nu_{[\frac{k-j}{4}]}^{\frac{(r-2)^3}{8(r-1)^3}}$$

So (5.83) doesn't exceed

$$\begin{aligned} & C_1 \sum_{j \neq k} \tilde{b}_j^3 \tilde{b}_k \vartheta_{[\frac{k-j}{4}]}^1 \leq C_1 \sum_{j \neq k} (\tilde{b}_j^4 + \tilde{b}_j^2 \tilde{b}_k^2) \vartheta_{[\frac{k-j}{4}]}^1 \\ &= C_1 \sum_{j=1}^{n-1} \sum_{p=1}^{n-j} (\tilde{b}_j^4 + \tilde{b}_j^2 \tilde{b}_{j+p}^2) \vartheta_{[\frac{p}{4}]}^1 \\ &= C_1 \left\{ \sum_{j=1}^{n-1} \sum_{p=1}^{n-j} \tilde{b}_j^4 \vartheta_{[\frac{p}{4}]}^1 + \sum_{j=1}^{n-1} \sum_{p=1}^{n-j} \tilde{b}_j^2 \tilde{b}_{j+p}^2 \vartheta_{[\frac{p}{4}]}^1 \right\} \\ &\leq C_1 \left\{ \sum_{j=1}^{n-1} \tilde{b}_j^4 \sum_{p=1}^{n-j} \vartheta_{[\frac{p}{4}]}^1 + M_1 \sum_{j=1}^{n-1} \sum_{p=1}^{n-j} \tilde{b}_j^2 \tilde{b}_{j+p}^2 \right\} \\ &C_1 \left\{ D_1 \sum_{j=1}^{n-1} \tilde{b}_j^4 + M_1 \sum_{j=1}^{n-1} \sum_{p=1}^{n-j} \tilde{b}_j^2 \tilde{b}_{j+p}^2 \right\} \leq C_2 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2 \end{aligned}$$

where $D_1 \equiv \sum_{p=1}^{\infty} \vartheta_{[\frac{p}{4}]}^1 < \infty$ and $M_1 \geq \vartheta_{[\frac{p}{4}]}^1$.

Similarly, following the same lines of proof we get that

$$\left| \sum_{j \neq k} E[\varepsilon_j b_j \varepsilon_k^3 b_k^3] \right| \leq C_2 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2$$

and therefore we achieve the desired bound for the third summand in (5.88).

We now bound the first summand in (5.81).

We first show that $\left| \sum_{j \neq k \neq l} E[\varepsilon_j^2 b_j^2 \varepsilon_k b_k \varepsilon_l b_l] \right|$ is bounded. Begin noticing that it doesn't exceed

$$\begin{aligned} & \sum_{j \neq k \neq l} |E[\varepsilon_j^2 b_j^2 \varepsilon_k b_k \varepsilon_l b_l]| = \sum_{j \neq k \neq l} |E[\varepsilon_j^2 \tilde{b}_j^2 d_j^2 \varepsilon_k \tilde{b}_k d_k \varepsilon_l \tilde{b}_l d_l]| \\ & \leq \sum_{j \neq k \neq l} \tilde{b}_j^2 \tilde{b}_k \tilde{b}_l \|d_j^2 d_k d_l\|_{\infty} (\|\varepsilon_j^2 \varepsilon_k\|_r (5\alpha_{[\frac{k-j}{4}]}^{1/2-1/r} \|\varepsilon_l\|_r + B_1 \nu_{[\frac{k-j}{4}]}) + \|\varepsilon_l\|_r B_2 \nu_{[\frac{k-j}{4}]}^{\frac{(r-2)^3}{8(r-1)^3}}) \\ & \leq C_1 \sum_{j \neq k \neq l} \tilde{b}_j^2 \tilde{b}_k \tilde{b}_l \vartheta_{[\frac{l-k}{4}]}^2 \end{aligned} \quad (\text{II.5.84})$$

where $\vartheta_{[\frac{l-k}{4}]}^2$ is defined similarly to $\vartheta_{[\frac{k-j}{4}]}^1$.

Now, (5.84) is bounded by

$$\begin{aligned} & C_1 \sum_{j \neq k \neq l} (\tilde{b}_j^2 \tilde{b}_k^2 + \tilde{b}_j^2 \tilde{b}_l^2) \vartheta_{[\frac{l-k}{4}]}^2 = C_2 \sum_{j=1}^{n-2} \sum_{p=1}^{n-j-1} \sum_{q=1}^{n-j-p} (\tilde{b}_j^2 \tilde{b}_{j+p}^2 + \tilde{b}_j^2 \tilde{b}_{j+p+q}^2) \vartheta_{[\frac{q}{4}]}^2 \\ & = C_2 \left\{ \sum_{j=1}^{n-2} \sum_{p=1}^{n-j-1} \sum_{q=1}^{n-j-p} \tilde{b}_j^2 \tilde{b}_{j+p}^2 \vartheta_{[\frac{q}{4}]}^2 + \sum_{j=1}^{n-2} \sum_{p=1}^{n-j-1} \sum_{q=1}^{n-j-p} \tilde{b}_j^2 \tilde{b}_{j+p+q}^2 \vartheta_{[\frac{q}{4}]}^2 \right\} \end{aligned} \quad (\text{II.5.85})$$

The first summand in (5.85) equals

$$\sum_{j=1}^{n-2} \sum_{p=1}^{n-j-1} \tilde{b}_j^2 \tilde{b}_{j+p}^2 \sum_{q=1}^{n-j-p} \vartheta_{[\frac{q}{4}]}^2 \leq D_2 \sum_{j=1}^{n-2} \sum_{p=1}^{n-j-1} \tilde{b}_j^2 \tilde{b}_{j+p}^2$$

$$\leq D_2 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2$$

where $D_2 \equiv \sum_{q=1}^{\infty} \vartheta_{[\frac{q}{4}]}^2$, which is finite under the theorem assumptions.

The second summand in (5.85) equals

$$\begin{aligned} \sum_{j=1}^{n-2} \tilde{b}_j^2 \sum_{p=1}^{n-j-1} \sum_{q=1}^{n-j-p} \tilde{b}_{j+p+q}^2 \vartheta_{[\frac{q}{4}]}^2 &\leq \sum_{j=1}^{n-2} \tilde{b}_j^2 \sum_{p=1}^{n-j} \tilde{b}_{j+p}^2 \sum_{q=1}^n \vartheta_{[\frac{q}{4}]}^2 \\ &\leq D_2 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2 \end{aligned}$$

By the same reasoning,

$$\left| \sum_{j \neq k \neq l} E[\varepsilon_j b_j \varepsilon_k^2 b_k^2 \varepsilon_l b_l] \right| \text{ and } \left| \sum_{j \neq k \neq l} E[\varepsilon_j b_j \varepsilon_k b_k \varepsilon_l^2 b_l^2] \right|$$

achieve the same bounds and therefore the first summand in (5.85) is bounded by $C_2 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2$.

We finally bound the second summand in (5.81). Consider first

$$\left| \sum_{j \neq k \neq l \neq m} E[\varepsilon_j b_j \varepsilon_k b_k \varepsilon_l b_l \varepsilon_m b_m] \right| \tag{II.5.86}$$

and the following cases:

- i) $m - l \geq (l - k, k - j)$
- ii) $l - k \geq m - l \geq k - j$
- iii) $l - k \geq k - j \geq m - l$
- iv) $k - j \geq m - l \geq l - k$
- v) $k - j \geq l - k \geq m - l$

Under case i) (5.86) doesn't exceed

$$\begin{aligned}
& \sum_{j < k < l < m} |E[\varepsilon_j b_j \varepsilon_k b_k \varepsilon_l b_l \varepsilon_m b_m]| \leq \sum_{j < k < l < m} \tilde{b}_j \tilde{b}_k \tilde{b}_l \tilde{b}_m \|d_j d_k d_l d_m\|_\infty |E\varepsilon_j \varepsilon_k \varepsilon_l \varepsilon_m| \\
& \leq C_1 \sum_{j < k < l < m} \tilde{b}_j \tilde{b}_k \tilde{b}_l \tilde{b}_m (\|\varepsilon_j \varepsilon_k \varepsilon_l\|_r (5\alpha_{[\frac{m-l}{4}]})^{1/2-1/r} \|\varepsilon_m\|_r + B_1 \nu_{[\frac{m-l}{4}]} + \|\varepsilon_m\|_r B_2 \nu_{[\frac{m-l}{4}]}^{\frac{(r-2)^3}{8(r-1)^3}}) \\
& \leq C_1 \sum_{j < k < l < m} \tilde{b}_j \tilde{b}_k \tilde{b}_l \tilde{b}_m \vartheta_{[\frac{m-l}{4}]}^3 \tag{II.5.87}
\end{aligned}$$

where $\vartheta_{[\frac{m-l}{4}]}^3 \equiv 5\alpha_{[\frac{m-l}{4}]}^{1/2-1/r} \|\varepsilon_m\|_r + B_1 \nu_{[\frac{m-l}{4}]} + B_2 \nu_{[\frac{m-l}{4}]}^{\frac{(r-2)^3}{8(r-1)^3}}$.

Thus (5.87) doesn't exceed

$$\begin{aligned}
& C_1 \sum_{j < k < l < m} (\tilde{b}_j^2 \tilde{b}_k^2 + \tilde{b}_l^2 \tilde{b}_m^2) \vartheta_{[\frac{m-l}{4}]}^3 = C_1 \sum_{j=1}^{n-3} \sum_{r=1}^{n-j-2} \sum_{p=1}^r \sum_{q=1}^r (\tilde{b}_j^2 \tilde{b}_{j+p}^2 + \tilde{b}_{j+p+q}^2 \tilde{b}_{j+p+q+r}^2) \vartheta_{[\frac{r}{4}]}^3 \\
& = C_1 \left\{ \sum_{j=1}^{n-3} \sum_{r=1}^{n-j-2} \sum_{p=1}^r \sum_{q=1}^r \tilde{b}_j^2 \tilde{b}_{j+p}^2 \vartheta_{[\frac{r}{4}]}^3 + \sum_{j=1}^{n-3} \sum_{r=1}^{n-j-2} \sum_{p=1}^r \sum_{q=1}^r \tilde{b}_{j+p+q}^2 \tilde{b}_{j+p+q+r}^2 \vartheta_{[\frac{r}{4}]}^3 \right\} \tag{II.5.88}
\end{aligned}$$

The first summand in (5.88) is bounded by

$$\begin{aligned}
& \sum_{j=1}^{n-3} \tilde{b}_j^2 \left(\sum_{p=1}^{n-j-2} \tilde{b}_{j+p}^2 \right) \sum_{r=1}^{n-j-2} \sum_{q=1}^r \vartheta_{[\frac{r}{4}]}^3 = \sum_{j=1}^{n-3} \sum_{p=1}^{n-j-2} \tilde{b}_j^2 \tilde{b}_{j+p}^2 \sum_{r=1}^{n-j-2} r \vartheta_{[\frac{r}{4}]}^3 \\
& \leq D_3 \sum_{j=1}^{n-3} \sum_{p=1}^{n-j-2} \tilde{b}_j^2 \tilde{b}_{j+p}^2 \leq D_3 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2
\end{aligned}$$

where $D_3 \equiv \sum_{r=1}^{\infty} r \vartheta_{[\frac{r}{4}]}^3 < \infty$ under the assumptions.

The second summand in (5.88) is bounded by

$$\sum_{j=1}^{n-3} \sum_{r=1}^{n-j-2} \sum_{p=1}^{n-j-2} \sum_{q=1}^r \tilde{b}_{j+p+q}^2 \tilde{b}_{j+p+q+r}^2 \vartheta_{[\frac{r}{4}]}^3 \leq \sum_{j=1}^{n-3} \left(\sum_{p=1}^{n-j} \tilde{b}_{j+p}^2 \right) \sum_{r=1}^{n-j-2} \sum_{q=1}^r \tilde{b}_{j+p+q+r}^2 \vartheta_{[\frac{r}{4}]}^3$$

$$\leq \sum_{j=1}^n \tilde{b}_j^2 \sum_{p=1}^n \tilde{b}_p^2 \sum_{r=1}^n r \vartheta_{[\frac{r}{4}]}^3 \leq D_3 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2$$

So we finally conclude that (5.86) under case i) is bounded by $C_1 \left(\sum_{j=1}^n \tilde{b}_j^2 \right)^2$.

It can be easily seen that the same bounds are attained for (5.86) under all the other cases by following the same lines of proof as above. The same conclusion holds for all the remaining different orderings of j, k, l and m .

II.6 References

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, 2nd International Symposium in Information Theory, Petrov, B.N. and Csaki, F. eds., Akademiai Kiado, Budapest, 267-281.

Andrews, D.W.K. (1991), Asymptotic Optimality of Generalized C_L , Cross-validation and Generalized Cross-validation in Regression with Heteroskedastic Errors, *Journal of Econometrics*, 47, 359-377.

Andrews, D.W.K. (1991), Asymptotic Normality of Series Estimators for Various Nonparametric and Semi-parametric Regression Models, *Econometrica*, 59, 307-345.

Burman, P., Chow, E. and Nolan, D. (1994), A Cross-validatory Method for Dependent Data, *Biometrika*, 84, 351-358.

Burman, P. and Nolan, D. (1992), Data-Dependent Estimation of prediction Functions, *Journal of Time Series Analysis*, 13, 189-207.

Burnham, K. and Anderson, D. (2003), *Model Selection and Multi-Model Inference*, Springer.

Candes, E. (1998), *Ridgelets: Theory and Applications*. Ph.D. Dissertation, Department of Statistics, Stanford University.

Candes, E. (1999), Harmonic Analysis of Neural Networks, *Applied and Computational Harmonic Analysis*, 6, 197-218.

- Candes, E. (2003), Ridgelets: Estimating with Ridge Functions, *Annals of Statistics*, 33, 1561-1599.
- Craven, P. and Wahba, G. (1979), Smoothing Noisy Data with Spline Functions, *Numerical Mathematics*, 31, 377-403.
- Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press.
- Frank, I.E. and Friedman, J. (1993), A Statistical View of Some Chemometrics Regression Tools, *Technometrics*, 35, 109-148.
- Gallant, R.A. and White, H. (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Oxford: Basil Blackwell.
- Fu, W. (1998), The Bridge versus the Lasso, *Journal of Computational and Graphical Statistics*, 7, 397-416.
- Hannan, E.J. and Quinn, B.G. (1979), The determination of the order of autoregression, *J.R. Statist. Soc. B*, 41, 190-195.
- Konishi, S. and Kitagawa, G. (1996), Generalised information criteria in model selection, *Biometrika*, 83, 875-890.
- Li, K.-C. (1986), Asymptotic Optimality of C_L and Generalized Cross-validation in Ridge regression with Applications to Spline Smoothing, *Annals of Statistics*, 14, 1101-1112.
- Li, K.-C. (1987), Asymptotic Optimality for C_p , C_L , Cross-validation and Generalized Cross-validation: Discrete Index Set, *Annals of Statistics*, 15, 958-975.
- Mallows, C. (1973), Some Comments on C_p , *Technometrics*, 15, 661-675.
- Newey, W.K. (1997), Convergence Rates and Asymptotic Normality for Series Estimators, *Journal of Econometrics*, 79, 147-168.
- Opsomer, J., Wang, Y. and Yang, Y. (2001), Nonparametric Regression with Correlated Errors, *Statistical Science*, 16, 134-153.
- Racine, J. (1997), Feasible Cross-validatory Model Selection for General Stationary Processes, *Journal of Applied Econometrics*, 12, 169-179.
- Shao, J. (1993), Linear Model Selection by Cross-validation, *Journal of the American Statistical Association*, 88, 486-494.

Shao, J. (1997), An Asymptotic Theory for Linear Model Selection, *Statistica Sinica*, 7, 221-264.

Stone, M. (1974), Cross-validatory Choice and Assessment of Statistical Predictions, *J.R. Statist. Soc. B*, 36, 111-147.

Schwarz, G. (1978), Estimating the Dimension of a Model, *Annals of Statistics*, 6, 461-464.

Takeuchi, K. (1976), Distribution of Informational Statistics and a Criterion of Model Fitting, *Suri-Kagaku (Mathematic Sciences)*, 153, 12-18.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, B*, 58, 267-288.

White, H. (2006), Approximate Nonlinear Forecasting Methods, *Handbook of Economic Forecasting*, Edited by Graham Elliott, Clive Granger and Allan Timmermann, Elsevier.

White, H. (2001), *Asymptotic Theory for Econometricians*, Academic Press.

Whittle, P. (1960), Bounds for the Moments of Linear and Quadratic Forms in Independent Variables, *Theory of Probability and its Applications*, 5, 302-305.

Yang, Y. (1999), Model Selection for Nonparametric Regression, *Statistica Sinica*, 9, 475-499.

III

Optimal Selection of Series Functions by h -Block Cross-validation

III.1 Introduction

This chapter constitutes an extension of chapter 2 and analyzes the asymptotic loss-efficiency of h -block cross-validation. This model selection procedure is a general version of leave-one-out cross-validation (which was analyzed extensively in the previous chapter) and is applicable when we are dealing with dependent data.

This model selection criteria was formally introduced by Burman et al. (1994) and studied from an empirical point of view by Racine (1997, 2000). It differs from leave-one-out cross-validation in that instead of removing 1 observation at each iteration, it removes $2h + 1$ (it actually removes at the t^{th} iteration not only the t^{th} observation but also the h observations preceding and following the t^{th}). The validation set continues to have one observation as the leave-1-out method but the estimation set only contains $n - 2h - 1$ observations (notice that

leave-1-out cross-validation arises when $h=1$).

The superiority of h -block over leave-one-out cross-validation arises when the errors of the data generating process are correlated. Remember that as we showed in the previous chapter, leave-one-out cross-validation is asymptotically optimal when applied to time series data as long as a condition implying the martingale difference assumption on the errors of the DGP is satisfied.

Remember that we formulated our data generating process as

$$Y_t = \mu(Z_t) + \varepsilon_t$$

where $\mu(Z_t) \equiv E[Y_t|Z_t]$ and Z_t is a $d \times 1$ vector of predictors (which may contain lag values of Y_t). We use series functions in order to approximate the unknown conditional expectation above given a number of predictors in our information set. When the set of conditioning predictors incorporates all the relevant lags of the variables we are using to predict, the error in the DGP will be a martingale difference (see section 3.5 of White (2001)). The omission of relevant past values of the predictors will make the errors correlated and thus the leave-1-out method will break down.

Since the models we are using to approximate the data generating process are nonparametric in nature (series estimators linear in the coefficients), the more predictors we include the slower will be the rate of convergence of the approximation to the unknown conditional expectation. There will be an additional cost in terms of estimation uncertainty. It is for these reasons that though a richer information set (in terms of current and lag values of the predictors) might lead to better predictive ability, it will clearly worsen the model approximation and estimation uncertainty. Thus we might be only interested in picking a few number of lags with each predictor (of course, whenever we have a very wide information set) and that will plausibly make the errors correlated.

Of course h -block cross-validation will perform worse than leave-1-out under uncorrelated errors (notice that a smaller estimation set is used at each iteration) and as it will be shown later in the chapter, it imply imposes more restrictions from the model selection perspective. A crucial question will be how to select the optimal h and we give some recipes in section 3 which are relevant whenever we work with large samples.

This is the first theoretical study (that we know) of h -block cross-validation for the selection of the optimal number of series functions and as we did in the previous chapters we generalized to consider ridge estimation (note that OLS arises as a particular case). We also generalize to consider heterogeneous near epoch-dependent observations, thus including very general dynamically stable processes that might contain breaks (as most arising in finance and macroeconomics).

In the next section we reintroduce the notation and concepts from chapter 2 that will be used throughout the chapter. Section 3 deals with the asymptotic loss-efficiency of h -block cross-validation and we state and discuss the assumptions under which our results hold. Section 4 contains a simulation in which we analyze the optimality of a range of model selection criteria (including the procedures analyzed in the previous chapters) for several time series processes, some of which include heterogeneity in the form of mean and variance shifts and error correlation. We then conclude and offer the proof of the main results.

III.2 Notation and definitions

In this section we reintroduce most of the notation and some definitions from chapter 2 for the sake of completeness. As we already stated there, we

want to obtain an optimal forecast of a target variable Y_t given a $d \times 1$ vector of predictors Z_t . We regard our forecasting model as an approximation to the conditional expectation of Y_t given Z_t (the optimal predictor in terms of PMSE) so we are implicitly assuming a quadratic loss. We assume that Y_t is generated as

$$Y_t = \mu(Z_t) + \varepsilon_t$$

where $\mu(Z_t) \equiv E[Y_t|Z_t]$ and ε_t is a zero mean error term representing the deviations between Y_t and $\mu(Z_t)$.

Given a sample of n observations on the Y_t 's and Z_t 's, we write $Y \equiv [Y_1, \dots, Y_n]'$, $\varepsilon \equiv [\varepsilon_1, \dots, \varepsilon_n]'$ and $Z^n \equiv [Z_1, \dots, Z_n]'$ to denote the vectors containing the n observations on Y_t and ε_t respectively and the matrix whose t^{th} row contains the $d \times 1$ vector of predictors Z_t' . We also write $Y = \eta + \varepsilon$ where $\eta \equiv [\mu(Z_1), \dots, \mu(Z_n)]'$ denotes the $n \times 1$ vector of conditional expectations corresponding to each t .

We now restate assumptions B.1 and B.2 from chapter 2 as C.1 and C.2 respectively, which refer to the data generating process (DGP) and the model respectively and will be used in the main results of the chapter.

C.1 (Data generating process)

- (a) Let $\{(Y_t, Z_t)\}$ be a sequence of dependent and possibly heterogeneous random vectors such that Y_t is real-valued and Z_t is a $\mathbf{Z}$ -valued vector, $\mathbf{Z} \subset \mathbb{R}^d$, $d \in \mathbb{N}$.
- (b) $\varepsilon_t \equiv Y_t - E[Y_t|Z_t] = Y_t - \mu(Z_t)$.

C.2 (Model) Let $\{\psi_j : \mathbf{Z} \rightarrow \mathbb{R}\}$ be a sequence of measurable functions. Let $\{p_n\}$ be a given non-decreasing sequence of integers and $\{A_n\}$ a given sequence of sets where each A_n is a set of subsets of $\mathbb{N}$ and each subset contains at most p_n elements, so that $\alpha \in A_n$ is a subset of $\mathbb{N}$ containing $q_n \equiv \#\alpha \leq p_n$

elements.

The model is given by

$$\mathfrak{M}_n = \{m : \mathbf{Z} \rightarrow \mathbb{R} \mid m(z) = z'\gamma + \sum_{j \in \alpha} \beta_j \psi_j(z) \quad \gamma \in \mathbb{R}^d, \beta_j \in \mathbb{R}, \alpha \in A_n\}$$

We next define the ridge estimator; for that purpose let $X_n(\alpha)$ be the $n \times (d + q)$ matrix whose t^{th} row has elements Z_t and $\psi_j(Z_t)$, $j \in \alpha$, $\alpha \in A_n$, $t = 1, \dots, n$, $n = 1, 2, \dots$. We consider the estimator given by

$$\hat{\eta}(\alpha, \lambda) \equiv [\hat{\eta}_1(\alpha, \lambda), \dots, \hat{\eta}_n(\alpha, \lambda)]' \equiv M_n(\alpha, \lambda)Y$$

where $M_n(\alpha, \lambda) \equiv X_n(\alpha)[X_n(\alpha)'X_n(\alpha) + \lambda I_q]^{-1}X_n(\alpha)'$ for given $\lambda \geq 0$ and $\alpha \in A_n$.

We also define $X_{p_n}(\cdot) \equiv [\psi_1(\cdot), \dots, \psi_{p_n}(\cdot)]'$ to denote a $p_n \times 1$ vector containing the maximum number of basis functions given a sample of size n and $X_{np_n} \equiv [X_{p_n}(Z_1), \dots, X_{p_n}(Z_n)]'$ to denote the corresponding $n \times p_n$ matrix which includes all the observations.

We next define the concept of near-epoch dependence, a memory requirement on the sequences of observations that will be used throughout the chapter:

Given a probability space (Ω, F, P) , let V_t be a F -measurable vector-valued sequence, let $F_{t-m}^{t+m} \equiv \sigma(V_{t-m}, \dots, V_{t+m})$ such that $\{F_{t-m}^{t+m}\}_{m=0}^\infty$ is an increasing sequence of σ -fields contained in F . Also let $\{S_{it}\}$ be a sequence of F -measurable random variables with $E|S_{it}| < \infty$ for all $i = 1, \dots, d$ and $t = 1, \dots, n$. Then $\{S_{it}\}$ is L_2 -NED on $\{V_t\}$ of size $-a$ iff

$$\nu_{mi} \equiv \sup_t \|S_{it} - E_{t-m}^{t+m} S_{it}\|_2 = O(m^{-a^*})$$

for $a^* > a$.

We define $L_n(\alpha, \lambda) \equiv n^{-1} \|\eta - \hat{\eta}(\alpha, \lambda)\|^2 = n^{-1} \sum_{t=1}^n (\mu_t - \hat{\mu}_t(\alpha, \lambda))^2$ as the average squared error loss, which can be decomposed as

$$L_n(\alpha, \lambda) = \Delta_n(\alpha, \lambda) + n^{-1} \varepsilon' M_n^2(\alpha, \lambda) \varepsilon - n^{-1} 2\eta' [I_n - M_n(\alpha, \lambda)] M_n(\alpha, \lambda) \varepsilon$$

where $\Delta_n(\alpha, \lambda) \equiv n^{-1} \|\eta - M_n(\alpha, \lambda) \eta\|^2$, and the unconditional risk is defined as its expectation; that is, $R_n(\alpha, \lambda) \equiv E(L_n(\alpha, \lambda))$.

We also defined $L_n^*(\alpha, \lambda) \equiv \Delta_n(\alpha, \lambda) + n^{-1} \varepsilon' M_n^2(\alpha, \lambda) \varepsilon$ which is the sum of the first two components of the average squared error loss. As we will show in the proofs, the difference between $L_n(\alpha, \lambda)$ and $L_n^*(\alpha, \lambda)$ is asymptotically negligible relative to average squared error loss.

III.3 Asymptotic loss-efficiency of h -block cross-validation

As we stated in the introduction, h -block cross-validation represents a generalization of leave-one-out. It splits the data in two parts. The first part contains $n - (2h + 1)$ observations used for fitting the model (model construction) and the observation left is used for assessing the predictive ability of the model (model validation). This procedure is repeated for each single observation and finally the model minimizing the average of all these predictive ability measures is the chosen one.

So it differs from the leave-one-out method in that we removed $2h + 1$ observations at each iteration (the t^{th} and h observations on either side of the t^{th}) instead of just one.

It seems to perform better than leave-1-out cross-validation when the errors in the data generating process are correlated for any block size h (see the simulations by Racine (1997), (2000)). Nevertheless we provide restrictions to achieve optimality that involve trade-offs between the block size, the rate at which we can increase the number of series functions with n and the sample size. Of course the block size will need to increase to infinity at a smaller rate than n to deliver asymptotic loss-efficiency in addition to further restrictions. We provide in this section some optimal ways of doing it.

Formally, the h -block cross-validation model selection procedure ($CV_{n,-1}^h$) selects $\alpha \in A_n$ and $\lambda \geq 0$ so as to minimize

$$CV_{n,-1}^h(\alpha, \lambda) \equiv (n - 2h)^{-1} \sum_{t=h+1}^{n-h} \left(Y_t - x'_t(\alpha) [X'_{(-t:h)}(\alpha) X_{(-t:h)}(\alpha) + \lambda I_{q+d}]^{-1} \right. \\ \left. \times X_{(-t:h)}(\alpha) Y_{(-t:h)} \right)^2 \quad (\text{III.3.1})$$

where $x_t(\alpha)$ denotes the t^{th} row of the $X_n(\alpha)$ matrix, $X_{(-t:h)}$ denotes the $X_n(\alpha)$ matrix with both the t^{th} observation (row) and h observations on either side of the t^{th} removed and $Y_{(-t:h)}$ denotes the Y vector with all its components except both the t^{th} and h observations preceding and following it.

Now assume the following regularity conditions:

$$\text{C.3 } \sum_{\alpha \in A_n} \left(E[\inf_{\lambda \geq 0} nL_n(\alpha, \lambda)] \right)^{-m} \longrightarrow 0$$

where m is some fixed positive integer such that the memory and moment restrictions below are satisfied.

$$\text{C.4 } \limsup_{n \rightarrow \infty} \sum_{\alpha \in A_n} [\iota_n(\alpha) - k] \left(E[\inf_{\lambda \geq 0} nL_n(\alpha, \lambda)] \right)^{-m^*} < \infty$$

where k is a positive integer not depending on n such that $\lambda_k(\alpha) > 0$ for all n (where $\lambda_k(\alpha)$ denotes the k^{th} eigenvalue of the $X_n(\alpha)'X_n(\alpha)$ matrix), $\iota_n(\alpha)$ is the largest t such that $\lambda_t(\alpha) > 0$, $t \leq q = \#\alpha \leq p_n$, and m^* is some fixed positive integer such that the memory and moment restrictions above are satisfied.

C.5 (Memory requirements)

- (a) The sequence of errors $\{\varepsilon_t\}$ is L_2 - NED on $\{V_t\}$ of size $-\max\{2^{2m-1}(\frac{r-1}{r-2})^{2m-1} + (m-1), 2^{2m^*-1}(\frac{r-1}{r-2})^{2m^*-1} + (m^*-1), 4(\frac{r-1}{r-2})^2\}$ for some $r > 2$ and where $\{V_t\}$ is an α -mixing sequence of size $-(\frac{2r}{r-2} + \max\{m-1, \kappa+1\})$ and κ is such that $p_n = O(n^\kappa)$.
- (b) $\{Z_{it}\}$ is a L_2 - NED sequence on $\{V_t\}$ of size $-8(\frac{r-1}{r-2})^3$ for some $r > 2$ and $i = 1, \dots, d$ and where $\{V_t\}$ is the same α -mixing sequence as before.

C.6 (Moment restrictions)

- (a) $\|\varepsilon_t\|_{\max\{(2m-1)r, (2m^*-1)r, 4r, 4p\}} \leq \Delta < \infty$, for some $r > 2$ and $p \geq 1$.
- (b) $\|\psi_j(Z_t)\|_{\max\{4r, 4p\}} \leq \Delta < \infty$, for all j and t .

C.7 The sequence of basis functions $\{\psi_j(\cdot)\}$ is such that $|\psi_j(Z^1) - \psi_j(Z^2)| \leq B(Z^1, Z^2)\rho(Z^1, Z^2)$ for all j , where $B(z^1, z^2) : \mathbf{Z} \times \mathbf{Z} \rightarrow \mathbb{R}^+$ is a non-negative $\mathbb{B}^{2d}$ -measurable function and $\rho(z^1, z^2) \equiv \sum_{i=1}^d |z_i^1 - z_i^2|$, where z_i is the i^{th} component of the $d \times 1$ vector z .

We further assume that $\|B(Z_t, E_{t-m}^{t+m}Z_t)\|_{q/(q-1)} < \infty$, $\|\rho(Z_t, E_{t-m}^{t+m}Z_t)\|_q < \infty$ and $\|B(Z_t, E_{t-m}^{t+m}Z_t)\rho(Z_t, E_{t-m}^{t+m}Z_t)\|_r < \infty$, where $1 \leq q \leq 2$ and $r > 2$.

C.8 For some $0 < \delta < 1$ and same p as in B.6,

$$(p_n^2 n)^{1/p} \left(\sum_{\tau=1}^{n-1} (\tau+2) \right)^{1/p} \left(\inf_{\alpha \in A_n} \inf_{\lambda \geq 0} nL_n(\alpha, \lambda) \right)^{-\delta} = O_p(1).$$

$$\text{C.9 } \lambda_{\min} \left\{ n^{-1} \sum_{t=1}^n EX_{p_n}(Z_t)X_{p_n}(Z_t)' \right\} = C_n$$

where $C_n(\inf_{\alpha \in A_n} \inf_{\lambda \geq 0} nL_n(\alpha, \lambda))^{-(1+\delta)} = O_p(1)$ for some $\delta > 0$.

C.10

1. $h \rightarrow \infty$ as $n \rightarrow \infty$.
2. $p_n h^{3/2} n^{-1/2} = O(1)$.
3. $d_n h = o_p(1)$, where $d_n \equiv \sup_{\alpha \in A_n} \sup_{\lambda \geq 0} \bar{\lambda}(M_n(\alpha, \lambda))$ with $\bar{\lambda}(\cdot)$ denoting the maximal diagonal element of a matrix.

Theorem III.3.1. *Under C.1-C.10 the pair $(\alpha^{CV_h}, \lambda^{CV_h})$ selected by h -block cross-validation is asymptotically loss-efficient.*

Remarks:

1. Conditions C.1-C.9 are the analogs of those for GC_L and leave-1-out cross-validation in chapter 2.
2. It is worth noticing that condition C.5 (a) might require stronger memory constraints on the mixing sequence on which the errors are near-epoch dependent than those needed in the earlier chapter (where we assumed that the errors in the DGP were a martingale difference sequence).
3. Condition C.10 (a) requires the block size to increase with the sample size (so the observations in the validation set and those used for estimation are approximately independent for large n). C.10 (b) implies a clear trade-off between the rate of increase of the block size and the rate at which we can incorporate larger models with additional basis functions to potentially achieve better approximations, as can be expected. C.10 (c) involves a similar trade-off and it will be weaker than C.10 (b) in those cases where p_n/n tending to zero also implies that d_n tends to zero (see Andrews, 1991).

III.4 Simulation

We use neural network models with the ridgelet activation function to approximate two data generating processes (referred to as DGP1 and DGP2 re-

spectively).

DGP1 is a nonlinear additive AR(2) model (NLAR(2)) with iid errors given by

$$Y_t = -0.4 \frac{3 - Y_{t-1}^2}{1 + Y_{t-1}^2} + 0.6 \frac{3 - (Y_{t-2} - 0.5)^3}{1 + (Y_{t-2} - 0.5)^4} + 0.1\varepsilon_t$$

where $\varepsilon_t \sim iid N(0, 1)$.

The second DGP is also the same nonlinear additive AR(2) model but were now the errors are given by

$$\varepsilon_t = 0.8\varepsilon_{t-1} + u_t$$

where $u_t \sim iid N(0, 0.1)$.

We applied a wide range of model selection criteria which includes GC_L , leave-1-out cross-validation (CV-1), generalized cross-validation (GCV), h -block cross-validation and BIC in order to compare their relative performances for choosing the number of ridgelet basis functions and the ridge parameter.

We follow the same methodology outlined in the first chapter for generating the ridgelet basis functions and that was proposed in White (2006). In this simulation we generated a total of 33 ridgelet basis functions.

For each DGP we compute the ratio of the average squared error loss evaluated at the model and ridge parameter (the latter from a grid of 20 values) chosen by the corresponding model selection procedure and average squared error loss evaluated at the pair (α, h) that minimize it. We did that for sample sizes of 100, 250, 500 and 1000 observations and always kept the number of models fixed at 33. For the case of h -block cross-validation the block size h was of the order of $n^{1/5}$.

Table III.1: Means of Ratios of Average Square Losses for DGP1

	n=100	n=250	n=500	n=1000
CV-1	1.2553	1.1234	1.0522	1.0043
GC_L	1.2345	1.1804	1.0322	1.0089
GCV	1.2335	1.1522	1.0903	1.0203
BIC	1.3391	1.2345	1.1523	1.1523
h-block	1.3567	1.2347	1.1543	1.0522

Table III.2: Standard Deviations of Ratios of Average Square Error Losses for DGP1

	n=100	n=250	n=500	n=1000
CV-1	0.0148	0.0085	0.0064	0.0047
GC_L	0.0119	0.01454	0.0064	0.0047
GCV	0.0136	0.0081	0.0106	0.007
BIC	0.0129	0.008	0.0064	0.0047
h-block	0.0232	0.0148	0.0064	0.0047

We compute 100 replications of the above ratio for each sample size and model selection criteria under consideration, reporting the mean and standard deviation of the ratios across replications as shown in tables 1 and 2 for DGP1 and 3 and 4 for DGP2. We would expect the mean of these ratios to converge to 1 with a decreasing variance for those criteria which are asymptotic loss-efficient.

We can clearly see from table 1 that for DGP1 the average of the ratios of average squared error losses for BIC doesn't have a tendency to converge, which can be explained as the BIC model selection procedure is not asymptotic loss-efficient. The ratio for h -block cross-validation clearly lies above that for the remaining selection procedures. This is due to the lack of correlation in the errors (they are actually independent) and seems natural that in this case h -block cross-

Table III.3: Means of Ratios of Average Square Losses for DGP2

	n=100	n=250	n=500	n=1000
CV-1	1.3356	1.1945	1.1413	1.1214
GC_L	1.3098	1.2034	1.1532	1.1342
GCV	1.3882	1.2544	1.1725	1.1388
BIC	1.3223	1.2532	1.2032	1.2243
h-block	1.2544	1.1123	1.0734	1.0032

Table III.4: Standard Deviations of Ratios of Average Square Error Losses for DGP2

	n=100	n=250	n=500	n=1000
CV-1	0.0167	0.00152	0.0035	0.0029
GC_L	0.0165	0.0154	0.003	0.0029
GCV	0.0184	0.0146	0.0056	0.0032
BIC	0.1633	0.0146	0.0032	0.0038
h-block	0.0618	0.0145	0.0036	0.0029

validation underperforms relative to the other procedures which are optimal.

In the case of DGP2 (where the errors are strongly correlated) h -block cross-validation clearly outperforms all the other model selection criteria.

III.5 Conclusion

In this chapter we analyzed the asymptotic loss-efficiency of h -block cross-validation under near-epoch dependent and heterogeneous processes. This procedure is the only one of those studied previously which is robust to error correlation in the data generating process.

It seems that under strong correlation h -block cross-validation is clearly superior though other selection procedures might perform better under weak cor-

relation. We showed in a small simulation its good performance relative to the other model selection criteria. It would be interesting to see a more ambitious and extensive Monte Carlo exercise for a wider range of correlation structures in the errors as well as data generating processes.

Of course it remains important for the performance of h -block cross-validation to optimally choose the block size h . Racine (1997, 2000) shows that under important error correlation h -block outperforms leave-1-out cross-validation even for small values of h . Here we gave some recipes about how to choose h optimally and through the proofs it can be seen a clear analogy between this and the problem of consistent estimation of variance-covariance matrices in the presence of autocorrelation and heterogeneity (see Andrews (1991) and Hansen (1992)). In fact h -block cross-validation is implicitly linked with the use of a flat kernel. We might borrow further insights from the consistent covariance estimation literature and extrapolate them to the present setting.

III.6 Proofs

In this section we proof theorem 3.1, which shows the conditions under which h -block cross-validation is asymptotic loss-efficient for selecting the number of terms in series estimators which are linear in the parameters. We start with the following decomposition (shown in Racine (1997)) and which is key for the analysis that follows.

$$\begin{aligned}
& \left(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n} - X'_{(t:h)}(\alpha)X_{(t:h)}(\alpha) \right)^{-1} = \\
& \left(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n} \right)^{-1} + \left(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha) \\
& \times \left[I_{1+2h} - X_{(t:h)}(\alpha) \left(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha) \right]^{-1} X_{(t:h)}(\alpha) \\
& \times \left(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n} \right)^{-1}
\end{aligned}$$

where I_{1+2h} denoted the $1 + 2h \times 1 + 2h$ identity matrix and $X_{(t:h)}$ denotes the $(1 + 2h) \times q_n$ matrix containing the $(1 + 2h)$ rows of observations that are missing in $X_{(-t:h)}$.

We now define the ridge estimator computed with all the observations except both the t^{th} and h observations on either side of it by

$$\hat{\eta}_{(-t:h)}(\alpha, \lambda) \equiv (X'_{(-t:h)}(\alpha)X_{(-t:h)}(\alpha) + \lambda I_{q_n})^{-1}X_{(-t:h)}(\alpha)Y_{(-t:h)} \quad (\text{III.6.1})$$

and (6.1) equals

$$\begin{aligned} & ((X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} + (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_{(t:h)}(\alpha) \\ & \times t[I_{1+2h} - X_{(t:h)}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_{(t:h)}(\alpha)]^{-1}X_{(t:h)}(\alpha) \\ & \times (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1})(X'_n(\alpha)Y - X_{(-t:h)}(\alpha)Y_{(-t:h)}) \end{aligned}$$

and therefore

$$\begin{aligned} Y_t - x'_t(\alpha)\hat{\eta}_{(-t:h)}(\alpha, \lambda) &= Y_t - x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_n(\alpha)Y \\ &- x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_{(t:h)}(\alpha)[I_{1+2h} - X_{(t:h)}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\ &\quad \times X'_{(t:h)}(\alpha)]^{-1}X_{(t:h)}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_n(\alpha)Y \\ &\quad + x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_{(t:h)}(\alpha)Y_{(t:h)} \\ &+ x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_{(t:h)}(\alpha)[I_{1+2h} - X_{(t:h)}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\ &\quad \times X'_{(t:h)}(\alpha)]^{-1}X_{(t:h)}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_{(t:h)}Y_{(t:h)} \quad (\text{III.6.2}) \end{aligned}$$

$$\begin{aligned} &= Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda) - x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}X'_{(t:h)}(\alpha)[I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} \\ &\quad \times X_{(t:h)}(\alpha)\hat{\eta}(\alpha, \lambda) \end{aligned}$$

$$\begin{aligned}
& + x'_t(\alpha) \left(X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha) \left[I_{1+2h} + [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} \right. \\
& \quad \left. \times H_{(t:h)}(\alpha, \lambda) \right) Y_{(t:h)} \quad (\text{III.6.3})
\end{aligned}$$

$$\begin{aligned}
& = Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda) - x'_t(\alpha) \left(X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha) [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} \\
& \quad \times X_{(t:h)}(\alpha) \hat{\eta}(\alpha, \lambda) \\
& + x'_t(\alpha) \left(X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha) [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} Y_{(t:h)} \quad (\text{III.6.4})
\end{aligned}$$

$$\begin{aligned}
& = Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda) + x'_t(\alpha) \left(X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha) [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} \\
& \quad \times (Y_{(t:h)} - X_{(t:h)}(\alpha) \hat{\eta}(\alpha, \lambda)) \quad (\text{III.6.5})
\end{aligned}$$

$$\text{where } H_{(t:h)}(\alpha, \lambda) \equiv X_{(t:h)}(\alpha) \left(X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha).$$

So we get that

$$\begin{aligned}
& CV_{n,-1}^h(\alpha, \lambda) \equiv (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha) \hat{\eta}_{(-t:h)}(\alpha, \lambda))^2 \\
& = (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda))^2 \\
& + (n-2h)^{-1} \sum_{t=h+1}^{n-h} \left(x'_t(\alpha) \left(X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha) [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} \right. \\
& \quad \left. \times (Y_{(t:h)} - X_{(t:h)}(\alpha) \hat{\eta}(\alpha, \lambda)) \right)^2 \\
& + 2(n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x'_t(\alpha) \left(X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n} \right)^{-1} X'_{(t:h)}(\alpha) \\
& \quad \times [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} (Y_{(t:h)} - X_{(t:h)}(\alpha) \hat{\eta}(\alpha, \lambda)) \quad (\text{III.6.6})
\end{aligned}$$

$$\begin{aligned}
&= (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))^2 \\
&+ (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_{(t:h)} - X_{(t:h)}(\alpha)\hat{\eta}(\alpha, \lambda))' [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} X_{(t:h)} \\
&\quad \times (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} x_t(\alpha)x'_t(\alpha) (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
&\quad \times X'_{(t:h)}(\alpha) [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} (Y_{(t:h)} - X_{(t:h)}(\alpha)\hat{\eta}(\alpha, \lambda)) \\
&+ 2(n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} X'_{(t:h)}(\alpha) \\
&\quad \times [I_{1+2h} - H_{(t:h)}(\alpha, \lambda)]^{-1} (Y_{(t:h)} - X_{(t:h)}(\alpha)\hat{\eta}(\alpha, \lambda)) \quad (\text{III.6.7})
\end{aligned}$$

$$\begin{aligned}
&= (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))^2 \\
&+ (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_{(t:h)} - X_{(t:h)}(\alpha)\hat{\eta}(\alpha, \lambda))' [I_{1+2h} + H_{(t:h)}(\alpha, \lambda) + H_{(t:h)}^2(\alpha, \lambda) + \dots] \\
&\quad \times X_{(t:h)} (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} x_t(\alpha)x'_t(\alpha) (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
&\quad \times X'_{(t:h)}(\alpha) [I_{1+2h} + H_{(t:h)}(\alpha, \lambda) + H_{(t:h)}^2(\alpha, \lambda) + \dots] (Y_{(t:h)} - X_{(t:h)}(\alpha)\hat{\eta}(\alpha, \lambda)) \\
&+ 2(n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} X'_{(t:h)}(\alpha) \\
&\quad \times [I_{1+2h} + H_{(t:h)}(\alpha, \lambda) + H_{(t:h)}^2(\alpha, \lambda) + \dots] (Y_{(t:h)} - X_{(t:h)}(\alpha)\hat{\eta}(\alpha, \lambda)) \quad (\text{III.6.8})
\end{aligned}$$

$$\begin{aligned}
&= (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))^2 \\
&+ (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_{(t:h)} - X_{(t:h)}(\alpha)\hat{\eta}(\alpha, \lambda))' X_{(t:h)} (X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} x_t(\alpha)
\end{aligned}$$

$$\begin{aligned}
& \times x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} X'_{(t;h)}(\alpha) (Y_{(t;h)} - X_{(t;h)}(\alpha) \hat{\eta}(\alpha, \lambda)) \\
& + (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_{(t;h)} - X_{(t;h)}(\alpha) \hat{\eta}(\alpha, \lambda))' X_{(t;h)} (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} x_t(\alpha) \\
& \times x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} X'_{(t;h)}(\alpha) H_{(t;h)}(\alpha, \lambda) (Y_{(t;h)} - X_{(t;h)}(\alpha) \hat{\eta}(\alpha, \lambda)) \\
& + \dots \\
& + (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_{(t;h)} - X_{(t;h)}(\alpha) \hat{\eta}(\alpha, \lambda))' H_{(t;h)}(\alpha, \lambda) X_{(t;h)} (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \times x_t(\alpha) x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} X'_{(t;h)}(\alpha) (Y_{(t;h)} - X_{(t;h)}(\alpha) \hat{\eta}(\alpha, \lambda)) \\
& + \dots \\
& + 2(n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} X'_{(t;h)}(\alpha) \\
& \times (Y_{(t;h)} - X_{(t;h)}(\alpha) \hat{\eta}(\alpha, \lambda)) \\
& + 2(n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} X'_{(t;h)}(\alpha) \\
& \times H_{(t;h)}(\alpha, \lambda) (Y_{(t;h)} - X_{(t;h)}(\alpha) \hat{\eta}(\alpha, \lambda)) \\
& + \dots \tag{III.6.9}
\end{aligned}$$

$$\begin{aligned}
& = (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda))^2 \\
& + (n-2h)^{-1} R_{1n}(\alpha, \lambda) + O_p \left(\frac{h_n(2h+1) R_{1n}(\alpha, \lambda)}{n-2h} \right) \\
& + 2(n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} x_t(\alpha) \\
& \times (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x_t(\alpha) (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) \\
& + 2(n-2h)^{-1} \sum_{l=1}^h \sum_{t=l+h+1}^{n-h} (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \times x_{t-l}(\alpha) (Y_{t-l} - x'_{t-l}(\alpha) \hat{\eta}(\alpha, \lambda))
\end{aligned}$$

$$\begin{aligned}
& + 2(n-2h)^{-1} \sum_{l=1}^h \sum_{t=h+1}^{n-h-l} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad x_{t+l}(\alpha)(Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda)) + O_p\left(\frac{d_n R_{2n}(\alpha, \lambda)}{n-2h}\right) \quad (\text{III.6.10})
\end{aligned}$$

where $d_n = \sup_{\alpha \in A_n} \sup_{h \geq 0} \bar{\lambda}(M_n(\alpha, \lambda))$ with $\bar{\lambda}(\cdot)$ denoting the maximal diagonal element of a matrix,

$$\begin{aligned}
R_{1n}(\alpha, \lambda) & \equiv \sum_{l=0}^h \sum_{t=h+1}^{n-h-l} (Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda))x'_{t+l}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_t(\alpha)x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}x_{t+l}(\alpha)(Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda)) \\
& \quad + \sum_{l=1}^h \sum_{t=h+1}^{n-h-l} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_{t+l}(\alpha)x'_{t+l}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}x_t(\alpha)(Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda)) \\
& + 2 \sum_{\tau=1}^h \sum_{l=0}^{h-\tau} \sum_{t=\tau+h+1}^{n-h-l} (Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda))x'_{t+l}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_t(\alpha)x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}x_{t+l-\tau}(\alpha)(Y_{t+l-\tau} - x'_{t+l-\tau}(\alpha)\hat{\eta}(\alpha, \lambda)) \\
& \quad + 2 \sum_{\tau=1}^h \sum_{l=1}^{h-\tau} \sum_{t=\tau+h+1}^{n-h-l} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_{t+l}(\alpha)x'_{t+l}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}x_{t-\tau}(\alpha)(Y_{t-\tau} - x'_{t-\tau}(\alpha)\hat{\eta}(\alpha, \lambda)) \quad (\text{III.6.11})
\end{aligned}$$

and

$$\begin{aligned}
R_{2n}(\alpha, \lambda) & \equiv \sum_{l=0}^h \sum_{t=h+1}^{n-h-l} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_{t+l}(\alpha)(Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda)) \\
& + \sum_{l=1}^h \sum_{t=h+1}^{n-h-l} (Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda))x'_{t+l}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1}
\end{aligned}$$

$$\begin{aligned}
& \times x_t(\alpha)(Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda)) \\
& + \sum_{\tau=1}^h \sum_{l=0}^{h-\tau} \sum_{t=\tau+h+1}^{n-h-l} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_{t+l-\tau}(\alpha)(Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda)) \\
& + \sum_{\tau=1}^h \sum_{l=0}^{h-\tau} \sum_{t=\tau+h+1}^{n-h-l} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_{t+l}(\alpha)(Y_{t+l-\tau} - x'_{t+l-\tau}(\alpha)\hat{\eta}(\alpha, \lambda)) \\
& + \sum_{\tau=1}^h \sum_{l=1}^{h-\tau} \sum_{t=\tau+h+1}^{n-h-l} (Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda))x'_{t+l}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_{t-\tau}(\alpha)(Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda)) \tag{III.6.12}
\end{aligned}$$

$$\begin{aligned}
& + \sum_{\tau=1}^h \sum_{l=1}^{h-\tau} \sum_{t=\tau+h+1}^{n-h-l} (Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda))x'_{t+l}(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_t(\alpha)(Y_{t-\tau} - x'_{t-\tau}(\alpha)\hat{\eta}(\alpha, \lambda)) \tag{III.6.13}
\end{aligned}$$

Thus we get that

$$\begin{aligned}
& CV_{n,-1}^h(\alpha, \lambda) = (n-2h)^{-1} \sum_{t=h+1}^{n-h} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))^2 \\
& + 2(n-2h)^{-1} \sum_{l=0}^h \sum_{t=l+h+1}^{n-h} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad \times x_{t-l}(\alpha)(Y_{t-l} - x'_{t-l}(\alpha)\hat{\eta}(\alpha, \lambda)) \\
& + 2(n-2h)^{-1} \sum_{l=1}^h \sum_{t=h+1}^{n-h-l} (Y_t - x'_t(\alpha)\hat{\eta}(\alpha, \lambda))x'_t(\alpha)(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})^{-1} \\
& \quad x_{t+l}(\alpha)(Y_{t+l} - x'_{t+l}(\alpha)\hat{\eta}(\alpha, \lambda)) + O_p() \tag{III.6.14}
\end{aligned}$$

$$\begin{aligned}
&= \frac{n}{n-2h} \left\{ n^{-1} \sum_{t=1}^n (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda))^2 - \frac{h}{n} h^{-1} \sum_{t=1}^h (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda))^2 \right. \\
&\quad \left. - \frac{h}{n} h^{-1} \sum_{t=n-h+1}^n (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda))^2 \right. \\
&\quad + 2n^{-1} \sum_{l=0}^h \sum_{t=l+1}^n (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} \\
&\quad \quad \times x_{t-l}(\alpha) (Y_{t-l} - x'_{t-l}(\alpha) \hat{\eta}(\alpha, \lambda)) \\
&\quad - 2 \frac{h}{n} h^{-1} \sum_{l=0}^h \sum_{t=l+1}^h (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} \\
&\quad \quad \times x_{t-l}(\alpha) (Y_{t-l} - x'_{t-l}(\alpha) \hat{\eta}(\alpha, \lambda)) \\
&\quad - 2 \frac{h}{n} h^{-1} \sum_{l=0}^h \sum_{t=l+n-h+1}^n (Y_t - x'_t(\alpha) \hat{\eta}(\alpha, \lambda)) x'_t(\alpha) (X'_n(\alpha) X_n(\alpha) + \lambda I_{q_n})^{-1} \\
&\quad \quad \times x_{t-l}(\alpha) (Y_{t-l} - x'_{t-l}(\alpha) \hat{\eta}(\alpha, \lambda)) \\
&\quad \quad \left. + \dots \right\} \tag{III.6.15}
\end{aligned}$$

$$\begin{aligned}
&= \frac{n}{n-2h} \left\{ n^{-1} (Y - \hat{\eta}(\alpha, \lambda))' (Y - \hat{\eta}(\alpha, \lambda)) + 2n^{-1} T_n(\alpha, \lambda) \right. \\
&\quad \left. + O_p \left(\frac{d_n h T_n(\alpha, \lambda)}{n} + \frac{h}{n} C V_{h,-1}^h \right) \right\} \tag{III.6.16}
\end{aligned}$$

where

$$\begin{aligned}
T_n(\alpha, \lambda) &\equiv (Y - \hat{\eta}(\alpha, \lambda))' \ddot{M}_n(\alpha, \lambda) (Y - \hat{\eta}(\alpha, \lambda)) \\
&\quad + (Y - \hat{\eta}(\alpha, \lambda))' \bar{M}_n(\alpha, \lambda) (Y - \hat{\eta}(\alpha, \lambda)),
\end{aligned}$$

$\ddot{M}_n(\alpha, \lambda)$ denotes a diagonal matrix whose t^{th} diagonal element is the t^{th} diagonal element of $M_n(\alpha, \lambda)$ and $\bar{M}_n(\alpha, \lambda)$ denotes a matrix identical in structure to $M_n(\alpha, \lambda)$ but with zeros along both its main diagonal and last $n - (h + 1)$ upper and lower diagonals; that is,

$$\bar{M}_n(\alpha, \lambda) \equiv \begin{bmatrix} 0 & m_{12}(\alpha, \lambda) & \dots & m_{1(h+1)}(\alpha, \lambda) & 0 & \dots & 0 \\ m_{21}(\alpha, \lambda) & 0 & \dots & m_{2(h+1)}(\alpha, \lambda) & m_{2(h+2)}(\alpha, \lambda) & \dots & 0 \\ m_{31}(\alpha, \lambda) & m_{32}(\alpha, \lambda) & \dots & m_{3(h+1)}(\alpha, \lambda) & m_{3(h+2)}(\alpha, \lambda) & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & m_{n(h+1)}(\alpha, \lambda) & m_{n(h+2)}(\alpha, \lambda) & \dots & 0 \end{bmatrix}$$

Now, (6.16) equals

$$\begin{aligned} & \frac{n}{n-2h} \left\{ n^{-1} \|\varepsilon\|^2 + L_n(\alpha, \lambda) + 2n^{-1} \varepsilon' (I_n - M_n(\alpha, \lambda)) \eta \right. \\ & + 4n^{-1} T_{1n}(\alpha, \lambda) + 2n^{-1} T_{2n}(\alpha, \lambda) + 2n^{-1} T_{3n}(\alpha, \lambda) + 4n^{-1} T_{4n}(\alpha, \lambda) \\ & \left. + 2n^{-1} T_{5n}(\alpha, \lambda) + 2n^{-1} T_{6n}(\alpha, \lambda) + O_p \left(\frac{d_n h T_n(\alpha, \lambda)}{n} + \frac{h}{n} C V_{h,-1}^h \right) \right\} \quad (\text{III.6.17}) \end{aligned}$$

where

$$\begin{aligned} T_{1n}(\alpha, \lambda) &\equiv \varepsilon' (I_n - M_n(\alpha, \lambda)) \ddot{M}_n(\alpha, \lambda) (I_n - M_n(\alpha, \lambda)) \eta \\ T_{2n}(\alpha, \lambda) &\equiv \eta' (I_n - M_n(\alpha, \lambda)) \ddot{M}_n(\alpha, \lambda) (I_n - M_n(\alpha, \lambda)) \eta \\ T_{3n}(\alpha, \lambda) &\equiv \varepsilon' [(I_n - M_n(\alpha, \lambda)) \ddot{M}_n(\alpha, \lambda) (I_n - M_n(\alpha, \lambda)) - M_n(\alpha, \lambda)] \varepsilon \\ T_{4n}(\alpha, \lambda) &\equiv \varepsilon' (I_n - M_n(\alpha, \lambda)) \bar{M}_n(\alpha, \lambda) (I_n - M_n(\alpha, \lambda)) \eta \\ T_{5n}(\alpha, \lambda) &\equiv \eta' (I_n - M_n(\alpha, \lambda)) \bar{M}_n(\alpha, \lambda) (I_n - M_n(\alpha, \lambda)) \eta \end{aligned}$$

and

$$T_{6n}(\alpha, \lambda) \equiv \varepsilon' [(I_n - M_n(\alpha, \lambda)) \bar{M}_n(\alpha, \lambda) (I_n - M_n(\alpha, \lambda)) \varepsilon]$$

Asymptotic loss-efficiency of $CV_{n,-1}^h$ will follow if we show that

$$\sup_{\alpha \in A_n} \sup_{\lambda \geq 0} |L_n(\alpha, \lambda) / L_n^*(\alpha, \lambda) - 1| = o_p(1), \quad (\text{III.6.18})$$

$$\sup_{\alpha \in A_n} \sup_{\lambda \geq 0} n^{-1} |\varepsilon' (I_n - M_n(\alpha, \lambda)) \eta| / L_n^*(\alpha, \lambda) = o_p(1) \quad (\text{III.6.19})$$

$$\sup_{\alpha \in A_n} \sup_{\lambda \geq 0} n^{-1} |T_{in}(\alpha, \lambda)| / L_n^*(\alpha, \lambda) = o_p(1) \quad (\text{III.6.20})$$

for $i = 1, 2, 4$ and 5 , and

$$\sup_{\alpha \in A_n} \sup_{\lambda \geq 0} n^{-1} |T_{3n}(\alpha, \lambda) + T_{6n}(\alpha, \lambda)| / L_n^*(\alpha, \lambda) = o_p(1) \quad (\text{III.6.21})$$

We already showed in chapter 2 that (6.18), (6.19) and (6.20) for $i = 1$ and 2 hold under the assumptions in section 3, so we only need to verify (6.20) for $i = 4$ and 5 and (6.21).

We now show (6.20) for $i = 4$.

For that purpose notice that

$$|T_{4n}(\alpha, \lambda)| \equiv |\varepsilon'(I_n - M_n(\alpha, \lambda))\bar{M}_n(\alpha, \lambda)(I_n - M_n(\alpha, \lambda))\eta| \quad (\text{III.6.22})$$

$$= |\bar{\varepsilon}'\bar{M}_n(\alpha, \lambda)\bar{\eta}| \quad (\text{III.6.23})$$

$$\leq d_n |\bar{\varepsilon}'\bar{D}_n\bar{\eta}| \quad (\text{III.6.24})$$

where $\bar{\varepsilon} \equiv (I_n - M_n(\alpha, \lambda))\varepsilon$, $\bar{\eta} \equiv (I_n - M_n(\alpha, \lambda))\eta$ and $\bar{D}_n$ denotes a $n \times n$ matrix with zeros along both its main diagonal and last $n - (h + 1)$ upper and lower diagonals and with ones elsewhere.

Now, (6.24) doesn't exceed

$$2d_n \sum_{l=1}^h \sum_{t=l+1}^n |\bar{\varepsilon}_t \bar{\mu}_{t-l}| \leq 2d_n h \left(\sum_{t=1}^n \bar{\varepsilon}_t^2 \right)^{1/2} \left(\sum_{t=1}^n \bar{\mu}_t^2 \right)^{1/2} \quad (\text{III.6.25})$$

$$= 2d_n h \left(\varepsilon'(I_n - M_n(\alpha, \lambda))^2 \varepsilon \right)^{1/2} \left(\eta'(I_n - M_n(\alpha, \lambda))^2 \eta \right)^{1/2} \quad (\text{III.6.26})$$

therefore we get that (6.20) for $i = 4$ doesn't exceed

$$\begin{aligned}
& 2d_n h \sup_{\alpha \in A_n} \sup_{\lambda \geq 0} \frac{(\varepsilon'(I_n - M_n(\alpha, \lambda))^2 \varepsilon)^{1/2} (\eta'(I_n - M_n(\alpha, \lambda))^2 \eta)^{1/2}}{nL_n^*(\alpha, \lambda)} \\
& \leq 2d_n h \sup_{\alpha \in A_n} \sup_{\lambda \geq 0} \frac{(\varepsilon'(I_n - M_n(\alpha, \lambda))^2 \varepsilon)^{1/2} (\eta'(I_n - M_n(\alpha, \lambda))^2 \eta)^{1/2}}{(\eta'(I_n - M_n(\alpha, \lambda))^2 \eta)} \\
& = 2d_n h O_p(1)
\end{aligned}$$

and the result will follow as long as $d_n h = o_p(1)$.

We now verify (6.20) for $i = 5$.

Notice that

$$\begin{aligned}
T_{5n}(\alpha, \lambda) & \equiv \eta'(I_n - M_n(\alpha, \lambda)) \bar{M}_n(\alpha, \lambda) (I_n - M_n(\alpha, \lambda)) \eta \\
& \leq 2d_n h (\eta'(I_n - M_n(\alpha, \lambda))^2 \eta)^{1/2} (\eta'(I_n - M_n(\alpha, \lambda))^2 \eta)^{1/2} \\
& \leq 2d_n h n L_n^*(\alpha, \lambda)
\end{aligned}$$

where the second inequality follows from Holder's inequality and the same reasoning as for the case $i = 4$ above. The result follows if $d_n h = o_p(1)$ once again.

We finally verify (6.21).

Now,

$$\begin{aligned}
|T_{3n}(\alpha, \lambda) + T_{6n}(\alpha, \lambda)| &= |\varepsilon'[(I_n - M_n(\alpha, \lambda))(\ddot{M}_n(\alpha, \lambda) + \bar{M}_n(\alpha, \lambda))(I_n - M_n(\alpha, \lambda)) \\
&\quad - M_n(\alpha, \lambda)]\varepsilon| \\
&\leq C|\varepsilon'(\ddot{M}_n(\alpha, \lambda) + \bar{M}_n(\alpha, \lambda))\varepsilon - \varepsilon'M_n(\alpha, \lambda)\varepsilon| \\
&= C|tr\hat{W}_n^{-1}(\alpha, \lambda)\hat{V}_n(\alpha) - trW_n^{-1}(\alpha, \lambda)V_n(\alpha) \\
&\quad + trW_n^{-1}(\alpha, \lambda)V_n(\alpha) \\
&\quad - tr\hat{W}_n^{-1}(\alpha, \lambda)\tilde{V}_n(\alpha)|
\end{aligned}$$

where

$$W_n(\alpha, \lambda) \equiv E[n^{-1}X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n}]$$

$$V_n(\alpha) \equiv E[n^{-1}X_n(\alpha)\varepsilon\varepsilon'X'_n(\alpha)]$$

$$\hat{W}_n(\alpha, \lambda) \equiv n^{-1}(X'_n(\alpha)X_n(\alpha) + \lambda I_{q_n})$$

$$\tilde{V}_n(\alpha) \equiv n^{-1}X_n(\alpha)\varepsilon\varepsilon'X'_n(\alpha)$$

$$\begin{aligned}
\hat{V}_n(\alpha) &\equiv n^{-1} \sum_{t=1}^n \varepsilon_t^2 x_t(\alpha) x'_t(\alpha) + n^{-1} \sum_{l=1}^h \sum_{t=l+1}^n \{ \varepsilon_t \varepsilon_{t-l} x_t(\alpha) x'_{t-l}(\alpha) \\
&\quad + \varepsilon_t \varepsilon_{t-l} x_{t-l}(\alpha) x'_t(\alpha) \}
\end{aligned}$$

and C is a finite constant.

Therefore

$$\sup_{\alpha \in A_n} \sup_{\lambda \geq 0} \frac{|T_{3n}(\alpha, \lambda) + T_{6n}(\alpha, \lambda)|}{nL_n^*(\alpha, \lambda)}$$

$$\leq C \sup_{\alpha \in A_n} \sup_{\lambda \geq 0} \frac{|tr\hat{W}_n^{-1}(\alpha, \lambda)\hat{V}_n(\alpha) - trW_n^{-1}(\alpha, \lambda)V_n(\alpha)|}{nL_n^*(\alpha, \lambda)} \quad (\text{III.6.27})$$

$$+ C \sup_{\alpha \in A_n} \sup_{\lambda \geq 0} \frac{|tr\hat{W}_n^{-1}(\alpha, \lambda)\tilde{V}_n(\alpha) - trW_n^{-1}(\alpha, \lambda)V_n(\alpha)|}{nL_n^*(\alpha, \lambda)} \quad (\text{III.6.28})$$

and we will have to check that both (6.27) and (6.28) are $o_p(1)$.

To verify that (6.27) is $o_p(1)$ it will be enough to show that

$$\left[\inf_{\alpha \in A_n} \inf_{\lambda \geq 0} nL_n(\alpha, \lambda) \right]^{-1} \sup_{\alpha \in A_n} \sup_{\lambda \geq 0} |\hat{W}_n(\alpha, \lambda) - W_n(\alpha, \lambda)| = o_p(1) \quad (\text{III.6.29})$$

and

$$\sup_{\alpha \in A_n} |\hat{V}_n(\alpha) - V_n(\alpha)| = O_p(1) \quad (\text{III.6.30})$$

Now, (6.29) will hold if

$$\|n^{-1}[X'_{np_n} X_{np_n} - E[X'_{np_n} X_{np_n}]]\| = O_p(1) \quad (\text{III.6.31})$$

and

$$\lambda_{\min} \left\{ n^{-1} \sum_{t=1}^n E X_{p_n}(Z_t) X_{p_n}(Z_t)' \right\} = C_n \quad (\text{III.6.32})$$

where $C_n(\inf_{\alpha \in A_n} \inf_{\lambda \geq 0} nL_n(\alpha, \lambda))^{-1} = o_p(1)$

Conditions (6.31) and (6.32) will hold under the assumptions in section 3 and it was proved in chapter 2, so here we will only verify condition (6.30).

To verify condition (6.30) we let

$$\begin{aligned} V_n^* \equiv E[n^{-1} X'_{np_n} \varepsilon \varepsilon' X_{np_n}] &= n^{-1} \sum_{t=1}^n E[\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)'] \\ &+ n^{-1} \sum_{l=1}^{n-1} \sum_{t=l+1}^n \{ E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})'] \\ &+ E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)'] \} \end{aligned}$$

$$\begin{aligned}
\bar{V}_n^* &= n^{-1} \sum_{t=1}^n E[\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)'] \\
&+ n^{-1} \sum_{l=1}^h \sum_{t=l+1}^n \{E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})'] \\
&+ E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)']\}
\end{aligned}$$

and

$$\begin{aligned}
\hat{V}_n^* &= n^{-1} \sum_{t=1}^n \varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)' \\
&+ n^{-1} \sum_{l=1}^h \sum_{t=l+1}^n \{\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})' \\
&+ \varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)'\}
\end{aligned}$$

and (6.30) will hold if $\tilde{V}_n^* - V_n^* = O_p(1)$.

Now,

$$\hat{V}_n^* - V_n^* = (\hat{V}_n^* - \bar{V}_n^*) + (\bar{V}_n^* - V_n^*) \quad (\text{III.6.33})$$

and the result will follow if we show that each of the summands in (6.33) are bounded in probability.

We first show that $\hat{V}_n^* - \bar{V}_n^*$ is $O_p(1)$. Notice for that purpose that

$$\begin{aligned}
\hat{V}_n^* - \bar{V}_n^* &= n^{-1} \sum_{t=1}^n \left(\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)' - E[\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)'] \right) \\
&+ n^{-1} \sum_{l=1}^h \sum_{t=l+1}^n \left\{ \left(\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})' - E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})'] \right) \right. \\
&\left. + \left(\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)' - E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)'] \right) \right\}
\end{aligned}$$

For what follows we denote by $\|\cdot\|$ the euclidian norm of a matrix and for a random matrix X we define $\|X\|_p \equiv (E\|X\|^p)^{1/p}$.

Now,

$$\|\hat{V}_n^* - \bar{V}_n^*\|_2 \leq \left\| n^{-1} \sum_{t=1}^n \varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)' - E[\varepsilon_t^2 X_{p_n}(Z_t) X_{p_n}(Z_t)'] \right\|_2 \quad (\text{III.6.34})$$

$$+ n^{-1} \sum_{l=1}^h \left\| \sum_{t=l+1}^n \left(\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})' - E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})'] \right) \right\|_2 \quad (\text{III.6.35})$$

$$+ n^{-1} \sum_{l=1}^h \left\| \sum_{t=l+1}^n \left(\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)' - E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)'] \right) \right\|_2 \quad (\text{III.6.36})$$

The right hand-side of (6.34) is bounded by $C_1 p_n n^{-1/2}$ as was shown in chapter 2 (5.59), where c_1 denotes a finite constant.

To bound (6.35) notice that

$$\left\| \sum_{t=l+1}^n \left(\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})' - E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})'] \right) \right\|_2^2$$

$$= \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} E \left[\left(\sum_{t=l+1}^n \varepsilon_t \varepsilon_{t-l} \psi_j(Z_t) \psi_h(Z_{t-l}) - E \varepsilon_t \varepsilon_{t-l} \psi_j(Z_t) \psi_h(Z_{t-l}) \right)^2 \right] \quad (\text{III.6.37})$$

The expectation term in (6.37) doesn't exceed $C_2(l+2)n$ by lemma 6.7(b) of Gallant and White (1988), which holds under the assumptions. Also notice that C_2 above denotes a finite constant.

Thus (6.37) is no greater than

$$C_2 \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} (l+2)n = C_2 p_n^2 (l+2)n = C_2 p_n^2 (h+2)n$$

So we get that

$$\|\hat{V}_n^* - \bar{V}_n^*\|_2 \leq C_1 p_n n^{-1/2} + 2C_2 p_n h (h+2)^{1/2} n^{-1/2}$$

and therefore $\|\hat{V}_n^* - \bar{V}_n^*\|_2$ will be $O_p(1)$ if $p_n n^{-1/2} h^{3/2} = O(1)$ by applying Chebychev's inequality.

Finally we show that $\bar{V}_n^* - V_n^*$ is $O(1)$.

First notice that

$$\bar{V}_n^* - V_n^* = n^{-1} \sum_{l=h+1}^{n-1} \sum_{t=l+1}^n \left\{ E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})'] \right\} \quad (\text{III.6.38})$$

$$+ E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)'] \Big\} \quad (\text{III.6.39})$$

and therefore

$$\|\bar{V}_n^* - V_n^*\| \leq n^{-1} \sum_{l=h+1}^{n-1} \left\{ \left\| \sum_{t=l+1}^n E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})'] \right\| \right\} \quad (\text{III.6.40})$$

$$+ \left\| \sum_{t=l+1}^n E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_{t-l}) X_{p_n}(Z_t)'] \right\| \right\} \quad (\text{III.6.41})$$

Now,

$$\begin{aligned} & \left\| \sum_{t=l+1}^n E[\varepsilon_t \varepsilon_{t-l} X_{p_n}(Z_t) X_{p_n}(Z_{t-l})'] \right\| \\ &= \left\{ \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} \left(\sum_{t=l+1}^n E|\varepsilon_t \varepsilon_{t-l} \psi_j(Z_t) \psi_h(Z_{t-l})| \right)^2 \right\}^{1/2} \end{aligned} \quad (\text{III.6.42})$$

and given that under the assumptions

$$E|\varepsilon_t \varepsilon_{t-l} \psi_j(Z_t) \psi_h(Z_{t-l})| \leq \Delta^2 (5\Delta^2 \alpha_{[l/4]}^{1/2-1/r} + K\nu_{[l/4]}^{\frac{(r-2)^2}{4(r-1)^2}})$$

we get that (6.42) doesn't exceed

$$\begin{aligned} & \left\{ \sum_{j=1}^{p_n} \sum_{h=1}^{p_n} \left(\sum_{t=l+1}^n \Delta^2 (5\Delta^2 \alpha_{[l/4]}^{1/2-1/r} + K\nu_{[l/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) \right)^2 \right\}^{1/2} \\ & \leq p_n n \Delta^2 (5\Delta^2 \alpha_{[l/4]}^{1/2-1/r} + K\nu_{[l/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) \end{aligned}$$

Therefore

$$\|\bar{V}_n^* - V_n^*\| \leq 2 \sum_{l=1}^{n-1} p_n \Delta^2 (5\Delta^2 \alpha_{[l/4]}^{1/2-1/r} + K\nu_{[l/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) \quad (\text{III.6.43})$$

$$- 2 \sum_{l=1}^h p_n \Delta^2 (5\Delta^2 \alpha_{[l/4]}^{1/2-1/r} + K\nu_{[l/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) \quad (\text{III.6.44})$$

and we need the sizes of both the mixing and near-epoch dependence coefficients to be such that $p_n \sum_{l=1}^{\infty} (\alpha_{[l/4]}^{1/2-1/r} + \nu_{[l/4]}^{\frac{(r-2)^2}{4(r-1)^2}}) < \infty$.

There only remains to proof that (6.28) is $o_p(1)$ and we can follow similar steps than those that verify condition (5.6) in chapter 2. By doing that it can be easily seen that (6.28) will hold if for some $0 < \delta < 1$

$$\begin{aligned} & \left[\inf_{\alpha \in A_n} \inf_{\lambda \geq 0} nL_n(\alpha, \lambda) \right]^{-(1-\delta)} \sup_{\alpha \in A_n} \sup_{\lambda \geq 0} \{ (n^{-1}[X_n(\alpha)'X_n(\alpha) + \lambda I_{q_n}])^{-1} \\ & - (n^{-1}E[X_n(\alpha)'X_n(\alpha) + \lambda I_{q_n}])^{-1} \} = o_p(1) \end{aligned} \quad (\text{III.6.45})$$

and

$$\begin{aligned} & \left[\inf_{\alpha \in A_n} \inf_{\lambda \geq 0} nL_n(\alpha, \lambda) \right]^{-\delta} \sup_{\alpha \in A_n} \{ n^{-1}X_n(\alpha)' \varepsilon \varepsilon' X_n(\alpha) \\ & - E[n^{-1}X_n(\alpha)' \varepsilon \varepsilon' X_n(\alpha)] \} = O_p(1) \end{aligned} \quad (\text{III.6.46})$$

As in chapter 2, (5.45) will hold if

$$\| n^{-1}[X'_{np_n} X_{np_n} - E[X'_{np_n} X_{np_n}]] \| = O_p(1) \quad (\text{III.6.47})$$

and

$$\lambda_{\min} \left\{ n^{-1} \sum_{t=1}^n EX_{p_n}(Z_t)X_{p_n}(Z_t)' \right\} = C_n \quad (\text{III.6.48})$$

where $C_n(\inf_{\alpha \in A_n} \inf_{\lambda \geq 0} nL_n(\alpha, \lambda))^{-(1-\delta)} = o_p(1)$.

Also by using the same argument than for showing (5.31) in chapter 2 we can conclude that (6.46) will hold if for some $p > 0$ and δ as specified above

$$(p_n^2 n)^{1/p} \left(\sum_{\tau=1}^{n-1} (\tau + 2) \right)^{1/p} \left(\inf_{\alpha \in A_n} \inf_{\lambda \geq 0} nL_n(\alpha, \lambda) \right)^{-\delta} = O_p(1)$$

III.7 References

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, 2nd International Symposium in Information Theory, Petrov, B.N. and Csaki, F. eds., Akademiai Kiado, Budapest, 267-281.

Andrews, D.W.K. (1991), Asymptotic Optimality of Generalized C_L , Cross-validation and Generalized Cross-validation in Regression with Heteroskedastic Errors, *Journal of Econometrics*, 47, 359-377.

Andrews, D.W.K. (1991), Asymptotic Normality of Series Estimators for Various Nonparametric and Semi-parametric Regression Models, *Econometrica*, 59, 307-345.

Andrews, D.W.K. (1991), Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica*, 59, 817-858.

Burman, P., Chow, E. and Nolan, D. (1994), A Cross-validatory Method for Dependent Data, *Biometrika*, 84, 351-358.

Burman, P. and Nolan, D. (1992), Data-Dependent Estimation of Prediction Functions, *Journal of Time Series Analysis*, 13, 189-207.

Burnham, K. and Anderson, D. (2003), *Model Selection and Multi-Model Inference*, Springer.

Candes, E. (1998), *Ridgelets: Theory and Applications*. Ph.D. Dissertation, Department of Statistics, Stanford University.

Candes, E. (1999), Harmonic Analysis of Neural Networks, *Applied and Computational Harmonic Analysis*, 6, 197-218.

Candes, E. (2003), Ridgelets: Estimating with Ridge Functions, *Annals of Statistics*, 33, 1561-1599.

Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press.

Droge, B. (1999), Asymptotic Optimality of Full Cross-validation for Selecting Linear Regression Models, *Statistics and Probability Letters*, 44, 351-357.

Gallant, R.A. and White, H. (1988), A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models, Oxford: Basil Blackwell.

Hansen, B.E. (1992), Consistent Covariance Matrix Estimation for Dependent Heterogeneous Processes, *Econometrica*, 60, 967-972.

Konishi, S. and Kitagawa, G. (1996), Generalised Information Criteria in Model Selection, *Biometrika*, 83, 875-890.

Li, K.-C. (1986), Asymptotic Optimality of C_L and Generalized Cross-validation in Ridge Regression with Applications to Spline Smoothing, *Annals of Statistics*, 14, 1101-1112.

Li, K.-C. (1987), Asymptotic Optimality for C_p , C_L , Cross-validation and Generalized Cross-validation: Discrete Index Set, *Annals of Statistics*, 15, 958-975.

Mallows, C. (1973), Some Comments on C_p , *Technometrics*, 15, 661-675.

Newey, W.K. (1997), Convergence Rates and Asymptotic Normality for Series Estimators, *Journal of Econometrics*, 79, 147-168.

Opsomer, J., Wang, Y. and Yang, Y. (2001), Nonparametric Regression with Correlated Errors, *Statistical Science*, 16, 134-153.

Racine, J. (1997), Feasible Cross-validatory Model Selection for General Stationary Processes, *Journal of Applied Econometrics*, 12, 169-179.

Racine, J. (2000), A Consistent Cross-validatory Method for Dependent Data: $h\nu$ -Block Cross-validation, *Journal of Econometrics*, 9, 39-61.

Shao, J. (1993), Linear Model Selection by Cross-validation, *Journal of the American Statistical Association*, 88, 486-494.

Shao, J. (1997), An Asymptotic Theory for Linear Model Selection, *Statistica Sinica*, 7, 221-264.

Stone, M. (1974), Cross-validatory Choice and Assessment of Statistical Predictions, *J.R. Statist. Soc. B*, 36, 111-147.

Schwarz, G. (1978) Estimating the Dimension of a Model, *Annals of Statistics*, 6, 461-464.

White, H. (2006), Approximate Nonlinear Forecasting Methods, Handbook of Economic Forecasting, Edited by Graham Elliott, Clive Granger and Allan Timmermann, Elsevier.

White, H. (2001), Asymptotic Theory for Econometricians, Academic Press.