Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2007/02

Flexible Time Series Forecasting Using Shrinkage Techniques and Focused Selection Criteria

Christian T. Brownlees,
Giampiero M. Gallo

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Flexible Time Series Forecasting Using Shrinkage Techniques and Focused Selection Criteria

Christian T. Brownlees* Giampiero M. Gallo*

This version: May 2007

Abstract

Nonlinear time series models can exhibit components such as long range trends and seasonalities that may be modeled in a flexible fashion. The resulting unconstrained maximum likelihood estimator can be too heavily parameterized and suboptimal for forecasting purposes. The paper proposes the use of a class of shrinkage estimators that includes the Ridge estimator for forecasting time series, with a special attention to GARCH and ACD models. The local large sample properties of this class of shrinkage estimators is investigated. Moreover, we propose symmetric and asymmetric focused selection criteria of shrinkage estimators. The focused information criterion selection strategy consists of picking up the shrinkage estimator that minimizes the estimated risk (e.g. MSE) of a given smooth function of the parameters of interest to the forecaster. The usefulness of such shrinkage techniques is illustrated by means of a simulation exercise and an intra-daily financial durations forecasting application. The empirical application shows that an appropriate shrinkage forecasting methodology can significantly outperform the unconstrained ML forecasts of rich flexible specifications.

Keywords: Forecasting, Shrinkage Estimation, FIC, MEM, GARCH, ACD
JEL: C22, C51, C53

*Dipartimento di Statistica "G. Parenti", Viale G.B. Morgagni 59, I-50134 Firenze, Italy, e-mail: ctb@ds.unifi.it gallog@ds.unifi.it.
Acknowledgments. We are grateful to Hal White, Francesco Maggina, Margherita Velucchi, Fabrizio Cipollini, Max Marinucci, Francisco Pascual, Livio Fenga and Corrado Pelizzari for comments. Financial support from the MIUR (PRIN 2006131140-004, FISR) is gratefully acknowledged. All mistakes are ours.

1 Introduction

Nonlinear time series models can exhibit components such as long range trends and seasonalities that may be modeled in a flexible fashion using splines, flexible functional forms, trigonometric polynomials, and so forth. Leading examples in the financial econometrics literature include the modelling of long run volatility trends (c.f. Engle & Rangel (2005)) and the analysis of intra–daily financial durations periodicity (c.f. Engle & Russell (1998)). The resulting unconstrained maximum likelihood estimator can sometimes be too expensively parameterised and suboptimal for forecasting purposes. In such cases it is possible to obtain some gains in terms of forecasting precision by appropriately restricting the specification. This is usually achieved by constraining the forecasting model using some model selection strategy. Shrinkage estimation techniques represent an alternative or complement to model selection strategies for many of these forecasting applications. In this work we use the term shrinkage to refer to penalized maximum likelihood estimation procedures. These methods consist of shrinking the maximum likelihood estimator in the attempt to obtain a new estimator with smaller risk (e.g. MSE).

This paper proposes the use of shrinkage estimation techniques for forecasting with flexible time series models in a general maximum likelihood framework. The class of shrinkage estimators we consider includes the Ridge and Generalized Ridge estimators (Hoerl & Kennard (1970)) as well as some variants of the Bridge estimator (Frank & Friedman (1993)) as special cases. Using the local misspecification framework developed in Hjort & Claeskens (2003), we show that in large samples shrinkage estimators produce estimators that are biased but have smaller variance than the maximum likelihood estimator. Shrinkage techniques can thus lead to a smaller expected loss under quite different loss functions. Moreover, as in Claeskens & Hjort (2003), the large sample analysis of the estimators' expected loss suggests a class of focused selection criteria. The term focused refers to the estimated expected loss of a given function of the parameter estimates which is of interest in the chosen application context. The focused information criterion selection strategy consist of choosing the model that minimizes the estimated risk (e.g. MSE) of a given smooth function of the parameters of interest to the forecaster. As examples, consider that precision in the estimation of a (nonlinear) function of the parameters (e.g. the persistence, the unconditional variance or the half-life of a shock in a GARCH model) may be more important than that of single parameters.

The discussion is developed with a special attention to the family of Multiplicative Error Models (MEM) (Engle (2002), Engle & Gallo (2006)), a model class that includes the GARCH and ACD families. The usefulness of such shrinkage techniques is illustrated by means of a simulation exercise and an intra-daily financial durations forecasting application. The simulation exercise consists of adopting a MEM where the conditional expectation mimics ultra–high frequency dynamics with a time-of-day periodic component specified with trigonometric polynomials. Cross-validation and focused information criteria with variants from different loss functions form the basis for choosing the amount

of shrinkage to adopt in the estimation. The resulting performance shows an improvement from the approach when contrasted against the MLE as a baseline. The empirical application refers to a MEM applied to financial durations exhibiting a seasonal pattern due to trading habits. In our prediction exercise, the model parameters are estimated all at once, rather than extracting the seasonal component first and then estimating the parameters for the dynamics of the conditional expectation.

The main contribution of the paper lies in suggesting shrinkage estimation techniques for flexible parametric MEM models and extending the results of Hjort & Claeskens (2003) to a class of shrinkage estimators that has not been previously considered. We also analyze the estimators' risk properties and develop selection criteria using asymmetric loss functions. The results of the intra-daily financial durations forecasting application show that shrinkage estimation is a promising method for prediction that performs better than the ML approach in rich flexible specifications.

There is a number of different contributions in the literature that relate to this work. Engle & Russell (1998) and Engle & Rangel (2005) contain examples of MEMs applications that resort to flexible modelling techniques (splines). Rodríguez-Poo, Veredas & Espasa (2007) propose a seminonparametric model for financial duration data. Fokianos & Tsolaki (2006) proposes Ridge estimators for INAR models. White (2006) reviews approximate nonlinear forecasting methods. Sen (1979) is an early contribution on the use of local asymptotics for the analysis of post selection estimators. Kiefer & Skoog (1984) investigate the effects of local misspecification on the maximum likelihood estimator. Knight & Fu (2000) use local asymptotics for the analysis of the large sample distribution of shrinkage type estimators to show how Bridge estimators can provide a risk improvement in the linear regression model framework. Hjort & Claeskens (2003) present some results regarding James–Stein type estimators. Hansen (2005) and Claeskens, Croux & Van Kerckhoven (2007) analyze the focused selection methods proposed in Claeskens & Hjort (2003) in a time series setting.

The rest of the paper is organized as follows. Section 2 outlines the shrinkage forecasting methodology within the context of a flexible parametric MEM. Section 3 presents the theoretical framework and results. Section 4 presents two forecasting applications on simulated data and intra-daily financial durations. Concluding remarks follow in Section 5.

2 Methodology

This sections describes a general shrinkage forecasting methodology that can be applied in many contexts: for the sake of clarity, we will consider a Multiplicative Error Model as a leading example.

Let $\{y_t\}$ denote a generic MEM process and let $\mathcal{F}_{t-1}$ be the information set at time $t-1$. The general definition of a MEM process is

$$y_t = \mu_t \, \epsilon_t \quad \epsilon_t | \mathcal{F}_{t-1} \sim Gamma(\psi, 1/\psi) \tag{1}$$

where, conditionally on $\mathcal{F}_{t-1}$, μ_t is the conditionally deterministic component of the process and ϵ_t is an i.i.d. innovation term with unit expectation. Let $\{x_t\}$ denote a predetermined variable that it is known to improve the forecasts of $\{y_t\}$ but for which no knowledge of the relationship with $\{y_t\}$ is available. A flexible specification of the conditional mean μ_t is given by

$$\mu_t = \omega + \alpha y_{t-1} + \beta \mu_{t-1} + \sum_{i=1}^{k} \eta_i \, h_i(x_{t-1}), \tag{2}$$

where $h_i : \mathbb{R} \to \mathbb{R}$ represents some appropriate linear basis expansion of x_{t-1}, for all i. A discussion on conditions which ensure stationarity and nonnegativity of the MEM process can be found in Engle (2002), Nelson & Cao (1992) and Doornik & Ooms (2000).

The list of possible choices of the $h_i(\cdot)$ basis functions is long: polynomials, trigonometric polynomials, wavelets, ridgelets and so forth. Different basis functions often have quite different properties which may turn out to be more or less useful depending on the problem at hand. We do not attempt to provide a detailed review of the possible choices of the basis functions (for more details we refer to White (2006)). We would only like to stress that bounded functions (e.g. trigonometric polynomials) are easier to handle than non bounded functions (e.g. splines), in that the latter can create more numerical difficulties in the MEM estimation using nonlinear optimization algorithms.

Typically, we would like the number of $h_i(\cdot)$ terms in Equation (2) to be reasonably large in order to be able to approximate sufficiently well the unknown link between y_t and x_t. However, this can lead to rather rich model parameterization that can inflate the estimator variance and turn out to be suboptimal for prediction. Shrinkage estimation methods allow one to handle this problem. Consider a partition of the model parameters in two vectors, say $\theta \in \mathbb{R}^p$ containing the parameters not to be shrunk (e.g. the ω, α, β and ψ parameters) and $\gamma \in \mathbb{R}^q$ containing the parameters to be shrunk (e.g. the η_i $i = 1, .., k$ parameters). Let $L_n(\theta, \gamma)$ denote the log-likelihood function of a sample of size n. For a given $\lambda \in \mathbb{R}^+$ the λ "ridge" shrinkage estimator of $(\theta', \gamma')'$ is the solution to the penalized likelihood maximization problem:

$$\begin{pmatrix} \hat{\theta}_{n,\lambda} \\ \hat{\gamma}_{n,\lambda} \end{pmatrix} = \arg\max \left\{ L_n(\theta, \gamma) - \lambda \, \|\gamma\|^2 \right\}. \tag{3}$$

The properties of the "ridge" shrinkage estimator depend on the regularizing parameter λ. Large values of λ will push the γ–parameters towards 0, increasing the bias of the estimator and reducing the variance. On the other hand, small values of λ will keep the estimator close to the unconstrained MLE, reducing the bias and increasing the variance. Therefore, there is a bias/variance trade–off that depends on the choice of the shrinkage parameter λ. By appropriately choosing the shrinking parameter λ, it is possible to obtain an estimator with smaller risk (e.g. MSE) than the MLE.

The success in beating the MLE relies in choosing λ appropriately. In the shrinkage literature the amount of shrinkage is often determined by cross-

validatory methods. In this work we propose a recently proposed criterion called the Focused Information Criterion (Claeskens & Hjort (2003)). Let some known function of the parameters $g : \mathbb{R}^{p+q} \to \mathbb{R}$ be denoted as the focus parameter. Let g_{true} denote the value of g evaluated at the true parameters and let $\hat{g}_\lambda$ denote the estimator of g_{true} using the λ shrinkage estimator. The focus parameter used in this paper is the unconditional mean of the process keeping the values of the predetermined variables fixed at x, that is

$$g \equiv \mu(x) = \frac{\omega + \sum_{i=m}^{k} \eta_m h_i(x)}{1 - \alpha - \beta}. \tag{4}$$

Using the *local misspecification framework* developed in Hjort & Claeskens (2003) it is possible to obtain the limiting distribution of $\hat{g}_\lambda$:

$$\sqrt{n}(g_{\text{true}} - \hat{g}_\lambda) \overset{a}{\sim} \Lambda_\lambda \equiv N(b_\lambda, \tau_\lambda^2),$$

where b_λ and τ_λ^2 respectively denote the bias and variance of the focus parameter shrinkage estimator. For some appropriate loss function L the asymptotic risk of the g_λ estimator is

$$r_L(g_{\text{true}}, \hat{g}_\lambda) \equiv \mathrm{E}(L(\Lambda_\lambda))) \,.$$

The loss functions considered in this work are both symmetric (square and absolute loss) as well as asymmetric (linex and linlin). The $\mathrm{FIC}_L(\lambda)$ turns out to be an estimator of such limiting risk

$$\mathrm{FIC}_L(\lambda) \equiv \hat{r}_L(g_{\text{true}}, \hat{g}_\lambda),$$

and the FIC shrinkage selection strategy consists of picking up the λ which minimizes the estimated risk. The appealing feature of such shrinkage selection strategy is that the forecaster can decide the most appropriate focus parameter for the context of his/her application and loss function.

3 Theory

This section provides the base assumptions and results for the derivation of the asymptotic distribution of the estimators of interest. This is achieved by using the local misspecification approach developed in Hjort & Claeskens (2003).

3.1 Local Misspecification Framework

Although, under appropriate regularity conditions more generic settings can be treated as well (nonlinear model using stochastic explanatory variables), the results of these section are more easily understood withing the original framework of Hjort & Claeskens (2003) of independent data $y_1, ..., y_n$. Their common density f is assumed to depend on the two previously defined parameter vectors, $\theta \in \Theta \subseteq \mathbb{R}^p$ and $\gamma \in \Gamma \subseteq \mathbb{R}^q$. The γ–parameter vector contains the parameters that may be attempted to constrain, while there is no such need for

the θ–parameter vector. It is assumed that there exist an unknown $\theta_0 \in \mathbb{R}^p$, a *known* $\gamma_0 \in \mathbb{R}^q$ and an unknown $\delta_0 \in \mathbb{R}^q$ so that the true density is

$$f_{\text{true}} = f_n \equiv f(y, \theta_0, \gamma_0 + \delta_0/\sqrt{n}), \tag{5}$$

this assumption is called the "local misspecification" assumption in that it states that the constrained model $f(y, \theta, \gamma_0)$ with $\theta \in \Theta$, also known as the *narrow* model, is locally misspecified. A central role in the large sample analysis is played by the *null* model, which is the density f in $(\theta_0', \gamma_0')'$, that is

$$f_0 \equiv f(y, \theta_0, \gamma_0). \tag{6}$$

Also, let $E(\cdot)$ and $Var(\cdot)$ indicate the expected value and variance with respect to the true model of Equation (5), while let $E_0(\cdot)$ and $Var_0(\cdot)$ denote the expected value and variance with respect to the null model of Equation (6).

As is customary, the average log–likelihood function determined by a sample $y^n \equiv (y_1, y_2, ..., y_n)'$ is denoted by

$$L_n(y^n, \theta, \gamma) \equiv n^{-1} \sum_{i=1}^{n} \log f(y_i, \theta, \gamma),$$

and the gradient of the log–likelihood function is

$$\nabla L_n = \begin{pmatrix} \nabla L_{n,1} \\ \nabla L_{n,2} \end{pmatrix} \equiv n^{-1} \sum_{i=1}^{n} \begin{pmatrix} s_1(y_i) \\ s_2(y_i) \end{pmatrix},$$

where $s(\cdot)$ is the score,

$$s(y) = \begin{pmatrix} s_1(y) \\ s_2(y) \end{pmatrix} \equiv \begin{pmatrix} \nabla_\theta \log f(y, \theta, \gamma) \\ \nabla_\gamma \log f(y, \theta, \gamma) \end{pmatrix}.$$

The subscripts 1 and 2 denote respectively the derivatives with respect to the θ and γ parameters.

An important ingredient of this large sample analysis is the variance–covariance matrix of the null gradient of the log–likelihood function under the null model, denoted by B_0. Let the score at the null point be

$$s_0(y) = \begin{pmatrix} s_{0,1}(y) \\ s_{0,2}(y) \end{pmatrix} \equiv \begin{pmatrix} \nabla_\theta \log f(y, \theta_0, \gamma_0) \\ \nabla_\gamma \log f(y, \theta_0, \gamma_0) \end{pmatrix};$$

then

$$B_0 \equiv \text{Var}_0 \left(n^{-1/2} \sum_{i=1}^{n} s_0(y_i) \right),$$

with the following structure

$$B_0 = \begin{pmatrix} B_{0,11} & B_{0,12} \\ B_{0,21} & B_{0,22} \end{pmatrix},$$

with blocks corresponding respectively to the θ and γ parameters.

Under appropriate regularity condition reported in the appendix, Hjort & Claeskens (2003) obtain the following result.

Lemma 1 *(Hjort-Claeskens Lemma) Consider the averages*

$$\nabla L_{0,n,1} = n^{-1}\sum_{i=1}^{n} s_{0,i,1}(y_i) \quad \text{and} \quad \nabla L_{0,n,2} = n^{-1}\sum_{i=1}^{n} s_{0,i,2}(y_i).$$

Under the local misspecification framework, then

$$\begin{pmatrix}\sqrt{n}\nabla L_{0,n,1}\\ \sqrt{n}\nabla L_{0,n,2}\end{pmatrix} \xrightarrow{d} \begin{pmatrix}B_{0,12}\ \delta_0\\ B_{0,22}\ \delta_0\end{pmatrix} + \begin{pmatrix}M\\ N\end{pmatrix} \quad \begin{pmatrix}M\\ N\end{pmatrix} \sim N_{p+q}(0, B_0).$$

This important lemma provides the large sample description of what happens to the distribution of the gradient of the log-likelihood function at the null model (θ_0, γ_0) under the truth within the local misspecified framework. This result allows one to derive the distribution of the estimators of interest using standard arguments.

3.2 Submodel, Shrinkage and Submodel Shrinkage Estimators

The local misspecification assumption implies that there is a known value of the parameter space γ_0 sufficiently close to the true value $\gamma_0 + \delta_0/\sqrt{n}$ in large samples. It may thus be advantageous to appropriately constrain the γ parameters of the model in order to construct estimators with better risk properties than the unconstrained maximum likelihood estimator. The family of constrained estimators that we consider in this work also comprises submodel as well as submodel shrinkage estimators.

We begin by defining the constraints that determine the estimators of interest. Let S be a subset of $\{1, 2, ..., q\}$ and let $v = (v_1, ..., v_q)'$ be a vector in $\mathbb{R}^q$. Denote by v_S the subvector of v of components v_j with $j \in S$. Analogously, denote by v_{S^c} the subvector of v of components v_j with $j \in S^c$, the complement of S with respect to $\{1, 2, ..., q\}$. Also denote by $\pi_S \in \mathbb{R}^{|S|\times q}$ the projection matrix mapping v to the subvector v_S, i.e. the matrix that $\pi_S v = v_S$. The set of constraints that determines a submodel estimator is defined as follows.

Definition 1 *(Submodel Constraint) For some subset $S \subseteq \{1, 2, ..., q\}$, the S submodel constraint of the model f is defined by the set*

$$\Gamma_S \equiv \{\gamma \in \Gamma : \gamma_{S^c} = \gamma_{0,S^c}\}.$$

The definition of the shrinkage constraint requires some more work. Broadly speaking, a shrinkage estimator imposes a bound on the deviation of γ from the null point γ_0, measured by some appropriate penalty function. We proceed by first providing an appropriate definition of penalty function for the scope of the current analysis.

Definition 2 *(Penalty Function) A function $\rho : \mathbb{R}^q \to \mathbb{R}^+$ is a* penalty function *if $\rho(\cdot)$ is continuously differentiable of order 2 on $\mathbb{R}^q$, $\rho(\mathbf{0}) = 0$, $\nabla\rho(\mathbf{0}) = \mathbf{0}$, $\nabla^2\rho(x)$ is positive definite for each x in $\mathbb{R}^q$.*

The type of penalties we take into account are smooth penalties like the Ridge penalty, while we are ruling out penalties that are non differentiable at the origin such as the LASSO (Tibshirani (1996)). The definition nevertheless allows considerable flexibility regarding the shape of the penalty, allowing one for instance to penalize different parameters in different ways.

We define the set of constraints that determine a shrinkage estimator as follows.

Definition 3 *(Shrinkage Constraint) For some penalty function $\rho(\cdot)$ and nonnegative real number c, the (ρ, c) shrinkage constraint of the model f is the set*

$$\Gamma_{\rho,c} \equiv \{\gamma \in \Gamma : \rho(\gamma - \gamma_0) \leq c\}.$$

The constraints of Definitions 1 and 3 can also be combined to achieve the constraint defining a submodel shrinkage estimator.

Definition 4 *(Submodel Shrinkage Constraint) For some subset $S \subseteq \{1, ..., q\}$, penalty function $\rho(\cdot)$ and nonnegative real number c the (S, ρ, c) submodel shrinkage constraint of the model f is defined by the set*

$$\Gamma_{S,\rho,c} \equiv \Gamma_S \cap \Gamma_{\rho,c}.$$

In what follows we will use the symbol m to denote a generic nested model.

Definition 5 *(Nested Model) Let S be a subset of $\{1, ..., q\}$, $\rho(\cdot)$ a penalty function and c a nonnegative real number. The nested model m of the model f is defined as the constrained specification satisfying the Γ_m constraints, where Γ_m is equal to either the Γ_S, $\Gamma_{\rho,c}$ or $\Gamma_{S,\rho,c}$ constraints.*

We introduce this definition not only for notational convenience but also because the class of submodel shrinkage constraints that we have defined does not nest the class of submodel constraints. A more general class of nested models has to be defined to include all possible cases of interest.

We can now establish a lemma that provides the asymptotic distribution of the estimators of a nested model m.

Lemma 2 *(Asymptotic Normality of the Nested Model Estimator) Let m be a nested model and let $(\hat{\theta}'_{n,m}, \hat{\gamma}'_{n,m})'$ be the nested model estimator. Under the local misspecification framework, then*

$$\begin{pmatrix} \sqrt{n}\,(\hat{\theta}_{n,m} - \theta_0) \\ \sqrt{n}\,\pi_S(\hat{\gamma}_{n,m} - \gamma_0) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} C_m \\ D_m \end{pmatrix}$$

with

$$\begin{pmatrix} C_m \\ D_m \end{pmatrix} = \begin{pmatrix} B_{0,11} & B_{0,12}\pi_S' \\ \pi_S B_{0,21} & \pi_S(B_{0,22} + \lambda\nabla^2\rho(\mathbf{0}))\pi_S' \end{pmatrix}^{-1} \begin{pmatrix} B_{0,12}\ \delta_0 + M \\ \pi_S B_{0,11}\ \delta_0 + \pi_S N \end{pmatrix}$$

where

- *if m imposes submodel constraints, S is a subset of $\{1, ..., q\}$, otherwise $S = \{1, ..., q\}$;*
- *if m imposes shrinkage constraints, for given $c \geq 0$ there corresponds a $0 < \lambda \leq +\infty$, otherwise $\lambda = 0$.*

Note that as the proof of Lemma 2 points out, in practice the shrinkage estimators are obtained by maximizing the unconstrained penalized likelihood for a given value of λ, that is

$$\begin{pmatrix} \hat{\theta}_{n,m} \\ \hat{\gamma}_{n,m} \end{pmatrix} = \arg\max_{\Theta\times\Gamma_S} \{L_{n,S}(\theta, \gamma_S) - \lambda\rho_S(\gamma_S - \gamma_{0,S})\}$$

where $L_{n,S}$ and ρ_S denote the log–likelihood function and the penalty functions as functions of the γ_S parameter only (with γ_{S^c} constrained to γ_{0,S^c}). In practice, maximizing the constrained log–likelihood function for a given c is usually avoided in that constrained maximization is much harder then unconstrained maximization.

Let us introduce some further notation to provide a more insightful expression for the asymptotic distribution of a nested model estimator given by Lemma 2. Let $B_{0,S}$ denote the variance–covariance matrix of the gradient of the log–likelihood function of the submodel S at the null point,

$$\mathrm{Var}_0 \begin{pmatrix} \sqrt{n}\nabla L_{0,n,1} \\ \sqrt{n}\nabla L_{0,n,2,S} \end{pmatrix} = B_{0,S} = \begin{pmatrix} B_{0,11} & B_{0,12}\pi_S' \\ \pi_S B_{0,21} & \pi_S B_{0,22}\pi_S' \end{pmatrix}$$

and denote its inverse by

$$B_{0,S}^{-1} = \begin{pmatrix} B_{0,S}^{11} & B_{0,S}^{12} \\ B_{0,S}^{21} & B_{0,S}^{22} \end{pmatrix},$$

which by using the matrix inversion formula for partitioned matrices can be represented as

$$\begin{pmatrix} B_{0,11}^{-1} + B_{0,11}^{-1}B_{0,12}\pi_S' K_S \pi_S B_{0,12}B_{0,11}^{-1} & -B_{0,11}^{-1}B_{0,12}\pi_S' K_S \\ -K_S\pi_S B_{0,12}B_{0,11}^{-1} & K_S \end{pmatrix}$$

where $K_S \equiv (\pi_S(B_{0,22} - B_{0,20}B_{0,11}^{-1}B_{0,12})\pi_S')^{-1}$.

Let K denote B_0^{22} and let us introduce $W \equiv K(N - B_{0,21}B_{0,11}^{-1}M)$, which is distributed as $N_q(0, K)$. Finally, let us define $D \equiv \delta_0 + W$, which shares the asymptotic distribution of the γ unrestricted maximum likelihood estimator

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \xrightarrow{d} D \sim N_q(\delta_0, K).$$

We can now state a corollary of Lemma 1 that delivers a simpler representation of the asymptotic distribution of the m nested model estimator.

Corollary 1 *(Nested Model Estimator) Let m be a nested model and let $(\hat{\theta}'_{n,m}, \hat{\gamma}'_{n,m})'$ be the nested model estimator. Under the local misspecification framework, then*

$$\begin{pmatrix} C_m \\ D_m \end{pmatrix} = \begin{pmatrix} B_{0,11}^{-1}M + B_{0,11}^{-1}B_{0,12}(\delta_0 - K^{1/2}(H_S - G_m)K^{-1/2}D) \\ (I_q - R_m)K_S\pi_S K^{-1}D \end{pmatrix}$$

where $H_S \equiv K^{-1/2}\pi'_S K_S \pi_S K^{-1/2}$ ($H_\emptyset \equiv \mathbf{0}_{q\times q}$), $G_m \equiv K^{-1/2}\pi'_S R_m K_S \pi_S K^{-1/2}$ and

- *if m imposes submodel constraints, S is a subset of $\{1, ..., q\}$ otherwise $S = \{1, ..., q\}$;*
- *if m imposes shrinkage constraints, $R_m \equiv K_S(K_S + \lambda^{-1}(\pi_S \nabla^2 \rho(\mathbf{0})\pi'_S)^{-1})^{-1}$, otherwise $R_m \equiv \mathbf{0}_{q\times q}$.*

3.3 Distribution and Risk of the Focus Parameter Estimator

In what follows we assume that a specific known scalar function of the parameters has been singled out, with a relevant interpretation within the application of interest (e.g. the persistence of shocks in a GARCH model). Such function will be referred as the focus parameter $g \equiv g(\theta_0, \gamma_0 + \delta_0/\sqrt{n})$, where $g(\cdot) : \mathbb{R}^{p+q} \to \mathbb{R}$. Following Hjort & Claeskens (2003), using Corollary 1 and the delta method it is possible to obtain the first order approximation of the asymptotic distribution of sufficiently smooth functions $g(\cdot)$ of the m nested model estimator. We will refer to the $\hat{g}_{n,m} \equiv g(\hat{\theta}_{n,m}, \hat{\gamma}_{n,m})$ as the m nested model estimator of the focus parameter g.

Lemma 3 *(Distribution and Moments of the m Nested Model Estimator of g) Let m be a nested model and let $(\hat{\theta}'_{n,m}, \hat{\gamma}'_{n,m})'$ be the nested model estimator. Let the function $g : \Theta \times \Gamma \to \mathbb{R}$ be continuously differentiable of order 2 on $\Theta \times \Gamma$ and let $\hat{g}_{n,m} \equiv g(\hat{\theta}_{n,m}, \hat{\gamma}_{n,m})$ and $g \equiv g(\theta_0, \gamma_0 + \delta_0/\sqrt{n})$. Under the local misspecification framework then*

$$\sqrt{n}(\hat{g}_{n,m} - g) \xrightarrow{d} \Lambda_m = \tfrac{\partial g}{\partial \theta}' B_{0,11}^{-1} M + \omega'(\delta_0 - K^{1/2}(H_S - G_m)K^{-1/2}D)$$

where $\omega \equiv B_{0,21}B_{0,11}^{-1}\frac{\partial g}{\partial \theta} - \frac{\partial g}{\partial \gamma}$ and $\frac{\partial g}{\partial \theta}, \frac{\partial g}{\partial \gamma}$ are the partial derivatives of $g(\cdot)$ with respect to θ and γ in $(\theta'_0, \gamma'_0)'$. The limiting distribution is a normal random variable with mean b_m and variance τ_m^2 equal to

$$b_m = \omega'(I_q - K_n^{1/2}(H_S - G_m)K^{-1/2})\delta_0$$

$$\tau_m^2 = \tau_0^2 + \omega' K^{1/2}(H_S - G_m)(H_S - G_m)K^{1/2}\omega$$

where $\tau_0^2 \equiv \frac{\partial g}{\partial \theta}' B_{0,11}^{-1}\frac{\partial g}{\partial \theta}$. Furthermore, let m' and m'' be two nested models, then, the covariance $\tau_{n,m',m''}$ between $\hat{g}_{n,m'}$ and $\hat{g}_{n,m''}$ is

$$\tau_{m',m''} = \tau_0^2 + \omega' K^{1/2}(H_{S'} - G_{m'})(H_{S''} - G_{m''})K^{1/2}\omega.$$

Once the asymptotic distribution a focus parameter estimator is obtained, it is straightforward to compute the corresponding expected loss for may loss functions. In the following proposition we provide the closed form expression of a number of expected losses of the $\hat{g}_{n,m}$ estimator using

- square loss: $L_s(x) = x^2$;
- absolute loss: $L_a(x) = |x|$;
- linex loss (Zellner (1986)): $L_{le}(x) = a_1(\exp(a_2 x) - a_2 x - 1)$ for $a_1 \in \mathbb{R}^+$, $a_2 \in \mathbb{R} - \{0\}$ and
- linlin loss (Granger (1969)): $L_{ll}(x) = a_1 \mathbf{1}_{\{x<0\}} x - a_2 \mathbf{1}_{\{x>0\}} x$, for $a_1, a_2 \in \mathbb{R}^+$.

Corollary 2 *(Risk of the m Nested Model Estimator of g) Let m be a nested model and let* $(\hat{\theta}'_{n,m}, \hat{\gamma}'_{n,m})'$ *be the nested model estimator. Let the function* $g : \Theta \times \Gamma \to \mathbb{R}$ *be continuously differentiable of order 2 on* $\Theta \times \Gamma$ *and let* $\hat{g}_{n,m} \equiv g(\hat{\theta}_{n,m}, \hat{\gamma}_{n,m})$ *and* $g \equiv g(\theta_0, \gamma_0 + \gamma_0/\sqrt{n})$*. Under the local misspecification framework then*

i. the asymptotic square risk of $\hat{g}_{n,m}$ *is*

$$r_{\text{sq}}(\hat{g}_{n,m}, g) \quad = \quad b_{n,m}^2 + \tau_{n,m}^2;$$

ii. the asymptotic absolute risk of $\hat{g}_{n,m}$ *is*

$$r_{\text{a}}(\hat{g}_{n,m}, g) \quad = \quad 2\tau_m \phi(b_m/\tau_m) + 2b_m[\Phi(b_m/\tau_m) - 1/2].$$

iii. the asymptotic linex risk of $\hat{g}_{n,m}$ *is*

$$r_{\text{le}}(\hat{g}_{n,m}, g) \quad = \quad a_1 \left(\exp\left\{ a_2 b_{n,m} + \frac{a_2^2 \tau_{n,m}^2}{2} \right\} - a_2 b_{n,m} - 1 \right);$$

iv. the asymptotic linlin risk of $\hat{g}_{n,m}$ *is*

$$r_{\text{ll}}(\hat{g}_{n,m}, g) \quad = \quad a_1 b_m + (a_1 + a_2)[\tau_m \phi(-b_m/\tau_m) - b_m \Phi(-b_m/\tau_m)]$$

Shrinkage, submodel and submodel shrinkage estimation may thus lead to a risk improvement in the estimation of the focus parameter over the unrestricted maximum likelihood estimator by appropriately selecting the specification restrictions.

3.4 The Focused Information Criterion

A focused selection criterion stemming from the local misspecification framework is the Focused Information Criterion (FIC) proposed by Claeskens & Hjort (2003). The FIC is an estimate of the focus parameter estimator risk. The following definition presents a generalization of the original FIC which takes into account shrinkage estimation techniques and asymmetric loss functions.

Definition 6 *(Focused Information Criterion)*

i. Extending Claeskens & Hjort (2003), the square FIC of the m nested model estimator is defined as

$$\mathrm{FIC}_{\mathrm{s}}(m) = (\hat{b}_m)^2 + 2\omega' K^{1/2}(H_S - G_m)(H_S - G_m)K^{1/2}\omega,$$

ii. Extending Claeskens, Croux & Van Kerckhoven (2006), the absolute FIC of the m nested model estimator is defined as

$$\mathrm{FIC}_{\mathrm{a}}(m) = 2\hat{\tau}_m\phi(\hat{b}_m/\hat{\tau}_m) + 2\hat{b}_m[\Phi(\hat{b}_m/\hat{\tau}_m) - 1/2],$$

iii. The linex FIC of the m nested model estimator is defined as

$$\mathrm{FIC}_{\mathrm{le}}(m) = \exp\left\{a_1\hat{b}_m + \frac{a_1^2}{2}\left(\hat{\tau}_m^2 - \mathrm{Var}\Big(\hat{b}_m\Big)\right)\right\} - a_1\hat{b}_m,$$

iv. The linlin FIC of the m nested model estimator is defined as

$$\mathrm{FIC}_{\mathrm{ll}}(m) = a_1\hat{b}_m + (a_1 + a_2)[\hat{\tau}_m\phi(-\hat{b}_m/\hat{\tau}_m) - b_m\Phi(-\hat{b}_m/\hat{\tau}_m)],$$

where

$$\hat{b}_m \quad = \quad \omega'(I - K^{1/2}(H_S - G_m)K^{-1/2})D.$$

The FIC selection strategy consists in picking up the model with lowest estimated risk for the focus parameter of interest. Details on the estimation of the FIC can be found in Hjort & Claeskens (2003) and Claeskens & Hjort (2003).

4 Applications

In this section we present two applications of the shrinkage–focused forecasting methodology on both simulated and real data, withe the goal to illustrate its usefulness in specific reference to 1-step ahead forecasting.

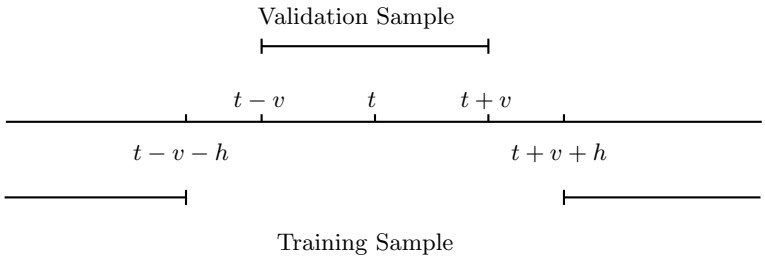


Figure 1: hv–block Cross Validation.

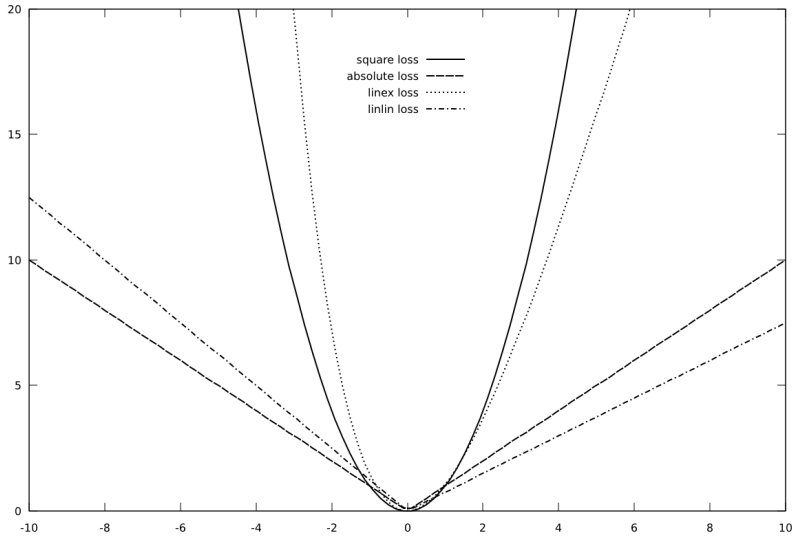


Figure 2: Loss functions graphs.

4.1 Forecasting with a Flexible MEM

The flexible MEM specification for $\{y_t\}$ adopted in the forecasting exercises is

$$y_t = \mu_t \, \epsilon_t, \quad \epsilon_t \sim Gamma(\phi, 1/\phi),$$

with

$$\mu_t = \omega + \alpha y_{t-1} + \beta \mu_{t-1} + \sum_{i=1}^{6} [\eta_{2i-1} \sin(i \, x_{t-1}) + \eta_{2i} \cos(i \, x_{t-1})]. \tag{7}$$

In order to apply the local misspecification framework results, it is further assumed that the θ–parameter vector of the model is $(\omega, \alpha, \beta, \eta_1, \eta_2, \phi)'$, the γ–parameter vector is $(\eta_3, \eta_4, ..., \eta_{11}, \eta_{12})'$ and γ_0 is $\mathbf{0}$. This implies that the first sine and cosine terms in Equation (7) are assumed to be relevant in explaining the relationship between y_t and x_t while, on the other hand, the relevance of the remaining terms is assumed to be marginal. The choice of a trigonometric function is by no means restrictive but is supported by three types of considerations: the possibility of decomposing the periodicity into components related to the frequencies present in the data, the need to adopt bounded functions for nonlinear approximations, and the translation of a limited range for such functions into a lighter burden on the optimization task.

For each period in the prediction sample the forecasting procedure consists of estimating a set of model estimators and then selecting one of them to produce the 1–step ahead forecast for the current period. The parameter estimates and selection criteria are computed each time using a rolling window scheme.

The set of model estimators considered comprises shrinkage estimators as well as the unrestricted maximum likelihood estimator. The penalty function of the shrinkage estimators is the square Euclidean norm of the γ–parameters, i.e. $\rho(\gamma - \gamma_0) = \|\gamma\|^2$.

FIC and Cross Validatory (CV) methods are employed for the selection of the shrinkage parameter λ.

The focus parameter of the FIC methods is the mean of the process conditional on the values of explanatory variable fixed at x_n, the last observation in the estimation sample, that is

$$g = \mu(x_n) = \frac{\alpha_0 + \sum_{i=1}^{6} [\eta_{2i-1} \sin(i \, x_n) + \eta_{2i} \cos(i \, x_n)]}{1 - \alpha_1 - \beta_1}. \tag{8}$$

The CV scheme employed for this simulation is hv–block CV, a cross validatory method for dependent data proposed in Racine (2000). As this CV method is computationally expensive, we resort to a cheaper multifold variant of the original proposal (e.g. Zhang (1993)). Figure 1 provides a graphical sketch of the way this cross–validatory scheme is implemented. For a given time period t, the *validation* sample is constructed using the v observations preceding and following t ($2v + 1$ data points) while the *training* sample is constructed using the observations from the beginning of the sample to the $(t - h - v)$–th

observation and from the $(t+h+v)$–th observation to the end of the sample ($n-2v-2h-1$ data points). The shrinkage estimate is then computed in the training sample and used to forecast in the validation sample. In the estimation step the model is estimated imputing the $2v+2h+1$ removed observations with their expected value. In the validation step the predictions are made using static forecasts. The forecast evaluation is then computed by averaging the prediction losses using the loss function $L : \mathbb{R}^2 \to \mathbb{R}^+$ of interest. The procedure is performed $r = n/(2v+2h+1)$ times so that the validation samples are not overlapping. hv–block CV is then defined as the average of the prediction losses. More compactly for a given loss function L, CV_L is defined as

$$CV_L = \frac{1}{r(2v+1)} \sum_{t \in \mathcal{T}} \sum_{i=-v}^{v+1} L(y_{t+i}, \hat{y}_{t+i\ /(t-v-h:t+h+v+1)}),$$

where $\mathcal{T} = \{v+h+j(2v+2h+1), j = 0, ..., r-1\}$ and $\hat{y}_{t+i\ /(t-v-h:t+h+v+1)}$ denotes the forecast of observation y_{t+i} produced using the parameter estimates obtained from the training sample. This CV measure is computationally demanding even with the multifold variant. For $n = 500$, $h = 10$ and $v = 39$, the model is estimated $r = 5$ times for each shrinkage level λ of the shrinkage estimator.

Both the FIC and CV are computed for the square, absolute, linex (with $a_1 = 10, a_2 = -0.5$) and linlin (with $a_1 = 0.75, a_2 = 1.25$) loss functions. Figure 2 displays the graphs of the loss functions. The parameters of the asymmetric loss functions are chosen in a way as to have the linex loss penalize positive errors more than the square loss and the linlin loss penalize positive errors more than the absolute loss (and viceversa).

The evaluation of the forecasting procedures is carried out by computing the same four loss functions on 1–step ahead forecasts.

4.2 Simulated Forecasting Exercise

The aim of the simulation exercise is to investigate the improvement of the shrinkage estimator forecasts over the MLE forecasts when the DGP deviates from the null model under two different parameter scenarios. In the first parameter setting (Design 1) the deviation of the DGP from the null model is mild, while it is more pronounced in the other case (Design 2). The explanatory variable x_t is assumed to be i.i.d as a $U(0, 2\pi)$. Table 1 reports the parameter values used under each design. Figure 3 displays the mean of y_t conditional on the values of explanatory variable fixed at x for each settings. The set of shrinkage λ values used in this application is $\{0.2\ k : k = 1, ..., 10\}$, with an upper limit of 2, since the corresponding estimated parameter values are virtually equal to 0. This simulation exercise consists of 100 simulated paths of 550 observations each, where the series $\{y_t\}$ has to be predicted from observations 501 to 550. The Monte Carlo experiment leads to a total of 5000 1–step ahead forecasts under each design.

Design 1				
ω	α	β	η_1	η_2
0.5	0.2	0.4	0.0	0.3
η_3	η_5	η_7	η_9	η_{11}
-0.08089	-0.00891	0.03232	0.01234	-0.00691
η_4	η_6	η_9	η_{10}	η_{12}
0.01873	-0.03023	-0.00887	0.01349	0.00481

Design 2				
ω	α	β	η_1	η_2
0.5	0.2	0.4	0.0	0.3
η_3	η_5	η_7	η_9	η_{11}
0.01213	0.01213	0.01339	0.01989	0.00128
η_4	η_6	η_9	η_{10}	η_{12}
0.19948	0.01083	-0.01320	0.01077	0.00233

Table 1: Parameter settings of the simulation exercise.

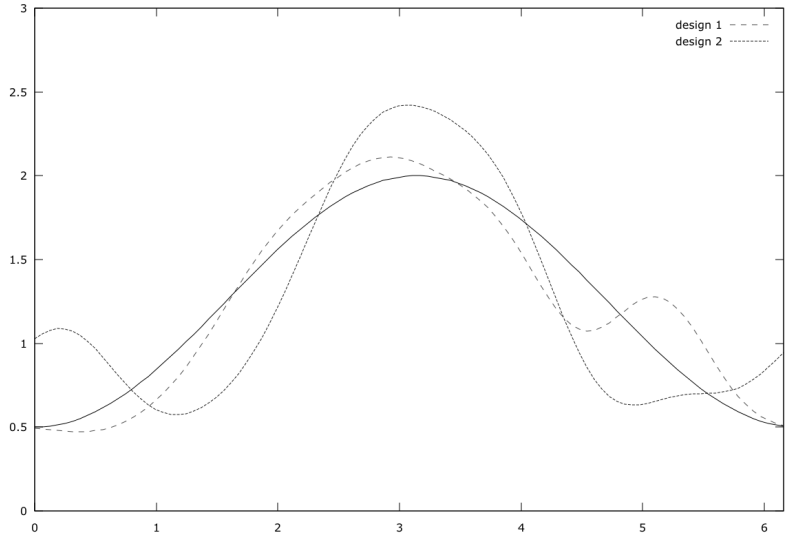


Figure 3: Mean of y_t under Design 1 (dashed line) and Design 2 (dotted line) and under the null model (continuous line).

Design 1

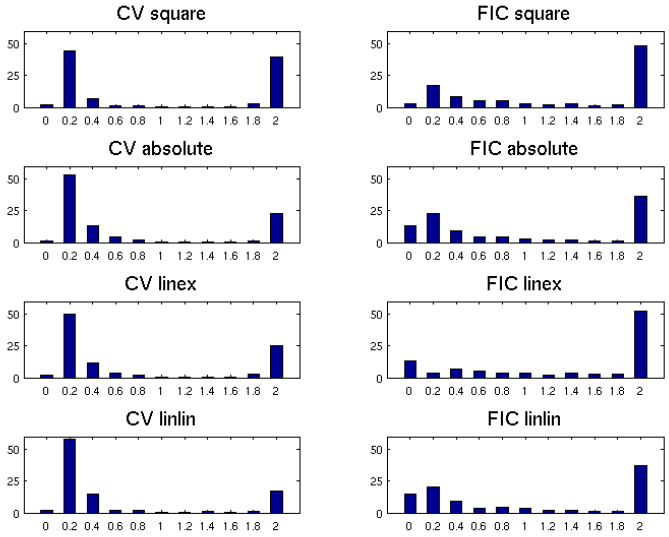
Strategy	Square	Absolute	Linex	Linlin	Norm Red.	Pers.
MLE	1.5657	0.8555	1.5055	0.8642		0.60
Percentage Gains						
CV_s	1.80***	0.72**	2.37***	0.66***	84.9	0.61
CV_a	1.83***	0.79***	2.49***	0.79***	83.9	0.61
CV_{le}	1.54***	0.71**	2.43***	0.70***	84.1	0.62
CV_{ll}	1.50***	0.77***	2.46***	0.76***	82.5	0.62
FIC_s	1.88***	0.81***	2.12***	0.79***	89.5	0.61
FIC_a	1.41***	0.60***	1.29***	0.62***	77.3	0.61
FIC_{le}	2.04***	1.59***	3.82***	2.16***	88.9	0.60
FIC_{ll}	1.32***	1.24***	2.50***	1.75***	73.5	0.61

Design 2

Strategy	Square	Absolute	Linex	Linlin	Norm Red.	Pers.
MLE	1.6580	0.8824	1.5680	0.8855		0.60
Percentage Gains						
CV_s	0.74	0.29	1.40	0.37	81.1%	0.62
CV_a	0.66	0.34	1.61	0.43	80.8%	0.62
CV_{le}	0.69	0.25	1.33	0.30	82.0%	0.62
CV_{ll}	0.57	0.32	1.57	0.42	79.9%	0.62
FIC_s	0.52	0.07	0.45	0.02	82.2%	0.62
FIC_a	0.18	0.07	0.08	0.06	65.8%	0.61
FIC_{le}	0.76	0.70	2.33	1.06	84.1%	0.61
FIC_{ll}	0.25	0.66**	1.17**	1.24**	62.4%	0.61

Table 2: Simulation results of each shrinkage selection strategy under different losses. Diebold Mariano Equal Predictive Ability test statistic is computed under the null hypothesis that the shrinkage forecasts have the same performance than the MLE forecasts. "Norm Red." refers to the average percentage norm reduction of the shrunk γ–parameters from the MLE. "Pers." refers to average estimated persistence $(\alpha + \beta)$.

Design 1



Design 2

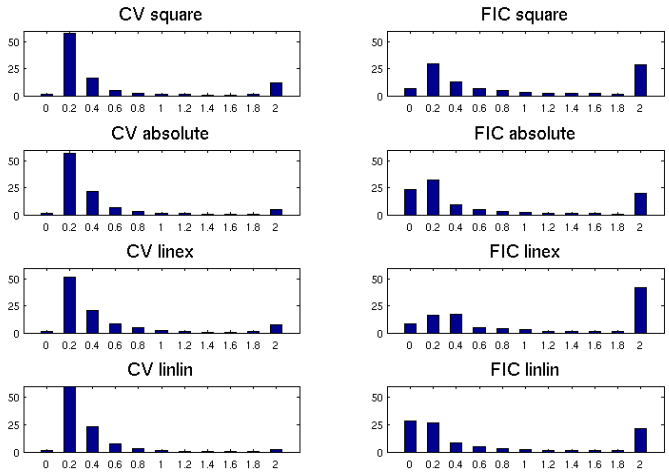
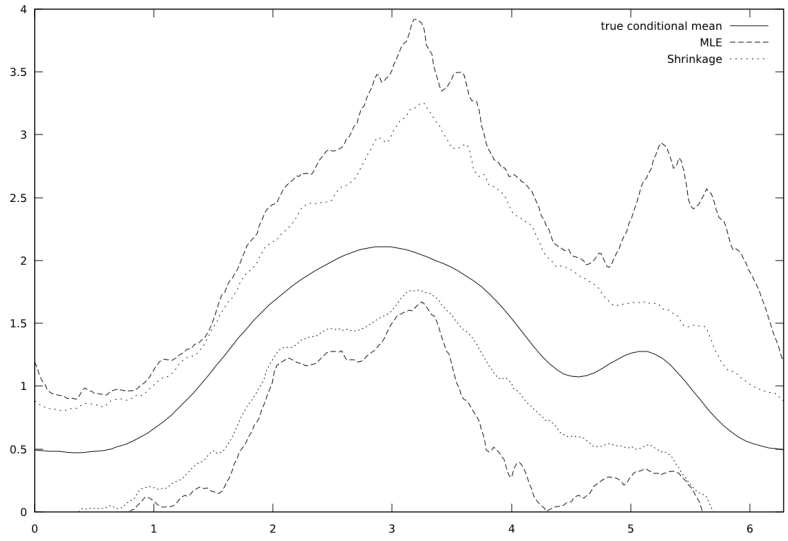


Figure 4: Empirical distribution of the optimal shrinkage parameter λ according to various selection strategies. Simulation Designs 1 and 2.

Design 1



Design 2

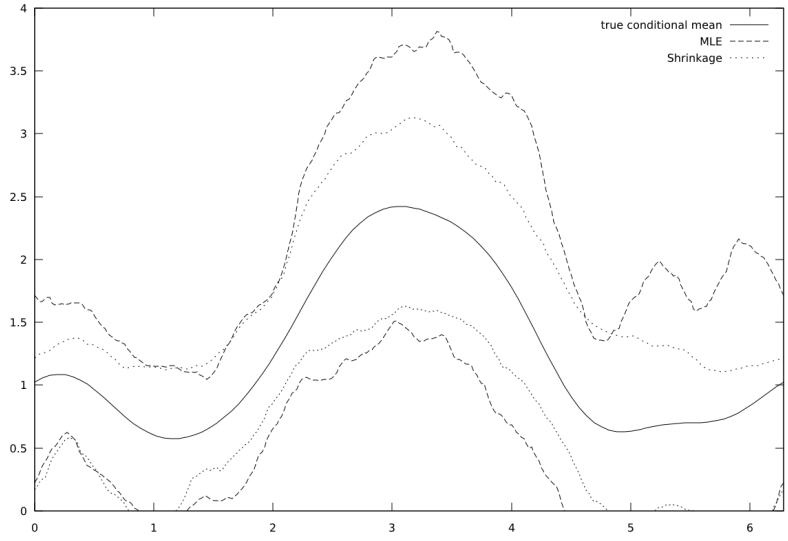


Figure 5: Maximum likelihood and shrinkage estimators of the mean of y_t. The lines below (above) the mean (solid) represent the 5% (95%) quantile of the MLE (dashed) and shrinkage (dotted) estimators.

Table 2 reports the Monte Carlo average prediction losses obtained with the unrestricted MLE as well as the percentage gains associated with each shrinkage selection strategy. Next to each value, we report one, two or three "*"'s according to the significance of a corresponding (10% 5% or 1% respectively) Equal Predictive Ability test statistic (Diebold & Mariano (1995), sign–test). For each shrinkage selection scheme, Table 2 also reports the average norm reduction of the γ–parameter and the average persistence (measured as the sum of the estimated α and β coefficients). Under Design 1, both Cross Validatory and FIC methods produce predictions which significantly outperform the MLE predictions. These averages are computed using the sequence of estimates which are selected at each step by the various shrinkage selection schemes. Furthermore, FIC methods achieve a better average performance than Cross Validatory methods, and the FIC based on a linex loss beats all other selection strategies. Under Design 2, Cross Validatory and FIC methods still produce forecasts with smaller average losses in comparison to the MLE predictions, but the evidence of a significant improvement is less strong, FIC_{ll} being an exception. Again, FIC methods achieve a better overall performance and asymmetric FIC losses seem to achieve the best results. For the cross validation methods, the choice of the loss function in the selection step seems to play a little role. On the other hand, for FIC methods the choice of the loss function does seem to have an impact on the forecasting performance. The criteria based on a linear loss tend to shrink the estimates more than others. Moreover, the asymmetric penalties tend to perform better than their symmetric analogs. A few words about the spike which appears in correspondence with the highest value in the grid of values of λ. It should be interpreted as evidence of the narrow model being chosen in such cases. This is not surprising under Design 1 since it corresponds to a choice of the parameter values very close to the null model. Under Design 2 the choice of a narrow model seems to more frequent with FIC than with CV. The gains from extending the λ grid to values greater than 2 seem to be outnumbered by the computational burden. Furthermore, values of λ deliver parameter estimates very close to zero and further refinements do not appear to be useful.

Figure 4 displays the empirical distribution of the optimal shrinkage parameter λ in according to various selection strategies. On average, the amount of shrinkage selected by the selection criteria is greater under Design 1 (closer to the narrow model) than 2. In both settings FIC methods seem to shrink much more than the cross validatory methods, the latter indicating a preference for a moderate amount of shrinkage ($\lambda = 0.2$. As far as cross validation is concerned, the use of different loss functions does not seem to dramatically change the behavior of the chosen shrinkage levels. On the other hand, the FIC methods seem to select very different shrinkage levels depending on the loss function: more specifically the absolute and linlin losses shrink less than the square and linex losses.

Figure 5 provides graphical evidence of the differences between shrinkage and maximum likelihood estimation. For both designs, the figure displays the plots of the mean of y_t conditional on the value of explanatory variable fixed at

Strategy	Square	Absolute	Linex	Linlin	Norm Red.	Pers.
MLE	1.1185	0.7389	1.1698	0.7546		0.65
			Percentage Gains			
CV_s	2.72	0.26	3.79	0.24**	82.4%	0.79
CV_a	2.76	0.08	3.80	-0.05	86.1%	0.78
CV_{le}	3.22	0.39	4.15	0.27**	82.9%	0.79
CV_{ll}	2.85	0.24	3.86	0.13**	84.5%	0.80
FIC_s	3.45*	0.42**	4.06	0.34***	84.2%	0.80
FIC_a	3.07	0.39	2.33	0.32**	68.7%	0.77
FIC_{le}	5.07*	1.46	6.28*	1.93**	85.6%	0.80
FIC_{ll}	2.65	1.07*	3.69**	1.55***	67.4%	0.77
2SMLE	2.76	-0.55	0.05	-0.48		

Table 3: Average prediction losses of the 1-step ahead forecasts with the CPA test significance (*: 10%; **: 5%; ***: 1%). "Norm Red." refers to the average percentage norm reduction of the shrunk γ–parameters from the MLE. "Pers." refers to average estimated persistence ($\alpha + \beta$).

x together with the 5% and 95% quantiles of the MLE and shrinkage estimator (computed for $\lambda = 1$) of the true mean of y_t. Visual inspection of the graphs clearly shows how shrinkage is generally associated with much more precise estimates.

4.3 Empirical Forecasting Exercise

The empirical application consists of a forecasting exercise of financial durations (e.g. Engle & Russell (1998)) using the General Electric (GE) stock data from the New York Stock Exchange in April 2005. The dataset consists of 766 intra-daily durations between transaction price changes above the threshold of USD 0.05. The procedures used to clean the data and construct the series are described in Brownlees & Gallo (2006).

Figure 6 displays the plot of the price durations. The series exhibits clustering and is affected by intra-daily periodicity, i.e. very short durations at the opening and closing of the trading day and longer durations around the middle of the trading day, with a maximum around lunch time. These stylized facts suggest that the flexible MEM of Equation (7) using the time of day as the predetermined variable x_t should be able to capture the dynamics adequately.

The 1–step ahead recursive prediction exercise starts from April 21, 2005 until the end of the month and using approximately the most recent 3 weeks of data (500 observations) to construct predictions. The set of values of the shrinkage parameter λ that characterizes the shrinkage estimators is $\{0.4\ k : k = 1, ..., 10\}$. The limit of the λ parameter is set to 4 in that for this level of shrinkage estimates are virtually equal to 0.

Table 3 reports the average prediction losses of the MLE as well as the gains

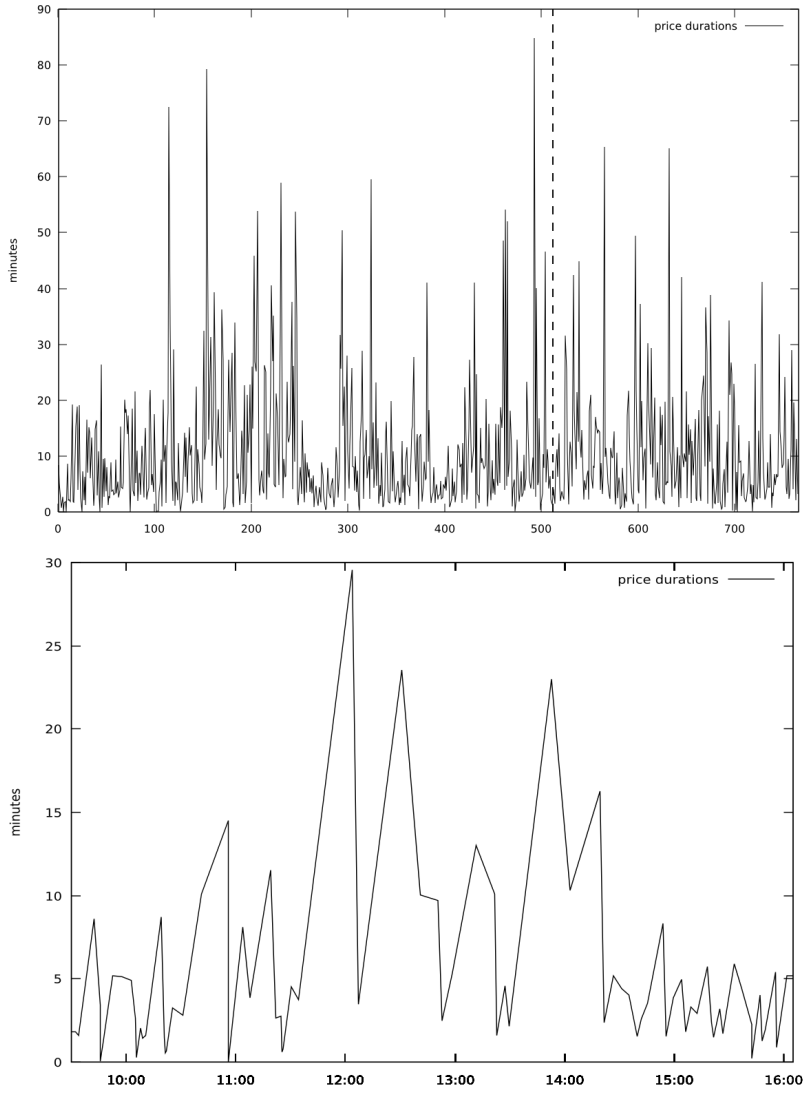


Figure 6: Price durations. GE Durations in the month of April, 2005 using a tick scale (top) GE Durations on the April 15, 2005 using a time scale (bottom).

Strategy	Opening	Mid	Closing	Strategy	Opening	Mid	Closing
CV_s	1.27	1.16	1.69	FIC_s	1.55	2.04	2.64
CV_a	1.58	1.53	1.47	FIC_a	0.97	1.47	2.12
CV_{le}	1.02	1.07	1.28	FIC_{le}	2.05	1.96	3.25
CV_{ll}	1.41	1.37	1.34	FIC_{ll}	1.24	1.37	2.00

Table 4: Average shrinkage level λ per time of day

obtainable with each shrinkage selection strategy. The "*"'s reported next to each value refer to the significance level (10%, 5% and 1% respectively) of the Conditional Predictive Ability test statistic (Giacomini & White (2006)) under the null of equal conditional predictive ability, which is more appropriate than DM in a conditional context. Table 3 also reports the average prediction gains of the 2 stage procedure estimator *a là* Engle and Russell (Engle & Russell (1998)) called 2SMLE. This estimation procedure consists of removing the multiplicative periodic component from the durations using cubic splines and then fitting a MEM(1,1) to the periodically adjusted durations. Predictions are then constructed by multiplicating the MEM forecast by the fitted periodic component. The various shrinkage selection strategies are able to improve upon the performance of the MLE predictions in all cases, and almost all strategies beat the 2 stage estimation procedure. Furthermore, all the FIC methods perform better than the cross validatory methods using the same loss in all cases but one.

The results show an interesting pattern that is worth pointing out. For almost all the cases and for both the FIC and cross validatory methods, using a loss function that penalizes proportionally bigger losses (square, linex) produces better forecasts than those (absolute, linlin) that do not. Furthermore, asymmetric loss functions perform better than their symmetric analog (linex and square, linlin and absolute). The best forecasting strategy appears to be the FIC based on linex loss. Such a method not only beats all other strategies and reference benchmarks, but judging from the CPA test it also produces significantly better forecasts than the MLE benchmark.

It is also interesting to provide some more details on the difference between the shrinkage and MLE estimates and predictions. Figure 7 displays the price durations against the time of day together with the 5% and 95% quantiles of the estimated mean of the process at each time of day using the set of rolling estimates obtained by the ML and Shrinkage estimator with $\lambda = 3$. The characterization of the intra-daily periodic patterns provided by the ML estimator appears to be quite rough while on the other hand the Shrinkage estimator for $\lambda = 3$ gives a much smoother representation. Figure 8 plots the forecasts of the MLE together with the forecasts of the shrinkage estimator for $\lambda = 3$. Again, MLE predictions appear to be rougher compared to the shrinkage forecasts.

Figure 9 displays the empirical distributions of the selected values of the shrinkage parameter λ in the forecasting exercise using the various strategies.

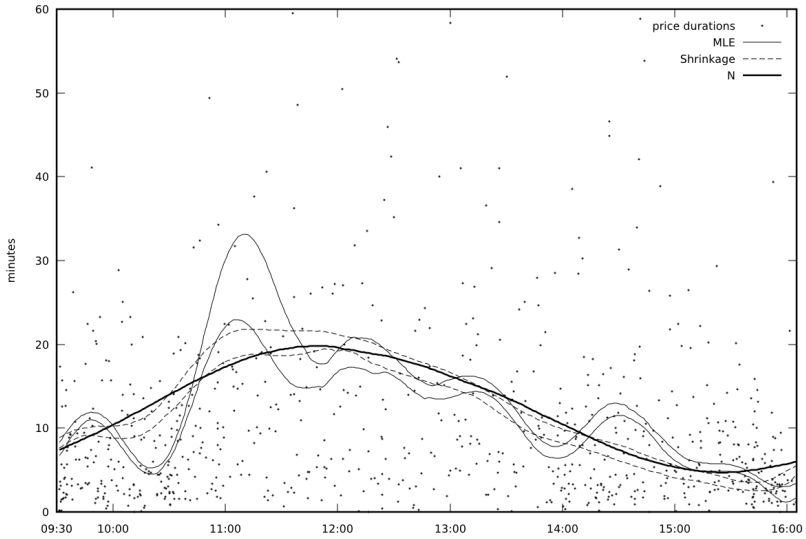


Figure 7: Price durations against time of day, 5% and 95% quantiles of the estimated unconditional mean of the process for each time of day using the set of rolling estimates obtained by the ML and shrinkage estimator with $\lambda = 3$

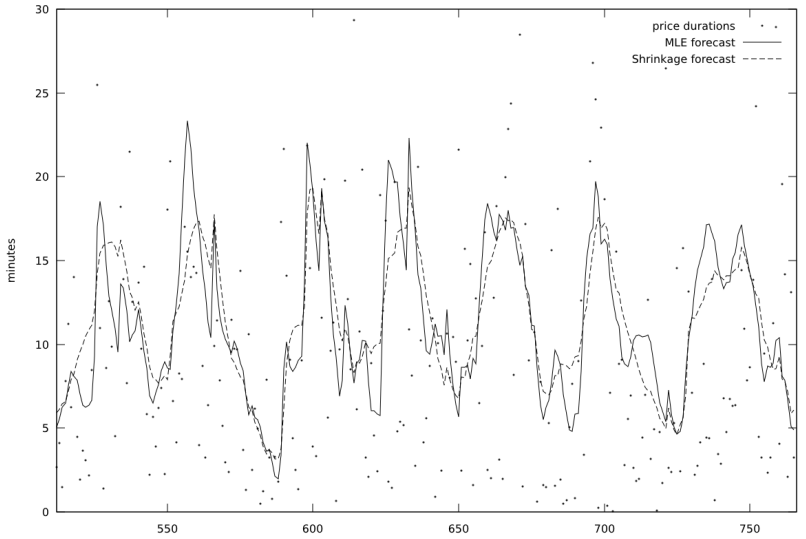


Figure 8: Price durations, MLE duration forecasts and Shrinkage duration forecasts ($\lambda = 3$) in the prediction sample.

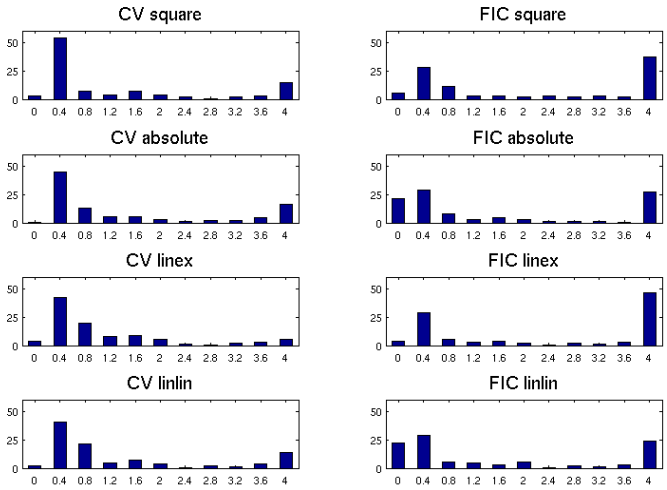


Figure 9: Empirical distribution of the optimal shrinkage parameter λ according to various selection strategies. GE data.

FIC methods have the tendency to penalize parameters more than cross validatory methods. The choice of the loss function seems to change the shape of the empirical distributions for the FIC cases and not so much for the cross validatory methods. We also compute the average level of the shrinkage parameter $\hat{\lambda}$ within (broad) time of day intervals. Table 4 reports such average levels classified by opening (9:30 to 10:45), mid–day (10:46 to 14:15) and closing (14:16 to closing). The amount of shrinkage appears to change according to the time of the day, as a consequence of the fact that the FIC selection strategy selects the most precise model depending on the time of day. On the other hand, the amount of shrinkage chosen by cross-validatory methods does not appear to change as much. Symmetric penalties tend to penalize progressively more across the day. Asymmetric penalties seem to penalize quite similarly at the opening and during the day and tend to penalize more severely at the closing.

5 Conclusions

The paper derives the local large sample distribution of a wide smooth class of shrinkage type estimators that contains Ridge–type estimators as a special case. Moreover, we extend the Focused Information Criterion family of model selection methods using asymmetric loss functions and this class of shrinkage estimators. The simulation exercise suggests that shrinkage estimation bonded with an appropriate selection strategy are able to improve upon MLE forecasts. In case

the deviation from the null model is not too severe, such forecasts outperform the maximum likelihood predictions. We favor FIC based selection strategies in view of their good performance and cheaper computational cost when compared with cross validatory schemes in this dependent and nonlinear framework. Such methods proved to be useful in improving the prediction performance in a real time forecasting exercise of financial durations where shrinkage techniques appear to perform better than the MLE. In expensively parameterised models MLE forecasts can be improved upon by using appropriate shrinkage estimation methodologies.

A Proofs

As shown by Hjort & Claeskens (2003), the local misspecification assumption allows to derive an alternative representation of the density function of the correct model as a function of the density at the null model

$$f_{\text{true}}(y) = f_0(y)(1 + s_2(y)'\delta/\sqrt{n} + R_2(y, \delta/\sqrt{n})) \tag{9}$$

where $R_2(y,t)$ is a remainder term. Such representation arises starting from a Taylor expansion of the log–likelihood ratio $\log\left(f(y,\theta_0,\gamma_0+t)/f(y,\theta_0,\gamma_0)\right)$ with respect to t. A set of regularity conditions is imposed on Equation (9) to get the results of interest.

(C1) The two integrals $\int f_0(y)s(y)s_1(y)R_2(y,t)dy$ and $\int f_0(y)s(y)s_2(y)R_2(y,t)dy$ are both $o(\|t\|)$.

(C2) The variance $|s_{1,i}^2 s_{2,j}|$ and $|s_{2,i}^2 s_{2,j}|$ have finite mean under f_0 for each i, j.

(C3) The two integrals $\int f_0(y) \parallel s_1(y) \parallel^2 R_2(y,t)dy$ and $\int f_0(y) \parallel s_2(y) \parallel^2 R_2(y,t)dy$ are both $o(1)$

(C4) The log–density has three continuous derivatives with respect to all the $p+q$ parameters in a neighbourhood around $(\theta_0', \gamma_0')'$ and there are dominated by function with finite means under f_0.

Proof of Lemma 1.
See Hjort & Claeskens (2003).

□

Proof of Lemma 2.

The proof is essentially the same as the proof of Lemma 3.2 of Hjort & Claeskens (2003) with a minor modification due to the presence of a shrinkage factor on the γ–parameters. As customary, the constrained maximization problem is reformulated as an unconstrained minimization problem. The solution of

$$\arg \max_{\Theta \times \Gamma_m} L_n(\theta, \gamma),$$

corresponds to

$$\arg \min_{\Theta \times \Gamma_S} Q_n(\theta, \gamma_S),$$

with

$$Q_n(\theta, \gamma_S) = -L_{n,S}(\theta, \gamma_S) + \lambda \rho_S(\gamma_S - \gamma_{0,S}),$$

where for given $0 \leq \lambda \leq +\infty$ there corresponds a $c \geq 0$. Note that

$$\nabla_0 Q_n = -\begin{pmatrix} \nabla_0 L_{n,1} \\ \pi_S \nabla_0 L_{n,2} \end{pmatrix}$$

and that

$$\nabla_0^2 Q_n = \begin{pmatrix} -\nabla_0^2 L_{n,11} & -\nabla_0^2 L_{n,12}\pi_S' \\ -\pi_S \nabla_0^2 L_{n,21} & -\pi_S \nabla_0^2 L_{n,22}\pi_S' + \lambda \pi_S \nabla^2 \rho(\mathbf{0})\pi_S' \end{pmatrix}.$$

The conclusion of Lemma 1 ensures that

$$\begin{pmatrix} \sqrt{n}\nabla Q_{0,n,1} \\ \sqrt{n}\nabla Q_{0,n,2} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} B_{0,12}\ \delta_0 \\ \pi_S B_{0,22}\ \delta_0 \end{pmatrix} + \begin{pmatrix} M \\ N_S \end{pmatrix} \quad \begin{pmatrix} M \\ N_S \end{pmatrix} \sim N_{p+|S|}(0, B_{0,S}),$$

and the local misspecification assumption together with the regularity conditions ensure that

$$\nabla_0^2 Q_n - \begin{pmatrix} B_{0,11} & \pi_S B_{0,12} \\ B_{0,21}\pi_S' & \pi_S B_{0,22}\pi_S' + \lambda \pi_S \nabla^2 \rho(\mathbf{0})\pi_S' \end{pmatrix} \xrightarrow{p} \mathbf{0}.$$

Therefore, the claim of the lemma follows using standard mean value theorem type of expansions for the proof of the asymptotic normality of the maximum likelihood estimator.

□

Proof of Corollary 1.

The conclusion of Lemma 2 is

$$\begin{pmatrix} C_m \\ D_m \end{pmatrix} = \begin{pmatrix} B_{0,11} & B_{0,12}\pi_S' \\ \pi_S B_{0,21} & \pi_S(B_{0,22} + \lambda \nabla^2 \rho(\mathbf{0}))\pi_S' \end{pmatrix}^{-1} \begin{pmatrix} B_{0,12}\delta_0 + M \\ \pi_S B_{0,22}\delta_0 + \pi_S N \end{pmatrix}, \quad (10)$$

where $(M', N')' \sim N_{p+q}(0, B_0)$. The first task is to find a simpler expression for the inverse matrix of Equation (10). Applying the matrix inversion formula for partitioned matrices, we get

$$\begin{pmatrix} B_{0,11}^{-1} + B_{0,11}^{-1} B_{0,12}\pi_S' T_S^{-1} \pi_S B_{0,21} B_{0,11}^{-1} & -B_{0,11}^{-1} B_{0,12}\pi_S' T_S^{-1} \\ -T_S^{-1}\pi_S B_{0,21} B_{0,11}^{-1} & T_S^{-1} \end{pmatrix} \quad (11)$$

where

$$T_S \equiv (\ \pi_S(B_{0,22} + \lambda \nabla^2 \rho(\mathbf{0}) - B_{0,21}B_{0,11}^{-1}B_{0,12})\pi_S'\ );$$

which can also be rearranged as

$$T_S = (\pi_S(B_{0,22} - B_{0,21}B_{0,11}^{-1}B_{0,12})\pi_S' + \lambda \pi_S \nabla^2 \rho(\mathbf{0})\pi_S').$$

The Sherman-Morrison-Woodbury formula allows one to express T_S^{-1} as

$$\begin{aligned} T_S^{-1} &= K_S - K_S(K_S + \lambda^{-1}(\pi_S \nabla^2 \rho(\mathbf{0})\pi_S')^{-1})^{-1} K_S \\ &= K_S - R_m K_S \end{aligned} \quad (12)$$

where $K_S = (\pi_S(B_{0,22} - B_{0,21}B_{0,11}^{-1}B_{0,12})\pi_S')^{-1}$ and $R_m = K_S(K_S + \lambda^{-1}(\pi_S \nabla^2 \rho(\mathbf{0})\pi_S')^{-1})^{-1}$.

Combining together the results of Equations (11) and (12) we decompose Equation (10) as

$$\begin{pmatrix} C_m \\ D_m \end{pmatrix} = \begin{pmatrix} C'_m \\ D'_m \end{pmatrix} - \begin{pmatrix} C''_m \\ D''_m \end{pmatrix}$$

where

$$\begin{pmatrix} C'_m \\ D'_m \end{pmatrix} = \begin{pmatrix} B_{0,11}^{-1} + B_{0,11}^{-1} B_{0,12} \pi'_S K_S \pi_S B_{0,21} B_{0,11}^{-1} & -B_{0,11}^{-1} B_{0,12} \pi'_S K_S \\ -K_S \pi_S B_{0,21} B_{0,11}^{-1} & K_S \end{pmatrix} \begin{pmatrix} B_{0,12}\delta_0 + M \\ \pi_S B_{0,22} \delta_0 + \pi_S N \end{pmatrix},$$

and

$$\begin{pmatrix} C''_m \\ D''_m \end{pmatrix} = \begin{pmatrix} B_{0,11}^{-1} B_{0,12} \pi'_S R_S K_S \pi_S B_{0,21} B_{0,11}^{-1} & -B_{0,11}^{-1} B_{0,12} \pi'_S R_S K_S \\ -R_S K_S \pi_S B_{0,21} B_{0,11}^{-1} & R_m K_S \end{pmatrix} \begin{pmatrix} B_{0,12}\delta_0 + M \\ \pi_S B_{0,22} \delta_0 + \pi_S N \end{pmatrix}.$$

We have decomposed C_m and D_m into the sum of two random quantities such that the first component only depends on the submodel constraint, while the second component depends on both the submodel and the shrinkage constraint.

We now go through a bit of algebra to obtain some nicer expressions for C'_m, D'_m, C''_m and D''_m. We begin by providing the simplified expression for C'_m following the steps outlined in Hjort & Claeskens (2003). Recall that $H_S = K^{-1/2} \pi'_S K_S \pi_S K^{-1/2}$.

$$\begin{aligned}
C'_m &= B_{0,S}^{11}(B_{0,12}\delta_0 + M) + B_{0,S}^{12}(\pi_S B_{0,22}\delta_0 + \pi_S N) \\
&= B_{0,11}^{-1} M + B_{0,11}^{-1} B_{0,12}\delta_0 \\
&\quad + B_{0,11}^{-1} B_{0,12} \pi'_S K_S \pi_S B_{0,21} B_{11}^{-1}(B_{0,12}\delta_0 + M) \\
&\quad - B_{0,11}^{-1} B_{0,12} \pi'_S K_S \pi_S (B_{0,22}\delta_0 + N) \\
&= B_{0,11}^{-1} M - B_{0,11}^{-1} B_{0,12} \pi'_S K_S \pi_S (N - B_{0,21} B_{0,11}^{-1} M) \\
&\quad + B_{0,11}^{-1} B_{0,12}(I + \pi'_S K_S \pi_S B_{0,21} B_{0,11}^{-1} B_{0,21} - \pi'_S K_S \pi_S B_{0,22})\delta_0 \\
&= B_{0,11}^{-1} M - B_{0,11}^{-1} B_{0,12} \pi'_S K_S \pi_S K^{-1} W \\
&\quad + B_{0,11}^{-1} B_{0,12}(I - \pi'_S K_S \pi_S K^{-1})\delta_0 \\
&= B_{0,11}^{-1} M - B_{0,11}^{-1} B_{0,12} K_S^{1/2} H_S K^{-1/2} W \\
&\quad + B_{0,11}^{-1} B_{0,12}(I - K_S^{1/2} H_S K^{-1/2})\delta_0 \\
C'_m &= B_{0,11}^{-1} M + B_{0,11}^{-1} B_{0,12}(\delta_0 - K_S^{1/2} H_S K^{-1/2} D)
\end{aligned}$$

Using similar steps we also obtain a nicer expression for C''_m. Recall that $G_m = K^{-1/2}\pi'_S R_m K_S \pi_S K^{-1/2}$.

$$\begin{aligned}
C''_m &= B^{-1}_{0,11}B_{0,12}\pi'_S R_m K_S \pi_S B_{0,21}B^{-1}_{0,11}(B_{0,21}\delta_0 + M) \\
&\quad -B^{-1}_{0,11}B_{0,12}\pi'_S R_m K_S \pi_S (B_{0,22}\delta_0 + N) \\
&= -B^{-1}_{0,11}B_{0,12}\pi'_S R_m K_S \pi_S (N - B_{0,21}B^{1}_{0,11}M) \\
&\quad -B^{-1}_{0,11}B_{0,12}\pi'_S R_m K_S \pi_S (B_{0,22} - B_{21}B^{-1}_{0,11}B_{0,12})\delta_0 \\
&= -B^{-1}_{0,11}B_{0,12}\pi'_S R_m K_S \pi_S K^{-1} W \\
&\quad -B^{-1}_{0,11}B_{0,12}\pi'_S R_m K_S \pi_S K^{-1}\delta_0 \\
&= -B^{-1}_{0,11}B_{0,12}\pi'_S R_m K_S \pi_S K^{-1} D \\
C''_m &= -B^{-1}_{0,11}B_{0,12}K^{1/2}G_m K^{-1/2} D
\end{aligned}$$

We now work on the expression for D'_m.

$$\begin{aligned}
D'_m &= B^{21}_{0,S}(B_{0,21}\delta_0 + M) + B^{22}_{0,S}(\pi_S B_{0,22}\delta_0 + \pi_S N) \\
&= (B^{21}_{0,S}B_{0,21} + B^{22}_{0,S}\pi_S B_{0,22})\delta_0 + B^{21}_{0,S}M + B^{22}_{0,S}\pi_S N \\
&= K_S\pi_S(B_{0,22} - B_{0,21}B^{-1}_{0,11}B_{0,21})\delta_0 + K_S\pi_S(N - B_{0,21}B^{-1}_{0,11}M) \\
&= K_S\pi_S K^{-1}\delta_0 + K_S\pi_S K^{-1} W \\
D'_m &= K_S\pi_S K^{-1} D
\end{aligned}$$

Lastly, we find the simplified expression for D''_m.

$$\begin{aligned}
D''_m &= R_m K_S\pi_S(B_{0,22} - B_{0,21}B^{-1}_{0,11}B_{0,21})\delta_0 + R_m K_S\pi_S(N - B_{0,21}B^{-1}_{0,11}M) \\
&= R_m K_S\pi_S K^{-1} D
\end{aligned}$$

By subtracting the two final expressions for C'_m and C''_m we get our first claim

$$C_m = B^{-1}_{0,11}M + B^{-1}_{0,11}B_{0,12}(\delta_0 - K^{1/2}(H_S - G_m)K^{-1/2}D),$$

and similarly by subtracting the final expressions for D'_m and D''_m we obtain our second claim

$$D_m = (I - R_m)K_S\pi_S K^{-1} D.$$

□

Proof of Lemma 3.
The proof is almost identical to Lemma 3.3 of Hjort & Claeskens (2003) with the only difference that bias and variance of the limiting approximation of $\sqrt{n}(\hat{g}_m - g_n)$ depends on some extra quantities that are related to the shrinkage estimation procedure. Using a delta method type of argument and the results of Lemma 2

$$\sqrt{n}(\hat{g}_m - g_n)) \xrightarrow{d} \Lambda_m = \tfrac{\partial g}{\partial \theta}' C_m + \tfrac{\partial g}{\partial \gamma_S}' D_m - \tfrac{\partial g}{\partial \gamma}' \delta_0.$$

Using the results of Corollary 1 we can find a nicer expression for Λ_m. In fact, noting that

$$\tfrac{\partial g}{\partial \gamma_S}' D_m - \tfrac{\partial g}{\partial \gamma}' \delta_0 = -\tfrac{\partial g}{\partial \gamma}' (\delta_0 - K^{1/2}(H_S - G_m)K^{-1/2}D),$$

by setting

$$\omega = B_{0,21}B_{0,11}^{-1}\tfrac{\partial g}{\partial \theta} - \tfrac{\partial g}{\partial \gamma},$$

we get

$$\Lambda_m = \tfrac{\partial g}{\partial \theta}' B_{0,11}^{-1} M + \omega'(\delta_0 - K^{1/2}(H_S - G_m)K^{-1/2}D).$$

It is now easy to derive the expressions for the mean and variance of the m estimator as well as the correlation between two generic nested models estimator m' and m''. The mean of the estimator m is

$$b_m = \mathrm{E}(\Lambda_m) = \omega'(I_{p+q} - K^{1/2}(H_S - G_m)K^{-1/2})\delta_0,$$

and its variance is

$$\begin{aligned}
\tau_m^2 &= \mathrm{Var}(\Lambda_m) \\
&= \tfrac{\partial g}{\partial \theta}' B_{0,11}^{-1} \mathrm{Var}(M)\, B_{0,11}^{-1} \tfrac{\partial g}{\partial \theta}' \\
&\quad + \omega' K^{1/2}(H_S - G_m)K^{-1/2}\mathrm{Var}(D)\, K^{-1/2}(H_S - G_m)K^{1/2}\omega \\
&= \tau_0^2 + \omega' K^{1/2}(H_S + G_m G_m - 2G_m)K^{1/2}\omega
\end{aligned}$$

where $\tau_0^2 = \tfrac{\partial g}{\partial \theta}' B_{0,11}^{-1} \tfrac{\partial g}{\partial \theta}$. Let m' and m'' be two nested models; then the covariance between the m' and m'' estimators is

$$\begin{aligned}
\tau_{m',m''} &= \mathrm{Cov}(\Lambda_{m'}, \Lambda_{m''}) \\
&= \tfrac{\partial g}{\partial \theta}' B_{0,11}^{-1} \mathrm{Var}(M)\, B_{0,11}^{-1} \tfrac{\partial g}{\partial \theta}' \\
&\quad + \omega' K^{1/2}(H_{S'} - G_{m'})K^{-1/2}\mathrm{Var}(D)\, K^{-1/2}(H_{S''} - G_{m''})K^{1/2}\omega \\
&= \tau_0^2 + \omega' K^{1/2}(H_{S'} - G_{m'})(H_{S''} - G_{m''})K^{1/2}\omega.
\end{aligned}$$

□

Proof of Corollary 2.
The asymptotic scaled square risk is

$$r_{\mathrm{sq}}(\hat{g}_{n,m}, g) = \mathrm{E}\big(\Lambda_m^2\big) = b_m^2 + \tau_m^2.$$

The asymptotic scaled linex risk is

$$\begin{aligned}
r_{\mathrm{le}}(\hat{g}_{n,m}, g) &= \mathrm{E}(a_1\,(\exp(a_2\Lambda_m) - a_2\Lambda_m - 1)) \\
&= a_1\left(\exp\left\{a_2 b_m + \frac{a_2^2\tau_m^2}{2}\right\} - a_2 b_m - 1\right).
\end{aligned}$$

The asymptotic scaled linlin risk is

$$\begin{aligned} r_{\mathrm{ll}}(\hat{g}_m, g_n) &= \mathrm{E}\big(a_1\Lambda_m \mathbf{1}_{\{\Lambda_m>0\}} - a_2\Lambda_m \mathbf{1}_{\{\Lambda_m<0\}}\big) \\ &= \frac{1}{\sqrt{2\pi}}\left[a_1\int_{-b_m/\tau_m}^{+\infty}(\tau_m z + b_m)e^{-z^2/2}\mathrm{d}z - a_2\int_{-\infty}^{-b_m/\tau_m}(\tau_m z + b_m)e^{-z^2/2}\mathrm{d}z\right] \\ &= a_1\tau_m\phi(-b_m/\tau_m) + a_1 b_m[1-\Phi(-b_m/\tau_m)] + a_2\tau_m\phi(-b_m/\tau_m) - a_2 b_m\Phi(-b_m/\tau_m) \\ &= a_1 b_m - b_m\Phi(-b_m/\tau_m)(a_1+a_2) + \tau_m\phi(-b_m/\tau_m)(a_1+a_2) \\ &= a_1 b_m + (a_1+a_2)[\tau_m\phi(-b_m/\tau_m) - b_m\Phi(-b_m/\tau_m)]. \end{aligned}$$

The asymptotic scaled absolute risk is

$$\begin{aligned} r_{\mathrm{a}}(\hat{g}_m, g_n) &= b_m - 2b_m\Phi(-b_m/\tau_m) + 2\tau_m\phi(-b_m/\tau_m) \\ &= 2\tau_m\phi(b_m/\tau_m) + 2b_m[\Phi(b_m/\tau_m) - 1/2]. \end{aligned}$$

□

References

Brownlees, C. & Gallo, G. M. (2006), 'Financial econometric analysis of ultra–high frequency: data handling concerns', *Computational Statistics and Data Analysis* **51**, 2232–2245.

Claeskens, G., Croux, C. & Van Kerckhoven, J. (2006), 'Variable selection for logistic regression using a prediction focussed information criterion', *Biometrics* **62**, 972–979.

Claeskens, G., Croux, C. & Van Kerckhoven, J. (2007), Prediction focussed model selection for autoregressive model, Technical report.

Claeskens, G. & Hjort, N. L. (2003), 'The focused information criterion', *Journal of the American Statistician Association* **98**, 900–916.

Diebold, F. X. & Mariano, R. (1995), 'Comparing predictive accuracy', *Journal of Business & Economic Statistics* **13**, 253–263.

Doornik, J. A. & Ooms, M. (2000), Multimodality and the garch likelihood.

Engle, R. F. (2002), 'New frontiers for arch models', *Journal of Applied Econometrics* **17**, 425–446.

Engle, R. F. & Rangel, J. G. (2005), The spline garch model for unconditional volatility and its global macroeconomic causes, Technical report, UCSD.

Engle, R. F. & Russell, J. R. (1998), 'Autoregressive conditional duration: A new model for irregularly spaced transaction data', *Econometrica* **66**, 987–1162.

Engle, R. & Gallo, G. (2006), 'A multiple indicator model for volatility using intra daily data', *Journal of Econometrics* **131**, 3–27.

Fokianos, K. & Tsolaki, E. (2006), Ridge estimation for inar(p) models, Technical report, Technical Report 17/2006, Department of Mathematics & Statistics, University of Cyprus.

Frank, I. E. & Friedman, J. H. (1993), 'A statistical view of some chemometrics regression tools', *Technometrics* **35**, 109–135.

Giacomini, R. & White, H. (2006), 'Tests of conditional predictive ability', *Econometrica* **74**, 1545–1578.

Granger, C. (1969), 'Prediction with a generalized cost of error function', *Operations Research* **20**, 199–207.

Hansen, B. E. (2005), 'Challenges for econometric model selection', *Econometric Theory* **21**, 60–68.

Hjort, N. L. & Claeskens, G. (2003), 'Frequentist model average estimators', *Journal of the American Statistician Association* **98**, 879–899.

Hoerl, A. E. & Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**, 55–67.

Kiefer, N. & Skoog, G. (1984), 'Local asymptotic specification analysis', *Econometrica* **52**, 873–886.

Knight, K. & Fu, W. (2000), 'Asymptotics for lasso–type estimators', *The Annals of Statistics* **28**, 1356–1378.

Lütkepohl, H. (1996), *Handbook of Matrices*, Wiley.

Nelson, D. B. & Cao, C. Q. (1992), 'Inequality constraints in the univariate garch model', *Journal of Business and Economic Statistics* **10**, 229–235.

Racine, J. (2000), 'Consistent cross–validatory model–selection for dependent data: *hv*–block cross-validation', *Journal of Econometrics* **99**, 39–61.

Rodríguez-Poo, J., Veredas, D. & Espasa, A. (2007), *Seminonparametric estimation for financial durations, Recent Developments in High Frequency Financial Econometrics*, Springer, p. forthcoming.

Sen, P. K. (1979), 'Asymptotic properties of maximum likelihood estimators based on conditional specification', *The Annals of Statistics* **7**, 1019–1033.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of The Royal Statistical Society. Series B* **58**, 267–288.

White, H. (2006), *Approximate Nonlinear Forecasting Methods, Handbook of Economic Forecasting*, Elsevier Science B.V.

Zellner, A. (1986), 'Bayesian estimation and prediction using asymmetric loss functions', *Journal of the American Statistical Association* **81**, 446–451.

Zhang, P. (1993), 'Model selection via multifold cross validation', *Annals of Statistics* **21**, 299–313.