THREE ESSAYS ON MODEL SELECTION, MODULATION ESTIMATORS AND HERD BEHAVIOR UNDER ASYMMETRIC BELIEFS

by

Ahmad R. Shahidi

MA, University of Pittsburgh, 2003

Submitted to the Graduate Faculty of

the School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Ahmad R. Shahidi

It was defended on

July 8, 2009

and approved by

Soiliou Namoro, Department of Economics

Mehmet Caner, Department of Economics (NCSU)

Irina Murtazashvili, Department of Economics

James Feigenbaum, Department of Economics

Dissertation Director: Soiliou Namoro, Department of Economics

THREE ESSAYS ON MODEL SELECTION, MODULATION ESTIMATORS AND HERD BEHAVIOR UNDER ASYMMETRIC BELIEFS

Ahmad R. Shahidi, PhD

University of Pittsburgh, 2009

This thesis is organized in three chapters. In the first two chapters, an econometric model selection procedure and a method to improve some existing estimators are proposed. In the third chapter, a theoretical microeconomic analysis of herd behavior is performed under a fairly new set of assumptions.

In chapter one, a model selection procedure based on the Penalized Empirical Likelihood (PEL) technique is developed, and guidelines are provided for the extension of the procedure to the setting of Generalized Empirical Likelihood (GEL). The procedure was initially applied to linear models and was called "Least Absolute Shrinkage and Selection Operator" (LASSO). It was subsequently extended to Generalized Method of Moments models in, and we now extend it to Empirical Likelihood (EL) models. Its main advantage over classical methods is in the combination of model selection and model estimation into a single step, while improving the post-selection properties of the resulting estimators. This procedure is easy to implement, and it remains computationally feasible even in models with a large number of parameters. A simulation study is performed to compare the newly proposed procedure to some classical methods such as AIC, BIC, and DT. The simulation results show a better

performance of the new procedure.

In chapter two, we define the modulation technique for the EL estimator modulation technique pertains to the class of methods generally known as “shrinkage methods”. Shrinkage methods are frequently used to improve the properties, in particular small-sample properties, of existing estimators. In this paper, a general theoretical analysis of modulation estimators is developed for EL models, along with a discussion of how they can be implemented in special cases.

In chapter three, a theoretical model of imitation and herd behavior is considered. It is assumed in that some participating agents have specific abilities to affect other peoples behavior. Results are provided on how “stars” or celebrity players can impact herd formation. In the particular setting of a financial market with a single traded asset, results are provided on the consequences of this celebrity effect on bubble formation in the financial market.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 MODEL SELECTION FOR MOMENT CONDITION MODELS USING THE PENALIZED EMPIRICAL LIKELIHOOD PROCEDURE

1.1 INTRODUCTION

Moment conditions are the basis for constructing estimators and making inferences in a large number of interesting economic problems. The generalized method of moments (GMM), along with new methods based on empirical likelihood theory (Owen 1988) are the major tools to construct estimators and make inferences in the framework of moment condition models. In this paper we address the problem of model selection when the available information is in the form of moment conditions This problem of model selection is a problem which practitioners face very often. We propose a method based on the penalized empirical likelihood procedure. This method, unlike other existing methods, selects and estimates the right model at the same time. As we will see in details, the proposed method is continuous in the sense that instead of including (1) or dropping (0) a particular coefficient, it shrinks the coefficients so that some of them will drop out. One problem with AIC, BIC, or more recent DT methods is that they are all discrete. They either include a parameter or drop it, this makes the procedure undesirably unstable. A small change in the data, which can be in the form of adding new information, will result in a completely different model to be selected. Another problem with the existing methods is that they are computationally

very expensive, specially when the number of parameters is very large. The proposed method addresses both of these two problems. Additionally, as we will see in the simulation results, compared to the existing methods, our method also has post-selection superiority, and it selects the right model more often, it is easier to implement, and computationally feasible in a model with a large number of parameters. It also has better variance results so that the final estimators obtained using this method are better compared to their counterparts in the RMSE (root mean squared error) sense. As a further contribution, we will show that the penalized empirical likelihood defined in this paper can be used to define other possibly useful procedures, and sometimes, enhance good properties of a given estimator. For example, we will define an estimator which is similar to EL estimator, but its implied probability measure has a larger Kullback-Leibler (KL)-entropy than the implied probability measure of EL. Furthermore, with our definition of penalized empirical likelihood, we are able to use the existing and advanced framework of the penalized maximum likelihood to investigate the asymptotic and convergence properties of the penalized empirical likelihood procedure in a general setting when a general penalty function is used.

In the remaining part of this introduction, I will elaborate on the heuristic origins of the topics which will be further analyzed in this paper.

1.1.1 GMM and GEL

The generalized method of moments estimator (GMM) has been the workhorse of econometric analysis since its introduction by Hansen (Hansen, 1982). Besides providing a unified framework to study different types of estimators, GMM extends the method of moments framework to include situations in which the number of moment conditions exceed the dimension of the parameter we want to estimate. Although

GMM is a very useful estimator and it is first-order asymptotically efficient, its small sample properties are relatively poor (Altonji and Segal, 1996; Tauchen, 1986). In addition, the two-step nature of GMM introduces a lot of arbitrariness to the estimator.

More recently, Owen's empirical likelihood method has provided other estimators, some of which overcome some of the shortfalls of GMM estimator. This family includes the EL estimator (Owen, 1988; Qin and Lawless, 1994; Imbens, 1997), Continuous Updating Estimator (CUE) (Hansen, Heaton, and Yaron, 1996), and the Exponentially Tilting Estimator (Kitamura and Stutzer, 1997; Imbens and Johnson, 1998). These estimators all belong to the class of Generalized Empirical Likelihood (GEL) estimators (Smith, 1997; Newey and Smith, 2004). [1] These estimators circumvent the need of estimating a weighting matrix in the two-step GMM by directly minimizing an information-theory-based concept of closeness between the estimated distribution and the empirical distribution.[2] While in theory these estimators, like GMM, all have the same first-order asymptotic efficiency, simulation studies, and Monte Carlo evidence have shown that, compared to GMM, some members of the GEL class have better finite-sample properties (see Hansen, Heaton, and Yaron, 1996; Ramalho, 2006 and references therein). Also, Newey and Smith (2004) have analytically shown, using a stochastic expansion argument, that while GMM and GEL share the same first-order asymptotic properties, their higher-order properties are different. Specifically, while the asymptotic bias of GMM often grows with the number of moment restrictions, the relatively smaller bias of EL does not. Moreover, a bias-corrected EL is higher-order

[1] There are other varieties, too. For example the Exponentially Tilted Empirical Likelihood estimator (ETEL) (SCHENNACH, 2007) which in essence is a combination of the two estimators, EL and ET, in hope to obtain an estimator that like EL has a smaller finite-sample bias, and at the same time inherits the better behavior of ET in the presence of mis-specification.

[2] The estimators mentioned so far are, like GMM, based on unconditional moment restrictions, using the empirical likelihood methods, we can construct estimators based on conditional moment restrictions see (ZHANG and GIJBELS, 2003), and Kitamura et al (2004)

efficient relative to any other regular method of moment estimator. In terms of inference, the empirical likelihood ratio test has some desirable features too. For example, The ELR test admits Bartlet correction (DiCiccio, Hall and Romano, 1991), which gives it the same accuracy rate as the parametric case. Kitamura (2001) has used the so called *Generalized Neyman-Pearson* approach to show that, for testing moment restrictions, the ELR test is uniformly most powerful in an asymptotic large deviation sense.

1.1.2 Model Selection

Let $\{M_\xi, \xi \in \Xi\}$ be a set of candidate models for a given observation. Based on the observed data we need to select a model from $\{M_\xi, \xi \in \Xi\}$ using an appropriate model selection criteria, or a justified procedure which selects the desired model. Model selection problems are encountered almost in every application . For instance, in linear regression analysis, it is often of interest to select the right number of nonzero parameters which have the most explanatory power. With a small model, interpretation is easier and statistical inferences can be carried out more efficiently. Also, in time series analysis, it is essential to know the true order of an ARMA. As another example, suppose we have two competing non-nested models with two different parameter vectors, and two sets of moment conditions. The two parameter vectors can be stacked together to yield a single parameter θ. Now we can select each model by setting the appropriate parts of θ to zero. A model selection method tells us what part of the parameter θ should be set to zero.

Different techniques and criteria have been developed to deal with model selection, each having its own advantage in a particular setting.[3] In the parametric likelihood-

[3]For a good survey of model selection literature see Rao, and WU (2001).

based model selection we have, alongside others, the famous AIC, and BIC criteria. When the information about the underlying density function of the data generating process is limited to moment conditions, Andrews (1999), and Andrews, and Lu (2001) provide downward testing (DT) and BIC-like criteria in the framework of GMM estimation. Also related to our work are the paper by Kolaczyk (1995), in which the author considers an analogue of AIC model selection criterion in the empirical likelihood context. Also, the paper by Houng, Preston, and Shum (2003), which extends the results of Andrews, and Lu (2001) to the setting of GEL.

As mentioned earlier, the classical methods of model selection usually involve a computationally heavy combinatorial search. Simple model selection via AIC and BIC, which can be applied to OLS, often select the wrong model (Breiman, 1996), and furthermore, these procedures are unstable, meaning small changes in the data can cause entirely different selections.[4]

To overcome these shortfalls Tibshirani (1996) introduced "Least absolute shrinkage and selection operator" (*LASSO*).[5] The lasso, which is based on the penalization technique, combines the selection and estimation steps and therefore reduces the variance of the final estimator while using less computation resources. Model selection in linear models is now being mostly carried out using lasso procedure. It is a computationally feasible alternative to the classical model selection methods. Furthermore, recent studies (Zhao, and Yu 2006) have shown that under very mild conditions, the lasso technique almost always selects the true model. In this paper we define the penalized empirical likelihood, and then use it to extend the lasso method of model selection to the framework of empirical likelihood. Although in this paper I restrict our attention to the EL estimation, I think that the extension of the proposed technique to the more

[4]For further information about model selection via AIC and BIC, and their shortfalls see (FAN and LI, 2001; FAN and LI, 2002) and the references therein.

[5]This method has also been extended to GMM setting, see Caner (2008).

general setting of GEL is possible.

Since, lasso is just one example of the numerous applications of the penalization method, it is important to perform a systematic study of the penalization method in the context of EL estimator. In this paper we use the parametric case of penalized maximum likelihood to study the nonparametric situation of penalized empirical likelihood procedure. We present asymptotic, and convergence rate results for the penalized EL with a fairly general penalty function.

The main contribution of this paper is to introduce a powerful method of model selection which can be used as an alternative to the existing procedures. As the simulation results will show, this method not only selects the right model more often, but it also has a better post-selection performances. In this paper, we also propose a general framework for defining and studying the penalized El and GEL estimators. We present results for this general case, and as an example we introduce an estimator similar to EL whose implied probabilities have a better entropy property.

The rest of this paper proceeds as follows. In section 2 we give a formal definition of penalized empirical likelihood estimator. In section 3 we study the problem of model selection via PEL. Section 4 presents asymptotic and convergence results for PEL with a general penalty function, in this section as an example of a general penalty function we introduce another potentially important estimator. Section 5 concludes the paper. All the proofs are collected in the appendix.

1.2 DEFINITION OF PEL

Let θ be the parameter we are interested to estimate. In general, when l_n is a functional which measures how well θ predicts the observed data set, $X_1, \ldots, X_n$, and $J(\theta)$

is a penalty functional which assesses the physical plausibility of θ, the method of penalization chooses a θ which optimizes

$$\ell_{n\lambda}(\theta) = l_n(\theta|data) - \lambda J(\theta), \quad \lambda > 0 \tag{1.2.1}$$

λ is called the regularization, or sometimes penalization parameter. Larger values of λ produces more regular estimators.

The maximum empirical likelihood procedure, much like maximum likelihood method, is based on maximizing a criterion functional over a parameter space. Therefore the method of penalization, should has a natural application in empirical likelihood estimation. Very often, specially when the parameter space is large or not well behaved, the optimization becomes difficult and the resulting estimators may have undesirable properties such as non-smoothness, inconsistency and so on. In some of these situations the maximization can be carried out based on the penalized version of the criterion function. In this subsection we formally introduce this idea and later in this paper, we present some of its most important applications, and investigate the properties of these procedures.

Definition 1:

(a) Let $X_1, \ldots, X_n$ be independently distributed random variables, with a common distribution (i.i.d). Let $l(\theta, X_i)$ be the criterion function evaluated at X_i, if $J(\theta)$ is the penalty function we define the ***penalized criterion function*** *to be*

$$\ell(\theta, X_i) = l(\theta, X_i) - \lambda_n J(\theta). \tag{1.2.2}$$

(b) Let $L_n(\theta) = L_n(\theta, X) = n^{-1}\sum_{i=1}^n \ell(\theta, X_i)$, and $l_n(\theta) = l_n(\theta, X) = n^{-1}\sum_{i=1}^n l(\theta, X_i)$. Maximizing $L_n(\theta)$ will produce an estimator for θ. We define an ***approximate maximizer*** *of $L_n(\theta)$ to be a $\hat{\theta}_n$ such that*

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} L_n(\theta) - \varepsilon_n, \tag{1.2.3}$$

where $\varepsilon_n \to 0$ *as* $n \to \infty$.

Now we can easily adapt definition 1 to obtain a definition for the *penalized empirical likelihood estimator*. Let

$$l_n(\theta) = -\max_{\gamma \in R^m} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \gamma' g(X_i, \theta)) \tag{1.2.4}$$

be the profile empirical likelihood function for θ. We define the *penalized empirical likelihood* as follows.

Definition 2:

The penalized empirical likelihood estimator for θ *is*

$$\hat{\theta}_{pel} = \arg\max_{\theta \in \Theta} \{l_n(\theta) - \lambda_n J(\theta)\}. \tag{1.2.5}$$

Notice that, if γ^* denotes the maximizer in (2.4), then the $l(\theta, X_i)$ in definition 1(a) is $l(\theta, X_i) = -log(1 + \gamma'^* g(X_i, \theta))$.

As an example, suppose that we know from external knowledge, that the true parameter is somewhere close to a linear subspace of the parameter space, Θ. In this case it is appropriate to try to shrink the estimator toward this linear subspace. For instance, if L is the following linear subspace

$$L = \{\theta : \theta_1 = \theta_2 = \ldots, \theta_r\} = \left\{\theta : \frac{1}{r} J\theta = \theta\right\}, \tag{1.2.6}$$

where J is a matrix of ones, $J = 11'$, then to shrink the estimator toward L we can use the penalty function $J(\theta) = \sum_{i=1}^{r} (\hat{\theta_1} - \theta_i)^2$.

In the following section we use the penalized EL defined in this section to construct the lasso-EL, and study its properties. A general theory for convergence, and asymptotic distribution of PEL will be developed in section 4.

1.3 MODEL SELECTION USING PEL

As a major example of penalization method, we introduce the *"Least Absolute Shrinkage and Selection Operator"* (LASSO) for the empirical likelihood setting. The easiest way to understand the purpose and usefulness of these type of estimators, is to take a look at the linear case. Consider the usual regression situation: we have data (x^i, y_i), $i = 1, 2, \ldots, N$, where $x^i = (x_{i1}, \ldots, x_{ip})'$ and y_i, are regressors and response for the i^{th} observation. The ordinary least squares (OLS) estimates are obtained by minimizing the residual squared error. There are two drawbacks to the OLS procedure. The OLS estimates often have low bias but large variance, resulting in a poor prediction accuracy. As we mentioned earlier in the introduction to this paper, prediction accuracy often can be improved by shrinking, or setting some of the coefficients to zero. By doing so we scarify a little bias to reduce the variance, which may improve the overall prediction accuracy. On the other, with a large number of predictors, we often prefer to use a smaller subset that exhibits the strongest effects, in statistical literature, this procedure is called *selection.* The traditional tools to deal with these problems, are ridge regression and model selection. Model selection provides interpretable models but can be extremely variable because it is a discrete process, regressors are either retained or dropped from the model. Small changes in the data can result in very different models being selected, which is obviously very undesirable. The ridge regression, in the other hand, is a continuous process, and therefore more stable, and it does shrink the coefficients. However, it does not set any coefficient to zero and hence does not give an easily interpretable model.

Tibshirani (1996) proposes a new technique, which he calls it lasso. It shrinks some coefficients and sets others to zero, therefore retaining the good features of both model selection and ridge regression. This method can be promising, particularly when

the econometrician needs to construct a model with a large number of parameters and then use model selection methods like BIC and AIC to select the desired model.

1.3.1 Definition and Assumptions

Let θ be a p-dimensional vector, and θ_0 represent the true value, which is in the interior of the compact set $\Theta \in \mathbb{R}^p$. As before, let the moment conditions provided by theory to be

$$E[g(X_i, \theta)] = 0. \tag{1.3.1}$$

After using the empirical likelihood set up let

$$\ell_n(\theta) = -\max_{\gamma \in R^m} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \gamma' g(x_i, \theta)) \tag{1.3.2}$$

be the profile empirical likelihood for θ. The lasso-type-EL estimator for θ_0 is a $\hat{\theta}$ that maximizes

$$\ell_n(\theta) - \lambda_n \sum_{j=1}^{p} |\theta_j|^{\gamma}, \tag{1.3.3}$$

where $0 < \gamma < 1$ and λ is a regularization parameter. Other penalty functions are also possible. Indeed some are proven to be more capable to achieve certain properties, see for instance Fan and Li (2001).

1.3.1.1 Properties of Lasso-EL Estimator: In this subsection, we analyze the consistency and large sample theory for the lasso-type-EL estimators. First we state the assumptions required for the results which will follow.

Assumptions:

A1: *(i)* $\frac{\partial g(x,\theta)}{\partial \theta}$ *is continuous in a neighborhood of the true parameter* θ_0*, and the rank of* $E[\frac{\partial g(x,\theta_0)}{\partial \theta}]$ *is p.*

(ii) In a neighborhood of θ_0, $||\frac{\partial g(x,\theta)}{\partial \theta}||$ *and* $||g(x,\theta)||^3$ *are bounded by some integrable function* $G(x)$.

(iii) The matrix $E[g(x,\theta_0)g'(x,\theta_0)]$ *is positive definite.*

A2: *(i)* $g_i(\theta)$ *is m-dependent for all i.*

(ii) $|g_i(\theta_1) - g_i(\theta_2)| \leq B_i|\theta_i - \theta_2|$, *with* $\lim_{n\to\infty}\sum_{i=1}^{n} E[B_i^d] < \infty$, *for some* $d > 2$.

(iii) $\sup_{\theta\in\Theta} E[|g_i(\theta)|^d] < \infty$, *for some* $d > 2$.

A3: *Define* $E[n^{-1}\sum_{i=1}^{n} g_i(\theta)] = m_{1n}(\theta)$

(i) $m_{1n}(\theta) \to m_1(\theta)$ *uniformly over* Θ, m_{1n} *is continuously differentiable in* θ *and* $m_1(\theta_0) = 0$, $m_1(\theta) \neq 0$ *for* $\theta \neq \theta_0$. *Also* $m_1(\theta)$ *is continuous in* θ.

(ii) Let $R_n(\theta) = \frac{\partial m_{1n}(\theta)}{\partial \theta'}$ *we assume that* $R_n(\theta) \xrightarrow{p} R(\theta)$, *uniformly in a neighborhood of* θ_0, $R(\theta_0)$ *is of full rank, and* $R(\theta)$ *is continuous in* θ.

A4: *Define* $W_n(\theta) = \left[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i'(\theta)\right]^{-1}$. *We assume that:* $W_n(\theta) \xrightarrow{p} W(\theta)$ *uniformly in* θ, *where* $W(\theta)$ *is a symmetric non-random positive definite matrix which is continuous for all* $\theta \in \Theta$

Assumptions $A1$ and $A2$ are the usual assumptions in the empirical likelihood literature. They guarantee that there is unique maximizer, $\hat{\theta}$, of the empirical likelihood ratio. Since we will use the empirical processes theory to prove some of the up coming results, we will be in need of assumption 3. For a good review of empirical processes and their econometrics' application consult Andrew (1986). Some of these assumptions are used by Caner (2008) to drive similar results for the GMM estimator.

The following proposition shows the consistency of the penalized estimator, under assumptions A1-A4, and some further conditions on λ_n.

Proposition 1:

If assumptions A1-A4, hold then:

I) If $\frac{\lambda_n}{n} \to \lambda_0 \geq 0$*, then*

$$\hat{\theta}_n \xrightarrow{p} \arg\min_{\theta \in \Theta} Z(\theta), \tag{1.3.4}$$

where

$$Z(\theta) = m_1(\theta)' W(\theta) m_1(\theta) + \lambda_0 \sum_{i=1}^{p} |\theta_i|^{\gamma}. \tag{1.3.5}$$

The convergence happens uniformly in θ*.*

II) If $\lambda_n = o(n)$ *then,*

$$\hat{\theta} \xrightarrow{p} \theta_0. \tag{1.3.6}$$

We notice that $Z(\theta)$ is the limiting process of $Z_n(\theta)$. And $Z_n(\theta)$ is obtained by manipulating $\ell_n(\theta)$ in definition 2.

Using (I) from proposition 1, it is clear that why we need to have $\lambda_n = o(n)$, in order to obtain the consistency of this estimator. But this rate still is too high to get any interesting result concerning the limiting distribution of $\hat{\theta}_n$. To get the $\sqrt{n}$-consistency it is required to have slower growth rate for λ_n. However, if λ_n grows too slowly then we won't get anything substantially different from the usual EL estimator. Our goal is, to get a limiting distribution for nonzero part of the parameters which is coincide with usual, non-penalized, EL estimator. And for the zero part of parameters the distribution should goes zero. To achieve this goal, we need a λ_n which grows with a right rate. The following proposition specifies the right conditions.

Proposition 2:

Suppose that $\frac{\lambda_n}{n^{\gamma/2}} \to \lambda_0 \geq 0$*, and assumptions A1-A4 satisfy then:*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{u}_n \Rightarrow \arg\min_{u \in K} V(u), \tag{1.3.7}$$

where

$$V(u) = 2u'R(\theta)'W(\theta_0)\Psi(\theta_0) + u'R(\theta_0)'W(\theta_0)R(\theta_0)u + \lambda_0 \sum_{j=1}^{p} |u_j|^{\gamma} 1_{\{\theta_{0j}=0\}}, \quad (1.3.8)$$

and K is a compact subset of $\mathbb{R}^p$, and $\Psi(\theta_0) \equiv N(0, \Omega(\theta_0))$, where $\Omega(\theta_0)$ is the variance-covariance matrix and

$$\Omega(\theta_0) = \lim_{n\to\infty} E\left[(n^{-1/2}\sum_{i=1}^{n} g_i(\theta_0))(n^{-1/2}\sum_{i=1}^{n} g_i(\theta_0))'\right]. \quad (1.3.9)$$

An interesting conclusion of proposition 2 is that, we can estimate nonzero parameters at the usual rate without introducing further asymptotic bias, while shrinking the estimates of zero parameters to 0 with positive probability. In fact when all parameters are non zero, $\theta_i \neq 0$ $i = 1, \ldots, n$, we have

$$V(u) = \arg\min_{u\in K}\{2u'R(\theta)'W(\theta_0)\Psi(\theta_0) + u'R(\theta_0)'W(\theta_0)R(\theta_0)u\}. \quad (1.3.10)$$

The solution to this minimization problem is

$$\hat{u} = -[R(\theta_0)'W(\theta_0)R(\theta_0)]^{-1}R(\theta_0)'W(\theta_0)\Psi(\theta_0). \quad (1.3.11)$$

This is the same as the limit distribution of the non-penalized EL estimator.

Now suppose that some of the parameters are indeed zero. In general when $R(\theta_0)'W(\theta_0)'R(\theta_0)$ is singular, $V(u)$ won't have a unique minimizer. If $u \in \arg\min V(u)$ and v lies in the null pace of $R(\theta_0)'W(\theta_0)'R(\theta_0)$, then for some t, $V(u) = V(u+tv)$. However, suppose that $\theta_{r+1} = \cdots = \theta_p = 0$, and the null space of $R(\theta_0)'W(\theta_0)'R(\theta_0)$ is spanned by the standard basis vectors $e_{r+1}, \ldots, e_p$; then we have

$$V(u) = V_0(u_1, \ldots, u_r) + \lambda_0 \sum_{j=r+1}^{p} |u_j|^{\gamma}, \quad (1.3.12)$$

which has a unique minimizer. In the other words, a larger specification of the model won't prevent us to estimate the non-zero part of the model and the redundant part

will be set to zero. Therefore we can, at the same time, estimate and select the correct model. If λ_n grows faster than specified by preposition 2, but not too fast, in such way that we have $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, and $\lambda_n/n^{\gamma/2} \to \infty$, we can prove an even more interesting result, at least asymptotically, which is usually called *oracle property*. To see this assume $\lambda_n/n^{\alpha/2} \to \lambda_0 \geq 0$ with $\gamma < \alpha < 1$. Suppose that $\theta_1, \ldots, \theta_r$ are nonzero while $\theta_{r+1}, \ldots, \theta_p$ are zero, and defining $V_n(u)$ as in the proof of proposition 4, it follows that $V_n(u) \xrightarrow{d} V(u)$ where

$$V(u) = \begin{cases} 2u'R(\theta)'W(\theta_0)\Psi(\theta_0) + u'R(\theta_0)'W(\theta_0)R(\theta_0)u, & \text{if } u_{r+1} = \cdots = u_p = 0, \\ \infty, & \text{otherwise.} \end{cases} \tag{1.3.13}$$

Applying the arguments given in the proof of proposition 4, it follows that

$$\sqrt(n)(\hat{\theta}_n - \theta) \xrightarrow{d} \arg\min(V), \tag{1.3.14}$$

where the last $(p-r)$ elements of $\arg\min(V)$ are exactly 0.[6] We can summarize the argument delivered above as an corollary, which usually is referred to as *oracle property* of the lasso estimator.

Corollary 1:

Suppose $\gamma < \alpha < 1$. *If* $\lambda_n/n^{\alpha/2} \to \lambda_0 \geq 0$, *and assumptions A1-A4 hold, we have*

$$\hat{u}_n \xrightarrow{d} \begin{pmatrix} \hat{u}_1 \\ 0_{p-r} \end{pmatrix}, \tag{1.3.15}$$

where

$$\hat{u}_1 \sim N\Big(0, \big(R(\theta^{r0})'W(\theta_0)R(\theta^{r0})\big)^{-1}\Big), \tag{1.3.16}$$

with $\theta_0 = (\theta^{r0'}, 0'_{p-r})'$. *Note that* θ_0 *is separated into nonzero and zero components.*

[6] Since V can be infinite, we can no longer define convergence of V_n to V via uniform convergence on a compact set, but instead we can define it via epiconvergence which allows for extended real-valued functions. See sections 3 and 4 of Geyer (1994) for more details

There are a host of other penalty functions available, some of which might be more appropriate for special circumstances, Fan and Li (2001), review some of these functions.

1.3.1.2 Monte Carlo Simulations The Monte Carlo simulations in this section, are aimed at providing an answer to two important questions that a practitioner faces. when doing applied work. First, in average which model selection method does the best job in selecting the right model. Second, what is the post selection performance of these methods. In this section, we compare our proposed LASSO-EL estimator with BIC, *"Downward Testing"* (DT) of Andrews and Lu (2001), and LASSO-GMM of Caner (2008). The simulation design is exactly the one in Caner (2008). I therefore refer the interested reader to that paper for a detailed description of the design. Here we review those aspects of the designs, which are essential for a reader to understand the simulation process, and the proceeding results.

We have the following data generating process.

$$y = \tilde{Y}\theta + \varepsilon, \tag{1.3.17}$$

$$\tilde{Y} = Z\Pi + V, \tag{1.3.18}$$

where Z is $N \times 6$, $\Pi : 6 \times 5$, $V : N \times 5$, $\tilde{Y} : N \times 5$ represent the endogenous regressors, and $\theta : 5 \times 1$. We set $N = 100$. The instruments $Z_i : 6 \times 1$ are i.i.d and we generate them according to $N(0, I_6)$. $u_i = (\varepsilon_i, V_i)$ is independent from Z_i, with ε_i a scalar and V_i

is 5×1 vector. We choose $u_i \sim N(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} 2 & 0.99 & 0.90 & 0.80 & 0.70 & 0.6 \\ 0.99 & 2 & 0 & 0 & 0 & 0 \\ 0.90 & 0 & 2 & 0 & 0 & 0 \\ 0.80 & 0 & 0 & 2 & 0 & 0 \\ 0.70 & 0 & 0 & 0 & 2 & 0 \\ 0.60 & 0 & 0 & 0 & 0 & 2 \end{pmatrix} \tag{1.3.19}$$

and

$$\Pi = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 3 & 0 \end{pmatrix} \tag{1.3.20}$$

and u_i, s are generated i.i.d.

In this experiment, we take the instruments as given and will try to select and estimate the right structural equation. Hence, all we seek is to select and estimate the true θ_0. There are two setups, in the first one $\theta_0 = (0.8 \ \ 0 \ \ 0.7 \ \ 0 \ \ 0.9)'$. The second one has the same effects as the first one with different magnitude, $\theta_0 = (2 \ \ 0 \ \ 1 \ \ 0 \ \ 0.5)'$. We compare the ability of each method to select the true model, and the small sample properties of the post-selection estimators.

For LASSO-EL, and LASSO-GMM we set $\gamma = 1/2$, $\alpha = 2/3$, and $\lambda_N = N^{1/3}\sqrt{2 \log p}$. This choice of λ_N has been suggested by Donoho, and Johnstone (1994) and has been further discussed in Fan, and Li (2001). For an in-depths discussion of how BIC, and DT methods work see Andrews and Lu (2001). Also, as mentioned before we use the

same design as Caner (2008), and therefore for a full discussion of how we calculate different properties of the estimators, used in this simulation study, consult that paper. Here we report the results and drive some conclusions based on these results.

Table 1: Bias, Standard Error (SE), and RMSE of Design 1

	LASSO-EL			LASSO-GMM			BIC			DT		
θ_i	SE	Bias	RMSE	SE	Bias	RMSE	SE	Bias	RMSE	SE	Bias	RMSE
θ_1	0.1290	-0.0322	0.13291	0.1286	-0.0638	0.1435	0.2059	0.0026	0.2059	0.2144	0.0036	0.2144
θ_2	0.0874	0.0009	0.0870	0.0860	-0.0017	0.0860	0.0029	-0.0003	0.0029	0.0035	-0.0002	0.0035
θ_3	0.1378	-0.0291	0.1379	0.1343	-0.0538	0.1446	0.1434	-0.1067	0.1787	0.1701	0.1113	0.2033
θ_4	0.0758	0.0003	0.0758	0.0758	0.0008	0.0758	0.0019	-0.0001	0.0019	0.0024	0.0003	0.0024
θ_5	0.1533	-0.0376	0.1578	0.1520	-0.0718	0.1676	0.1701	-0.0716	0.2000	0.0631	0.2613	0.2688

Table 2: Bias, Standard Error (SE), and RMSE of Design 2

	LASSO-EL			LASSO-GMM			BIC			DT		
θ_i	SE	Bias	RMSE	SE	Bias	RMSE	SE	Bias	RMSE	SE	Bias	RMSE
θ_1	0.1752	-0.0791	0.1922	0.1731	-0.1622	0.2372	0.2073	0.0009	0.2073	0.2141	0.0011	0.2141
θ_2	0.0991	-0.0001	0.0991	0.0986	-0.0002	0.0986	0.0032	0.0001	0.0032	0.0037	0.0007	0.0037
θ_3	0.1580	-0.0385	0.1626	0.1573	-0.0813	0.1770	0.1804	-0.0597	0.1900	0.1711	-0.1016	0.1989
θ_4	0.0991	-0.0003	0.0991	0.0984	-0.0006	0.0984	0.0021	0.0001	0.0021	0.0024	-0.0009	0.0025
θ_5	0.1383	-0.0273	0.1409	0.1368	-0.0552	0.1475	0.0923	-0.1722	0.1953	0.0641	-0.2595	0.2673

We summarize the findings as follows: LASSO-EL picks the right model as often as LASSO-GMM, which is very superior in choosing the right model compared to BIC, and DT methods. While in terms of choosing the right models, LASSO-EL, and LASSO-GMM have almost the same power, LASSO-EL almost always yields a smaller RMES.

Remark: In this experiment our goal was to compare the lasso-El with lasso-GMM of Caner (2008). Because GMM perform the best when the errors are distributed according to the normal distribution, this setting is favorable to GMM. I expect that

lasso-EL will perform even better, compared to lasso-GMM, if we consider a bad behaved distribution. It is well known that GMM has very poor bias, and variance when the underlying distribution is a *bad behaved* distribution. For instance when a thicker-tailed or long-tailed skewed distribution (t_5 and log-normal are two examples) are used, EL does a much better job in comparison to GMM (see Ramalho 2005).

Table 3: Percentage of Correct Model

Estimators	Design 1	Design 2
LASSO-EL	85.24	75.15
LASSO-GMM	84.39	74.83
BIC	67.33	45.88
DT	29.08	28.70

1.4 PEL WITH A GENERAL PENALTY FUNCTION

In this section, I investigate the large sample theory of the penalized empirical likelihood estimator with a fairly general penalty function. We will use the framework developed by Cox and O'Sullivan (1990), and Shen and Wang (Shen 1994; 1997; and Wang and Shen, 1995) to derive the asymptotic distribution of the PEL estimator, and establish some exponential bounds on the convergence rate of $\hat{\theta}_n$, when it is converging toward θ_0. We see that, there are two forces in play. The size of local parameter space, and the degree of penalization. To get a reasonable convergence rate, which

also guarantee the asymptotic normality, we need to increase the degree of penalization, λ_n, when the size of the local parameter space grows large. For an in-depth study of penalization method in statistics consult the references above.

1.4.1 Asymptotic Normality

Like most cases of asymptotic analysis, we try to obtain a linearized version of the penalized criterion function, L_n. Informally let

$$S_n(\theta) = \frac{\partial L_n(\theta)}{\partial \theta} \tag{1.4.1}$$

we want to expand S_n around the true parameter θ_0 and then study its behavior when $n \to \infty$. Off course we hope the limiting score function, $S(\theta)$ exists and we have

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} - \lambda \frac{\partial J(\theta)}{\partial \theta} \tag{1.4.2}$$

where $l(\theta)$ is the limiting version of $l_n(\theta)$. To formally develop an asymptotic theory for PEL, we accept the framework of Shen 1997, and use the empirical process theory to find the limiting distribution of our estimator. Before we be able to do all of that, we need to introduce some notations, and regularity conditions.

Suppose, for all $\theta \in \Theta$ and all x, there exists $l'_{\theta_0}(\theta - \theta_0, x)$ such that the remainder in the linear approximation can be written as

$$r(\theta - \theta_0, x) = l(\theta, x) - l(\theta_0, x) - l'_{\theta_0}(\theta - \theta_0, x), \tag{1.4.3}$$

where $l'_{\theta_0}(\theta - \theta_0, x)$ is defined as

$$\lim_{t \to 0} \frac{l(\theta(\theta_0, t), x) - l(\theta_0, x)}{t}, \tag{1.4.4}$$

and $\theta(\theta_0, t) \in \Theta$ is a path in t connecting θ_0 and θ such that $\theta(\theta_0, 0) = \theta_0$ and $\theta(\theta_0, 1) = \theta$. A good choice for $\theta(\theta_0, t)$ is $\theta_0 + t(\theta - \theta_0)$, which is linear in t. In this case, $l'_{\theta_0}(\theta -$

$\theta_0, x)$ becomes the directional derivative of $l(\theta)$ at θ_0. Here we consider the general case because, in some cases we don't have any other choice but facing a nonlinear form of $\theta(\theta_0, t)$. Let $\|\cdot\|_s$ be a norm different from $\|\cdot\|$, (it is often chosen to be the Sobolev norm when it is appropriate to do so) such that $\|\cdot\| \leq \alpha\|\cdot\|_s$, and assume that the convergence rate of the PEL estimator under $\|\cdot\|$ and $\|\cdot\|_s$, be $o_p(\delta_n)$ and $o_p(\delta_n^s)$ respectively.

Suppose f is a functional with the following smoothness property: for all $\theta \in \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta_n^s\}$,

$$|f(\theta) - f(\theta_0) - f'_{\theta_0}(\theta - \theta_0)| \leq O(\|\theta - \theta_0)\|^w) \quad as \ \|\theta - \theta_0\| \to 0, \tag{1.4.5}$$

where $w > 0$ is the degree of smoothness of f at θ_0, and

$$f'_{\theta_0}(\theta - \theta_0) = \lim_{t \to 0} \frac{f(\theta(\theta_0, t), x) - f(\theta_0, x)}{t} \tag{1.4.6}$$

in this way $f'_{\theta_0}(\theta - \theta_0)$ is linear in $(\theta - \theta_0)$ and $\|f'_{\theta_0}\| < \infty$, where

$$\|f'_{\theta_0}\| = \sup_{\{\theta \in \Theta : \|\theta - \theta_0\| > 0\}} \frac{|f'_{\theta_0}(\theta - \theta_0)|}{\|\theta - \theta_0\|}. \tag{1.4.7}$$

Let V be the space spanned by $\Theta - \theta_0$, and suppose that $\|\cdot\|$ induces an inner product, $\langle \cdot, \cdot \rangle$, on the completion of V, which we show it be $\bar{V}$. By the Rise representation theorem, there exists $v^* \in \bar{V}$ such that, for any $\theta \in \Theta$, $f'_{\theta_0}(\theta - \theta_0) = \langle \theta - \theta_0, v^* \rangle$. Furthermore, let $\varepsilon_n = o(n^{-1/2})$ and for all $\theta \in \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta_n^s\}$

$$\theta^*(\theta, \varepsilon_n) = (1 - \varepsilon_n)\theta + \varepsilon_n(u^* + \theta_0) \in \Theta, \quad with \ u^* = \pm v^* \tag{1.4.8}$$

Let $K(\theta_0, \theta) = n^{-1} \sum_{i=1}^n E\left[l(\theta_0, X_i) - l(\theta, X_i)\right]$, which is the Kullback-Leibler information measure based on n observation when $l(\theta, X)$ is a likelihood function, and let

$$\nu_n(g) = n^{-1/2} \sum_{i=1}^n \left(g(X_i) - Eg(X_i)\right) \tag{1.4.9}$$

be the empirical process induced by g.

Now we are in a position to formulate some regularity conditions, under which we can derive the asymptotic distribution of $f(\hat{\theta}_n)$.

Assumptions:

A5: *(Stochastic equicontinuity). For the reminder function, $r(.,.)$, defined above we have:*

(i)

$$\sup_{\{\theta\in\Theta:\|\theta-\theta_0\|_s\leq\delta_n^s\}} n^{-1/2}\nu_n\Big(r(\theta-\theta_0,X)-r(\theta^*(\theta,\varepsilon)-\theta_0,X)\Big)=O_p(\varepsilon_n^2). \quad (1.4.10)$$

(ii)

$$\sup_{\{\theta\in\Theta:\|\theta-\theta_0\|_s\leq\delta_n^s\}} n^{-1/2}\nu_n\big(r(\theta-\theta_0,X)\big)=O_p(\varepsilon_n). \quad (1.4.11)$$

A6:

$$\sup_{\{\theta\in\Theta:\|\theta-\theta_0\|_s\leq\delta_n^s\}} \Big[K(\theta_0,\theta^*(\theta,\varepsilon_n))-K(\theta_0,\theta)\Big]-\frac{1}{2}\Big[\|\theta^*(\theta,\varepsilon)-\theta_0\|^2-\|\theta-\theta_0\|^2\Big]=O(\varepsilon^2). \quad (1.4.12)$$

A7: *For some constant $c>0$ and any $\theta_i\in\{\theta\in\Theta:\|\theta-\theta_0\|_s\leq\delta_n^s\}$, $i=1,2$, we have*

$$J(\theta_1+\theta_2)\leq c\big(J(\theta_1)+J(\theta_2)\big). \quad (1.4.13)$$

In addition, $\lambda_n=O(\varepsilon_n)$ and $J(v^*)<\infty$.

A8: *We have:*

$$\sup_{\{\theta\in\Theta:\|\theta-\theta_0\|_s\leq\delta_n^s\}} = n^{-1/2}\nu_n\big(l'_{\theta_0}(\theta-\theta_0)\big)=O_p(\varepsilon). \quad (1.4.14)$$

The following result proves the asymptotic normality of the PEL estimator.

Proposition 3:

Suppose assumptions A5-A8 are satisfied and f is a function which satisfies (4.5) with $\delta_n^s = O(n^{-1/2})$ and $Var_0(l'_{\theta_0}(v^*, X)) < \infty$. Then, for the approximate plug-in penalized estimator $f(\hat{\theta})$ we have

$$n^{1/2}\big(f(\hat{\theta}) - f(\theta_0)\big) \xrightarrow{p} N\Big(0, Var_0(l'_{\theta_0}(v^*, X))\Big). \tag{1.4.15}$$

The following corollary is a direct consequence of proposition 1.

Corollary 2:

If assumptions A1-A4 hold, then for the approximate penalized estimator, $\hat{\theta}$, we have

$$n^{1/2}\langle \hat{\theta} - \theta_0, s\rangle \xrightarrow{p} N\Big(0, Var_0(l'_{\theta_0}(s, X)\Big), \tag{1.4.16}$$

where $s \in \Theta - \theta_0$.

Typically, $Var(l'_{\theta_0}(\hat{\theta}_n - \theta_0)) = \| f'_{\theta_0}\|^2$.

1.4.2 Rate of Convergence

In this subsection of the paper, we use the results of Shen (1998) to obtain some probability bounds for the convergence of penalized EL estimator.

We first introduce some notation and list the regularity assumptions, which we will need to obtain the results of this section. Let $l_n(\theta|data)$ be the criterion function that we discussed earlier, which measures how well a model with parameter θ predicts the observed data. We define $K_n(\theta, \theta_0) = |E\big[l_n(\theta) - l_n(\theta_0)\big]|$. Now we define $\rho_n(\theta, \theta_0) = K_n^{1/2}(\theta, \theta_0)$, $\rho_n(\theta, \theta_0)$ will be used to measure the distance between two parameter points. In this context, which $l_n(\theta)$ represents the log empirical likelihood function, $K(\theta, \theta_0)$ becomes the Kullback-Leiber information criteria. Let $V(\theta, \theta_0) = Var\big(l(\theta) -$

$l(\theta_0)\big)$, where $l(\theta)$ is the limit of $l_n(\theta)$ when the sample size grows large. Also we define for any $k_i > 0$,

$$A(k_1,k_2) = \{\theta \in \Theta : k_1 \leq \rho(\theta_0,\theta) \leq 2k_1,\ J(\theta) \leq k_2\}, \tag{1.4.17}$$

and

$$B(k_1,k_2) = \{l(\theta) - (\theta_0) : \theta \in A(k_1,k_2)\}. \tag{1.4.18}$$

Let P_i be the probability measure on a measurable space $\mathcal{X}_i$ induced by the density $p_i(\theta_0,x)$. Define $P = n^{-1}\sum_{i=1}^{n} P_i$. Expectation E and E_i are evaluated under P and P_i respectively. Now we are in a position to state the required assumptions.

Assumptions:

A9: *For some* $0 \leq \beta < 1$ *and* $c_1 > 0$,

$$\sup_{A(k_1,k_2)} V(\theta_0,\theta) \leq c_1 k_1^2\big(1 + (k_1^2,k_2)^{\beta}\big). \tag{1.4.19}$$

A10*: There exists a random variable* $W(Z_i)$*, such that*

$$|l(\theta,Y_i) - l(\theta_0,Y_0)| \leq |\theta(X_i) - \theta_0(X_i)|W(Z_i), \tag{1.4.20}$$

where $\{X_i\}$ *and* $\{Z_i\}$ *are independent. Also,* $\sup_i E_i\big[exp(t_0 W(Z_i))\big] < \infty$ *and* $E\big[(\theta(X) - \theta_0(X))^2\big] \leq c_2 V(\theta_0,\theta)$*, with* $t_0 > 0$ *and* $c_2 > 0$*. Furthermore,*

$$\sup_{A(k_1,k_2)} \|\theta - \theta_0\| \leq c_3(k_1^2 + k_2)^{\gamma}. \tag{1.4.21}$$

For $0 \leq \gamma < 1$*, and* $c_3 > 0$*, the norm is the supermom norm on* Θ*.*

A11: *We have*

$$\sup_{\{k_1 \geq 1, k_2 \geq 1\}} \Psi(k_1,k_2) \leq c_4 n^{1/2}, \tag{1.4.22}$$

where $\Psi(k_1,k_2) = \int_L^U H^{1/2}(u,B(k_1,k_2))du/L$ *with* $U = c_5\varepsilon(k_1^2 + k_2)^{(1+max(\beta,\gamma))/2}$*, and* $L = c_6\lambda_n(k_1^2 + k_2)$*, and* $c_5, c_6 > 0$.[7]

[7] H(u,B) is called the Hellinger metric entropy. For a definition see the appendix. For more information consult Kolomogorov and Tihomirov (1959)

The following results establish some exponential probability bounds on the rate of convergence for the penalized EL estimator.

Proposition 4:

If assumptions A5-A7 are satisfied. Then there exists a constant $c_8 > 0$ *such that for any* ε *stisfying assumption A7, and* $max(J(\theta_0),1)\lambda_n \leq c_7\varepsilon^2$. *We have*

$$P^*\Big(\sup_{\{pho(\theta_0,\theta)\geq\varepsilon,\theta\in\Theta\}} n^{-1}\sum_{i=1}^{n}\big(\ell(\theta,Y_i)-\ell(\theta_0,Y_i)\big)\geq -\varepsilon^2/2\Big) \tag{1.4.23}$$

$$\leq 7exp(-c_8 n\, min(\lambda_n^2/\varepsilon^2,\lambda_n)),$$

where the P^* *is the outer measure (see for example Pollard (1984)).*

The following corollary gives the bounds for the estimator $\hat{\theta}$.

Corollary 3:

Suppose assumptions A9-A11 are satisfied. Then for the penalized estimator defined in definition (1b) with $a_n = o(\varepsilon_n^2)$, *we have*

$$P\big(\rho(\theta_0,\hat{\theta})\geq\zeta\big)\leq 7exp(-c_8 n\zeta_n^2), \tag{1.4.24}$$

where $\zeta_n = max(\varepsilon_n,\lambda_n^{1/2})$ *with* ε_n *the smallest* ε *satisfying assumption A11. The best possible rate can be obtained by setting* $\lambda_n \sim \varepsilon_n^2$.

Proposition 4 essentially says that, the rate of convergence is determined by equation (4.22) of assumption $A11$, which relates the size of the parameter space, the local behavior of the profile empirical likelihood function, and the degree of penalization (λ_n). We clearly see that when ε_n is large, which is an indicator of a large parameter space in a neighborhood of θ_0, we need to increase λ_n, the degree of penalization, in order to get an acceptable convergence rate.

Remarks 1: Method of *sieve* is another important statistical method, which is very

close to the method of penalization. In sieve approximation, like the penalization method, we often have a very large parameter space, and optimization on the whole space does not produces any meaningful estimator. In penalization technique, we restrict the optimization to a manageable subspace and then carry out the optimization. In sieve method, we carry out the optimization within a subset which is dense in the original parameter space. More formally, if Θ_n is a sequence of spaces dense in Θ, for every $\theta \in \Theta$ there exists $\theta_n \in \Theta_n$ such that $\|\theta_n - \theta\| \to 0$. A sieve estimate $\hat{\theta}_n$, is an optimizer of the criterion over Θ_n.

Remark 2: Another very important method which has close connection with penalization method, is the Bayesian method. We can interpret the penalty function as formulating prior knowledge about the unknown parameters. More specifically, constructing a prior such that the posterior distribution is supported on a desirable set with large probability. This suggests that one way to construct a Bayesian empirical likelihood estimator, is to try to do it via penalization method.[8]

1.4.3 Example: Penalized Minimum Distance

Since using the empirical likelihood methods, we estimate two sets of parameters, the unknown parameter θ_0, and the probability distribution $p = (p_1, \ldots, p_n)$, we can use the penalization method to construct better distributions. In this subsection we try do that.

Penalty functions an be design to take care of unnecessary small p_i in implied probability distribution, or just to take account of external information that the econometrician might have. In this section we study the penalized empirical likelihood, in which the penalty function is designed to regulate the implied probability measure in

[8]There has been attempt to construct Bayesian EL estimators, (S. Schennach 2005, and N. Lazar 2004), although the authors have taken other roots.

order to get a measure as close as possible to the maximum entropy measure. As it is mentioned earlier in this paper, people have used a combination of empirical likelihood and exponential tilting by embedding the implied probabilities of exponential tilting procedure in the criterion function of empirical likelihood estimator, see for instance Jing and Wood (1996), Corcoran (1998), Schennach (2007), and Smith (2005). The penalized method, introduced here, attempt to combine EL and ET methods too. Here we use the implied probability measures of the EL procedure, but use the exponential tilting criterion to penalize those $\hat{p}_i(\theta)$ which are not in agreement with ET criterion. More studies need to investigate the properties of this new estimator, but it seems to me that this procedure is more in line with statistical theory. There is a big literature studying the penalized methods, but simply plug in the implied probabilities of one procedure to the criterion function of another procedure might seem a little ad hoc.

For a data set $X_n = \{x_1, \ldots, x_n\}$, let P_n be the empirical distribution which assigns equal weights to each x_i. For a given distribution P let $d(P, P_n)$ be a distance defined on the space of probability measures. Furthermore, assume that

$$\ell_n(\theta) = -\max_{\gamma \in R^m} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \gamma' g(x_i, \theta)) \tag{1.4.25}$$

be the profile empirical likelihood, obtained after accounting for the moment conditions $E[g(X_i, \theta)]$ in the following definition, we define the penalized minimum distance estimator.

Definition 3:

The penalized minimum distance empirical likelihood estimator for θ *is* $\hat{\theta}_n$ *such that*

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \{\ell_n(\theta) - \lambda_n d(P, P_n)\} \tag{1.4.26}$$

where $P = (p_1, \ldots, p_n)$ *and*

$$p_i = \frac{1}{1 + \gamma' g(x_i, \theta)} \tag{1.4.27}$$

In definition 3, $J(\theta) = d(P, P_n)$. For instance, KL is a distance measure which if used in the above definition, will penalise $p_i's$ in such way that the final estimator will have implied probabilities with higher entropy. There are various distances, like Hollinger distance, Kolmogorov-Smimov distance and so on. Depending on what we expect the estimator to achieve, different measures can be used. the most commonly used distance measure is the KL which was introduced earlier.

Schennach (2007) has investigated the first and second order properties of the ELET estimator. In a nutshell this estimator is a compromise between EL and ET estimators, and therefore one should expect to see that, ETEL has a better behavior under miss-specification compared to EL, and at the same time has better second order bias properties compared to ET. In fact ELET has the same higher order bias and variance properties as EL. I expect the penalized estimator, introduce here has the same higher order properties too. The source of these better performance is the EL criterion function which our estimator is based on it, too. I intend to do a more in-depth study of the first and higher order properties of this estimator.

1.5 CONCLUSIONS

This paper extends the "least absolute shrinkage and selection operator" to the framework of empirical likelihood estimation. It also provides a guideline to implement it for the more general setting of GEL. We show how this procedure is able to consistently select the best possible model. The simulation results show the better performance of LASSO-EL compared to the classical AIC, BIC, and DT criteria. Also, we see from the presented simulation results that better bias property of EL estimator (compared to GMM) is carried out to the LASSO-EL too, in a way that the LASSO-EL has better

post-selection preference than LASSO-GMM has.

As by a product, we investigate the large sample properties of the penalized empirical likelihood in setting with a fairly general penalty function. One interesting conclusion is that, the rate of convergence depends on the complexity of the parameter space, as measured by the Hellinger metric entropy (HME), around θ_0, and the degree of penalization λ_n. We saw that, the higher the HME, the bigger the degree of penalization has to be in order to get a faster rate of convergence. In other words, while the consistency of the penalized estimator is determined by the global behaviour of the criterion, the rate of that convergence and therefore the asymptotic normality of the estimator is determined by the local behavior of the criterion. Finally, We presented other forms of penalty functions which they might be able to produce estimators with possibly important properties. Studying the properties of these estimators is a subject of future studies.

2.0 MODULATION METHOD FOR EMPIRICAL LIKELIHOOD ESTIMATOR

2.1 INTRODUCTION

Empirical likelihood (El) (Owen 1988) is regarded as the non-parametric version of parametric likelihood procedure. Its robustness against distributional assumptions on one hand, and its good properties analogous, to the parametric likelihood, on the other hand, make it a very powerful tool when it is applied to the moment condition models in econometric applications. GMM (Hansen 1982) and other recently developed techniques based on Empirical Likelihood (Owen, 1988; Qin and Lawless, 1994; Imbens, 1997) use a set of given moment conditions to construct estimators for the unknown parameters. In this paper, and a companion paper (Shahidi 2008) we study the use of *shrinkage* techniques in improving the empirical likelihood procedure. While this paper introduces the *modulation* method, the other paper deals with penalization technique. Modulation, and *penalization* methods belong to the wider class of shrinkage procedures. Shrinkage methods enable us to use extra information, and incorporate prior beliefs into the estimation. For instance, in the penalization method one can construct the penalty function based on the external information she wants to take into account. Shrinkage methods are also useful in correcting some undesirable features of some class of estimators. In this paper we develop a general framework, in which, one

can study different estimators using the modulation techniques. Using this framework, we introduce several examples of new estimators, and examine their properties.

The remaining part of this introduction, is devoted to the heuristic origins of the topics which will be further analyzed in the following sections.

2.1.1 GMM and GEL

Generalized method of moments estimator (GMM) has been the workhorse of econometric analysis since its introduction by Hansen (1982). Besides providing a unified framework to study different types of estimators, GMM extends the method of moments framework to include situations in which the number of moment conditions exceed the dimension of the parameter we want to estimate. Although the GMM estimator has desirable properties, such as being first-order asymptotically efficient, its small sample properties are relatively poor (Altonji and Segal, 1996; Tauchen 1986). More recently, Owen's empirical likelihood method has provided other estimators, some of which overcome some of the shortfalls of the GMM. From this family we have the EL estimator (Owen, 1988; Qin and Lawless, 1994; Imbens, 1997), Continuous Updating Estimator (CUE) (Hansen and Yaron, 1996), and the Exponentially Tilting Estimator (Kitamura and Stutzer, 1997; Imbens and Johnson, 1998) which all belong to the class of Generalized Empirical Likelihood (GEL) estimators (Smith, 1997; Newey and Smith, 2004). [1] These estimators circumvent the need of estimating a weighting matrix in the two-step GMM by directly minimizing an information-theory-based concept of closeness between the estimated distribution and the empirical distribution.[2] While

[1] There are other varieties, too. For example the Exponentially Tilted Empirical Likelihood estimator (ETEL) (SCHENNACH, 2007) which in essence is a combination of the two estimators, EL and ET, in hope to obtain an estimator that like EL has a smaller finite-sample bias, and at the same time inherits the better behavior of ET in the presence of mis-specification.

[2] The estimators mentioned so far are, like GMM, based on unconditional moment restrictions, using the empirical likelihood methods, we can construct estimators based on conditional moment restrictions,

in theory these estimators, like GMM, all have the same first-order asymptotic efficiency, simulation works and Monte Carlo evidences have shown that, compared to GMM, some members of the GEL class have better finite-sample properties (Hansen and Yaron, 1996; Ramalho, 2006) and references therein. Also, Newey and Smith (2004) have analytically shown, using a stochastic expansion argument, that while GMM and GEL share the same first-order asymptotic properties, their higher-order properties are different. Specifically, while the asymptotic bias of GMM often grows with the number of moment restrictions, the relatively smaller bias of EL does not. Moreover, a bias-corrected EL is higher-order efficient relative to any other regular method of moment estimator. In term of inferences, the empirical likelihood ratio test has some desirable features too. For example ELR test admits Bartlet correction, DiCiccio, Hall, Romano (1991), which gives the same accuracy rate as the parametric case. Kitamura (2001) used the so called *Generalized Neyman-Pearson* approach to show that for testing moment restriction the ELR test is uniformly most powerful in an asymptotic large deviation sense.

2.1.2 Shrinkage and Modulation

Shrinkage is a general method in statistics for improving an estimator and regularizing ill-posed inference problems. Commonly used procedures like Bayesian inference, and penalized likelihood inference, implicitly use the shrinkage technique.

In this part of the introduction, I will use the simple ordinary least square (OLS) to demonstrate how the shrinkage method works, and also we hope to justify its usefulness. The Gauss-Markov theorem states that among all linear unbiased estimators, the OLS has the smallest variance, but this property which sometimes is called BLUE

too. See (ZHANG and GIJBELS, 2003), and Kitamura et al (2004)

"Best linear Unbiased Estimator" does not yield the best estimator in the sense of MSE "Mean Squared Errors". In the other words, if we drop the unbiased restriction we can do better in MSE sense. To demonstrate it, assume $\hat{\beta}_i$ is the OLS estimator of β_i, and define $\tilde{\beta}_i = \frac{1}{1+\lambda}\hat{\beta}_i$. We notice that if $\lambda = 0$; we get the OLS estimator back, and when λ is too large, $\tilde{\beta}_i$ shrinks to zero. Furthermore, $E\tilde{\beta}_i = \frac{1}{1+\lambda}E\hat{\beta}_i = \frac{1}{1+\lambda}\beta_i$, therefore $\tilde{\beta}_i$ is a biased estimator of β_i.

The MSE of $\tilde{\beta}_i$ can be written as

$$K\sigma^2(\frac{1}{1+\lambda})^2 + (\frac{\lambda}{1+\lambda})^2 \sum_{i=1}^{K} \beta_i^2. \tag{2.1.1}$$

The first part is the variance component, which is the largest when λ is zero. The second part is the squared bias and it grows with λ. In principal with the right choice of λ, we can get an estimator which does better than the OLS in MSE sense. This new estimator is not unbiased, but what we pay for in bias, we make up for in variance. The first order condition gives the optimal choice for λ as

$$\lambda = \frac{K\sigma^2}{\sum \beta_i^2}. \tag{2.1.2}$$

Although this choice of λ is not feasible, it is possibler to find good estimations for it. For example we can replace σ with an unbiased estimation of variance, and also replacing β with some appropriate estimation of β. Two of the most mentioned feasible estimators of λ are James-Stein estimator (James and Stein, 1961), and Sclove estimator (Sclove, 1968).

We can summarize the construction of this new estimator as follows: First, we used the scalar parameter λ to obtain an equation in which a new parameter, $\tilde{\beta}$, depends on the OLS estimator through the parameter λ; this step is called modulation. Second, we have a criterion which we are looking to optimize; here we want to minimize the MSE risk of the estimator. Third, this minimization yields us an optimal choice for

the newly introduced parameter λ, which in turn results in a new estimator which has a better MSE compared to the OLS estimator. In Beran, and Dümbgen (1998) terminology, the parameter λ, is called a modulator. They extended this argument roughly in the following manner. Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_n)$, then the modulated parameter is $\tilde{\beta} = (\lambda_1\hat{\beta}_1, \ldots, \lambda_n\hat{\beta})$, and therefore the modulator is the n-dimensional vector $\lambda = (\lambda_1, \ldots, \lambda_n)$

In this paper, we will use a method similar to "modulation estimator" (Beran and Dümbgen, 1998) to obtain a new modulation of the old estimator. Then we use an appropriate criterion to pick the best "modulator" which gives the best estimator, judged by the chosen criterion . The procedure that we are trying to implement can be summarized as follow:

1. Modulation: Use modulators to modulate the estimator to a family of estimators, depending on the modulator.
2. Selecting a criterion: Use a criterion to choose the best modulator. The criterion is usually a risk function, evaluating the risk associated with a given estimator. Because in El estimation method we estimating two entities, the unknown parameter θ, and the multinominal distribution $p = (p_1, \ldots, p_n)$, we can use a criterion which measures the goodness of a distribution like p. Both of these methods are discussed in the section two of this paper.
3. Adaptation: Find a modulator that optimizes the given criterion.

2.1.3 Other Interpretations for λ

So far we have considered the modulator, $\lambda = (\lambda_1, \ldots, \lambda_n)$, to be a purely mathematical tool which helps to change, construct and choose form, the already known estimators. Another possible interpretation of the vector $\lambda = (\lambda_1, \ldots, \lambda_n)$ is to consider λ as a vec-

tor of weights. Weighting is commonly used in econometrics and statistics to account for specific structure of a data set. Interpreting w as weights will result in a *weighted empirical likelihood*, which we hope to take account of heteroskedasticity in the data. If the data under consideration possesses an unknown structure, for instance, we are aware of a heteroskedasticity in the data, but the exact structure of it is unknown. In this case, we use an unknown vector of weights $w = (w_1, \ldots, w_n)$, and then by using some criteria, we try to choose the best weights according to that criteria. The weighting vector w can be considered as kernel weights. Instead of maximizing the empirical distribution we might want to maximize a smoothed version of the empirical distribution, which requires weighting by a smoothing kernel.

This paper contributes to the EL literature in two ways. First, we introduce the modulation method in the empirical likelihood framework. This method enables us to study several different estimators using the same theoretical framework. For example, when the modulators are interpreted as weighting vectors, we can define and study the weighted empirical likelihood procedure. Second, we use the modulation method in some well known econometrics and statistical models, like GLM model. We study them analytically, and conduct Monte Carlo simulations, which shows the improved estimators indeed work better then the original ones, specially when the sample size is very small. Although, we focus our attention on EL procedure in this paper, extending the results to the more general setting of GEL is not very far off.

The rest of this paper proceeds as follows. In section 2, we introduce the modulation technique and use it to construct new EL estimators. In section 3 we study GLM as an example, we use this technique to obtain an estimator for the *generalized linear model, GLM*, (Kolaczyk, 1994; Chen and Cui, 2003). We will show that this estimator has a lower variance than the traditional quasi-likelihood estimator of the GLM models. Section 4 reports the outcome of some Monte Carlo simulations, and section 5

concludes the paper, while all proofs are collected in the appendix.

2.2 MODULATION METHOD

In this section, we try to implement the three steps which we discussed in the previous section. We first present an example, which shows how the method works in a linear setting. This example is a slightly modified version of Beran (2000). We construct an estimation for a density function f, which we hope to show how modulation method works.

2.2.1 How Modulation Works

As we explained above, this subsection serves as an illustration of the modulation method. We hope a reader who might be unfamiliar with this method can gain enough insight from this example to follow the rest of this paper.

Definition 4:

A modulator is a vector $w = (w_1, \ldots, w_n)$*, where* $w_i \in [0,\ 1]$ *for* $i = 1, \ldots, n$*.*

Now we define a modulation estimator.

Definition 5:

A modulation estimator is a component-wise linear estimator of the form

$$\hat{\theta}(w) = (w_1\hat{\theta}_1, \ldots, w_n\hat{\theta}_n), \tag{2.2.1}$$

where w is a modulator.

Suppose that

$$Y_i = f(x_i) + \sigma\varepsilon_i \tag{2.2.2}$$

, where $\varepsilon_i \sim N(0,1)$ and $x_i = 1/n$.

Assume $f \in L_2[0,1]$ therefore we can expand it as

$$f(x) = \sum_{i=1}^{\infty} \theta_i \phi_i(x) \quad and \quad \theta_i = \int (f(x)\phi_i dx \tag{2.2.3}$$

where $\{\phi_i\}_1^\infty$ is an orthonormal basis for $L_2[0,1]$. Define $\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^n Y_i \phi_j(x_i)$ therefore

$$E(\hat{\theta}_j) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i)\phi_j(x_i) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\phi_j(x_i) \tag{2.2.4}$$

$$\approx \int f(x)\phi_i(x)dx = \theta_j \tag{2.2.5}$$

and

$$Var(\hat{\theta}_j) = \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i)\phi_j^2(x_i) = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\phi_j^2(x_i) \approx \frac{\sigma^2}{n^2}\int \phi_j^2(x)dx = \frac{\sigma^2}{n^2}. \tag{2.2.6}$$

Considering the dimensionality of the data set, $\hat{f}_n(y) = \sum_{i=1}^n \hat{\theta}_i \phi_i(y)$ is a good estimator for $f(y)$. The estimator $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ often results in a $\hat{f}_n(y)$ which has very poor risk. Using modulators we can improve the risk of $\hat{f}_n(y)$. Let $\hat{\theta}(w) = (w_1\hat{\theta}_1, \ldots, w_n\hat{\theta}_n)$. By Parsevel equality, the loss function is

$$L(\hat{f}_n, f_n) = \int (\hat{f}_n(y) - f_n(y))^2 \tag{2.2.7}$$

$$= \sum_{i=1}^{n} (w_i\theta_i - \theta_i)^2 \tag{2.2.8}$$

therefore the risk function is

$$R(\hat{f}_n, f_n) = E[L(\hat{f}_n, f_n)] \tag{2.2.9}$$

$$= \sum_{i=1}^{n} \left(w_i \frac{\sigma^2}{n} + (1-w_i)^2 \theta_i^2 \right) \tag{2.2.10}$$

An unbiased estimator for the risk function can be obtained, by replacing θ_i by $\hat{\theta}_i$ and σ^2 by an unbiased estimator of $\hat{\sigma}^2$.

$$\hat{R}(w) = \sum_{i=1}^{n} \left(\hat{\theta}_i - \frac{\hat{\sigma}^2}{n} \right) (1-w_i)^2 + \frac{\hat{\sigma}^2}{n} \sum_{i=1}^{n} w_i^2 \tag{2.2.11}$$

now the minimum risk estimator for $f(y)$ is obtained by using $\hat{\theta}(w^*)$, where w^* is the minimizer of $\hat{R}(w)$.

Therefore to obtain the modulation estimator, first we derived an estimator for the density function f. Then in the second stage, we modulated this estimator and obtained a family of estimators $\hat{f}_n^w$ for f. And finally in the third stage, we used a criteria to compare the members of this family and choose the best one which we show it by $\hat{f}_n^{w^*}$. We now perform this three-steps procedure for the empirical likelihood estimator. The main difference, from the setting discussed above, is that the modulations we consider here are no longer necessarily linear.

2.2.2 Modulated EL

Definition 3 defines both what a modulator is, and what we mean by a modulation estimator in a nonlinear setting.

Definition 6:

We define:

1. *A "modulator" is a vector* $w = (w_1, \ldots, w_n)$ *where* $w_j \in \mathbb{R}$, *for* $j = 1, 2, \ldots, n$

2. A "modulation estimator" is a component-wise estimator of the form

$$\hat{\theta}(w) = (\hat{\theta}_1(w), \ldots, \hat{\theta}_n(w)) \tag{2.2.12}$$

where w is a modulator.

The idea is, to derive a class of estimators in a manner that all of them satisfy the desired sample moment conditions, and furthermore, each one depends on the modulator w. To achieve this, we change the objecting function used in empirical likelihood estimation in such a way that the new objective function depends on the modulator w. Obviously there are more than one way to do this, but we should be able to provide reasonable interpretations for any selected procedure. Below, we propose one of these ways, which we think has a very natural interpretation as *weighted empirical likelihood*. In lemma 2, we show that this procedure in equivalent to another one which is easy to interpret too, and therefore we can use them interchangeably. Later in this paper we discuss the weighting interpretation in details.

Definition 7:

For a modulator $w = (w_1, w_2, \ldots, w_n)$ *define*

$$\hat{p}(w) = \underset{p_1,\ldots,p_n}{\arg\min} \sum_{i=1}^{n} -w_i \log p_i \tag{2.2.13}$$

subject to:

$$\sum_{i=1}^{n} p_i g_i(\theta) = 0 \quad \textit{and} \quad \sum_{i=1}^{n} p_i = 1 \tag{2.2.14}$$

notice that here $g_i(\theta) = g(x_i, \theta)$.

The following lemma shows that every solution to the minimization in definition 2, can be manipulated to get a solution for another minimization problem, which sometimes is easier to implement.

Lemma 1:

Let $\hat{p}(w)$ be the solution obtained from definition 2, then there exits a solution $\hat{q}(\varpi)$ to the following minimization problem

$$\min_{q_1,\ldots,q_n} \sum_{i=1}^{n} -\log q_i \tag{2.2.15}$$

subject to:

$$\sum_{i=1}^{n} q_i g_i(\theta)\varpi_i = 0 \quad \textit{and} \quad \sum_{i=1}^{n} q_i = 1, \tag{2.2.16}$$

where $\varpi = (\varpi_1,\ldots,\varpi_n)$ is a new modulator.

Proof. See the appendix □

Since adding extra constraints does not increase the variance of the EL estimators, Qin and Lawless (1994), Newey and Smith (2004) we might be able to achieve better estimators by adding some extra constrains which help us to use more information or more efficiently the same information, to estimate the parameters. Using the idea of modulators, introduced by definition 2 and lemma 1 we can develop an extended version of EL estimator, by adding extra moment conditions to the set of original moment conditions. The following definition introduce this modification and later in this paper we show, through an example using GLM (generalized linear models), how to use this modification, along with a modulator, to construct better estimators.

Definition 8:

Let $g(X_i,\theta) = g_i = (g_i^1,\ldots,g_i^k)$, $h_i^1 = (g_i^{l_1},\ldots,g_i^{l_j})$, $h_i^2 = (g_i^{t_1},\ldots,g_i^{t_{j'}})$, $\{l_1,\ldots,l_j\} \cup \{t_1,\ldots,t_j'\} = \{1,\ldots,k\}$ and $h_i = (h_i^1,h_i^2)$. Let $w = (w_1,\ldots,w_n)$ be a modulator as in definition 1, the extended EL estimator for θ is the estimator obtained from EL procedure by replacing the constrain $\sum_{i=1}^n p_i g_i = 0$ with the new constrain $\sum_{i=1}^n p_i(h_i^1 + w_i h_i^2) = 0$

Remark: This definition can be considered as a generalization of definition 2. By setting $h_i^1 = 0$ and $h_i^2 = g_i$ we get definition 2. Also, it should be noticed that, if $j + j' > k$, then the number of original moment conditions, k, has been extended, and some new moment conditions have been added to the original set. This proves to be useful specially in cases that, the number of moment conditions are the same as the number of parameters.

All the procedures introduced so far, have the important feature of linking the estimator of θ to the vector $w = (w_1, \ldots, w_n)$. In the other words all the maximization procedures, introduced above, yield us an estimator for the empirical probability measure $p = (p_1, \ldots, p_n)$, $\hat{p}(w)$, and as a by product, we obtain an estimator for the unknown parameter θ, which we show it by $\hat{\theta}(w)$. To see this, we set up the Lagrangian

$$\mathcal{L} = -\sum_{i=1}^{n} \log(p_i) + \lambda' \sum_{i=1}^{n} p_i g_i(\theta) w_i + \mu\big(\sum_{i=1}^{n} p_i - 1\big) \tag{2.2.17}$$

where $\mu \in \mathbb{R}$ and $\lambda \in \mathbb{R}^p$ are the Lagrange multipliers. It takes some simple algebra to show that the first order conditions are solved by

$$\hat{\mu} = n, \quad \lambda(\theta) = \arg\min_{\lambda \in \mathbb{R}^p} -\sum_{i=1}^{n} \log(1 + \lambda' g_i(\theta) w_i) \tag{2.2.18}$$

and

$$\hat{p}(\theta) = \frac{1}{n(1 + \hat{\lambda}(\theta)' g_i(\theta) w_i)} \tag{2.2.19}$$

therefore the likelihood profile will be

$$\ell(\theta) = \min_{\lambda \in \mathbb{R}^p} -\sum_{i=1}^{n} \log(1 + \lambda' g_i(\theta) w_i) - n \log n \tag{2.2.20}$$

finally the empirical likelihood estimator for θ is

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\max_{\theta \in \Theta} \min_{\lambda \in \mathbb{R}^p} -\sum_{i=1}^{n} \log(1 + \lambda' g_i(\theta) w_i) \tag{2.2.21}$$

As we notice, both estimators, the estimator for θ, $\hat{\theta}(w)$, and the estimator for the empirical distribution P, $\hat{P} = (\hat{p_1}(\hat{\theta}(w)), \ldots, \hat{p_n}(\hat{\theta}(w)))$, depend on the modulator w. In this manner we have a class of estimators for the empirical measure, and a class of estimators for the unknown parameter θ

$$\hat{P} = \{\hat{p}(w) | w \in W\} \quad \textit{and} \quad \hat{\Theta} = \{\hat{\theta}(w) | w \in W\} \tag{2.2.22}$$

where W is the set of all allowable modulators.

Definition 2, resembles the definition of empirical likelihood estimator, with the exception of coefficients w_i. These w_i's can be interpreted as weights or just a mathematical device to modulate and construct new estimators. Our goal is, to show that this device is indeed a useful one which helps us to find estimators with better desired properties. Later in this paper we discuss weighting and other interpretations of the w_i's.

The following lemma shows that every member of $\hat{\Theta}$ along with the corresponding $\hat{p}(w) \in \hat{P}$ satisfies the moment conditions.

Lemma 2:

Let modulator $w = (w_1, \ldots, w_n)$ be given, then the estimator $\hat{\theta}(w)$, and its corresponding implied probabilities $\hat{p}(w)$, given by definition 2, exist and satisfy the sample moment conditions.

Proof. See the appendix □

Now that we have a set of estimators, $\hat{\Theta}$, and the sample moment conditions are satisfied using whatever member of this set, we need some criteria to choose from this set of estimators. This leads us to the second step of the procedure introduced earlier in the introduction, "*risk estimation*". Since, we are estimating two unknowns, the empirical distribution and the models' parameter θ, we have several options for using an

appropriate criteria. First, we will explain how to choose the best empirical measure from the set $\hat{P}$. As we will see, this is much simpler than trying to estimate the risk of a given estimator. Because choosing the appropriate empirical distribution is achieved by choosing a suitable modulator, $w^0 = (w_1^0, \ldots, w_n^0)$, and because given the modulator w^0 we can pick the estimator $\hat{\theta}(w^0)$, the procedure of choosing an appropriate empirical measure yields us an estimator for θ with the best implied probabilities judged by the given criteria.

2.2.3 The Minimum distance Criteria

There are a host of metrics available to quantify the distance between two given measures. Although some are not metrics in the mathematical sense of the word, but possess a notion of "distance" which have been proven to be useful. Among many such distance measures we restrict ourselves to the forward Kullback-Leibler divergence, also known as "relative entropy". Kullback-Leibler (KL) divergence is one of the fundamental concepts in statistics and information theory. From many interpretations, are measuring goodness of fit, and measuring lose of power in a likelihood ratio test. Just as likelihood measures how well a model explains the data, we can think of KL as measuring the lack of fit between model and data relative to a perfect fit. Also we can think of KL divergence from P_A to Q as measuring how much power we lose with the likelihood ratio test if we mis-specify the alternative hypothesis P_A as Q.

For two measures P and Q the forward Kullback-Leibler (KL) divergence between P and Q is defined to be

$$K(P,Q) = \int \log \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}P \tag{2.2.23}$$

when the state space,Ω, is discreet we can write it as

$$K(P,Q) = \sum_{\omega \in \Omega} P(\omega) \log \frac{P(\omega)}{Q(\omega)} \tag{2.2.24}$$

The empirical likelihood minimizes the forward KL divergence between the empirical measure μ_n and the measure obtained by enforcing the moment condition. Let's

$$\mathcal{P}(\theta) = \left\{ P \in M \middle| \int g(x,\theta) \mathrm{d}P = 0 \right\} \tag{2.2.25}$$

where M is the set of all probability measures on $\mathbb{R}^P$ and define

$$\mathcal{P} = \bigcup_{\theta \in \Theta} \mathcal{P}(\theta) \tag{2.2.26}$$

then

$$\inf_{\theta \in \Theta} \inf_{P \in \mathcal{P}(\theta)} K(\mu_n, P) = \inf_{P \in \mathcal{P}} K(\mu_n, P) \tag{2.2.27}$$

If $\hat{p}(w) = (\hat{p_1}(w), \ldots, \hat{p_n}(w)) \in \hat{P}$ then

$$K(\mu_n, p(w)) = \sum_{i=1}^{n} -\frac{1}{n} \log n p_i(w) \tag{2.2.28}$$

now we are in a position to present an example of how the modulation method works. In the following example, we construct an estimator for θ and then definition 6, defines it as an special case of the general method we introduced earlier.

Example 1:

Let the data set $X_n = \{x_1, \ldots, x_n\}$ satisfy the moment condition

$$E[g(x_i, \theta] = 0 \tag{2.2.29}$$

setting up the EL estimation for θ, using the modulator $\lambda = (\lambda_1, \ldots, \lambda_n)$, yield the following empirical measure $\hat{P} = (\hat{p_1}, \ldots, \hat{p_n})$ such that

$$\hat{p_i}(\theta, \lambda) = \frac{1}{n(1 + \hat{\gamma}(\theta, \lambda)' g(x_i, \theta)\lambda_i)} \tag{2.2.30}$$

and

$$\hat{\gamma}(\theta) = \arg\min_{\gamma \in \mathbb{R}^p} - \sum_{i=1}^{n} \log(1 + \gamma' g(x_i, \theta)\lambda_i) \tag{2.2.31}$$

Let λ^* be the solution to the following maximization problem

$$\lambda^*(\theta) = \arg\max_{\lambda \in \mathbb{R}^n} \sum_{i=1}^{n} \hat{p}_i(\theta,\lambda)\log(\hat{p}_i(\theta,\lambda)) \tag{2.2.32}$$

the desired estimator for θ is

$$\hat{\theta_*} = \arg\max_{\theta \in \Theta} -\sum_{i=1}^{n} \log \hat{p}_i(\theta,\lambda^*(\theta)) \tag{2.2.33}$$

Using the framework developed in definition 2, we pick the vector λ^* in order to maximize $K(p(\lambda),\mu_n)$.

Definition 9:

Let

$$w^* = \arg\max_{w_1,\ldots,w_n} \sum_{i=1}^{n} p_i(w)\log p_i(w) \tag{2.2.34}$$

we call the corresponding $\hat{\theta}(w^*)$ the KL-adapted empirical likelihood estimation of the parameter, θ_0.

This estimator has the special property that its implied probability distribution is the *maximum entropy distribution* and at the same time it maximizes the empirical likelihood too.[3] The following proposition shows that the implied probabilities for KL-adapted-EL are indeed better than those of EL estimator, as long as the KL criteria concerns.

Proposition 5:

If $\hat{p} = (\hat{p}_1,\ldots,\hat{p}_n)$ be the implied empirical probability measure of EL and $\hat{p}_{w^*} = (\hat{p}_{1w^*},\ldots,\hat{p}_{nw^*})$ be the implied empirical probability measure of KL-adapted-EL then

$$K(\mu_n,\hat{p}_{w^*}) \leq K(\mu_n,\hat{p}) \tag{2.2.35}$$

[3] According to the *maximum entropy principal,* the least biased distribution that encodes certain given information, is the one which maximizes the information entropy

Proof. $\hat{p}_{w^*}$ minimizes $K(\mu_n, p)$ for all $\hat{p} \in P(w)$, if $w^0 = (1, 1, \ldots, 1)$ then $\hat{p}_{w^0} = \hat{p}$, the implied probabilities obtained from empirical likelihood estimation, therefore we have

$$\min_{w \in W} K(\mu_n, \hat{p}_w) \leq K(\mu_n \hat{p}_{w^0}) \tag{2.2.36}$$

The left hand side is $K(\mu_n, \hat{p}_{w^*})$, and the right hand side is $K(\mu_n, \hat{p})$, therefore

$$K(\mu_n, \hat{p}_{w^*}) \leq K(\mu_n, \hat{p}) \tag{2.2.37}$$

□

An interesting exercise is, to compare θ_{ael} and θ_{etel}. θ_{etel} is designed to take advantage of both empirical likelihood, maximized empirical likelihood ratio, and exponentially tilted empirical likelihood, maximized entropy. As we will see, while it does it to some extend, (Schennach 2007), it does not produce an empirical measure that has the two impotent property to see this let define θ_{etel} as it is done in Schennach paper.

Using *minimum empirical discrepancy*, (MED) (Corcoran, 1998; Cressie and Read, 1984) we have

$$\hat{\theta}_{etel} = \arg\min_{\theta} \sum_i \tilde{h}(\hat{p}_i(\theta)) \tag{2.2.38}$$

where $\hat{p}_i(\theta)$ is the solution to

$$\min_{\{p_i\}_{i=1}^n} \sum_i h(p_i) \tag{2.2.39}$$

subject to

$$\sum_i p_i g(x_i, \theta) = 0 \quad \textit{and} \quad \sum_i p_i = 1 \tag{2.2.40}$$

so that

$$\tilde{h}(p_i) = -\log(p_i) \quad \textit{and} \quad h(p_i) = p_i \log(p_i) \tag{2.2.41}$$

to ease comparison, we can re-define the estimator of example 1 as

$$\hat{\theta}(w) = \arg\min_{\theta \in \Theta} \sum_i w_i h(\hat{p}_i(\theta, w)) \tag{2.2.42}$$

where $\hat{p}_i(\theta, w)$ is the solution to

$$\min_{\{p_i\}_{i=1}^n} \sum_i w_i h(p_i) \tag{2.2.43}$$

subject to

$$\sum_i p_i g(x_i, \theta) = 0 \quad and \quad \sum_i p_i = 0 \tag{2.2.44}$$

in this way we obtain

$$\hat{\Theta} = \{\hat{\theta}(w) | w \in W\} \tag{2.2.45}$$

and

$$\hat{P} = \{\hat{p}(w) = \hat{p}(\hat{\theta}(w), w) | w \in W\} \tag{2.2.46}$$

if we define $w^p = (-p_1, \ldots, -p_n)$ then

$$\hat{\theta}_{ael}(w^p) = \hat{\theta}_{etel} \tag{2.2.47}$$

this implies that $\hat{\theta}_{etel} \in \hat{P}$ and therefore

$$K(\mu_n, \hat{p}_{ael}) \leq K(\mu_n, \hat{p}_{etel}) \tag{2.2.48}$$

at least in theory we get an improvement upon θ_{etel}.

2.2.4 Minimum Risk Criteria

In the previous subsection, we tried to pick the best empirical measure, $\hat{p}$, in the set $\hat{P}$. Here in this subsection, we try to define a criteria which helps us pick the best estimator for the unknown parameter $\hat{\theta}$. The very large literature in statistical theory which deals with this problem is commonly know as "statistical decision theory".[4]

Here, we give a very short overview in hope to further facilitate the understanding of this paper. Let $\mathcal{A}$ be the set of allowable decisions, usually is called the *action space*, and Θ is the parameter space characterizing the set of models under consideration. A loss function $L(\theta, a)$, $a \in \mathcal{A}$ and $\theta \in \Theta$, gives the loss or dis-utility suffered from taking action a when the parameter is θ. In the context of point estimation the set $\mathcal{A}$ represents the set of all relevant estimators, and therefore $L(\theta, a)$ measures the loss incurred when the true parameter is θ and $a(x_1, \ldots, x_n)$ is chosen as an estimator of θ, when the observation of the random variable X is $X = (x_1, \ldots, x_n)$. The risk, or expected loss, of a decision rule a under θ is defined as

$$R(\theta, a) = E_\theta[L(\theta, a(X))]. \tag{2.2.49}$$

An example of a loss function is the *squared error loss*, $L(\theta, a) = (a - \theta)^2$, the risk associated with this loss function is the famous *Mean Squared Error*, (MSE), criterion. Since the value of the true parameter, θ, is not known we might like to use an estimator that has a small risk, $R(\theta, a)$, for all possible values of θ. Therefore we expect between two estimators a_1 and a_2, if $R(\theta, a_1) \leq R(\theta, a_2)$ for all $\theta \in \Theta$ and inequality is strict for some θ then estimator a_1 is preferred to the estimator a_2.

[4] For a introductory treatment of decision theory see "Theory of Point Estimation" by E.L. Lehmann and G. Casella 1998. A more advanced treatment can be found in "Statistical Decision Theory" by S. French and D.R. Insua. For a survey of applications of decision theory in econometrics see "Decision Theory in Econometrics" K. Hirano 2006. Also, "Econometrics and Decision Theory" by Chamberlian 2000

While it seems promising, but except in some very special, and mostly linear, cases it is almost impossible to estimate the risk function. While more studies need to be done to distinguish the appropriate risk functions and ways to estimate them, we still can find other criteria to choose the best estimator. An example is the empirical Bayes implementation.[5]

Suppose, we have a prior belief that $\theta_0 \sim N(\theta_*, \sigma^2 I)$. For any given $a > 0$ the probability that $\theta_0 \in (\theta - a, \theta + a)$ is the greatest, when $\theta = \theta_*$. Therefore, intuitively, we want our estimator be as close as possible to θ_*. Let's call this property, the "interval property". The following estimator achieves this goal.

Definition 10:

Let

$$\lambda^* = \arg\min_{\lambda \in \Lambda} ||\hat{\theta}(\lambda) - \theta_*|| \tag{2.2.50}$$

which Λ is the set of allowable modulators. Estimator $\hat{\theta}(\lambda^*)$, is the estimator with best interval property.

2.3 GENERALIZED LINEAR MODEL (GML) AS AN EXAMPLE

In this subsection we try to apply the previous results and derive the optimal weights for a *generalized linear model*. This class of models include the famous frameworks like log-linear models, logit models, probit models, and many more. For an in depth review of GLM and its applications see McCullagh and Nedler (1990) and James Lindsey (1997). This example is derived from Chen and Cui (2003). Here we, briefly,

[5]For more information about empirical Bayes inference and its applications, including economics applications like "revenue sharing", "insurance rate and risk evaluation" and other applications, see Morris 1983 and references therein.

introduce the general framework of GLMs.

Suppose data $(Y_1, X_1), \ldots, (Y_n, X_n)$ are observed, where $Y_i \in \mathbb{R}$ independent random variables and $X_i \in \mathbb{R}^p$, random variable Y is the response of the random vector X, a GLM specification is the model with following representation

$$E[Y|X] = G(X'\beta) \quad and \quad Var[Y|X] = \sigma^2 V[G(X'\beta)] \tag{2.3.1}$$

where $\beta \in \mathbb{R}^p$ is a vector of parameters, G is a known smooth link function and V is a known variance function. The standard estimation tool in this framework, is the quasi-likelihood (Wedderburn 1974). Let $\mu(\beta) = G(X'\beta)$, the log quasi-likelihood ratio of β is defined as

$$Q\{y; \mu(\beta)\} = \int_y^{\mu(\beta)} \frac{y-u}{V(u)} du \tag{2.3.2}$$

Now suppose that $(x_1, y_1), \ldots, (x_n, y_n)$ be an $i.i.d$ data set and $\mu_i(\beta) = G(X_i'\beta)$. The joint quasi-likelihood ratio of the data is

$$Q(\mu, Y) = \sum_{i=1}^{n} Q(Y_i, \mu_i(\beta)) \tag{2.3.3}$$

differentiating with respect to β and doing some algebra, the quasi-score function can be written as

$$\frac{\partial}{\partial \beta} Q(\mu_i, Y_i) = \frac{Y_i - \mu_i}{V[(\mu_i(\beta))]} \frac{\partial \mu_i}{\partial \beta} \tag{2.3.4}$$

since $E[\frac{\partial}{\partial \beta} Q(\mu, Y)] = 0$, we have

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i(\beta)) G'(X_i'\beta) X_i}{V[(\mu_i(\beta))]} = 0 \tag{2.3.5}$$

the same but more demanding argument will show that[6]

$$\sum_{i=1}^{n} \left(\frac{(Y_i - G(X_i'\beta))^2}{\sigma^4 V[(X_i'\beta)]} - \frac{1}{\sigma^2} \right) = 0 \tag{2.3.6}$$

[6]For a complete derivation see Eric D. Kolaczyk 1994.

To use empirical likelihood we need moment conditions which these two equations can provide it for us. For $i = 1, \ldots, n$, define

$$g_i^1(\beta) = \frac{(Y_i - \mu_i(\beta))G'(X_i'\beta)X_i}{V[(\mu_i(\beta))]} \tag{2.3.7}$$

and

$$g_i^2(\beta, \sigma^2) = \Big(\frac{(Y_i - G(X_i'\beta))^2}{\sigma^4 V[(X_i'\beta)]} - \frac{1}{\sigma^2}\Big) \tag{2.3.8}$$

remembering definition 3, let $g_i(\beta, \sigma^2) = (g_i^1, g_i^2)$, $h_i^1 = (g_i^1, g_i^2)$, $h_i^2 = g_i^2$ and $h_i = (h_i^1, h_i^2)$, now we can define the adapted empirical likelihood for the pair (β, σ^2), given the modulator $w = (w_1, \ldots, w_n)$

$$L(\beta, \sigma^2) = \max_{\{p_i\}_{i=1}^n} \sum_{i=1}^n \log p_i \tag{2.3.9}$$

subject to

$$\sum_{i=1}^n p_i(h_i^1 + h_i^2 w_i) = 0 \quad \textit{and} \quad \sum_{i=1}^n p_i = 1. \tag{2.3.10}$$

The common method to estimate β is to use quasi-likelihood (QL) estimators, MacCullagh and Nedler (1990). It is easy to set up the EL procedure for this problem, because the number constrain is equal to the number of equations. In this case we get $p_i = 1/n$ and the estimator is the same as QL estimator. If we use the procedure introduced in definition 3 we can obtain an estimator which has better variance than QL estimator.

Let $\hat{\beta}_{ql}$ be the estimator obtained by using QL method, and $\hat{\beta}(w)$ is the estimator obtained by the method introduced in this paper for a given w.[7] As we discussed earlier in this paper, we need a criteria in order to choose the best modulator w. Here we compare the variance of $\hat{\beta}(w)$ to the variance of $\hat{\beta}_{ql}$. We try to find a w^* such that

$$\forall w \qquad \Sigma_{\hat{\beta}_{ql}} - \Sigma_{\hat{\beta}(w)} \leq \Sigma_{\hat{\beta}_{ql}} - \Sigma_{\hat{\beta}(w^*)} \tag{2.3.11}$$

[7] Here we keep the two original constrains and add a weighted version of the constrain related to the variance. As we will see this help us to use the data more efficiently, and results in an estimator with reduced variance.

In the other words, we choose w so that $\Sigma_{\hat{\beta}_{ql}} - \Sigma_{\hat{\beta}(w)}$ is maximized.

Remark: For two positive semi definite matrices, A and B, we say $A > B$ if $A - B$ is a positive semi-definite matrix (see "Mathematics for Econometrics" by P. Dhrymes for further discussion).

The following result establishes the desired modulator, or weights depending on the interpretation we might have.

Assumptions:

The following assumptions, which are standard in GLM estimation, are required in the proof of proposition 2

A1: *$G(.)$ is twice continuously differentiable, and $V(.)$ is continuously differentiable.*

A2: *$E[Z_1(\beta,\sigma^2)Z_1'(\beta,\sigma^2)]$ is non-singular.*

A3: *For some $\delta > 0$, $E[||\varepsilon|^2 + ||X||]^{2+\delta} < \infty$, $E[|G'(X'\beta)| + V^{-1} + w]^{2+\delta} < \infty$ and $E[|G''(X'\beta)| + |V'|]^{1+\delta} < \infty$.*

A4: *The matrix $\left(E[\frac{\partial Z_1(\beta,\sigma^2)}{\partial \beta}], E[\frac{\partial Z_1(\beta,\sigma^2)}{\partial \sigma^2}]\right)$ has full rank.*

Proposition 6:

If $E[\varepsilon^3|X] = 0$, $E[\varepsilon^4|X] = \kappa\sigma^4 V^2$ for some $\kappa > 1$ and $Cov(\frac{V'GX}{V}) > 0$, then the optimal weights so that maximize $\Sigma_{\hat{\beta}_{ql}} - \Sigma_{\hat{\beta}(w)}$ is

$$w^*(X'\beta, X) = w_i^* = \frac{V'[G(X_i'\beta)]G'(X_i'\beta)X_i}{V[G(X_i'\beta)]} \tag{2.3.12}$$

Notice that, V' and G' are the first order derivatives for V and G, and X_i' is the matrix transpose of X_i.

Proof. See the appendix. □

2.4 IMPLEMENTATION AND MONTE CARLO SIMULATIONS

In this section, to evaluate the methods developed in the previous parts, we design and preform sets of Monte Carlo simulations. At the moment, from five simulation problems, which I am working on, I will only report two of them. This is both for keeping this paper in an acceptable size range, and some technical difficulties with some of the other simulations. Therefore, I consider this part as an incomplete section, and I am working to complete it by designing viable algorithms. The main computational problem is optimization with respect to the modulators. The lack of closed from solution in most cases, makes this optimization a very computationally intensive procedure. Although this is a very big draw back, but one can argue that the modulation, or weighting, has better efficiency than the unweighted EL only when the sample size is small. Therefore there is not much gain from applying the method of weighting when the sample size is large, because both methods are asymptotically equivalent. Therefore, the hope is that, when the sample size is small the optimization with respect to the modulation would work. In this section, I present the result obtained from two simulations.

2.4.1 GLM Estimation

Here we report some simulation results, using the GLM model

$$Y_i = G(X_i'\beta) + \sigma V\big(G(X'i\beta)\big)\varepsilon_i \tag{2.4.1}$$

where $\varepsilon_i \sim N(0,1)$, $V(t) = t$, and $G(t) = \log(X_i'\beta)$. The parameters used in this model are $\beta_1 = 0.5$, $\beta_2 = 1$, and $\sigma^2 = 0.25$. $X_i = (X_{i1}, X_{i2})'$ are generated from uniform distribution on $[0\ \ 2] \times [0\ \ 2]$. I have used the quasi-likelihood procedure of R, which is the

main tool of estimating GLM models in R, to obtain $(\beta_{ql}, \sigma_{ql})$. To derive the weighted EL estimation we use the usual empirical likelihood procedure augmented with the optimal weights obtained from proposition 2. Table 1 summarizes the quasi-likelihood estimation results, and table 2 summarizes the results obtained from weighted EL. As these results suggest there is a sensible improvement, though small, in the variance of WEL estimator compared with QL estimator, and very important, this improvement comes at no cost in bias.

Table 4: Standard Error(SD) and Bias of the QL Estimator

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\sigma}$	
Sample Size	SD	Bias	SD	Bias	SD	Bias
40	0.41	0.053	0.49	0.057	0.10	0.0041
60	0.38	0.032	0.45	0.039	0.086	0.0035
100	0.32	0.025	0.35	0.028	0.051	0.0020
200	0.24	0.013	0.24	0.010	0.035	0.0010

2.4.2 Heteroskedastic Data

The data set, (Y_i, X_i), for this experiment is generated from

$$Y_i = \beta_1 X_i + \beta_2 X_i^2 + |X_i|^{1/2} \varepsilon_i. \tag{2.4.2}$$

where $\varepsilon_i \sim N(0,1)$, drawn $i.i.d$. To generate X, we draw X_i from the $N(1,1)$ distribution. The moment condition is $E[g(\beta_1, \beta_2, X_i)] = 0$ where

$$g(\beta_1, \beta_2, X_i) = \frac{Y_i - \beta_1 X_i + \beta_2 X_i^2}{|X_i|^{1/2}}. \tag{2.4.3}$$

Table 5: Standard Error(SD) and Bias of the WQL Estimator

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\sigma}$	
Sample Size	SD	Bias	SD	Bias	SD	Bias
40	0.38	0.046	0.44	0.054	0.095	0.0045
60	0.35	0.030	0.41	0.037	0.085	0.0036
100	0.30	0.025	0.34	0.028	0.051	0.0021
200	0.23	0.013	0.23	0.010	0.035	0.0010

We obtain the empirical likelihood estimate of two parameters β_1 and β_2. We compare these estimates with an estimator in which the weighting vector $w = (w_1, \ldots, w_n)$ with $w_i = 1/|X_i|$, is used alongside the empirical likelihood estimate, as it was describe in section 2. These weights are driven from the same argument as optimum weight are obtained in *GLE*, the information coming from densities with higher variances should weighted less compare to information coming from densities with lower variances. This was we avoid optimizing the objective function with respect to the weighting vector. The results indicate that, when the sample size is small, the weighting helps to improve both the quality of the estimator in one hand and the tests biased on weighted EL ratio are more reliable than tests based on the usual EL ratios. Table one compares the bias property of the two estimator and in table two we compare the tests based on EL ratio and weighted EL ratio. As we expect the importance of weighting drops as the sample size grows.

Table 2 summarizes the probability of rejecting the null hypothesis $H_0 : \beta_1 = 1$ and

Table 6: Bias comparison of the EL and EL using Weights

	Estimated bias for $\hat{\beta}_1$ in %		Estimated bias for $\hat{\beta}_2$ in %	
Sample Size	EL Method	W-EL Method	EL Method	W-EL Method
10	56.0	51.8	48.2	41.9
20	41.7	35.4	39.0	37.4
50	26.4	25.6	21.2	20.8
100	18.6	18.2	14.0	14.0

$H_0 : \beta_2 = 1$ at the normal 95% confidence level. It is interesting to see that, when in computing empirical likelihood ratio, heteroskedasticity is accounted for, test statistics are more reliable.

2.5 CONCLUSIONS

This, and a companion paper (Shahidi 2008), investigate the use of shrinkage methods in empirical likelihood framework. We introduce two of the most widely used of these methods, adaptation and penalization, and then extend the empirical likelihood procedure to encompass these methods. Shrinkage methods not only help to improve the EL estimator, but also can be used to regularize some ill-posed inference problems. We define modulation and use it to construct adapted empirical likelihood procedure. This estimator can be regarded as a weighting method which weight the data points ac-

Table 7: Bias comparison of the EL and EL using Weights

	Probability of rejecting $\beta_1 = 1$		Probability of rejecting $\beta_2 = 1$	
Sample Size	EL Method	W-EL Method	EL Method	W-EL Method
10	0.13	0.06	0.12	0.06
20	0.08	0.06	0.06	0.06
50	0.06	0.05	0.06	0.05
100	0.05	0.05	0.05	0.05

cording to their importance. We see that this is a very useful tool when we are dealing with a small sample heteroskedastic data set. simulation results confirm the superiority of our proposed estimator to the plain empirical likelihood estimator. Also, in the presence of heteroskedasticity, specially in small samples, the test statistics based on adapted empirical likelihood ratios are more reliable than their EL ratio counterparts.

While modulation method in theory improves the empirical likelihood estimator, the computation difficulties limit its usefulness to very special cases. We have studied the modulation method in a generalized linear model framework, in which the Monte Carlo simulations suggest promising results. For future studies, I plan to develop and design more efficient algorithms to implement the estimators introduced in this paper. Another area which needs more study, is the risk function estimation. Choosing and estimating an appropriate risk function is the subject of statistical decision theory, and is usually a hard problem. Choosing and estimating an appropriate risk function is the key to practical use of some of the results presented in this paper.

3.0 CELEBRITY EFFECTS: HOW FAMOUS TRADERS IMPACT THE FINANCIAL MARKET

3.1 INTRODUCTION

The actions and opinions of celebrities in particular, and public opinion leaders in general, have a special effect on their fans and on the society they live in. Indeed, attempts have been made to benefit from the popularity of these celebrities. These days we see more and more celebrities becoming candidates for political offices, while many politicians try to get endorsement from athletes and other kinds of celebrities. For example, It is now acceptable for a serious candidate for a high electoral office to submit to interviews by celebrities such as David Letterman, Jay Leno, and Jon Stewart on their daily late night talk shows.

At the same time celebrities are becoming more aware of the power they have and try to use it more often. Actors and musicians, in increasing numbers, are endorsing and campaigning for candidates and making political statements with the obvious goal of influencing the opinions and the behavior of their fans. To mention but a few, we can name the U.N. celebrity diplomacy, and Bono's involvement in raising aid money for poverty reduction and health care initiatives in Africa and so on. In a nutshell, all of these increased activities by celebrities and their fans in political and public life suggest that celebrity endorsements have the ability to make certain statements more

palatable while increasing the level of agreement for already popular opinions.[1]

Aside from the realm of politics and public opinion, celebrity endorsement is a big business in the marketing industry. Advertisement campaigns have been paying great sums of money to celebrities to endorse, or even just to use, their products. The best sign that these kinds of endorsements are beneficial is the amount of money that companies spend on celebrity endorsement, a practice that shows no sign of slowing down. For instance, in Forbes magazine's (2004) lists of the top 100 celebrities Golfer Tiger Woods, ranks number 3 and has a $105 million dollar contract with Nike. "Several studies have examined consumers response to celebrity endorsements in advertising, findings show that celebrities make advertising believable." (Jagdish & Wagner 1995) and "advertising uses celebrities as pioneers in order to dictate trends". Also, studies have shown positive relationships between the stock price and the usage of celebrity endorsement in the advertising strategies of a company.[2]

One of the questions which I try to answer in this paper is the effect of imitation in financial markets. In other words, is the price mechanism in stock and other financial markets able to convey information efficiently in such a way that diminishes the celebrity status of famous traders? Numerous cases can be mentioned as evidence that prices lack such ability. For example on Wednesday September 16, 1992, a day that is remembered as *Black Wednesday,* George Soros almost single-handedly forced the British government of the day to abandon the European Exchange Rate Mechanism. Besides yielding him almost one billion US dollars, this incident hugely enhanced his

[1] Another example of the effect that celebrities' actions and behaviors can have on the society they live, even when there is no intention of having that effect, is the former first lady Nancy Reagan's mastectomy, instead of breast-conserving surgery in October 1987.Studies show that compared to women undergoing surgery for breast cancer in the third quarter of 1987-just prior to the Mrs. Reagan's surgery-wo men were 25 percent less likely to undergo BCS in the fourth quarter of 1987 and first quarter of 1988. In subsequent quarters the rate returns to the base line. (JAMA 1998)

[2] For example see "Srivastava et al" Journal of Marketing 1998 and the references therein.

reputation too, so that in April 1993, when he bought around 3 million ounces of gold at $ 345 per ounce and invested $ 400 million in Newmont Mining-a gold mining company, as soon as the traders learned of Soros' purchase, gold rose $ 5 after a long period of decline, a trend that continued to 1996 and lifted the price of gold to $ 405. His investment in British real estate, which subsequently skyrocketed the price of real estate, and the Malaysian prime minister's accusation that Gorge Soros has ruined the East Asian economies-in reference to the 1997 crisis in East Asia - are other examples of how much influence a single trader can have on other traders' behavior and subsequently the market as a whole.

More recently, after the market crash of 2000, the United States Congress held hearings entitled "Analyzing the Analyst" aimed at addressing stock analysts and their recommendations, suggesting that words and recommendations can have a huge impact on the behavior of other participants. Also in 2002 the NYSE and NASDAQ issued new regulations, which were primarily aimed at the top ten investment banks, usually called big tens, to curb the conflicting interests on the analysis and recommendations issued by the big banks and famous analysts. Some even suggested that there has been a conspiracy to push the market up by frequently issuing very positive recommendations. Titles like "Wall Street treachery: leading the lambs to the slaughter" or "The betrayed investors: American bought to the idea that stocks would only make them richer" (both from Business Week) suggest a more intentional misleading.

The question we intend to ask and try to answer in this paper is: what mechanism causes the agents acting in an economical environment to follow the "*popular figures*"? The argument made by Banerjee (1992) and Bikhchandani et al (1992), from now on BHW, shows that herding is not necessarily an irrational phenomenon. These papers argue that, if people act in sequence and observe the actions of their predecessors without accessing the actual information received by them, the information

contained in the history of actions eventually will overwhelm the private information of every agent forcing them to abandon their own private information and follow the actions of their predecessors. BHW also argue that their model can be a base for understanding the uniformity of social behaviors and the creation of norms and fashions. Avery and Zemsky (1998) have shown that while it might be the case when the cost of choosing different actions is fixed, the argument breaks down in the presence of an adjustable price. Therefore the price mechanism in financial markets will adjust in such a way that every participant will be better off following his own private signal. They show that in order for herding to happen we need what they call multidimensional uncertainty.

While Avery & Zemsky (1998) suggest that informational herding is a very rare phenomenon, other sources of herd behavior might still exist. There is a large literature in reputation-based herding. Scharfstein and Stein (1990), Trumen (1994), Zweibel (1995), Graham (1999) and others provide another theory of herding in financial market based on the reputational concerns of fund managers or analysts.[3]

In this study, I will try to expand the BHW model based on the central idea that not all agents in an economic or social environment carry the same weight when it comes to influencing other peoples' actions. Although some agents have the ability to reach out to a larger portion of the population, and their actions are highly influential, there are other agents-the majority of agents-where their actions go largely unnoticed and they don't have any influence on other's opinion and actions.

The contribution of this paper is two fold. First, I extend the BWH model to include agents with "celebrity status", providing a potential framework to study and design different advertising policies. Using this framework, we can better understand

[3]For a survey of herding in financial market see "Herd Behavior in Financial Markets" by Bikhchandani and Sharma.

the disproportional effects of celebrities statements, and the ability of famous traders in financial and other markets to influence market activities. I believe there is a large host of social, political, and economical phenomena which fit in this framework. Therefore, our model in this paper, can be a good starting point to study these phenomena. The second contribution is providing a framework to help understand how some bubbles form and burst, and what role major traders have in creating them.

The remainder of this paper proceeds as follows: In section 2, we construct a model to incorporate the notion of celebrity or what we will call *"The Star"* agent. There, we study the model and its implications. In section 3, as an example, we study a model of the stock market in which there is a star trader. This model will be similar to the model used by Avery-Zemsky (1998). The main difference is that, we use the model of herding developed in this paper instead of BHW. Also, in section 3, we will show that the star trader has a limited ability to pull the market in her/his direction. Section 4 concludes the paper. All proofs are collected in the appendix.

3.2 THE MODEL

In this paper, we assume that an individual can only see the actions of his or her predecessors. The crucial point here is that the agents cannot observe the actual signals of their predecessors. If they were able to do so, then the pieces of information available to individuals would effectively aggregate and talking about the effect of somebody's action on somebody else's behavior wouldn't make much sense. Because agents can't observe their predecessor's signals, it is possible that they believe some of their predecessors had access to better information. This helps towards the rise of some of those predecessors to the "star" status.

3.2.1 A Simple Model

1. There is a sequence of exogenously ordered individuals, each deciding to adopt or reject some action based on the information they have, and in order to maximize their value. If the information they have cannot distinguish between the two alternatives, they chose to adopt with probability $1/2$.
2. Each individual observes the decisions of all those ahead of him.
3. All individuals have the same cost of adopting, c. For simplicity, we assume $c = 1/2$. The gain of adopting, V, is also the same for everyone. Again, for simplicity, we assume V is 0 or 1 with equal probability.
4. Each individual privately observes a conditionally independent signal about the true value, V. Each individual i's signal is either H or L. H is observed with probability $p_i > 1/2$ if the true value is 1 and, likewise, L is observed with probability $p_i > 1/2$ if the true value is zero. Again, for simplicity we assume that

$$p_i = p, \qquad \forall i. \tag{3.2.1}$$

5. There is a special individual, whom we call "star", such that when he acts a portion of other agents, who are distributed randomly between the whole population of agents, will view his decision as more informative than the decisions of other agents, including their own signal. This randomly distributed part of the population who consider the actions of the star to be more informative, or in fact more influentially, are called *"fans"*.[4]

[4]We notice that there is no assumption indicating that the "star" has indeed access to better information nor that his signal is more accurate than others. Although it might be the case in the real world that famous people have such information, fans, anyway, frequently put too much weight on the star's actions. This model can be considered an attempt to capture such over reactions by the fans.

6. To clearly define the difference between fans and non-fans we have to consider two different probability measures according to which they associate different probabilities to the same event. Suppose that the "star" appears at time t; if a fan acts at time $t+1$ he assigns

$$P_f(V=1)=\pi^*, \tag{3.2.2}$$

as the probability while if a non-fan acts at the time $t+1$ he assigns

$$P_{nf}(V=1)=\pi \tag{3.2.3}$$

such that $\pi^* > \pi$.

If H_t is the history of actions up until time t, and h_t^* is the piece of information at time t capturing the star's action, we can interpret π_t^* and π_t as $P_f(V=1|H_{t-1},h_t^*)$ and $P_{nf}(V=1|H_t)$.[5] For further simplicity, we assume that the "star" enters at $t=0$ and therefore we set $t=0$ to obtain

$$P_f(V=1)=\pi^* \qquad \textit{and} \qquad P_{nf}(V=1)=\pi, \tag{3.2.4}$$

such that $\pi^* > \pi$, right after the star's entry.[6]

7. We assume that the population of agents is a continuum and every agent has a label in [0,1]. A portion of this population accounts for the fans and the set of labels corresponding to fans is of measure μ. To choose an agent at time t, a random number, r, will be chosen from a uniform distribution on $[0,1]$. If $r<\mu$, then a fan is chosen, otherwise a non-fan. The law of large numbers guarantees that in each date, t, the probability that a fan is chosen is μ.

[5] Note that for the non-fan we have $H_t=\{H_{t-1},h_t^*\}$. However, this doesn't hold for the fans' information sets.

[6] With this assumption π is the signal accuracy p, and we can calculate $P_f(V=1|H_t^*)$ and $P_{nf}(V=1|H_t)$ for every subsequent t using Bayes' rule.

Note: We assume that the agents don't take into account the presence of the other fans. If they were to do so, it will make the inferences intractable.

3.2.2 Some Observations:

In this subsection I mention some of the results that can be derived from the model which was introduced above.

First, we define a naive fan:

Definition 11:

A naive fan is a fan who thinks every other fan is following her/his own signal. In other words, a naive fan doesn't take to consideration the possibility that previous agents might be herding. When we talk about fans we mean this naive kind of fans except if we state it otherwise.

Second, using Bayes' rule we define the belief update operator f by

$$f(x) = \frac{(1-p)x}{(1-p)x + p(1-x)}, \tag{3.2.5}$$

and define n to be

$$n = min\{m | f^m(\pi) \leq 1/2\} \tag{3.2.6}$$

. As proposition 1 will show, this number will help us to transfer a fan's belief in a star to the number of signals opposing the star's choice that are needed in order for this fan to "abandon the star". Note that $0 < \pi < 1$ and $f^k(x)$ is a k times composition of f with itself. In the following lemma we show that *n* indeed exists and is finite. We will also explore some properties of f that will be used later in this paper.

Lemma 3:

For any $0 < \pi < 1$*, n exists and is finite. Furthermore n increases with an increase in* π.

Proof. see the appendix. □

When is a fan ready to abandon the star and instead, use his own information? A fan who favors the star would like to follow her, but if he keeps getting signals indicating that others are receiving information suggesting the star is wrong, the fan will reach a point in which he finds the accumulated evidence compelling enough to abandon the star and choose a different action instead. The following result, which can be proven using lemma 1 formalizes this intuition.

Proposition 7:

Define n^* to be

$$n^* = min\{m| f^m(\pi_0^*) \leq 1/2\}.$$

At least n^* consecutive opposing actions to the star's action are needed for a fan to abandon the star.

Proof. Without loss of generality, suppose that the star acts at $t = 0$ and, therefore, $P_f(V = 1|H_0) = \pi^*$. If the fan receives a negative signal then he will update his beliefs to

$$P_f(V = 1|h^*, x = 0) = \frac{(1-p)\pi^*}{(1-p)\pi^* + p(1-\pi^*)} = f(\pi^*). \tag{3.2.7}$$

In addition,

$$\begin{aligned} &P_f(V = 1|h^*, x_1 = 0, x_2 = 0, \ldots x_n = 0) = \\ &P_f((V = 1|h^*, x_1 = 0, x_2 = 0, \ldots x_{n-1} = 0), x_n = 0) \\ &= f(f^{n-1}(\pi^*)). \end{aligned} \tag{3.2.8}$$

We also have $E_f[V|H_t] = P_f(V = 1|H_t)$. Thus, the fan follows the star as long as $P_f(V = 1|H_t) > 1/2$. This implies that a fan abandons the star if

$$P_f(V = 1|H_t) = f^n(\pi^*) < 1/2.$$

When $f^n(\pi^*) = 1/2$, the fan abandons the star with probability $1/2$. □

Using proposition 1 we can construct a simple optimal decision rule. This decision rule is the basis for proposition 2, which greatly enhances our understanding of this model and simplifies the calculations.

Let a be the number of predecessors who have adopted and r the number of those predecessors who have rejected and set $d = a - r$. We have the following optimal decision rule for a fan:

If n is the number obtained from proposition 1, the star has adopted, and $d > -n$ then a fan should adopt regardless of his private signal. If $d = -n$, the fan should adopt if the private signal is high and otherwise reject with probability 1/2. If $d < -n$, the fan should reject regardless of his private signal. Similarly, for a non-fan we have the following rule. If $d > 1$, the non-fan should adopt regardless of his private signal. If $d = 1$, the agent should adopt if the private signal is high and reject with probability 1/2 if the private signal is low. If $d < 1$, should reject regardless of his private signal.

If we define s_n to be the state in which $d = n$ and let S to be the set of all such s_n we have the following proposition:

Proposition 8:

The subsequent actions of the agents entering after "a star" form a Markov chain which has only two absorbing states : (a) *A cascade in the direction of the star's choice.* (b) *A cascade in the opposite direction of the star's choice.*

Proof. See the appendix for a proof. □

Example 2:

Consider the simplest case in which there is no star in the model, (this is the original model studied in BHW). In this model we have 5 different states. Two of them are

absorbing sates and, therefore, (from Markov chain theory) the process will eventually absorb to one of these states as the number of agents goes to infinity. Here $p = \pi^*$*, which implies:*

$$f(\pi^*) = \frac{(1-p)\pi^*}{(1-p)\pi^* + p(1-\pi^*)} = 1/2.$$

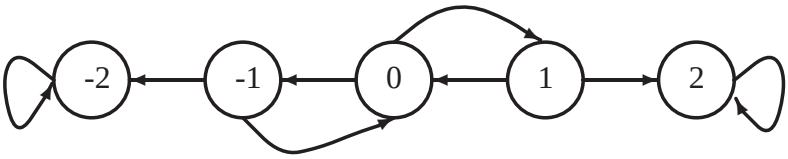
Therefore,

$$n = min\{m | f^m(\pi^*) \leq 1/2\} = 1$$

and

$$S = \{s_1 = -2, s_2 = -1, s_3 = 0, s_4 = 1, s_5 = 2\}.$$

The following figure shows the Markov diagram of the resulting Markov chain.



Example 3:

Now suppose that there is a star in the model who acts in time $t = 0$ *and chooses to adopt. Suppose the fans' initial faith on the star is* $\pi^* = 0.60$*, and the signal accuracy is* $p = 0.56$*. We have:*

$$f(\pi^*) = 0.541 \qquad \textit{and} \qquad f^2(\pi^*) = f(0.541) = .480 < 1/2.$$

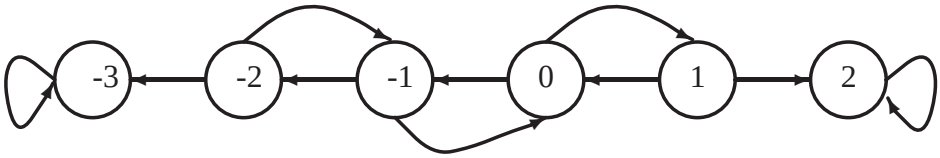
Hence,

$$n = min\{m | f^m(\pi^*) \leq 1/2\} = 2,$$

and

$$S = \{s_1 = -3, s_2 = -2, s_3 = -1, s_4 = 0, s_5 = 1, s_6 = 2\}.$$

We, thus, have a Markov chain with 6 different states. Again, this process has two absorbing states, although the probability of being absorbed to the cascade in the direction of the star's choice (state $s_6 = 2$) is much greater than the probability of being absorbed to the cascade in the opposite direction of the star's choice (state $s_1 = -3$). The following figure shows the Markov diagram associated with this Markov chain:



3.2.3 Fragility:

In a model without stars any kind of cascade is very fragile. Indeed, when participants in such an environment find themselves in a cascade, they can realize that the cascade is based on little information. For example, in an up cascade, where everybody adopts, they know for sure that the first person had a high signal and there is a probability of $1/2$ that the second actor also have had a high signal. Now if one agent gets a low signal plus another piece[7] of negative information, she will be in a position where her own private signal is more informative than the information that comes from observing their predecessors' actions.

This fragility is somehow counter intuitive, in the sense that it suggests that after the appearance of the first signs of a problem with an existing norm, tradition, or fashion, the public will abandon it and the participants will start to use their own private information. This is, off course, somewhat different from what we observe in reality,

[7] For instance suppose this particular agent gets two signals vs. others who just get one.

where it is usually hard to break an established norm or social tradition. Some even suggest that the biggest obstacle in some developing societies are certain existing and traditions and convincing the members of those societies to abandon them. Although many people in those societies understand the devastating consequences of their traditions and social norms, it is still difficult to convince the population to change their "old ways". This study suggests that we should at least investigate for the role of stars, opinion leaders, and so forth, in order to understand the rigidity and of some of these norms.

In the presence of a star, any cascade which favors her choice is not so fragile and will resist defections, although a cascade that is not in her direction will be equally fragile as in the model without the star. Let's first clearly define what we mean when we say that a cascade is broken.

Definition 12:

i) A cascade has been broken at time t if and only if the actor at time $t+1$ *ignores the cascade and follows her own information.*

ii) A defection from a cascade at time t is successful if it breaks the cascade at time $t+k$.

If we assume that after time t, which is after the emergence of a cascade in the star's direction, every participant receivers a signal opposing the star; then the following proposition applies with regards to the fragility of this cascade.

Proposition 9:

Suppose that after a defection at time t, every other agent in subsequent times receives a negative signal (a signal pointing to the opposite direction of the ongoing cascade). The probability that the defection at time t will be successful is $(1-\mu)^{(n-1)}$ *where n is the same as in proposition 1.*

Proof. See the appendix. □

So far, we have assumed that all fans are *"naive"*. If we drop this assumption and assume that fans take to account the possibility of herding by their predecessors. We formally call this kind of agents *"sophisticated"* agents. The sophisticated agents will end up following the "star" regardless of their own signal (given they have a strong enough belief in the "star"). This is somehow counter intuitive, since more sophisticated agents are aware of the possibility that the actions of their predecessors might be the result of herding behavior. Still, they end up ignoring all the previous information. Formally, we have the following:

Proposition 10:

Assume that the fans are "sophisticated" and let

$$n = min\{m | f^m(\pi_0^*) \leq 1/2\}$$

and $n > 2$. *Then, these fans always follow the "star" regardless of their own signal.*

Proof. See the appendix. □

3.2.4 Possible Extensions

In this subsection I discuss possible extensions of the model we just introduced. The main intuition in the previous model was the fact that not all individuals are equally important, but rather special individuals exist, who have the power to influence others. We can extend this intuition by asking "Is it possible that the actions of all individuals are equally visible by all other participants." I believe the answer to this question is *no*. In most real world cases, not only are the individuals different in their ability to influence other people's decisions and actions, but also they are different in their ability

to reach out to other people. For example, a decision, opinion or action by somebody like myself will most likely go unnoticed by the majority of the population, while actions, opinions or decisions by, say Tom Hanks can catch the eyes of the world. To this end, we can define a network of connections, in the sense that $a \to b$ means that "b will notice a", but not *vise versa*. A natural definition of a star in this framework is as the agent who can be observed by every (or a large portion) of the other agents.

Other possibilities like a system with two or more stars or even opposing stars can be exploited as well. What is the dynamic of behaviors in a polarized society in which two opposing stars have their fans and "anti-fans" and what role do the independents play in such a society? I will not study these issues here. However, in the next section I will use a very simple network to study the mis-pricing of a stock in a simple financial market. I will show that, under special circumstances, mispricing and bubbles can occur. Furthermore, rational traders won't be able to realize or correct such phenomena.

3.3 AN EXAMPLE: FINANCIAL MARKETS

In this section we study a simple market with one asset. We will see that when there are enough traders, like individual investors who don't necessarily have much skills or knowledge about the market, who are ready to trust the star investors and follow them, two different probability measures will emerge which deter the ability of prices to convey information efficiently to prevent bubbles.

3.3.1 Stating the problem

Although rational approaches to asset pricing have been considerably successful, it is hard to believe that imitative behavior in such markets are totally erased. In fact, there has been a resurgence of interest in the study of such behavior in recent years, with behavioral finance gaining popularity. In this section we first illustrate the idea using an example derived in part from Avery-Zemsky. First, I will show why in a BHW framework, rationality prevents herd behavior. I will then use the framework built in section 2 of this paper to investigate a market in which there is a star investor who is noticed by everyone, and where the normal investors (non stars) believe that the information from the star investor is more accurate than their own. We will investigate how herd behavior becomes a possibility under these conditions. Furthermore, we believe that these conditions are not plausible. For example, there have been times when big investment firms issued positive recommendations on stocks, thereby causing the mass of inexperienced or even experienced investors to buy and push the price of the stock very high. If we interpret the combination of these investment banks as "the star investor", we believe the model introduce in the section 2 of this paper can be used to understand such issues.

3.3.2 A Simple Example

First, let's review the original BHW model in light of this example. Agents face a choice of whether or not to adopt a new technology. The cost of adoption is $c = 1/2$. The value of the new technology is V, which is either 1 or 0 with equal probability. Each agent gets an independent, but not perfect, signal abut V, denoted by x, $x \in \{0,1\}$, where $P(x = V) = p > 1/2$. Agents act sequentially and observe H_t, the history of actions up until time t. Let $\pi_1^t = P(V = 1|H_t)$. The choice made by an agent depends

on whether the expected value for adopting is greater or less than c.

The expected value of an agent with bad news at time t is:

$$V^t(x=0) = E[V|x=0,H_t] = P(V=1|x=0,H_t) \tag{3.3.1}$$

$$= \frac{(1-p)\pi_1^t}{(1-p)\pi_1^t + p(1-\pi_1^t)}.$$

The expected value for an agent with good news at time t is:

$$V^t(x=1) = E[V|x=1,H_t] = P(V=1|x=1,H_t) \tag{3.3.2}$$

$$= \frac{p\pi_1^t}{p\pi_1^t + (1-p)(1-\pi_1^t)}.$$

Therefore π_1^t increases in the difference between the number of prior agents who adopted and those who did not. When there are two more adopters than non-adopters we will have $\pi_1^t > p$, which implies

$$V^t(x=1) > V^t(x=0) > 1/2. \tag{3.3.3}$$

In this situation every agent who acts at time t will adopt regardless of his signal, in the words of BHW an informational cascade will arise.

In financial markets the price mechanism suppresses this imitative effect and prevents the cascades from occurring. To see how, suppose that in the above example, agents are traders in a financial market and their choice is whether to buy or to sell a unit of an asset whose value is given by V. Furthermore, suppose that the financial market is informationally efficient, which implies that the price reflects all publicly available information (here we interpret the cost in previous examples as being the

price of the asset). Therefore, unlike the previous case, here, the cost will adjust when new information arrives. More precisely, we have :

$$\hat{c} = E[V|H_t] = P(V = 1|H_t) = \pi_1^t, \tag{3.3.4}$$

which implies:

$$V^t(x = 1) > \hat{c} > V_t(x = 0). \tag{3.3.5}$$

Therefore, an agent with good news will buy while an agent with bad news will not adopt (in this case buy) and, thus, no herding occurs.

Now suppose a competitive group of market makers, or equivalently a market maker who makes zero profit, determine the prices, by setting bids and asks prices as

$$B^t = E[V|h_t = S, H_t], \tag{3.3.6a}$$

and

$$A^t = E[V|h_t = B, H_t]. \tag{3.3.6b}$$

Here, S stands for selling orders and B stands for buying ones. We only analyze the buying activities (selling is similar). Therefore, we focus our attention on the prices at which the agents are willing to buy the asset. (See Lawrence Glosten and Milgrom (1985)). Suppose that there is a star investor such that his decisions are observed by all other investors. There are also regular investors (non-stars) who do not observe each others decisions. These assumption have been made to simplify the calculations and the computer simulations we perform. We also assume that all regular investors consider the actions of the star investor to be more informative than their own, and that the star investor enters at the beginning. Every buyer receives a private signal $x \in \{0, 1\}$, s.t. $P(x = V) > 1/2$. Suppose that the prior probability of $V = \{0, 1\}$ is

$P(V=1)=P(V=0)$. Given this information, we can find the probability of the value being equal to one if the star investor buys.

$$P(V=1|h_s=B) \tag{3.3.7}$$

$$=\frac{P(h_s=B|V=1)P(V=1)}{P(h_s=B|V=1)P(V=1)+P(h_s=B|V=0)P(V=0)}$$

$$=\frac{P(h_s=B|V=1)}{P(h_s=B|V=0)+P(h_s=B|V=1)}$$

$$=P(h_s=B|V=1)=\pi_1.$$

Here, π_1 is the probability that $V=1$ if the star investor buys. We have assumed that $\pi_1 > p$, which implies that other agents consider the star's information more accurate.

Now, suppose that at time $t=0$ the star investor buys. The market marker will set the price for $t=1$ to be

$$V_1^m = E_m[V|h_0=B] = P(V=1|h_0=B) = p. \tag{3.3.8}$$

At the same time, a fan buyer who gets a negative signal at time $t=1$ will evaluate the price as:

$$V_1^A = E[V|h_s=B,x=0] = P(V=1|h_s=B,x=0) \tag{3.3.9}$$

$$=\frac{(1-p)\pi_1}{(1-p)\pi_1+p(1-\pi_1)}=\pi_2$$

Now, if $\pi_2 > p$, the fan investor will buy despite receiving a negative signal. The important observation is that this situation can indeed happen. Figure 1 describes a simulation with $p=.52$, $\pi_1=.75$, and $\pi_1=P(V=1|h_s=B)$. The probability that

the fan investor initially assigns to the event that $V = 1$ when he sees the action of the star is assumed to be $\pi_1 > p$. As we see it takes a while (6 periods in this case) for the agents with negative signals to stop buying.

To illustrate this point better we repeat the process one more period. Now suppose that at time $t = 2$ the agent whose turn is to act again receives a negative signal ($x = 0$). The market marker will set the price:

$$V_2^m = E[V|h_0 = B, h_1 = B] = \frac{pV_1^m}{pV_1^m + (1-p)(1-V_1^m)}. \tag{3.3.10}$$

While the agent's value is:

$$V_2^A = \frac{(1-p)\pi_2}{(1-p)\pi_2 + p(1-\pi_2)}. \tag{3.3.11}$$

Again, if $V_2^A > V_2^m$, the agent will buy even thought he has a negative signal. Thus, herding can happen in this situation. However, it will be short lived. The important point to notice is that the market maker and agents use two different measures for evaluating the relevant probabilities. [8]

3.3.3 A General Model

Here,we consider a more general model in which the market is for just a single asset with true value V in such a way that $V \in \{0,1\}$. Like the example we studied above, prices are set by a competitive market maker who interacts with an infinite sequence of individual traders who are chosen from a continuum population. This assumption guarantees that no trader appears in the sequence more than one time. Thus, we need not to worry about strategic considerations. Each trader is risk neutral and has the

[8] We conjecture that the price that market marker sets is still a martingale with respect to the market maker's measure. This is intuitively obvious since if it was not a martingale, then his assessment of V_t would be systematically mistaken in a manner which should be predictable to him.

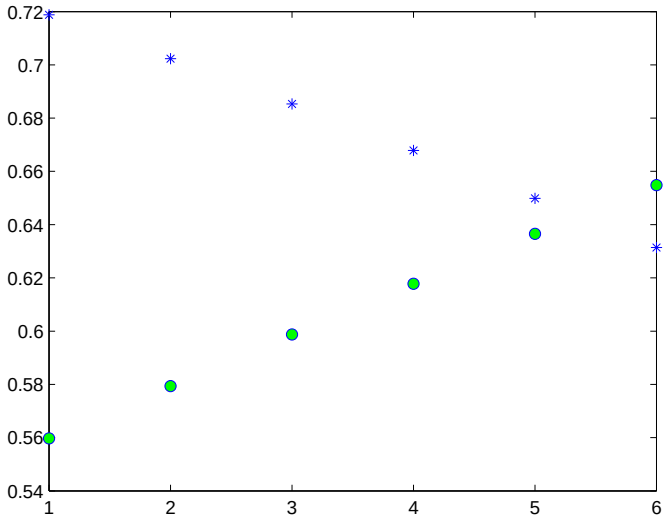


Figure 1: The stars show the prices as they are set by the fan agent. The circles represent the prices set by the market maker. The horizontal axes shows the number of periods.

option to buy, sell, or hold onto one unit of stock. Trades occur at dates $t = 0, 1, 2, \ldots$. The publicly available information up until time t is denoted by H_t and is referred to as "the history of trades up until time t".

There are two classes of traders in our model. Informed traders who receive private information and try to maximize their profit using their private, and public information, H_t. This class divides into two subclasses. "Normal traders" who follow strict Bayesian reasoning without putting any special weight on any particular traders, and "fan traders" who also use Bayesian reasoning, but put more weight on the action of a particular trader who we shall call *the star* trader. The second class of traders are "noise traders" acting for liquidity considerations. [9]

We let $\mu < 1$ denotes the probability of an informed trader arriving at any given time t. Therefore, $1 - \mu$ is the probability of a noise trader arriving. Furthermore, and for further convenience, we assume that noise traders buy, sell, or do nothing, with equal probability: $\lambda = (1 - \mu)/3$.

Finally, there is a special trader whose action is considered more informative by some other traders. We assume that she trades at $t = 0$ and that the portion of traders who "believe in her" is γ.

3.3.3.1 A Definition of Herd Behavior: We want to define herd behavior in such a way that it rules out the situations in which everybody is buying because all have positive signals, or everybody is selling because each trader gets a negative signal. By "herd behavior", we mean a situation in which everybody is ignoring his signal in favor of public information. For instance, a trader is in the buying herd if, based on her private information she should sell the asset, but after observing the public information

[9] In the absence of noise traders, the no-trade theorem of Milgrom-Stocky(1982) applies and the market breaks down.

H_t she decides to buy. We have the following definition.

Definition 13:

A trader with private information, x, engages in herd behavior at time t if he buys when $V_x < V_m < V_{x,H_t}$ *or sells when* $V_x > V_m > V_{x,H_t}$*; and buying (selling) is strictly preferred to other actions.*

3.3.4 Some Observations:

Given the model in the last section, here we investigate whether if mispricing and bubbles can occur. To this ends we define

$$f(x) = \frac{px}{px + (1-p)(1-x)}, \tag{3.3.12}$$

and

$$g(x) = \frac{(1-p)x}{(1-p)x + p(1-x)}. \tag{3.3.13}$$

Let

$$n = min\{m | g(\pi^*) \leq f^m(p)\}. \tag{3.3.14}$$

Then we have the following.

Proposition 11:

Let $\bar{\beta} = f^n(p)$ *and* $\beta = g^n(p)$*, where n is given as above. Then, the size of any bubble is bounded from above by*

$$\delta = | \bar{\beta} - \beta |$$

Another question that arises is that of how long it takes for the price of the asset reach to its highest level. The next proposition attempts to answer this question.

Proposition 12:

Let $\pi^* = P_f(V = 1|H_0)$, $p = P_{nf}(V = 1|H_0)$, *and n taken from proposition 3. Let T denote the time it takes for the price of the asset to reach* δ. *We have the following.*

(a) If $\gamma \leq 1/6 + 1/3\mu$, *then Prob*$(T < \infty) = 1$, *but* $\mathbb{E}[T] = \infty$.

(b) If $\gamma < 1/6 + 1/3\mu$, *then*

$$Prob(T < \infty) = (\frac{\gamma - 1/3\mu + 1/3}{2/3 - \gamma + 1/3\mu})^n < 1. \tag{3.3.15}$$

(c) If $\gamma > 1/6 + 1/3\mu$. *then*

$$\mathbb{E}[T] = \frac{3n}{6\gamma - 2\mu - 1}. \tag{3.3.16}$$

The difference between $f(p)$ and $g(p)$ in proposition 5 is not very large. This implies that $(\bar{\beta} - \beta)$ won't grow too large. Therefore, when π^* (the primary faith of fans on the star) is not too high, the size of any bubble won't grow very large. Furthermore, proposition 6 suggests that it would be difficult for the price to "grow out of control". Additionally, when there are enough traders who don't follow the star, it is almost impossible to obtain a bubble in which the asset is substantially mispriced. The only time that we can expect these kind of bubbles to appear is when fan traders are dominating the market, so that a substantial portion of market participants are positively biased toward the star trader.

I have simulated the model discussed in this section. Figure 1 shows a sample path of the real price as implied by the model. We can observe from figure 2 that there won't be any substantial mispricing when we have enough normal traders to *"time"* the market. However, as figure 3 shows, in times when the fan traders dominate the market, 60% in this case, there is a good chances that we see bubbles particularly in bad times when the actual price should be falling. Both in this paper and in the simulations I have assumed that there is no changes of opinion, and that the fan traders have a fixed biased toward the star. A good exercise would be to alter the model so that in every

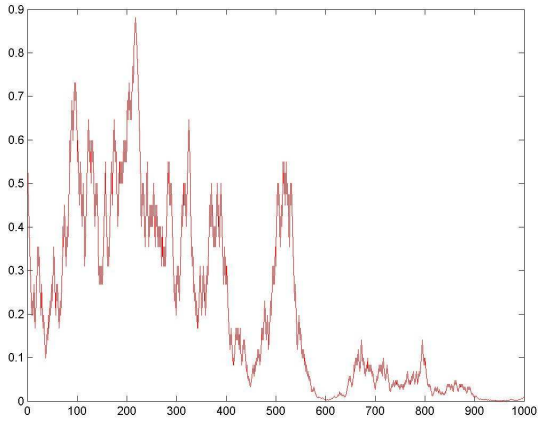


Figure 2: A sample path of the real price as implied by the model

period a participant is assigned a type which indicates whether the participant is a fan, and if she is, how biased she is towards the star. In this case, we can study situations in which the fan traders eventually will alert their trust on the star if the market is not going well in the direction that the star recommends. In order to do so, we need a model for this alternation. In other words, we need a theory that tells us how people alter their beliefs in critical times.[10]

3.3.5 A Possible Extension

In the previous section, we studied a case in which the star appears once at the beginning and, because some of the other agents consider her action to be more informative, they are willing to pay more for the asset than what their own signal recommends. This

[10] If we just assume a random alternation of beliefs, I suspect that we won't get substantially different results from the simulations presented in this section.

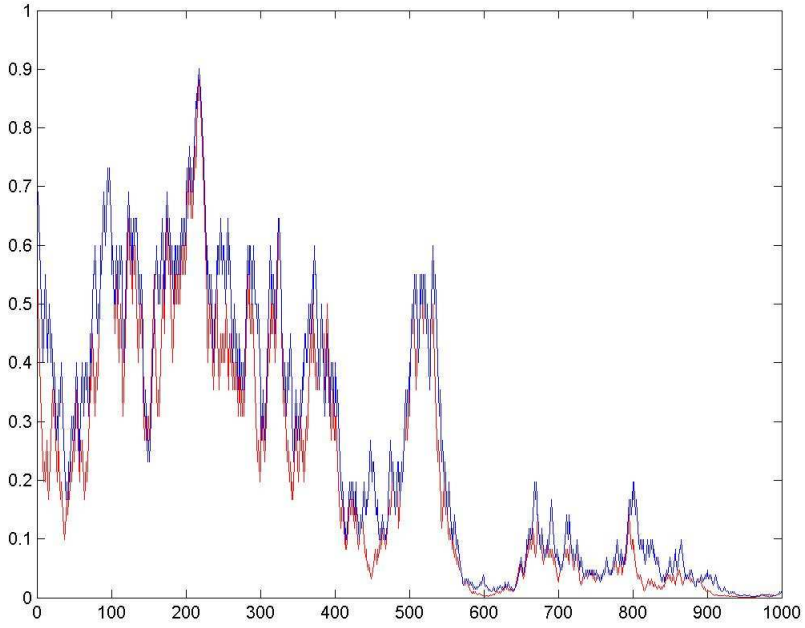


Figure 3: When fans are 30%, noise tarders are 10%, and normal traders are 60% of the total market

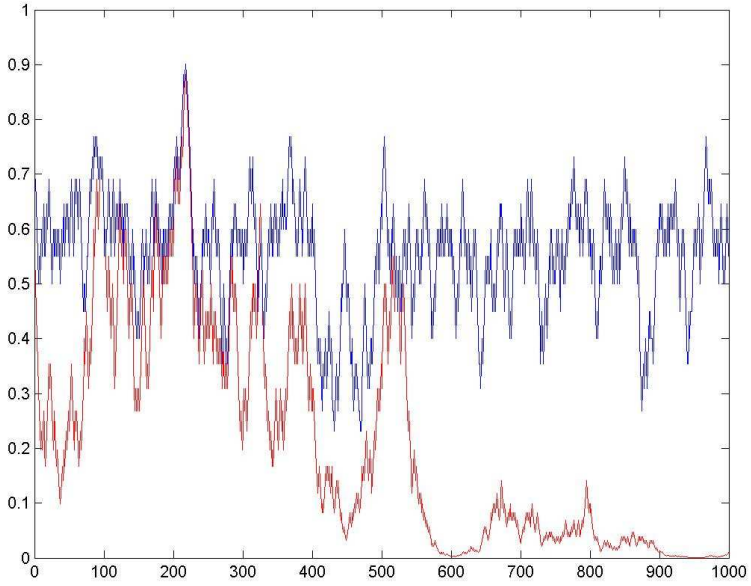


Figure 4: When fans are 60%, noise traders are 10%, and normal traders are 30% of the market

causes the price to be higher and a bubble is created.

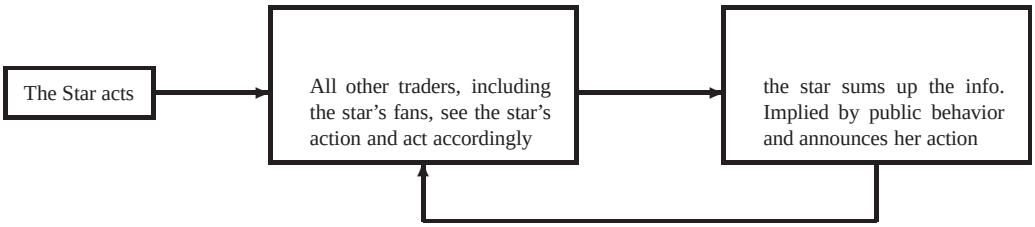
It is worth noting that so far we have not assumed that the star investor has indeed access to special information which gives her the actual ability to make better decisions. While it might be the case in the real world that big investment firms have both better information and better ability to process this information, this model can be taken to suggest that inexperienced traders may exaggerate those abilities and subsequently put more weight on the stars' actions, more weight than the star action actually deserve.

An interesting question arises. What would happen if the star investor in our model can trade more than once? Is it possible that she starts to follow the herd which she herself has helped to create, and if so, what will be the size of a possible bubble created in this manner?

To answer these questions, we assume that, unlike other traders, the star trader can indeed enter the market frequently. Furthermore, we assume that the trust of her fans won't decrease nor increase after each entry. [11]

Now suppose that, for some exogenous reason, the star investor starts following the herd. For instance, we can think of a situation in which the star trader indeed does not get any informative signal, but is just summing up the information which is being revealed by the price and announces her choice to the public. I conjecture that large bubbles can exists in this scenario. This would be an example of a situation in which already publicly available information can have a large impact. Simply because the information is being announced by the star, her fans overreact to that information. The diagram below explains this idea.

[11] In real world cases the trust or belief in the star will change from time to time. Imagine, for example, an investor who follows a recommendation and makes good money. It is quiet possible that next time around he will follow the star's recommendations with more confidence.



3.4 CONCLUSION

In the first part of this paper, we studied cases where the population of agents or a part of that population is positively biased toward a special agent whom we called "the star agent". We showed how in a BHW framework this phenomenon will affect the other agents' behavior, and how imitative behavior can produce herd behavior and informational cascades. In the second part of the paper we showed, that while the market mechanism can prevent herd behavior from happening in a very simple setting, it will fail to do so when the herd behavior is the result of a more complex belief system.

One of the implicit implications of our study is that it suggests that a rise or fall in prices of stocks of big investment banks may have a broader impact on the entire market. This is because, besides the real effects that change in the price of a particular stock might have on the market, a rise or fall in the price of stocks of the investment banks will have the additional effect that the investors who have been following these firms (being fans in our terminology) will revise their belief on the accuracy of the information of these firms. For example, in the case of a price fall, the fan investors might put much less weight on the recommendations given by their star or even re-visit their previous investment decisions which were done in accordance to the actions previously taken by the star, resulting in a further decline. To give a measure of herd

behavior or to determine when herding is happening, is difficult.[12] However, it is possible to measure and test the correlation of stock prices with the movements in the stock price of big financial firms, specially in times of bubbles.

This study also might be able to shed some light on the question of why announcements of already published information sometimes have a substantial effect on the stock prices. Another implication of our study suggest that when there are a lot of inexperienced traders in the market, and the sources who are trusted by the public fail to provide carefully crafted and implied analysis, and instead they themselves are being driven by the public's actions, the probability of crisis is very high.

[12] See Bikhchandani and Sharma (2000) for references.

A.0 PROOFS AND SUPPLEMENTAL MATERIALS FOR CHAPTER 1

Proof of proposition 1:

Proof. First we derive an expression for $\ell_n(\theta)$ and then use that to prove the theorem. using Lagrange multipliers, and setting up the optimization problem we arrive at

$$\mathcal{L}(\theta,\lambda,\mu) = \sum_{i=1}^{n} \log(p_i) - \mu\Big(\sum_{i=1}^{n} p_i - 1\Big) - n\lambda' \sum_{i=1}^{n} p_i g_i(\theta) \tag{A.0.1}$$

doing the optimization we get

$$p_i = \frac{1}{n(1+\lambda' g_i(\theta))} \tag{A.0.2}$$

applying the moment conditions and we have

$$0 = \sum_{i=1}^{n} p_i g_i(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\lambda' g_i(\theta)} g_i(\theta), \tag{A.0.3}$$

because of condition $0 \le p_i \le 1$, it is necessary for λ and θ to satisfy $1+\lambda' g_i(\theta) \ge 1/n$ for each i. For fix θ, let

$$D_\theta = \{\lambda : 1+\lambda' g_i(\theta) \ge 1/n\}, \tag{A.0.4}$$

D_θ is convex and closed, and it is bounded if 0 is inside the convex hull of the $g_i(\theta)'s$. Furthermore

$$\frac{\partial}{\partial\lambda}\Big[-\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\lambda' g_i(\theta)} g_i(\theta)\Big] = -\frac{1}{n}\sum_{i=1}^{n} \frac{g_i(\theta) g_i'(\theta)}{\big(1+\lambda' g_i(\theta)\big)^2} \tag{A.0.5}$$

is negative definite for every λ in D_θ, provided that $\sum_{i=1}^{n} g_i(\theta)g_i'(\theta)$ is positive definite. Therefore, by inverse function theorem, $\lambda = \lambda(\theta)$ is a continuous differentiable function of θ.

Now for every $\theta \in \{\theta : ||\theta - \theta_0|| = n^{-1/3}\}$, let $\theta = \theta_0 + un^{-1/3}$, where $||u|| = 1$. When $E[||g(x,\theta)||^3] < \infty$ and $||\theta_0 - \theta|| \leq n^{-1/3}$ we have

$$\lambda(\theta) = \Big[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i'(\theta)\Big]^{-1}\Big[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)\Big] + o(n^{-1/3}) \quad (a.s) \tag{A.0.6}$$

uniformly around $\theta \in \{\theta : ||\theta - \theta_o|| \leq n^{-1/3}\}$. Doing a Taylor series expansion, and plug in the expression we derived for λ we get uniformly for u

$$\ell_n(\theta) = \sum_{i=1}^{n} \lambda'(\theta)g_i(\theta) - \frac{1}{2}\sum_{i=1}^{n}[\lambda'(\theta)g_i(\theta)]^2 + o(n^{1/3}) \tag{A.0.7}$$

plug in $\lambda(\theta)$, which we calculated before, to this equation we get

$$\ell_n(\theta) = \frac{1}{2}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n} g_i(\theta)\Big]'\Big[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i'(\theta)\Big]^{-1}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n} g_i(\theta)\Big] + o(n^{1/3}) \quad (a.s). \tag{A.0.8}$$

Using

$$W_n(\theta) = \frac{1}{2}\Big[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i'(\theta)\Big]^{-1} \tag{A.0.9}$$

and for large enough n we can rewrite the objective function of the Lasso estimator as

$$\mathcal{L}_n(\theta) = \Big[n^{-1/2}\sum_{i=1}^{n} g_i(\theta)\Big]' W_n(\theta)\Big[n^{-1/2}\sum_{i=1}^{n} g_i(\theta)\Big] + \lambda_n\sum_{j=1}^{n}|\theta_j|^\gamma \tag{A.0.10}$$

Now we can use this expression to prove the proposition 3:

I) First notice that if $\hat{\theta}$ minimizes $\mathcal{L}_n(\theta)$, the it will minimize $n^{-1} \times \mathcal{L}_n(\theta)$ too, and therefore we can choose this object function to work with. We will denote it by $Z_n(\theta)$. First we realize that,

$$n^{-1}\sum_{i=1}^{n} g_i(\theta) = n^{-1}\sum_{i=1}^{n}\big(g_i(\theta) - E[g_i(\theta)]\big) + n^{-1}\sum_{i=1}^{n} E[g_i(\theta)] \tag{A.0.11}$$

under assumption $A2$ we can use a well known result in empirical process theory, see Andrews (1994) and obtain

$$\frac{1}{\sqrt{(n)}}\sum_{i=1}^{n}\big(g_i(\theta)-E[g_i(\theta)]\big)=O_p(1) \tag{A.0.12}$$

furthermore, by assumption $A3-(i)$

$$E\Big[n^{-1}\sum_{i=1}^{n}g_i(\theta)\Big]\xrightarrow{p}m_1(\theta). \tag{A.0.13}$$

Putting all of these together and using assumption $A4$, and the fact that $\frac{\lambda_n}{n}\to\lambda_0\geq 0$, we have

$$Z_n(\theta)=\Big[n^{-1}\sum_{i=1}^{n}g_i(\theta)\Big]'W_n(\theta)\Big[n^{-1}\sum_{i=1}^{n}g_i(\theta)\Big]+\frac{\lambda_n}{n}\sum_{j=1}^{n}|\theta_j|^{\gamma}= \tag{A.0.14}$$

$$\Big[n^{-1}\sum_{i=1}^{n}\big(g_i(\theta)-E[g_i(\theta)]\big)+n^{-1}\sum_{i=1}^{n}E[g_i(\theta)]\Big]'W_n(\theta)\Big[n^{-1}\sum_{i=1}^{n}\big(g_i(\theta)-E[g_i(\theta)]\big)+n^{-1}\sum_{i=1}^{n}E[g_i(\theta)]\Big]$$

$$+\frac{\lambda_n}{n}\sum_{j=1}^{n}|\theta_j|^{\gamma}\xrightarrow{p}m_1(\theta)'W(\theta)m_1(\theta)+\lambda_0\sum_{j=1}^{p}|\theta_j|^{\gamma}=Z(\theta).$$

This finishes the proof of the first part of proposition 3.

II) When $\lambda_n=o(n)$, and all the assumptions are satisfied, uniformly in θ we have $\frac{\lambda_n}{n}\to 0$, when n goes to infinity. Therefore, we have uniformly in θ,

$$Z_n(\theta)\xrightarrow{p}m_1(\theta)'W(\theta)m_1(\theta) \tag{A.0.15}$$

Since by assumption $A3-(ii)$ there exist a unique minimizer for the last expression, using Corollary 3.2.3 of Van der Vaart and Wellner (1996), we have the consistency result:

$$\hat{\theta}_n=\arg\min_{\theta\in\Theta}Z_n(\theta)\xrightarrow{p}\arg\min_{\theta\in\Theta}\big[m_1(\theta)'W(\theta)m_1(\theta)\big]=\theta_0. \tag{A.0.16}$$

For the sake of completeness, bellow is Corollary 3.2.3 from Van der Vaan and Wellner (1996). □

Proof of Proposition 2:

Proof. As we showed in the proof of proposition 3, when $E\left[\parallel g(x,\theta)\parallel^3\right] \leq infty$, and $\parallel \theta-\theta_0 \parallel \leq n^{-1/3}$, uniformly in θ we have

$$\ell_n(\theta) = \frac{1}{2}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n} g_i(\theta)\Big]'\Big[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i'(\theta)\Big]^{-1}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n} g_i(\theta)\Big] + o(n^{1/3}) \quad (a.s). \tag{A.0.17}$$

Since $\ell_n(\theta)$ is a continuous function around θ for every θ belonging to the ball $\parallel \theta-\theta_0 \parallel \leq n^{-1/3}$, $\ell_n(\theta)$ has a minimum value in the interior of this ball, which we denote it by $\hat{\theta}$. Now let's define

$$\begin{aligned} V(u) = {} & \frac{1}{2}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n} g_i(\theta_0+\frac{u}{n^{1/2}})\Big]'\Big[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0+\frac{u}{n^{1/2}})g_i'(\theta_0+\frac{u}{n^{1/2}})\Big]^{-1}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n} g_i(\theta_0+\frac{u}{n^{1/2}})\Big] \\ & -\frac{1}{2}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n} g_i(\theta_0)\Big]'\Big[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)g_i'(\theta_0)\Big]^{-1}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n} g_i(\theta_0)\Big] \\ & +\lambda_n\sum_{j=1}^{n}\Big[|\theta_{j0}+\frac{u_j}{n^{1/2}}|^{\gamma}-|\theta_{j0}|^{\gamma}\Big]+o(n^{1/3}). \end{aligned} \tag{A.0.18}$$

We can do this because

$$\{\theta :\parallel \theta-\theta_0 \parallel \leq n^{-1/2}\} \subseteq \{\theta :\parallel \theta-\theta_0 \parallel \leq n^{-1/3}\} \tag{A.0.19}$$

which implies that

$$(\theta_0+\frac{u}{n^{1/2}}) \in \{\theta :\parallel \theta-\theta_0 \parallel \leq n^{-1/3}\} \tag{A.0.20}$$

Now, we can notice that $V_n(u)$ is minimized at $n^{1/2}(\hat{\theta_n}-\theta_0) = hatu_n$. Therefore we can write

$$\hat{u}_n = \arg\min_{u\in K} V_n(u) \tag{A.0.21}$$

where K is a compact subset of $\mathbb{R}^p$. In order to obtain the asymptotic distribution of our estimator we first need to show the following convergence results.

$$V_n(u) \Longrightarrow V(u) \tag{A.0.22}$$

and also

$$\hat{u} = O_p(1). \tag{A.0.23}$$

Using assumption $A2$, we can use theorem one in Andrews (1994) to obtain

$$n^{-1/2}\sum_{i=1}^{n}\left[g_i(\theta_0+\frac{u}{n^{1/2}})-Eg_i(\theta_0+\frac{u}{n^{1/2}})\right] \Rightarrow \Psi(\theta_0) \equiv N(0,\Omega(\theta_0)) \tag{A.0.24}$$

Also, expanding $g_i(\theta_0+\frac{u}{n^{1/2}}$ around $u=0$ using Taylor series expansion, and using assumption $A3-(ii)$, and noticing that $Eg_i(\theta_0)=0$, uniformly in u we have

$$n^{-1/2}\sum_{i=1}^{n}E\left[g_i(\theta_0+\frac{u}{n^{1/2}})\right] \Longrightarrow R(\theta_0)u. \tag{A.0.25}$$

Combining the last two equations we arrive at

$$n^{-1/2}\sum_{i=1}^{n}g_i(\theta_0+\frac{u}{n^{1/2}}) \Longrightarrow \Psi(\theta_0)+R(\theta_0)u. \tag{A.0.26}$$

Since in the theorem, we assumed that $\lambda_n/n^{\gamma/2} \to \lambda_0 \geq 0$ we have, in other words, $\lambda_n = O(n^{\gamma/2}) = o(n^{1/2})$. Therefore it follows that

$$\lambda_n\left[|\theta_{j0}+\frac{u_j}{n^{1/2}}|^{\gamma}-|\theta_{j0}|^{\gamma}\right] \to 0 \tag{A.0.27}$$

whenever $\theta_{j0} \neq 0$ and

$$\lambda_n\left[|\theta_{j0}+\frac{u_j}{n^{1/2}}|^{\gamma}-|\theta_{j0}|^{\gamma}\right] \to \lambda_0|u_j|^{\gamma} \tag{A.0.28}$$

which means

$$\lambda_n\sum_{j=1}^{p}\left[|\theta_{j0}+\frac{u_j}{n^{1/2}}|^{\gamma}-|\theta_{j0}|^{\gamma}\right] \to \lambda_0\sum_{j=1}^{p}|u_j|^{\gamma}1_{\{\theta_{j0}=0\}} \tag{A.0.29}$$

combining all these equation we get

$$V_n(u) \Rightarrow \big[\Psi(\theta_0)+R(\theta_0)u\big]'W(\theta_0)\big[\Psi(\theta_0)+R(\theta_0)u\big]-\big[\Psi(\theta_0)\big]'W(\theta_0)\big[\Psi(\theta_0)\big] \tag{A.0.30}$$

$$+\lambda_0\sum_{j=1}^{p}|u_j|^{\gamma}1_{\{\theta_{j0}=0\}}$$

$$=u'R(\theta_0)'W(\theta_0)R(\theta_0)u+2u'R(\theta_0)'W(\theta_0)\Psi(\theta_0)+\lambda_0\sum_{j=1}^{p}|u_j|^{\gamma}1_{\{\theta_{j0}=0\}}\equiv V(u).$$

This proves that$V_n(u)\Longrightarrow V(u)$. To complete the proof we notice that, on the space of functions with a topology in which convergence on compact sets implies uniform convergence on these sets, To prove that $\arg\min(V_n)\xrightarrow{d}\arg\min(V)$, it suffices to show that $\arg\min(V_n)=O_p(1)$, see Kim and Pollard (1990). To prove this, let $\delta>0$ be a positive constant such that $\lambda_n/n^{\gamma/2}\leq(\lambda_0+\delta)$, the we have for all u, if n is sufficiently large

$$V_n(u)\geq\frac{1}{2}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n}g_i(\theta_0+\frac{u}{n^{1/2}})\Big]'\Big[\frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0+\frac{u}{n^{1/2}})g_i'(\theta_0+\frac{u}{n^{1/2}})\Big]^{-1}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n}g_i(\theta_0+\frac{u}{n^{1/2}})\Big] \tag{A.0.31}$$

$$-\frac{1}{2}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n}g_i(\theta_0)\Big]'\Big[\frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0)g_i'(\theta_0)\Big]^{-1}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n}g_i(\theta_0)\Big]-\lambda_n\sum_{j=1}^{n}|\frac{u_j}{n^{1/2}}|^{\gamma}$$

$$\geq\frac{1}{2}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n}g_i(\theta_0+\frac{u}{n^{1/2}})\Big]'\Big[\frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0+\frac{u}{n^{1/2}})g_i'(\theta_0+\frac{u}{n^{1/2}})\Big]^{-1}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n}g_i(\theta_0+\frac{u}{n^{1/2}})\Big]$$

$$-\frac{1}{2}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n}g_i(\theta_0)\Big]'\Big[\frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0)g_i'(\theta_0)\Big]^{-1}\Big[\frac{1}{\sqrt{(}n)}\sum_{i=1}^{n}g_i(\theta_0)\Big]-(\lambda_0+\delta)\sum_{j=1}^{n}|\frac{u_j}{n^{1/2}}|^{\gamma}$$

$$=V_n^l(u)$$

now define the empirical process

$$\Psi_n(\theta_0+\frac{u}{n^{1/2}})=n^{-1/2}\sum_{i=1}^{n}\Big[g_i(\theta_0+\frac{u}{n^{1/2}})-Eg_i(\theta_0+\frac{u}{n^{1/2}})\Big] \tag{A.0.32}$$

also let

$$\frac{1}{2}\Big[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)g_i'(\theta_0)\Big]^{-1} = W(\theta_0 + \frac{u}{n^{1/2}}), \tag{A.0.33}$$

then we can rewrite $V_n^l(u)$ as

$$V_n^l(u) = \left[\Psi_n(\theta_0 + \frac{u}{n^{1/2}})'W(\theta_0 + \frac{u}{n^{1/2}})\Psi_n(\theta_0 + \frac{u}{n^{1/2}})\right] \tag{A.0.34}$$

$$+\left[2u'R(\theta_0)'W(\theta_0 + \frac{u}{n^{1/2}})\Psi_n(\theta_0 + \frac{u}{n^{1/2}})\right]$$

$$+\left[u'R(\theta_0)'\Psi_n(\theta_0 + \frac{u}{n^{1/2}})R(\theta_0)u\right]$$

$$-\left[\Psi_n(\theta_0)'W_n(\theta_0)\Psi_n(\theta_0)\right] + o(1)$$

$$-(\lambda_0 + \delta)\sum_{j=1}^{p} |u_j|^{\gamma}.$$

The first term converges to the fourth term in the equation for $V_n^l(u)$ also, when n is large the second term is linear. Therefore, we have a quadratic term and the $|u_j|^{\gamma}$ and, because $0 < \gamma < 1$, the quadratic term dominate all other terms, which implies that $\arg\min V_n^l(u) = O_p(1)$, and from the inequality we get

$$\arg\min V_n(u) = O_p(1). \tag{A.0.35}$$

Because our assumptions guarantee the uniqueness of $\arg\min V_n(u)$, we can apply theorem 3.2.2 of Van der Vaat and Wellner (1996) to get the results. □

Proof of proposition 3:

Proof. Let $L_n(\theta)$ be the same as definition 1. The key idea is that to find an appropriate linear approximation for $L_n(\theta) - L_n(\theta_0)$ characterized by stochastic equicontinuity. To give a road map of our proof we notice that from the Rise representation theorem, there exists $v^* \in \bar{V}$ such that $f'_{\theta_0}(\hat{\theta} - \theta_0) = \langle \hat{\theta} - \theta_0, v^* \rangle$, by screening the definition of $f'_{\theta_0}(\hat{\theta} - \theta_0)$ we see that $f(\hat{\theta}) - f(\theta_0)$ can be linearly approximated by $\langle \hat{\theta} - \theta_0, v^* \rangle$. It is possible to derive a linear approximation for $l(\hat{\theta}, X_i) - l(\theta_0, X_i)$, linear in $\hat{\theta} - \theta_0$. Since $L(\theta)$ is just a summation of $l(\theta, X_i)$ we have a bridge between f and L. Now a linear approximation of $L(\hat{\theta}) - L(\theta_0)$ will gives a linear approximation of $f(\hat{\theta}) - f(\theta_0)$. The last step is to use the central limit theorem on this linear approximation. Since

$$l(\hat{\theta}, X_i) = r(\hat{\theta} - \theta_0, X_i) + l(\theta_0, X_i) + l'_{\theta_0}(\hat{\theta} - \theta_0, X_i) \tag{A.0.36}$$

a simple summation and some algebraic manipulation yields

$$l_n(\hat{\theta}) = l_n(\theta_0) - K(\theta_0, \theta) + n^{-1/2}\nu_n\big(r(\hat{\theta} - \theta_0, X)\big) + n^{-1/2}\nu_n\big(l'_{\theta_0}(\hat{\theta} - \theta_0, X)\big). \tag{A.0.37}$$

Now we notice that by definition 1 $-O(\varepsilon_n^2) \leq L_n(\hat{\theta}_n - L_n(\theta_0)$. Combining this with assumption $A5 - ii$, $A6$ $A7$ gives

$$\begin{aligned} -O(\varepsilon_n^2) \leq L_n(\hat{\theta}_n - L_n(\theta_0) \leq -\frac{1}{2}\|\hat{\theta}_n - \theta_0\|^2 + n^{-1/2}\nu_n\big(r(\hat{\theta} - \theta_0, X)\big) \\ + n^{-1/2}\nu_n\big(l'_{\theta_0}(\hat{\theta} - \theta_0, X)\big) \\ -\lambda_n(J(\hat{\theta}_n) - J(\theta_0)) + O_p(\varepsilon_n) \\ \leq -\lambda_n(J(\hat{\theta}_n) - J(\theta_0)) + O_p(\varepsilon_n). \end{aligned} \tag{A.0.38}$$

Therefore $\lambda_n\big(J(\hat{\theta}_n) - J(\theta_0)\big) \leq O_p(\varepsilon_n)$. Because $J(u^*) < \infty$ and using assumption $A6$ we have

$$\lambda_n(J(\theta^*(\hat{\theta}_n, \varepsilon_n)) - J(\hat{\theta}_n)) \leq c\lambda_n J\big(\varepsilon_n[-\hat{\theta}_n + \theta_0 + u^*]\big) \tag{A.0.39}$$

$$\leq c\lambda_n\varepsilon_n\big(J(\hat{\theta}_n-\theta_0)+J(u^*)\big)=O_p(\varepsilon_n^2)$$

for some $c>0$. Now that we have controlled the penalty part and obtained a bound on that we can turn our attention to $L_n(\hat{\theta}_n)$. From equation (2) we get

$$L_n(\hat{\theta}_n)=L_n(\theta_0)-K(\theta_0,\theta)+n^{-1/2}\nu_n\big(r(\hat{\theta}-\theta_0,X)\big) \tag{A.0.40}$$

$$+n^{-1/2}\nu_n\big(l'_{\theta_0}(\hat{\theta}-\theta_0,X)\big)+\lambda_n J(\hat{\theta}).$$

Noticing that $\|\theta^*(\hat{\theta}_n,\varepsilon_n)-\theta_0\|=\|(1-\varepsilon_n)(\hat{\theta}_n-\theta_0)+\varepsilon_n u^*\|\leq\delta_n$ the equation holds if we replace $\hat{\theta}_n$ with $\theta^*(\hat{\theta}_n,\varepsilon_n)$. If we do so and subtract the two equations we get

$$\begin{aligned}L_n(\hat{\theta}_n)&=L_n(\theta^*(\hat{\theta}_n,\varepsilon_n))-\big[K(\theta_0,\hat{\theta})-K(\theta_0,\theta^*(\hat{\theta}_n,\varepsilon_n))\big] \qquad (A.0.41)\\ &\quad+n^{-1/2}\nu_n\big(l'_{\theta_0}(\hat{\theta}_n-\theta^*(\hat{\theta}_n,\varepsilon_n),X)\big)+n^{-1/2}\nu_n\big(r(\hat{\theta}_n-\theta^*(\hat{\theta}_n,\varepsilon_n),X)\big)+O_p(\varepsilon_n^2)\\ &=L_n(\theta^*(\hat{\theta}_n,\varepsilon_n))-\frac{1}{2}\big[\|\hat{\theta}_n-\theta_0\|^2-\|\theta^*(\hat{\theta}_n,\varepsilon_n)-\theta_0\|^2\big]\\ &\quad+n^{-1/2}\nu_n\big(l'_{\theta_0}(\hat{\theta}_n-\theta^*(\hat{\theta}_n,\varepsilon),X)\big)+O_p(\varepsilon_n^2).\end{aligned}$$

Using definition 1 and assumptions $A6$ and $A7$ we get

$$\begin{aligned}-O_p(\varepsilon_n^2)&\leq-\frac{1}{2}(1-(1-\varepsilon_n)^2)\|\hat{\theta}_n-\theta_0\|^2+(1-\varepsilon)\langle\hat{\theta}_n-\theta_0,\varepsilon_n u^*\rangle \qquad (A.0.42)\\ &\quad-n^{-1/2}\nu_n\big(l'_{\theta_0}(\varepsilon_n(u^*-(\hat{\theta}_n-\theta_0)),X)\big)+O_p(\varepsilon_n^2)\\ &\leq-\varepsilon_n\|\hat{\theta}_n-\theta_0\|^2+(1-\varepsilon)\langle\hat{\theta}_n-\theta_0,\varepsilon_n u^*\rangle\\ &\quad-n^{-1/2}\nu_n\big(l'_{\theta_0}(\varepsilon_n u^*,X)\big)+O_p(\varepsilon_n^2)\\ &\leq(1-\varepsilon)\langle\hat{\theta}_n-\theta_0,\varepsilon_n u^*\rangle-n^{-1/2}\nu_n\big(l'_{\theta_0}(\varepsilon_n u^*,X)\big)+O_p(\varepsilon_n^2)\end{aligned}$$

Therefore

$$\begin{aligned}-(1-\varepsilon_n)\langle\hat{\theta}_n-\theta_0,u^*\rangle+n^{-1/2}\nu_n\big(l'_{\theta_0}(u^*,X)\big)&=O_p(\varepsilon_n)+O_p(\varepsilon_n) \qquad (A.0.43)\\ &=o_p(n^{-1/2}).\end{aligned}$$

If we replace u^* with $-u^*$ in the last equation and then put them together we arrive at the following equation

$$|\langle\hat{\theta}_n-\theta_0,u^*\rangle-n^{-1/2}\nu_n\big(l'_{\theta_0}(u^*,X)\big)|=o_p(n^{-1/2}). \tag{A.0.44}$$

Therefore $\langle\hat{\theta}_n-\theta_0,v^*\rangle=n^{-1/2}\nu_n\big(l'_{\theta_0}(v^*,X)\big)+o_p(n^{-1/2}$. From this equation and (4.22) we have

$$f(\hat{\theta}_n)-f(\theta_0)=f'_{\theta_0}(\hat{\theta}_n-\theta_0)+o_p(u_n\|\hat{\theta}_n-\theta_0\|^w) \tag{A.0.45}$$

$$\begin{aligned}&=\langle\hat{\theta}_n-\theta_0,v^*\rangle+o_p(n^{-1/2})\\&=n^{-1}\sum_{i=1}^{n}l'_{\theta_0}(v^*,X_i)+o_p(n^{-1/2}).\end{aligned}$$

Therefore $n^{1/2}\big(f(\hat{\theta}_n)-f(\theta_0)\big)=n^{-1/2}\sum_{i=1}^{n}l'_{\theta_0}(v^*,X_i)+o_p(1)$ and the result follows by applying the central limit theorem on $n^{-1/2}\sum_{i=1}^{n}l'_{\theta_0}(v^*,X_i)$. □

Proof of corollary 2:

Proof. If we replace v^* with s in proposition 3, the result is corollary 2. □

Proof of proposition 4:

Proof. The following lemma is needed in the proof of proposition 2. The proof can be find in Shen and Wong (1994). Also here we define the Hellinger metric entropy with bracketing, which we are using the assumption *A*?.

Definition 14:

Let $f : \Theta \times \mathcal{X} \to \mathcal{R}$ *with* $E[f^2(\theta, X)] < \infty$ *for all* $\theta \in \Theta$ *and let* $\|\cdot\|_2$ *be the usual* L^2 *norm. Let*

$$\mathcal{F} = \{f(\theta, .) : \theta \in \Theta, \ \|f\|_2 < \infty\}. \tag{A.0.46}$$

For any given $\varepsilon > 0$*, if there exists*

$$S(\varepsilon, n) = \{f_1^l, f_1^u, \ldots, f_n^l, f_n^u\} \subset \mathcal{L}_2 \tag{A.0.47}$$

with $\max_{1 \le j \le n} \|f_j^u - f_j^l\|_2 \varepsilon$ *such that for every* $f \in \mathcal{F}$ *there exists a* j *such that* $f_j^l \le f \le f_j^u$ *a.s., then* $S(\varepsilon, n)$ *is called a bracketing* ε*-covering of* $\mathcal{F}$ *with respect to* $\|\|_2$*.* $H(\varepsilon, \mathcal{F}) = \log N(\varepsilon, \mathcal{F})$ *is called the Hellinger* L_2 *metric entropy of* $\mathcal{F}$ *with bracketing, where*

$$N(\varepsilon, \mathcal{F}) = \min\{n : S(\varepsilon, n) \text{ is a bracketing } \varepsilon - \text{covering of } \mathcal{F}\}. \tag{A.0.48}$$

The Hellinger metric entropy of $\mathcal{F}$ *with bracketing is the logarithm of the cardinality of* $\varepsilon -$ *cover of* $\mathcal{F}$ *of smallest size. when appropriately defined, it provides a measure of the size of parameter space. For more discussions about metric entropy see Kolmogorov and Tihomirov (1961).*

Lemma 4:

*Suppose assumption A*11 *is satisfied, and let* $v^2 \ge \sup_{\theta \in \mathcal{A}} n^{-1} \sum_{i=1}^n V(\theta_0, \theta)$ *and* $b \ge \sup_{\theta \in \mathcal{A}} \|\theta - \theta_0\|$*. Also assume that*

$$\int_L^U H^{1/2}(u, \mathcal{A}) du \le (n^{1/2} M a^{3/2})/2^{10} \tag{A.0.49}$$

where $U = H^-(\Psi(M, v), \mathcal{A}$ *and* $L = aM/2^8 \ (0 < a < 1)$*, and* $\Psi(M, v) = (1-a)nM^2/[2(v^2 + bM/3)]$*. Then*

$$P^*\Big(\sup_{\theta \in \mathcal{A}} \nu_n\big(l(\theta, X) - l(\theta_0, X)\big)\Big) \le 3exp(-\Psi(M, v)). \tag{A.0.50}$$

If $U \le L$ *the above inequality continues to hold with 1 replacing 3.*

The idea of proof is to control and bound the mean and variance of the criterion differences when it is evaluated at θ_0 and $\theta \in \Theta$. Without loss of generality we can assume that $max(\lambda_n, \varepsilon) \leq 1$. For any $i, j \geq 1$ we have

$$\inf_{A_{i,j}} \Big\{ K(\theta, \theta_0) + \lambda_n (J(\theta) - J(\theta_0)) \Big\} \geq (2^{i-1}\varepsilon)^2 + \lambda_n (2^{j-1} - 1) J(\theta_0), \tag{A.0.51}$$

and

$$\inf_{A_{i,0}} \Big\{ K(\theta, \theta_0) + \lambda_n (J(\theta) - J(\theta_0)) \Big\} \geq (2^{i-1}\varepsilon)^2 - \lambda_n J(\theta_0). \tag{A.0.52}$$

Since $max(J(\theta_0), 1) \leq c_7 \varepsilon^2$, we have

$$I = P^* \Big(\sup_{\{\rho(\theta_0,\theta) \geq \varepsilon, \theta \in \Theta\}} n^{-1} \sum_{i=1}^{n} \big(\ell(\theta, X_i) - \ell(\theta_0, X_i) \big) \geq -\varepsilon^2/2 \Big) \tag{A.0.53}$$

$$\begin{aligned} &= \sum_{i,j=1}^{\infty} P^* \big(\sup_{A(i,j)} \nu_n (l(\theta, X) - l(\theta_0)) \geq M(i,j) \big) \\ &+ \sum_{i,j=1}^{\infty} P^* \big(\sup_{A(i,0)} \nu_n (l(\theta, X) - l(\theta_0)) \geq M(i,0) \big) \\ &= I_1 + I_2, \end{aligned}$$

where

$$M(i,j) = \frac{1}{2} \lambda_n \big[(2^{i-1})^2 + (2^{j-1} - 1) J(\theta_0) \big]. \tag{A.0.54}$$

Now we separately bound I_1 and I_2. To do this we use lemma 2. Because it is very similar to establish the bounds for I_1 and I_2, we just show the it for I_1. To bound I_1 we verify that lemma 2 is indeed applicable. By assumption $A6$, when

$$Mb/v^2 \leq 3, \ \ and \ \ \Psi(M, v) \geq (1 - a) n M^2 / 4 v^2 \tag{A.0.55}$$

we have

$$\sup_{A(i,j)} V(\theta_0, \theta) \leq v^2(i,j) = c_1 (2^i \varepsilon)^2 (1 + ((2^i)^2 + 2^j J(\theta_0))^{\beta}). \tag{A.0.56}$$

Similarly, when $Mb/v^2 \leq 3$, and $U \leq M^{1/2}(i,j)B^{1/2}(i,j)$ we have $H^{-}(\Psi(M,v),\mathcal{A}) \leq v(i,j)$. By assumption $A7$ we have

$$\int_{aM(i,j)}^{max\left(v(i,j),M^{1/2}(i,j)B^{(i,j)}\right)} H^{1/2}\big(u,B(2^i\varepsilon,2^j)\big)du/M(i,j) \leq c_5 n^{1/2}. \tag{A.0.57}$$

Therefore the requirement of the lemma 2 is satisfied and we have: (using the inequality $(a+b)^c \leq a^c + b^c$ for $a,b>0$, and $0<c<1$.)

$$I_1 \leq 3\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} exp\Big(-c_8 n\, min\big(M^2(i,j)/v^2(i,j), M(i,j)/B(i,j)\big)\Big) \tag{A.0.58}$$

$$\leq 3\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} exp\Big(-c_8 n\, min\big((\lambda_n^2/\varepsilon^2)[(2^{i-1})^2+2^{j-1}]^{1-\beta}, \lambda_n[(2^{i-1})^2+2^{j-1}]^{1-\gamma}\big)\Big).$$

A similar reasoning bounds I_2. Putting them together yields

$$I \leq 6exp\big(-c_8 n\, min(\lambda_n^2/\varepsilon^2,\lambda_n)\big)/[1-exp(-c_8 n\, min(\lambda_n^2/\varepsilon^2,\lambda_n))] \tag{A.0.59}$$

$$\leq 7exp(-c_8 n\, min(\lambda_n^2/\varepsilon^2,\lambda_n)).$$

This finishes the proof. □

Proof of corollary 3:

Proof. By definition 1, for every $\varepsilon_n > 0$, which satisfies (4.22), there exists $c>0$ such that:

$$P\big(\rho(\theta_0,\hat{\theta}\geq\varepsilon_n\big) \leq P^*\Big(\sup_{\{\rho(\theta_0,\hat{\theta})\leq\varepsilon_n,\theta\in\Theta\}} (L_n(\theta)-L_n(\theta_0)) \leq -a_n\Big) \tag{A.0.60}$$

$$P^*\Big(\sup_{\{\rho(\theta_0,\hat{\theta})\leq\varepsilon_n,\theta\in\Theta\}} (L_n(\theta)-L_n(\theta_0)) \leq -c\varepsilon_n^2\Big).$$

By proposition 4, $\rho(\theta_0,\hat{\theta}_n) = O_p(\varepsilon_n$ whenever we have $max\big(J(\theta_0),1\big)\lambda_n \leq c_?\varepsilon_n^2$, and ε_n is the smallest ε which satisfies (4.22). Therefore replacing ε_n with $\lambda_n^{-1/2}$ whenever $max\big(J(\theta_0),1\big)\lambda_n \leq c_?\varepsilon_n^2$ results in $rho(\theta_0,\hat{\theta}) = O_p(\lambda_n^{1/2})$. Now the results directly follows from proposition 4. □

B.0 PROOFS AND SUPPLEMENTAL MATERIALS FOR CHAPTER 2

In this appendix we provide the proofs for the lemmas and propositions which appeared earlier in chapter 2.

Proof of lemma 1:

Proof. The proof is simple and somewhat mechanical. We find a transformation that, for every w_i and p_i produce a ϖ_i and a q_i in such a way that $\sum_{i=1}^{n} q_i = 1$, this proves lemma 1.

Let $p = (p_1, \ldots, p_n)$ and $w = (w_1, \ldots, w_n)$ are give; define $P_N = \sum_{i=1}^{N} p_i$ and let $q_i = \frac{p_i^{w_i}}{P_N}$. Now, the problem

$$\min_{p_1,\ldots,p_n} \sum_{i=1}^{n} -w_i \log p_i \tag{B.0.1}$$

subject to:

$$\sum_{i=1}^{n} p_i g_i(\theta) = 0 \quad \textit{and} \quad \sum_{i=1}^{n} p_i = 1 \tag{B.0.2}$$

can be transformed to:

$$\min_{q_1,\ldots,q_n} \sum_{i=1}^{n} -\log q_i \tag{B.0.3}$$

subject to:

$$\sum_{i=1}^{n} q_i g_i(\theta)\varpi_i = 0 \quad \textit{and} \quad \sum_{i=1}^{n} q_i = 1 \tag{B.0.4}$$

with $q_i = \frac{p_i^{w_i}}{P_N}$ and $\varpi_i = \frac{p_i^{1-1/w_i}}{P_N^{w_i}}$.

Because this transformation is one to one, if the first problem has a solution the second

problem will have one and vice versa. □

Proof of lemma 2:

Proof. The existence of a solution, $\hat{\theta}(w)$ and $\hat{p}(w)$, is a consequence of maximization of a convex function on a compact sent. Virtually the same reasons that grantee the existence of a solution to the EL procedure, as long as we maintain the same assumptions like the compactness of the Θ. Obviously any solution of this problem will depend on the $w = (w_1, \ldots, w_n)$.

let $\hat{\theta}(w)$ and $\hat{p}(w)$ be a pair that minimize the objective function with the given constrains. The constrains are $\sum_{i=1}^{n} p_i g_i(\theta) = 0$ and $\sum_{i=1}^{n} p_i = 1$, because any solution has to satisfy these constrains, $\hat{\theta}(w)$ and $\hat{p}(w)$ satisfy these constrains too. Therefore

$$\sum_{i=1}^{n} \hat{p}_i(w) g_i(\hat{\theta}(w)) = 0 \tag{B.0.5}$$

which is the sample moment conditions. □

Proof of proposition 2:

Proof. Let

$$\ell(\beta, \sigma^2) = -2 \log \left(n^n L(\beta, \sigma^2) \right) \tag{B.0.6}$$

be the log empirical likelihood ratio. Using Lagrange multipliers to optimize $\ell(\beta, \sigma^2)$ we get

$$\mathcal{L}(\beta, \sigma^2, \lambda, \mu) = \sum_{i=1}^{n} \log(p_i) - \mu \Big(\sum_{i=1}^{n} p_i - 1 \Big) - n\lambda' \sum_{i=1}^{n} p_i g_i(\beta, \sigma^2) \tag{B.0.7}$$

doing some algebra we see that

$$p_i = \frac{1}{n(1+\lambda' g_i(\beta,\sigma^2))} \tag{B.0.8}$$

and $\lambda(\beta,\sigma^2)$ minimizes $\ell(\beta,\sigma^2)$ therefore the log empirical likelihood ratio, which we seek to minimize is:

$$\ell(\beta,\sigma^2) = 2\sum_{i=1}^{n} \log\{1+\lambda' g_i(\beta,\sigma^2)\} \tag{B.0.9}$$

and $\lambda \in \mathbb{R}^q$ satisfies

$$\frac{\partial \ell(\beta,\sigma^2)}{\partial \lambda} = \sum_{i=1}^{n} \frac{g_i(\beta,\sigma^2)}{1+\lambda' g_i(\beta,\sigma^2)} = G_{1n}(\beta,\sigma^2,\lambda) = 0 \tag{B.0.10}$$

differentiating $\ell(\beta,\sigma^2)$ with respect to β and σ^2, we have:

$$\frac{\partial \ell(\beta,\sigma^2)}{\partial \beta} = \lambda' \sum_{i=1}^{n} \frac{\partial g_i(\beta,\sigma^2)/\partial\beta}{1+\lambda' g_i(\beta,\sigma^2)} = G_{2n}(\beta,\sigma^2,\lambda) \tag{B.0.11}$$

$$\frac{\partial \ell(\beta,\sigma^2)}{\partial \sigma^2} = \lambda' \sum_{i=1}^{n} \frac{\partial g_i(\beta,\sigma^2)/\partial\sigma^2}{1+\lambda' g_i(\beta,\sigma^2)} = G_{3n}(\beta,\sigma^2,\lambda) \tag{B.0.12}$$

let's denote

$$A = E\left[g_1(\beta,\sigma^2) g_1'(\beta,\sigma^2)\right] \tag{B.0.13}$$

and

$$B = \left(E\left[\frac{\partial g_1(\beta,\sigma^2)}{\partial \beta}\right], E\left[\frac{\partial g_1(\beta,\sigma^2)}{\partial \sigma^2}\right]\right) \tag{B.0.14}$$

Under assumptions $A1 - A4$, the solution $(\hat{\beta},\hat{\sigma}^2,\hat{\lambda})$ to this problem is triple such that, see for example Qin and Lawless (1994), $G_{1n}(\hat{\beta},\hat{\sigma}^2,\hat{\lambda}) = 0$, $G_{2n}(\hat{\beta},\hat{\sigma}^2,\hat{\lambda}) = 0$, $G_{3n}(\hat{\beta},\hat{\sigma}^2,\hat{\lambda}) =$

0. In this case the empirical likelihood ration $\ell(\beta,\sigma^2)$ attains its minimum at $(\hat{\beta},\hat{\sigma^2})$ and $\hat{\lambda}=\lambda(\hat{\beta},\hat{\sigma^2})$. This results in a an asymptotic limit

$$\begin{pmatrix} \sqrt{(}n)(\hat{\beta}-\beta_0) \\ \sqrt{(}n)(\hat{\sigma}^2-\sigma_0^2) \end{pmatrix} \xrightarrow{d} N(0,\Sigma)^{1} \tag{B.0.15}$$

where $\Sigma=(B'A^{-1}B)^{-1}$. Therefore we can drive the asymptotic variance of the empirical likelihood estimators $\hat{\beta}$ and $\hat{sigma}^2$. After some simple algebra we can convince ourselves that $\Sigma_{\hat{\beta}}=(I_p,0)\Sigma(I_p,0)'$ and $\Sigma_{\hat{\sigma}^2}=(1,0)\Sigma(1,0)'$. The corresponding asymptotic variance of the usual EL estimator, without using the modulation method is the same as the asymptotic variance of quasi-likelihood estimator which for $\breve{\beta}_{ql}$ is:

$$\Sigma_{\breve{\beta}_{ql}}=\sigma^2\Big(E\left[G'(X'\beta)^2XX'/V\right]\Big)^{-1} \tag{B.0.16}$$

A standard estimator for σ^2 is

$$\breve{\sigma}^2=n^{-1}\sum_{i=1}^{n}\Big((Y_i-G(X_i'\breve{\beta}_{ql}))/V(G(X_i'\breve{\beta}_q))\Big) \tag{B.0.17}$$

which we denote its asymptotic variance by $\Sigma_{\breve{\beta}_{ql}}$. Now we are in a position to compare these variances and drive the optimum weights. We need the following definitions:

$$\mu_3=E\left[\big(\varepsilon/(\sigma\sqrt{(}V))\big)^3|X\right] \tag{B.0.18}$$

$$\mu_4=E\left[\big(\varepsilon/(\sigma\sqrt{(}V))\big)^4|X\right] \tag{B.0.19}$$

$$A_{11}=\sigma^4\Sigma^{-1}_{breve\beta_{ql}}=\sigma^2E\left[\frac{G'(X'\beta)^2}{V}XX'\right] \tag{B.0.20}$$

[1]The derivation is a standard practice in the literature, for example see Qin and Lawless (1994) for detailed derivations and proofs of the asymptotic limit theorems.

$$A_{22} = \begin{pmatrix} E\left[\frac{\mu_4-1}{\sigma^4}\right] & E\left[\frac{\mu_4-1}{\sigma^4}w'\right] \\ E\left[\frac{\mu_4-1}{\sigma^4}w'\right] & E\left[\frac{\mu_4-1}{\sigma^4}ww'\right] \end{pmatrix} \tag{B.0.21}$$

$$A_{12} = \left(E\left[\frac{\mu_3 G'(X'\beta)}{\sigma\sqrt{(V)}}\right], E\left[\frac{\mu_3 G'(X'\beta)}{\sigma\sqrt{(V)}}Xw'\right] \right) \tag{B.0.22}$$

$$B_1 = \left(E\left[\frac{V'G'(X'\beta)}{\sigma^2 V}X\right], E\left[\frac{V'G'(X'\beta)}{\sigma^2 V}Xw'\right] \right) \tag{B.0.23}$$

and finally

$$B_2 = \left(\frac{1}{\sigma^4}, E\left[\frac{w'}{\sigma^4}\right]\right). \tag{B.0.24}$$

Doing some tidies algebra we obtain,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A'_{12} & A_{22} \end{pmatrix} \tag{B.0.25}$$

and

$$B' = -\begin{pmatrix} \sigma^{-2}A_{11} & B_1 \\ 0 & B_2 \end{pmatrix}. \tag{B.0.26}$$

Under the assumptions $A1 - A4$ we can calculate to get

$$B'A^{-1}B = \begin{pmatrix} \sigma^{-4}A_{11} & 0 \\ 0 & 0 \end{pmatrix} + B_1\left(A_{22.1}^{-1} - \frac{A_{22.1}^{-1}B_2'B_2A_{22.1}^{-1}}{B_2A_{22.1}^{-1}B_2'}\right)B_1' \tag{B.0.27}$$

whit

$$A_{22.1}^{-1} = \frac{\sigma^4}{k-1}\begin{pmatrix} 1+E(w')D^{-1}E(w) & -E(w')D^{-1} \\ -D^{-1}E(w) & D^{-1} \end{pmatrix} \tag{B.0.28}$$

where $D = cov(w)$. Doing some algebra reveals that

$$B_2'B_2 = \frac{1}{\sigma^8}\begin{pmatrix} 1 & E(w) \\ E(w') & E(w)E(w') \end{pmatrix} \tag{B.0.29}$$

and

$$\left(A_{22.1}^{-1}-\frac{A_{22.1}^{-1}B_2'B_2A_{22.1}^{-1}}{B_2A_{22.1}^{-1}B_2'}\right)=\frac{\sigma^4}{k-1}\begin{pmatrix}1&0\\0&0\end{pmatrix}. \tag{B.0.30}$$

Therefore,

$$B_1\left(A_{22.1}^{-1}-\frac{A_{22.1}^{-1}B_2'B_2A_{22.1}^{-1}}{B_2A_{22.1}^{-1}B_2'}\right)B_1'=\frac{\sigma^4}{k-1}B_1\big(E(w'),-1\big)D^{-1}\big(E(w'),-1\big)B_1'=C(w)C'(w)/(k-1) \tag{B.0.31}$$

and doing some algebra we have

$$C(w)=\left(E\big[\frac{V'G'(X'\beta)Xw'}{V}\big]-E\big[\frac{V'G'(X'\beta)}{V}Xw'\big]E(w')\right)D^{-1/2}=E\left[\frac{V'G'(X'\beta)X}{V}(w'-E(w'))D^{-1/2}\right] \tag{B.0.32}$$

Using the above calculations we can show that,

$$\Sigma_{hat\beta(w)}=(I_p,0)\Sigma^{-1}(I_p,0)'=\Big(\frac{A_{11}}{\sigma^4}+\frac{C(w)C'(w)}{k-1}\Big)^{-1}. \tag{B.0.33}$$

For any two any two positive definite and symmetric $n\times n$ matrices A and B we have $A-B>0$ if and only if $B^{-1}-A^{-1}>0$, Therefore finding a w to maximize $\Sigma_{\check{\beta}}-\Sigma_{\hat{\beta}(w)}$ is equivalent to finding w to maximize $C(w)C'(w)$. Let define $\eta=\frac{V'G'(X'\beta)X}{V}$ and $\xi=(w'-E(w'))$ form the equation for $C(w)$ we obtained above we have $C(w)=E[\eta\xi]$. Now we can write $C(w)C'(w)=E(\eta\xi')[E(\xi\xi')]^{-1}E(\xi\eta)$ by the lemma which will follow this proof we have

$$C(w)C'(w)\leq E\big[\frac{\big(V'G'(X'\beta)\big)^2XX'}{V^2}\big] \tag{B.0.34}$$

and the equality holds if and only if

$$\eta=\Big(E(\eta\xi')[E(\xi\xi')]^{-1}\Big)\xi \tag{B.0.35}$$

It can be directly check that the equality is hold when $w=\frac{V'G'(X'\beta)X}{V}$, and this finishes the proof. □

The following lemma was used in the last step of the proof of proposition 2.

Lemma 5:

If ξ *and* η *are to n-dimensional and m-dimensional random variables and* $n \leq m$, $E[||\xi||^2 + ||\eta||^2] < \infty$ *and* $E[\xi\xi'] > 0$, *then* $E(\eta\xi')[E(\xi\xi')]^{-1}E(\xi\eta') \leq E(\eta\eta')$. *Furthermore, equality holds if and only if* $\eta = \Big(E(\eta\xi')[E(\xi\xi')]^{-1}\Big)\xi$.

Proof. Let $c = \Big(E(\eta\xi')[E(\xi\xi')]^{-1}\Big)$, because $E[||\xi||^2 + ||\eta||^2] < +\infty$, we have $E[(c\xi - \eta)(c\xi - \eta)'] \geq 0$ implies that $cE(\eta\eta')c' - c\xi\eta' - \eta\xi'c' + \eta\eta' \geq 0$. Replacing c with $\Big(E(\eta\xi')[E(\xi\xi')]^{-1}\Big)$ we get $E(\eta\xi')[E(\xi\xi')]^{-1}E(\xi\eta') \leq E(\eta\eta')$. Equality holds if and only if $E[(c\xi - \eta)(c\xi - \eta)'] = 0$, which implies $\eta = c\xi$. □

C.0 PROOFS AND SUPPLEMENTAL MATERIALS FOR CHAPTER 3

Proof of Lemma 1:

Proof. First, notice that

$$f'(x) = \frac{p(1-p)}{((1-p)x + p(1-x))} > 0 \tag{C.0.1}$$

which implies that this function is increasing for $0 \leq p \leq 1$.

Second, we have $f(x) < x$. To see this, notice that

$$f(x) < x \Leftrightarrow (1-p)x < (1-p)x^2 + p(1-x)x \Leftrightarrow x(1-x) > 0.$$

The last statement is always true because $0 < x = \pi^* < 1$.

Third, if

$$\{a_n = f^n(\pi^*)\}_{n=1}^{n=\infty}, \qquad then \lim_{n\to\infty} a_n = 0$$

This is so because $f(x) < x$ and $f(x)$ is increasing together. These imply that $\{f^n(\pi^*)\}_0^\infty$ is a bounded and decreasing sequence of real numbers and, therefore, has a limit. Let $\lim_{n\to\infty} a_n = a_0 > 0$. Then, $f(a_0) < a_0$, which is a contradiction. Therefore, a_n has to converge to a fixed point of $f(x)$, which is zero.

The above argument shows that

$$\exists n \, s.t. \quad f^n(\pi^*) \leq 1/2 \tag{C.0.2}$$

and, therefore,

$$\{m|\, f^n(\pi^*) \leq 1/2\} \neq \emptyset. \tag{C.0.3}$$

Hence the minimum exists.

Now suppose that $\pi < \pi'$. We have:

$$\pi < \pi' \Rightarrow f(\pi) < f(pi') \Rightarrow \forall m f^m(\pi) < f^m(\pi') \Rightarrow (f^m(\pi') \leq 1/2 \rightarrow f^m(\pi) \leq 1/2) \tag{C.0.4}$$

this implies that

$$\{m|\, f^m(\pi) \leq 1/2\} \subseteq \{m|\, f^m(\pi') \leq 1/2\}. \tag{C.0.5}$$

Therefore,

$$min\{m|\, f^m(\pi) \leq 1/2\} \leq min\{m|\, f^m(\pi') \leq 1/2\}, \tag{C.0.6}$$

which is to say, $n \leq n'$.

□

Proof of proposition 3:

Proof. Suppose that at time t an agent defects and chooses the opposite outcome of the cascade. With probability $1-\mu$ the next actor is a fan who, by assumption, receives a negative signal (here negative means a signal which points to the opposite direction of the cascade). His updated belief is

$$f^2(\pi^*) > 1/2, \qquad \text{since we have} n > 2. \tag{C.0.7}$$

Therefore

$$\mathbb{E}[V = 1|H_{t+1}] = P_f(V = 1|H_{t+1}) > 1/2, \tag{C.0.8}$$

so is optimal for him to follow the cascade. If the next agent is a fan, he will perform the same calculation and will defect only if $f^3(\pi^*) \leq 1/2$ and that is so if $n = 3$. Continuing this argument, a fan with an n given by lemma 1 will defect only if $f^n(\pi^*) \leq 1/2$,

which requires that the last $n-1$ agents are defectors and that is so if all of them are non-fans which happens with probability $(1-\mu)^{(n-1)}$. In this case, the cascade breaks at time t+n. Therefore the defection is successful with probability

$$(1-\mu)^{(n-1)}. \tag{C.0.9}$$

□

Proof of the proposition 4:

Proof. Let's suppose there are two actions a and b to be chosen and a fan, f, sophisticated enough to take in to account the possibility of herd, resulting from the action he is about to choose. Also, suppose the star has chosen action a. Because $n > 2$ by proposition 1 more than 2 opposite signals are needed for this fan to choose b. Now suppose that every of the k predecessors has chosen b, and k is arbitrary large. The only information that f can extract from this chain of actions is that the star received a signal. The first two non-stars had b signals, and the rest of the population is in an informational cascade. Since f needs more than 2 b signals in order to choose action b, she will follow the star and choose a. □

Proof of Porposition 5:

Proof. Because non-fan traders fallow their own signal, their participation helps to control any miss-pricing. Therefore, in order to find an upper bound for any possible bubble, we can assume that all traders are fans.

Suppose everybody receives a negative signal but after weighting in her/his initial belief decides to buy. How long can this process continue? As soon as $g(\pi^*) \leq f^m(p)$, the m^{th} trader stops buying. Therefore, the length of the buying process is

$$n = min\{m|g(\pi^*) \leq f^m(p)\}. \tag{C.0.10}$$

The next step is to investigate how much a bubble can grow during these n periods. If the market maker could see the actual signals he would have set the price according to $\beta = g^n(p)$. Since, he cannot see the actual signals and he only observes the "buy" and "sell" actions, he increases the price according to $\bar{\beta}$. Therefore, the size of the bubble is

$$\bar{\beta} - \beta \tag{C.0.11}$$

□

Proof of Proposition 6:

Proof. In the proof of proposition 3 we assumed that all traders are fans, which implies that no correction takes place and the size of any possible bubble rapidly grows until it reaches the established upper bound. Now, if we take into consideration the presence of noise traders and non-fans, we are going to have an asymmetric random walk on $\mathbb{R}$ which moves up and down with different probabilities depending on the combination of fans, non-fans, and the noise traders. The following lemma is the core part of the proof.

Lemma 6:

Let $X_1, X_2, \ldots$ *be i.i.d with*

$$P(X_i = 1) = p \quad \textit{and} \quad P(X_i = -1) = 1 - p \quad p > 1/2$$

and let

$$S_n = X_1 + X_2 + \cdots + X_n \quad \alpha = inf\{n : S_n > 0\} \quad \beta = inf\{n : S_n < 0\}.$$

Then,

$(i)\, P(\alpha < \infty) = 1 \quad \textit{and} \quad P(\beta < \infty) < 1.$

$(ii)\, If \quad Y = \inf S_n, \quad then\, P(Y \leq -k) = P(\beta < \infty)^k.$

$(iii) \mathbb{E}\alpha = \frac{1}{2p-1}.$

Proof. Sketch of a proof:

(i): We need the following result for the proof of this part this can be found as theorem in "*Probability: Theory and Examples* by Richard Durrett."

Theorem 1:

For a random walk on $\mathbb{R}$ there are only four possibilities, one of which has probability one.

(1) $S_n = 0$, for all n.

(2) $S_n \to \infty$.

(3) $S_n \to -\infty$.

(4) $-\infty = liminf S_n < limsup S_n = \infty$.

We also need the following statement in the proof.

Let α and β be the same as above. Then the four possibilities of the theorem correspond to the following four combinations $P(\alpha < \infty) < 1 or = 1$ and $P(\beta < \infty) < 1 or = 1$.

Part (i) of the lemma can easily be derived from the fact that

$$P(\beta < \infty) < P(\alpha < \infty). \tag{C.0.12}$$

(ii): This part is obvious when we consider that the S_i, s are independent, and $Y \leq S_i$, $\forall i$.

(iii) A result in stopping time theory -sometimes referred to as Wald's equation- states

that:

If $X_1, X_2, \ldots$ are i.i.d with $\mathbb{E}|X_i| < \infty$, and if τ is a stopping time with $\mathbb{E}\tau < \infty$, then:

$$\mathbb{E}S_\tau = \mathbb{E}X_1\mathbb{E}\tau. \tag{C.0.13}$$

Apply Wald's equation to the stopping time $\alpha \wedge n$ and let $n \to \infty$ to obtain:

$$\mathbb{E}\alpha = \frac{1}{\mathbb{E}X_1} = \frac{1}{2p-1}. \tag{C.0.14}$$

□

The only thing that remains is to calculate the probability of a "buy" which moves the price up. This probability is $1/3(1-\mu)+\gamma$. Now to prove part (a), notice that when $\gamma = 1/6 + 1/3\mu$ the $1/3(1-\mu)+\gamma = 1/2$, and we have a symmetric random walk in which $Prob(T < \infty) = 1$, and $\mathbb{E}[T] = \infty$.

For part (b), if $\gamma < 1/6 + 1/3\mu$, then $1/3(1-\mu)+\gamma < 1/2$ and, therefore, we have an asymmetric random walk, thus, by part (i) of lemma 3, $P(T < \infty) < 1$. In additions, by part (ii) of the lemma 3,

$$P(T < \infty) = P(\beta < \infty)^n. \tag{C.0.15}$$

For part (c), notice that if $\gamma > 1/6 + 1/3\mu$, we have an asymmetric random walk with the probability going up greater than the probability of going down. By part (i) of lemma 3, $Prob(T < \infty) = 1$ and by part (iii) of lemma 3, we have

$$\mathbb{E}[T] = \frac{n}{2p-1}, \tag{C.0.16}$$

where p is the probability of going up. □

D.0 BIBLIOGRAPHY

ALTONJI, J. and L.M. SEGAL, "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal Business and Economic Statistics*, 1996, *14*, 353–366.

ANDERSON, D. W. K., "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 1999, *67*, 543–564.

ANDREWS, D.W. and B. LU, "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, 2001, *101*, 123–165.

ANDREWS, D.W.K., "Empirical Process Methods in Econometrics," in R.F. Engle and D. McFadden, eds., *Handbook of Econometrics*, 1 ed., Vol. 4, Elsevier, 1986, chapter 37, pp. 2247–2294.

AVERY, P. and P. ZEMSKY, "Multidimentional Uncertainty and Herd Behavior in Financial Markets," *American Economic Review*, 1998, *88*, 724–48.

BACK, K. and D. BROWN, "Implied Probabilities in GMM Estimators," *Econometrica*, 1993, *61*, 971–976.

BALA, V. and S. GOYAL, "Learning from Neighbours," *The Review of Economic Studies*, 1998, *65* (3), 595–621.

BANERJEE, A., "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 1992, *107*, 797–818.

BERAN, R. and L. DÜMBGEN, "Modulation of estimators and confidence sets," *The Annals of Statistics*, 1998, *26*, 1826–1856.

BIKHCHANDANI, S. and S. SHARMA, "Herd Behavior in Financial Markets," *IMF Staff Papers*, 2000, *47* (3), 279–310.

____ , D. HIRSHLEIFER, and I. WELCH, "A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades," *Journal of Political Economy*, 1992, *100*, 992–1026.

BREIMAN, L., "Heuristics of Instability and Stabilization in Model Selection," *Annals of Statistics*, 1996, *24*, 2350–2383.

CANER, M., "LASSO Type GMM Estimator," *Econometric Theory, Forthcoming*, 2008.

CHAMBERLIAN, G., "Econometrics and Decision Theory," *Journal of Econometrics*, 2000, *95*, 255–283.

CHEN, X.S. and H. CUI, "An Extended Empirical Likelihood For Generalized Linear Models," *Statistica Sinica*, 2003, *13*, 69–81.

CORCORAN, S.A., "Bartlett Adjustment of Empirical Discrepancy Statistics," *Biometrika*, 1998, *85*, 967–972.

COX, D. D. and F. O'SULLIVAN, "Asymptotic Analysis of Penalized Likelihood and Related Estimators," *The Annals of Statistics*, 1990, *21*, 903–924.

CRESSIE, N. and T. READ, "Read, Multinomial Goodness of Fit Tests," *Journal of the Royal Statistical Society, Series B*, 1984, *46*, 440–464.

DURRETT, A. R., *Probability: Theory and Examples*, Thomson Brooks/Cole, 2005.

FAN, J. and R. LI, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 2001, *96*, 1348–1360.

____ and ____ , "Variable Selection For Cox's Proportional Hazard Model and Frailty Model," *Annals of Statistics*, 2002, *30*, 74–99.

FRANK, I.E. and J.H. FRIEDMAN, "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 1993, *35*, 109–148.

FRENCH, S. and D.R. INSUA, *Statistical Decision Theory*, A Hodder Arnold Publication, 2000.

GEYER, C.J., "On the Asymptotics of Constrained Estimations," *Annals of Statistics*, 1994, *22*, 1993–2010.

GLOSTEN, R. L. and P. R. MILGROM, "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 1985, *14*, 71–100.

GRAHAM, J. R., "Herding among Investment Newsletters: Theory and Evidence," *Journal of Finance*, 1999, *54*, 237–68.

HANSEN, L.P., "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, 1982, *50*, 1029–1054.

HEATON, J. HANSEN L.P. and A. YARON, "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 1996, *14*, 262–280.

HIRANO, K., "Decision Theory in Econometrics," *Dept. of Economics, University of Arizona Working Paper*, 2006.

IMBENS, G.W., "One-Step Estimators for Over-Identified Generalized Method of Moments Models," *Review of Economic Studies*, 1997, *64*, 369–408.

IMBENS, G.W. SPADY R.H. and P. JOHNSON, "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 1998, *66*, 333–357.

JAMES, W. and C. STEIN, "Estimation With Quadratic Loss," in "Berkeley Symposiume on Mathematical Statistics and Probability" Univ. of Calif. Press. 1961, pp. 361–379.

JING, B.Y. and A.T.A. WOOD, "Exponential Empirical Likelihood is Not Bartlett Correctable," *Annals of Statistics*, 1996, *24*, 365–369.

KABAILA, P., "The Effect of Model Selection on Confidence Regions and Prediction Regions," *Econometric Theory*, 1995, *11*, 537–549.

KIM, J. and D. POLLARD, "Cube root asymptotics," *The Annals of Statistics*, 1990, *18*, 191–219.

KITAMURA, Y. and M. STUTZER, "An Information Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 1997, *65*, 861–874.

KNIGHT, K., "Epi-Convergence in Distribution and Stochastic Equi-Semicontinuity," *University of Toronto, Department of Statistics, Working Paper*, 2003.

____ **and W. FU**, "Asymptotics for Lasso-type Estimators," *Annals of Statistics*, 2000, *28*, 1356–1378.

KOLACZY, E. D., "An Information Criterion for Empirical Likelihood with General Estimating Equations," 1995. unpublished manuscript-Department of Statistics, University of Chicago.

KOLACZYK, E.D., "Empirical Likelihood For Generalized Linear Model," *Statistics Sinica*, 1994, *4*, 199–218.

KOLMOGOROV, A. N. and V. M. TIHOMIROV, "ε-entropy and ε-capacity of Sets in Function Spaces," *Uspekhil Math. Nauk.*, 1959, *14*, 3–86. English translation, American Math. Soc. Transl. 277-364 (1961).

LAZAR, A.N., "Bayesian Empirical Likelihood," *Biometrika*, 2003, *90*, 319–326.

LEHMANN, E.L. and G. CASELLA, *Theory of Point Estimation*, Springer-Verlag, 1998.

LINDSEY, J., *Applying Generalized Linear Models*, Springer-Verlag, 1997.

McCULLAGH, P. and J.A. NELDER, *Generalized Linear Models*, London: Chapman and Hall, 1990.

MILGROM, P. R. and N. STOKEY, "Information, Trade and Common Knowledge," *Journal of Economic Theory*, 1982, *26*, 17–27.

MORRIS, C.N., "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 1983, *78*, 47–55.

NATTINGER, A. B. and OTHERS, "Effect of Nancy Reagan's Mastectomy on Choice of Surgery for Breast Cancer by US Women," *JAMA*, 1998, *279*, 762–766.

NEWEY, W.K. and McFADDEN, "Large Sample Estimation and Hypothesis Testing," in R. F. Engle and D. McFadden, eds., *Handbook of Econometrics*, Vol. 4, Elsevier, 1986, chapter 36, pp. 2111–2245.

____ **and R.J. SMITH**, "Higher order properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 2004, *72*, 219–255.

OWEN, A.B., "Empirical Likelihood Ratio Confidence Intervals for a single Functional," *Biometrika*, 1988, *75*, 237–249.

____ , *Empirical Likelihood*, London: Chapman and Hall, 2001.

POLLARD, D., *Convergence of Stochastic Processes*, New York: Springer-Verlag, 1984.

PRESTON, B. HOUNG H. and M. SHUM, "Generalized Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models," *Econometric Theory*, 2003, *19*, 923–943.

QIN, J. and J. LAWLESS, "Empirical Likelihood and Generalized Estimating Equations," *Ann. Statist.*, 1994.

RAMALHO, J.S., "Small Sample Bias of Alternative Estimation Methdos for Moment Condition Models: Monte Carlo Evidence for Covariance Structures," *Studies in Nonlinear Dynamics & Econometrics*, 2005, *9* (1).

____ , "Bootstrap Bias-Adjusted GMM Estimators," *Economics letters*, 2006, *92*, 149–155.

RAO, C. R. and Y. WU, "On Model Selection," IMS Lecture Notes - Monograph Series (2001) Volume 38 2001.

SCHARFSTEIN, D. S. and J. C. STEIN, "Herd Behavior and Investment," *American Economic Review*, 1990, *80* (3).

SCHENNACH, S.M., "Bayesian Exponentiaally Tilted Empirical Likelihood," *Boimetrika*, 2005, *92*, 31–46.

____ , "Point Estimation with Exponentially Tilted Empirical Likelihood," *Ann. Statist.*, 2007, *35*, 634–672.

SCLOVE, S.L., "Improved Estimators for Coefficients in Linear Rergression," *Journal of American Statist. Assco.*, 1968, *63*, 597–606.

SHEN, X., "On Method of Sieve and Penalization," *The Annals of Statistics*, 1997, *25*, 2555–2592.

____ , "On the Method of Penalization," *Statistica Sinica*, 1998, *8*, 337–357.

SIRVASTAVA, R. K. and OTHERS, "Market-Based Asset and Shareholder Value: A Framework for Analysis," *Journal of Marketing*, 1998.

SLATER, I. R., *SOROS: The Unauthorized Biography, the Life, Times and Trading Secrets of the World's Greatest Investor*, McGraw-Hill, 1997.

SMITH, R.J., "Alternative semi-parametric likelihood approaches to generalized method of moments estimation," *Economic Journal*, 1997, *107*, 503–519.

____ , "Weak Instruments and Empirical Likelihood," *Working Paper, University of Cambridge*, 2005.

TAUCHEN, G., "Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained from Financial Market Data," *Journal of Business and Economic Statistics*, 1986, *4*, 397–416.

TIBSHIRANI, R.J., "Regression Shrinkage and Selection Via Lasso," *Journal of The Royal Statistical Society Series B*, 1996, *58*, 267–288.

TRUEMAN, B., "Analyst Forecasts and Herding Behavior," *eview of Financial Studies*, 1994, *7*, 97–124.

VAART, A.W. VAN DER and WELLNER, *Weak Convergence and Empirical Processes*, New York: Springer Verlag, 1996.

WELCH, I., "Herding Among Security Analysts," *Journal of Financial Economics*, 2000, *58* (3), 369–96.

WONG, W. H. and X. SHEN, "Probability Inequalities for Likelihood Ratios and Convergence Rate of Sieve MLEs," *The Annals of Statistics*, 1995, *18*, 339–362.

ZHANG, J. and I. GIJBELS, "Sieve Empirical Likelihood and Extensions of the Generalized Least Squares," *Scandinavian Journal of Statistics*, 2003, *30*, 1–24.

ZHAO, P. and B. YU, "On Model Selection Consistency of Lasso.," *Journal of Machine Learning Research*, 2006, *7*, 2541–2563.

ZWIEBEL, J., "Corporate Conservatism and Relative Compensation," *Journal of Political Economy*, 1995, *103* (1), 1–25.