Density Estimation and Combination under Model Ambiguity via a Pilot Nonparametric Estimate: an Application to Stock Returns.

Stefania D'Amico*

First Version: May 2003, This Version: September 2003
Primary Job Market Paper

Abstract

This paper proposes a method to estimate the probability density of a variable of interest in the presence of model ambiguity. In the first step, each candidate parametric model is estimated minimizing the Kullback-Leibler 'distance' (KLD) from a reference nonparametric density estimate. Given that the KLD represents a measure of uncertainty about the true structure, in the second step, its information content is used to rank and combine the estimated models. The paper shows that the resulting parameters estimator is root-n consistent and asymptotically normally distributed. The KLD between the nonparametric and the parametric density estimates is also shown to be asymptotically normally distributed. This result leads to determine the weights in the model combination, using the distribution function of a Normal centered on the average performance of all plausible models. As such, this combination technique does not require that the true structure belongs to the set of competing models and is computationally simple. I apply the proposed method to estimate the density function of daily stock returns under different phases of the business cycle. The results indicate that the double Gamma distribution is more adequate than the Gaussian distribution in modeling stock returns, and that the models combination outperforms in- and out-of-sample each individual candidate model. I also explore the model's implications for the optimal share to invest in the risky asset.

*I am indebeted to Phoebus Dhrymes for his guidance and many interesting discussions. I wish to thank Jean Boivin, Xiaohong Chen, Mitali Das, Rajeev Dehejia, Marc Henry and Alexei Onatski for valuable comments. I would also like to thank Stefano Eusepi and Mira Farka, without their constant support this paper would have never existed.
Address: Department of Economics, Columbia University. E-mail: sd445@columbia.edu

1 Introduction

> "Prediction may be regarded as a special type of decision making under uncertainty : the acts available to the predictor are the possible predictions, and the possible outcomes are success (for a correct prediction) and failure (for a wrong one). In a more general model, one may also rank predictions on a continuous scale, measuring the proximity of the prediction to the eventuality that actually transpires, allow set-valued predictions, probabilistic predictions, and so forth."[1]

Econometric models are implemented in order to deal with uncertainty and guide decisions. Nevertheless, very often econometric models are developed without any reference to the "uncertainty about the model" that characterizes the decision context. To this end, because of the complexity of the decision-setting and the level of approximation embodied in a simple model, I contemplate the presence of model ambiguity. In other words, instead of specifying a unique statistical structure and treat it as the true model, I consider a set of competing models.

Empirical models are based on the idea that the occurrence of events (i.e. the data) reveals information. Typically, although the available database is not sufficient to choose a unique well-defined model, it still provides relevant knowledge that can be used to differentiate among priors. In this study, a pilot nonparametric density, summarizing all the information contained in the data, is used to estimate and rank candidate parametric models.

Furthermore, since the model classes can be large due to high uncertainty, it is necessary to develop a tool to combine the different models in a weighted predictive distribution, where the weights are determined by the ignorance about the true structure. This model combination provides an explicit representation of uncertainty across models and allows to extract information from 'all' plausible ones.

It is sensible to think that, since we do not know the true model and we approximate it by choosing among a set of candidate models, at most we can aspire to estimate its best approximation. Because parsimony and computational simplicity are desirable characteristics of an econometric model, typically the set of competing models consists of simple parametric alternatives, even when a better infinite-dimensional approximation is available. This implies that most likely, the true model does not even belong to the set of candidates and that more than one model can perform fairly well, such that it can be hard to discard one of them. In these cases, the models combination could provide a better hedge against the lack of knowledge of the correct structure and outperform each competing model including the best one.

This modelling approach will permit to study and exploit model misspecification which is defined as the discrepancy between the candidate and the actual model. Since probabilistic models are often used as the belief of an "expected utility maximizer", ignoring this misspecification will cause a higher risk of the optimal decision. For this reason, this study focuses on the formation of an econometric model as a general-purpose tool: to quantify the plausibility of different probabilistic models, to combine them in a unique distribution and to explore the impact of the latter on the derivation of optimal choice under uncertainty.

The selection of parametric candidate models in combination with the simple device developed to determine their probability of being correct, provides a closed form solution of the optimal choice even when the predictive density is the models combination. This simplicity has little cost in terms of information, since through the models' weights we are still able to account for model misspecification and to extract information from a nonparametric estimate.

[1] Gilboa I. and D. Schmeidler ; "A theory of Case-Based Decisions", 2001, pp 59-60.

I develop a method of prediction that ranks different probabilistic models and determines the optimal value of their parameters, maximizing the sum of their similarities to relevant past cases. The similarity is measured by the opposite of the distance, that is the Kullback-Leibler Information (KI), between the probabilistic model and the estimated nonparametric density. The final weights used to combine models are function of these distances which embody the ignorance about the true structure. The cognitive plausibility of my methodology is founded on case-based decision theory (CBDT). In particular the behavioral axioms of Inductive Inference[2] developed by Gilboa and Schmeilder (2001) provide support for my prediction method.

This estimation approach, being based only on an objective measure of the proximity between multiple candidate models and actual data, aims to overcome the necessity to have a specific prior over the set of models and about parameters belonging to each of the models under consideration. It refers only to the analogy between past samples (actually encountered cases) and models at hand[3]. This requires a limited amount of hypothetical reasoning since it relies directly on data that are available to any observer without ambiguity.

I apply the proposed method to determine the predictive density of daily stock returns under different phases of the business cycle and I use the latter to investigate the implications of the model on portfolio choice under uncertainty. This empirical application is motivated both by the difficulty in estimating the probability law of asset returns which usually are modelled with misspecified density function, and by the large availability of data for financial series which facilitates the use of nonparametric techniques.

This way of implementing probabilistic prediction is essential to improve econometric modeling and to decision making. In fact, my method like others in the literature, can be considered as a preliminary step to account explicitly for model ambiguity in econometrics. One of the first studies that uses information criteria to identify the most adequate regression model among a set of alternatives is due to Sawa (1978). Later contributions by White (1980,'82) examine the detection and consequences of model misspecification in Nonlinear Regression Model and MLE. A Subsequent work by Sin and White (1996) uses information criteria for selecting misspecified parametric models. More recently, a paper by Skouras (2001), investigates the determination of a predictive density by exploiting the discrepancy between expected utilities under the true and the misspecified model. Nevertheless, none of these studies makes use of a preliminary nonparametric estimation to estimate and distinguish among alternative models[4]. On the other hand, a study by Cristobal, Roca and Manteiga (1987) which describes linear regression parameter estimators based on preliminary nonparametric estimation does not incorporate the assumption of model uncertainty. To my knowledge, this is the first work which develops an estimation technique via a pilot nonparametric estimate under the assumption of model ambiguity. Furthermore and more importantly, none of these papers focuses on model combination.

There are three additional strands of literature related to this work. The first includes Bayesian Model Averaging (BMA) and its application to stock returns predictability and to the investment opportunity set, see for example Avramov (2002) and Cremers (2002). Differently from the Bayesian approach, in this study it is not necessary to assume that the true structure belongs to the set of candidate models. Further, the implementation of the model combination is computationally extremely easy because it does not require numerical integration to obtain for each model the 'probability' of being correct. The second vein, though characterized by a completely different approach, represents the studies about forecast evaluation

[2] As shown in Gilboa-Schmeidler (2001) this is also the same principle at the base of Maximum Likelihood Estimation.

[3] However also these models are suggested by past experience or by economic theory.

[4] The literature on nonparametric testing provides me the technical machinery to derive the asymptotic distribution of the KLI. See for example Hall(1984, 1987), Robinson(1991), Fan(1994), Zheng (1996, 2000), and Hong and White(2000).

and combination: Diebold and Lopez (1996), Hendry and Clements (2001) and Giacomini (2003) among others. Finally, the third strand consists of the vast literature on dynamic portfolio choice under model misspecification where investors try to learn from historical data, see for example Uppal and Wang (2002) and Knox(2003).

The paper is organized as follows: Section II describes the estimation and selection method; section III illustrates the models combination technique; Section IV analyzes the asymptotic properties of the parameters estimator and the asymptotic distribution of the uncertainty measure; Section V discusses the finite sample performance of the parameters estimator; Section VI contains the empirical application to stock returns; Section VII investigates the model's implications for the optimal asset allocation; and Section VIII concludes. Analytical proofs and technical issues are discussed in the Appendix.

2 Description of the estimation and selection method

I consider a prediction problem for which a finite set of candidate models $\mathcal{M} \equiv \{M_j, j = 1, ..., J\}$ is given. In particular, these models M_j are defined as probability density functions $f_j(x;\theta)$ $\{f : \mathcal{R} \rightarrow [0,\infty]\}$ of a random variable of interest X $\{X : \Omega \rightarrow \mathcal{R}\}$ defined on the probability space $(\Omega, \mathcal{A}, \mathcal{P})$ taking values in $(\mathcal{R}, \mathcal{B}(\mathcal{R}), P_x)$. The goal of the predictor is to estimate and rank these models according to their similarity to past observations, and finally to combine them in a similarity-weighted probability distribution. Given the set $\mathcal{M}$, we define the set of elements that have to be ranked as $\Theta = \left\{\theta_{M_j} : f_j(x;\theta) \equiv M_j \in \mathcal{M}\right\}$, and $\Theta \subset \mathcal{R}^k$.

Since in the empirical analysis, I want to allow the random variable of interest to follow a different distribution over different regimes, I define an additional finite set $\mathcal{S}$ which is the set of the states of nature. Define the state s $\{s : \mathcal{S} \rightarrow Z_+,\ Z_+$ is the set of positive integers$\}$ a random variable defined on the probability space $(\mathcal{S}, \sigma(\mathcal{S}), p)$, taking on only discrete values. Further, in order to focus the attention only on the uncertainty about the model, let me assume that s can be observed. Thus, the model's definition is equal to $f_j(x/s;\theta)$ and Θ equals $\left\{\theta_{M_{js}} : f_j(x/s;\theta) \equiv M_{js} \in \mathcal{M}\right\}$.

The information set Ω is a finite set of Q samples of N_q independent realizations of the random variable X. Given the set Ω, its information content is processed estimating a nonparametric density $\widehat{f_n}\ (x/s)$ for each sample $q = 1, ...Q$. Subsequently, from the set Ω, I derive the set of past cases $\mathcal{C} = \left\{\widehat{f_{nq}}(x/s) : x \in \Omega \text{ and } s \in \mathcal{S}\right\}$, which is the final information that the predictor posses to judge the different models. I assume that, given a regime, all the subsamples derive from the same fixed distribution. The problem is then to describe how to process and recall this information to assess the similarity of past observations to the set of candidate models.

Lets define the weight a map $w : \Theta \times \mathcal{C} \rightarrow \mathcal{R}$, it assigns a numerical value w_{qj} to each pair of past case $\widehat{f_{nq}}(x/s)$ and parameter $\theta_{M_{js}}$, representing the support that this case lends to the model $f_j(x/s;\theta)$ in $\mathcal{M}$.

The sum of weights w_{qj} represents the tool through which the predictor judges the similarity of a particular model to the estimated distributions which his knowledge is equipped with. More precisely these weights represent the degree of support that past distributions lend to the specific model at hand. However, they also embody the misspecification contained in each model, that being just an approximation of the reality still preserves a distance from the actual data. It seems reasonable that the model with the lowest distance from the nonparametric densities, is also the model with the highest similarity to past observations. As such, it has to be the model characterized by the highest sum of weights.

For these reasons, it seems natural to determine w_{qj} by the opposite of the distance between the non-parametric density $\widehat{f_{nq}}(x/s)$ and the model $f_j(x/s;\theta)$:

$$w_{qj} = -KI(\widehat{f_{nq}}(x), f_j(x/s;\theta)), \tag{1}$$

where $KI(\widehat{f_{nq}}(x), f_j(x/s,\theta))$ is the Kullback-Leibler distance[5], whose empirical version in this study is defined as follows:

$$\widehat{KI}_{qj} = \sum_{i=1}^{N_q} \widehat{f_{nq}}(x_i) \log\{\frac{\widehat{f_{nq}}(x_i)}{f_j(x_i,\theta)}\}. \tag{2}$$

where i is the index for all observations contained in a sample q. For simplicity I dropped the index relative to the regime s.

If the values of the optimal parameters were known, the prediction rule - ranking the plausibility of each model through the sum of their weights (over the past cases) - will lead us to choose as predictive density f_1 rather than f_2 if and only if:

$$\sum_{q\in C_s} w_{q1} > \sum_{q\in Cs} w_{q2}, \tag{3}$$

(where C_s is a partition of C and represents the set of past cases relative to regime s) or equivalently:

$$\sum_{q\in C_s} KI(\widehat{f_{nq}}(x), f_1(x/s;\theta)) < \sum_{q\in C_s} KI(\widehat{f_{nq}}(x), f_2(x/s;\theta)). \tag{4}$$

The sum of the weights relative to model f_1 can be interpreted as in Gilboa and Schmeilder (2001) as the "aggregate similarity or plausibility" of model f_1. However, as the values of the optimal parameters are unknown, it is necessary to estimate them. Since the model with the largest aggregate similarity to past cases is the most appropriate to achieve a good prediction, the candidate model's parameters $\theta_{M_{js}}$ are obtained in the following way:

$$\max_{\theta_{M_{js}}} \sum_{q\in Cs} w_{qj} = \min_{\theta_{M_{js}}} \sum_{q\in Cs} KI(\widehat{f_{nq}}(x), f_j(x/s;\theta)). \tag{5}$$

The minimization of the sum of these pseudo-distances allows us to obtain the optimal minimum contrast (MC) estimates[6] of the parameters that characterize the a priori distributions. This method gives us the opportunity to extract the information contained in a nonparametric estimate, while preserving the simplicity of a parametric model. This goal can be achieved by density-matching: the optimal model is derived to be consistent with the observed distribution of the data[7].

It follows then that the rank of the competing models is obtained as follows:

$$f_1 \succ f_2 \text{ IFF } \min_{\theta_{M_1}\in\Theta} \sum_{q\in Cs} KI(\widehat{f_{nq}}(x), f_1(x/s;\theta)) < \min_{\theta_{M_2}\in\Theta} \sum_{q\in Cs} KI(\widehat{f_{nq}}(x), f_2(x/s;\theta)), \tag{6}$$

which in turn implies that the best model can be represented by the following prediction rule:

$$\inf_{\{\ j:1,\ldots,J\}} \left\{ \min_{\theta_{M_j}\in\Theta} \sum_{q\in Cs} KI(\widehat{f_{nq}}(x), f_j(x/s;\theta)) \right\}. \tag{7}$$

[5] We can choose many other distances, on this purpose see Ullah A.(1996).
[6] See Dhrymes P. J. (1994) p. 282 .
[7] See Aït Sahalia Y. (1996).

It is easy to note that if we have a unique sample for each regime ($q = 1$), or alternatively under the assumptions that all samples derive from the same fixed distribution, parameter estimation is reduced to the minimization of a unique KI where the true model is approximated by only one nonparametric density, that is the nonparametric equivalent of quasi maximum likelihood estimation (NPQMLE). In this case the similarity weight is defined as follows:

$$w_{ij} = \log f_j(x_i/s;\theta)\widehat{f_n}(x_i), \tag{8}$$

and hence the maximization problem is given by:

$$\max_{\theta_{M_j}} \sum_{i=1}^{N} w_{ij} = \max_{\theta_{M_j}} \sum_{i=1}^{N} \log f_j(x_i/s;\theta)\widehat{f_n}(x_i), \tag{9}$$

since this is the only part of KI that depends on the parameters[8].

It is easily observable that in this approach as in QMLE the criterion functional to be maximized is

$$\int \log f_j(x/s;\theta) dF_n(x).$$

But, while in QMLE the weighting function $F_n(x)$ is chosen to be equal to

$$F_n(x) = \frac{1}{N}\sum_{i=1}^{N} 1_{(-\infty,x]}(x_i),$$

such that the empirical criterion becomes:

$$\frac{1}{N}\sum_{i=1}^{N} \log f_j(x_i/s;\theta);$$

in NPQMLE, $F_n(x)$ is chosen to have the following form

$$F_n(x) = \int_{-\infty}^{x} \widehat{f_n}(x)dx = \frac{1}{nh}\sum_{i=1}^{n\leq N}\int_{-\infty}^{x} K_h(x_i - x)dx,$$

such that the empirical criterion becomes equation (2). This implies that not all observations will obtain a mass equal to $\frac{1}{n}$. In contrast each observation will be rescaled by a smoother weight that depends on the bandwidth parameter h and the kernel function K. It is this different weighting scheme that allows in finite samples a 'better' performance than QMLE estimation. In particular, as it is documented through a limited set of Monte Carlo experiments, the parameter estimates obtained using NPQMLE deliver a KI between the true model and the parametric candidate which is smaller than that obtained using QMLE.

Finally, in this context the best model attains solving the following problem:

$$\inf_{\{j:1,\ldots,J\}} \left\{ \min_{\theta_{M_j}\in\Theta} \sum_{i=1}^{N} \left(\log \widehat{f_n}(x_i) - \log f_j(x_i/s;\theta)\right)\widehat{f_n}(x_i) \right\}. \tag{10}$$

[8] Minimizing the functional $\int \left(\log \widehat{f_n}(x_i) - \log f_j(x_i/s;\theta)\right) dF_n(x)$ or maximizing $\int \log f_j(x_i/s;\theta) dF_n(x)$ w.r.t θ provides the same results.

3 Description of the combination method

Selecting a single model as described in the previous section, even if implicitly recognizes the presence of misspecification, does not account explicitly for model ambiguity. More importantly, it does not consider that the true structure may not belong to the initial set of candidate models, as such to use only the best minimizer is not necessarily the ultimate solution. This implies that in order to incorporate the information contained in the KI, the combination of all plausible models in a similarity-weighted predictive distribution is needed, where the weights are function of $\widehat{KI}_j\left(\widehat{f_n}(x), f_j(x/s;\widehat{\theta})\right)$.

The intuition is the following : KI_j, can be interpreted as a measure of uncertainty or ignorance about the true structure**.** When computed at the optimal value of the parameter $\widehat{\theta}_{M_j}$, it can be considered as a measure of the goodness of the model, since it represents the margin of error of this model in a particular sample. If it is different from zero for each candidate distribution and/or there are many models that exhibit a similar loss, then the econometrician fearing misspecification will explicitly account for it combining the models in the predictive distribution $M(\widehat{\theta}_{M_j}) = \sum_j p_j(\widehat{KI}) f_j(x/s, \widehat{\theta})$. The similarity-weight $p_j(\widehat{KI})$ can be loosely interpreted as the probability of model M_j to be correct. In contrast, if the predictor selected a single distribution M_j, he would overestimate the precision of this model, since he would implicitly assign to the model probability $(p_j(\widehat{KI}))$ of being correct equal one.

In order to better appreciate the importance of the information contained in the model's misspecification and subsequently in $M(\widehat{\theta}_{M_j})$, it is necessary to give a brief description of the spaces in which we operate when the statistical structural assumptions are not necessarily true. Define G the space of functions to which the true unknown model $g(x/s)$ belongs: by assumption $g(x/s)$ minimizes the KI over G. $F_{\Theta_{M_j}} \subseteq G$ represents the finite dimensional space to which the parametric candidate models belong, we can call it the approximation space and it is also the space where the estimation is carried out. The best approximation $f_j(x/s, \theta^*)$ in $F_{\Theta_{M_j}}$ to the function $g(x/s)$ is the p.d.f. that minimizes the KI over $F_{\Theta_{M_j}}$, while $f(x/s, \widehat{\theta}) \in F_{\Theta_{M_j}}$ minimizes the sample version of the KI. The distance between $f(x/s, \widehat{\theta})$ and $f(x/s, \theta^*)$ represents the estimation error which vanishes as $n \to \infty$. Instead, the approximation error[9] given by the distance between $f(x/s, \theta^*)$ and $g(x/s)$, can be reduced only if the dimension of $F_{\Theta_{M_j}}$ grows with the sample size[10]. Model combination can therefore be considered as a method to increase the dimension of the parameter space accounting for the approximation error.

Only if $F_{\Theta_{M_j}} \equiv G$, then $g(x) = f(\theta_0, x) = f(\theta^*, x)$ and $\widehat{\theta}$ is a consistent estimator of the true parameter θ_0. Typically, because of the advantages[11] offered by parsimonious models, $F_{\Theta_{M_j}}$ is a small subset of G and hence model misspecification can be a serious problem also affecting the asymptotic results. Furthermore, in finite sample the $\widehat{KI}_j$ embodies information about both the estimation and approximation errors relative to M_j, and as such it can not be ignored.

Once it is decided to use the combinations of p.d.f. $M(\widehat{\theta}_{M_j})$ as predictive density, the main task consists in determining the probability $p_j(\widehat{KI})$. For this purpose I show that (see section IV and the Appendix for more details) $\widehat{KI}_j$ minus a correction term (m_n), mainly due to the approximation error, is asymptotically distributed Normal $N(0, 2\sigma^2)$, where a consistent estimate of σ^2 is determined only by the nonparametric density. Then, the probability of being the correct model can be determined by the probability of obtaining

[9] See Chen X. and J.Z. Huang (2002).

[10] For example a countable mixture of Normals (Ferguson (1983)) or the kernel density estimator (Silverman (1986)) can approximate arbitrarly close any well-behaving density function. We can view these models as infinite-dimensional parameter alternatives.

[11] Closed form solution, ease of manipulation and low computational costs.

a misspecification $\widehat{KI}_j$ worse than the one actually obtained. That is:

$$p_j(\widehat{KI}) = 1 - P(\widehat{KI}_j \leq ki). \tag{11}$$

Since it is well known that $KI(g, f_j(\theta)) \geq 0$, where the equality attains if and only if $g = f_j$, then $p_j(\widehat{KI}) = 1$ if and only if $ki = 0$. This follows trivially from the fact that $P(\widehat{KI}_j \leq 0) = 0$. Consequently, $p_j(\widehat{KI})$ will be less than one for any positive realization of $\widehat{KI}_j$. Accordingly, if the ki is very small, then the probability $(P(\widehat{KI}_j \leq ki))$ of obtaining a realization of the misspecification even smaller than a such low value will be very little; it then follows that the probability $p_j(\widehat{KI})$ of having a good model will be very high.

It is clear that to determine the weight it is just sufficient to compute the c.d.f of a Normal with mean m_n and variance $2\sigma^2$ for the realized value ki. Nevertheless, in the implementation of this methodology, it is necessary to pay attention to the mean m_n that, being affected by the approximation error, varies with the candidate model. In the next section and in the appendix, the device to fix this problem and the measurement of m_n are described in more details.

4 Asymptotic results

Before proceeding with the theorems let me state first all the assumptions[12]:

A1:$\{X_i\}$ are i.i.d with compact support S, their marginal density g exists, is bounded away from zero, and is twice differentiable. Its first order derivative is also bounded and moreover $|g''(x_1) - g''(x_2)| \leq C\,|x_1 - x_2|$ for any $x_1, x_2 \in S$ and for some $C \in (0, \infty)$.

A2: The kernel K is a bounded symmetric probability density function around zero, s.t :(i) $\int K(u)du = 1$; (ii) $\int u^2 K(u)du < \infty$; (iii) $h = h_n \to 0$ as $n \to \infty$; (iv) $nh_n \to \infty$ as $n \to \infty$.

A3: Depending on the application, it is possible to select a kernel K that satisfies A2 and such that the tail-effect terms involved in the use of the KI are negligible.

A4: Θ is a compact and convex subset of R^k, the family of distributions $F(\theta_{M_j})$ has density $f_j(\theta, x)$ which are measurable in x for every $\theta_{M_j} \in \Theta$ and continuous in θ_{M_j} for every $x \in \Omega$; $E_g[\log g(x) - \log f_j(\theta, x)]$ exists and has a unique minimum at an interior point $\theta^*_{M_j}$of Θ; $\log f_j(\theta, x)$ is bounded by a function $b(x)$ for all $\theta_{M_j} \in \Theta$, where $b(x)$ is integrable w.r.t. the true distribution G.

A5: the first and second derivative of $\log f_j(\theta, x)$ w.r.t. θ_{M_j} and $\left|\frac{\partial \log f_j(\theta,x)}{\partial \theta} \times \frac{\partial \log f_j(\theta,x)}{\partial \theta}\right|$ are also dominated by $b(x)$; $B(\theta^*_{M_j})$ is non singular and $A(\theta^*_{M_j})$ has a constant rank in some open neighborhood of $\theta^*_{M_j}$; where $B(\theta^*_{M_j}) = E\left[\left(\frac{\partial \log f_j(\theta^*,x)}{\partial \theta} \times \frac{\partial \log f_j(\theta^*,x)}{\partial \theta}\right) g^2(x)\right]$ and $A(\theta^*_{M_j}) = E\left[\frac{\partial^2 \log f_j(\theta^*,x)}{\partial \theta_i \partial \theta_j} g(x)\right]$.

Assumption A1 requires that X_i are continuously distributed and imposes regularity conditions on the unknown density g. A2 represents the standard assumptions on the kernel function and the smoothing parameter used in the nonparametric literature. Assumption A3 is a practical assumption that we need in order to simplify the proofs and ignore the tail-effects due to the use of the Kullback-Leibler distance. As indicated by Hall(1987) it is important that K is chosen such that its tails are sufficiently thick with respect to the tails of the underlying function $f_j(\theta, x)$. Since we know the candidate parametric models it is always possible to choose an adequate Kernel. Furthermore, Hall suggested a practical alternative which is given by the Kernel $K(u) = 0.1438 * \exp[-\frac{1}{2}\{\log(1 + |u|)\}^2]$ whose tails decrease more slowly than the tails of

[12]It is important to note that from now on, for simplicity, I drop the index indicating the regime s.

the Gaussian Kernel and that allows in most cases to neglect the tails-effect terms. Finally, the last two assumptions A4 and A5 are standard to ensure the consistency and asymptotic normality of QMLE (White (1982)).

4.1 Consistency and Asymptotic Normality of the NPQMLE estimator

For the NPQMLE parameter estimator $\widehat{\theta}_{M_j}$ we have the following results:

THEOREM 1 (Consistency): *Given Assumptions* A1-A4, *as* $n \longrightarrow \infty$, $\widehat{\theta_{n}}_{M_j} \longrightarrow \theta^*_{M_j}$ *with probability 1.*
Proof: See the Appendix

The main idea is that if KI_{nj} is a contrast[13] relative to the contrast function $KI(g,\ f_j(\theta,x))$, that is it converges at least in probability to $KI(g, f_j(\theta,x))$, and $\theta^*_{M_j}$ is the unique minimizer in Θ of $KI(g, f_j(\theta,x))$, then the sequence $\widehat{\theta_{n}}_{M_j}$in Θ that minimizes $\widehat{KI}_{nj}$ will converge to $\theta^*_{M_j}$.

This implies the following: given that each candidate parametric model is potentially misspecified, since we do not know the true model and we do not even know if it belongs to the set of candidate models, the NPQMLE estimation procedure, as QMLE, will provide an estimator that converges to the best approximation $\theta^*_{M_j}$. In other words, it converges to the best we can attain given that we are minimizing the Kullback-Leibler information over a space $F_{\Theta_{M_j}} \subseteq G$ rather than G. From now on, for simplicity $\widehat{\theta_{n}}_{M_j} = \widehat{\theta_n}$ and $\theta^*_{M_j} = \theta^*$.

Next I establish that NPQMLE has a limiting normal distribution with mean zero and variance-covariance matrix $C(\theta^*)$, and that it is root-n consistent, that is it has the same rate of convergence of parametric method as QMLE. In particular, similarly to Powell-Stock and Stoker(1989) this estimator converges faster than the nonparametric density $\widehat{f_n}$ exploited in the estimation technique, therefore avoiding the necessity for extremely large dataset. It is much easier to understand the rationale for this convergence rate by observing the U-statistic representation of the first order condition to derive the optimal value of the parameters:

$$\binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{h} K(\frac{x_j - x_i}{h}) \left[s(\theta, x_i) - s(\theta, x_j) \right], \text{ where } s(\theta, x) = \frac{\partial \log f(\theta, x)}{\partial \theta}.$$

As in Powell-Stock and Stoker(1989), it follows by the averaging of the nonparametric density estimate $\widehat{f}_n$, which appears in the previous formula in the particular form:

$$\widehat{f}_n(x_i) = \frac{1}{n-1} \sum_{j=1}^{n} \frac{1}{h} K(\frac{x_j - x_i}{h}).$$

Thus, I have the next result

THEOREM 2: (Asymptotic Normality): *Given Assumptions* A1-A5, *and given that* $E\left[\|H_n(x_i, x_j)\|^2\right] = o(n)$, *where*

$$H_n(x_i, x_j) = \frac{1}{h} K(\frac{x_j - x_i}{h}) \left[s(\theta^*, x_i) - s(\theta^*, x_j) \right]$$

then

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \sim^A N(0, C(\theta^*))$$

[13] See Definition 3 and 4 Dhrymes (1998).

Furthermore,

$$C_n(\widehat{\theta}_n) \longrightarrow^{a.c} C(\theta^*)$$

(Proof: See the Appendix)

where

$$C(\theta^*) = A(\theta^*))^{-1} B(\theta^*) A(\theta^*))^{-1} \tag{12}$$

$$A_n(\theta) = \left(\frac{1}{n} \sum_i \frac{\partial^2 \log f(\theta, x_i)}{\partial \theta} \widehat{f}_n(x_i) \right) \tag{13}$$

$$B_n(\theta) = \left(\frac{1}{n} \sum_i \frac{\partial \log f(\theta, x_i)}{\partial \theta_i} \cdot \frac{\partial \log f(\theta, x_i)}{\partial \theta_j} \widehat{f}_n^2(x_i) \right) \tag{14}$$

and

$$A(\theta) = E\left[\frac{\partial^2 \log f(\theta, x_i)}{\partial \theta_i \partial \theta_j} g(x_i) \right] \tag{15}$$

$$B(\theta) = E\left[\frac{\partial \log f(\theta, x_i)}{\partial \theta_i} \cdot \frac{\partial \log f(\theta, x_i)}{\partial \theta_j} g^2(x_i) \right] \tag{16}$$

It is important to point out that also in this framework, similar to White(1982) in the context of QMLE, in the presence of misspecification the covariance matrix $C(\theta^*)$ no longer equals the inverse of Fisher's Information (FI).

4.2 Asymptotic distribution of KI: heuristic approach

In order to obtain the weights in the models combination, as indicated by the formula (11), we need to derive the asymptotic distribution of $\widehat{KI_j}$, the random variable that measures the ignorance about the true structure.

The purpose of this section is to provide a sketch of the proof (developed in the Appendix), in order to give the main intuition and to convey two main pieces of information. First, the effect of estimating the true model g by $f_j(\widehat{\theta}, x)$ on the limiting distribution of $\widehat{KI}_j$. Second, how and which of the different components of the KI affect the mean and variance of the asymptotic distribution.

To simplify the notation I drop the index j and we rewrite $f_j(\widehat{\theta}, x) = f_{\widehat{\theta}}$, $\widehat{f_n}(x) = \widehat{f_n}$ and $g(x) = g$, then $\widehat{KI}$ is given by the following formula:

$$\widehat{KI} = KI(\widehat{f_n}, f_{\widehat{\theta}}) = \int_x (\ln \widehat{f_n} - \ln f_{\widehat{\theta}}) \widehat{f_n} \, dx =$$

$$= \int_x (\ln \widehat{f_n} - \ln g) d\widehat{F}_n - \int_x (\ln f_{\widehat{\theta}} - \ln g) d\widehat{F}_n = \widehat{KI}_1 - \widehat{KI}_2, \tag{17}$$

where the definition of $\widehat{KI}_1$ and $\widehat{KI}_2$ is clear from the previous expression.

1) $\widehat{KI}_1$ can be approximated in the following way[14]:

[14] This can be easily seen by rewriting $\frac{\widehat{f}_n}{g}$ in the following way: $\frac{\widehat{f}_n - g + g}{g} = 1 + \frac{\widehat{f}_n - g}{g} = 1 + \gamma$, then $\ln(1+\gamma) \simeq \gamma - \frac{1}{2}\gamma^2$.

$$\widehat{KI}_1 \simeq \int_x \left(\frac{\widehat{f_n} - g}{g}\right) d\widehat{F}_n - \frac{1}{2}\int_x \left(\frac{\widehat{f_n} - g}{g}\right)^2 d\widehat{F}_n = \widehat{KI}_{11} - \frac{1}{2}\widehat{KI}_{12}, \tag{18}$$

where $\widehat{KI}_{11}$ is a stochastic element that will affect the asymptotic distribution of $\widehat{KI}$, while $\widehat{KI}_{12}$ is roughly[15] the sum of squared bias and variance of $\widehat{f_n}$. It is $O((nh)^{-1} + h^4)$ and it will contribute to the asymptotic mean of $\widehat{KI}$.

2) $\widehat{KI}_2$ has a different nature: it represents the part of the KI that is affected by the parameters estimation. $\widehat{KI}_2$ can be rewritten in the following way:

$$\widehat{KI}_2 = \int_x (\ln f_{\hat{\theta}} - \ln f_{\theta^*}) d\widehat{F}_n + \int_x (\ln f_{\theta^*} - \ln g(x)) d\widehat{F}_n = \widehat{KI}_{21} + \widehat{KI}_{22}. \tag{19}$$

Although in this case, the first term $\widehat{KI}_{21}$ is stochastic, it will not affect the asymptotic distribution of $\widehat{KI}$. In fact, since it is $O_p\left(\frac{1}{n}\right)$ when rescaled by the appropriate convergence rate $d_n = nh^{1/2}$ it converges to zero:

$$d_n \widehat{KI}_{21} \longrightarrow^p 0. \tag{20}$$

The second term $\widehat{KI}_{22}$ has the following behavior:

$$\widehat{KI}_{22} \longrightarrow^p E_g\left[\ln f_{\theta^*} - \ln g(x)\right] = (-KI(g, f_{\theta^*})) \leq 0, \tag{21}$$

as such its presence is due to the approximation error. It is important to note that $\widehat{KI}_{22}$ varies with the underlying candidate model and it can not be observed. This implies that a term of the $\widehat{KI}$'s asymptotic mean will depend on the specific model M_j, then in order to determine and estimate a limiting distribution that is the same for all candidate models the following assumption is needed:

$$\textbf{A6}\text{: } \widehat{KI}_{22} \simeq \alpha h^{1/2} \widehat{KI}_{12}. \tag{22}$$

A6 requires that the mean of the approximation error is proportional to a quantity $\left(\widehat{KI}_{12}\right)$ whose estimation depends only on $\widehat{f_n}$, consequently it will not be influenced by any specific model $f_j(\widehat{\theta}, x)$. Further, when $h \propto n^{-\beta}$ with $\beta \geqslant \frac{1}{5}$, $\widehat{KI}_{12} \sim C(nh)^{-1}$, then we obtain that:

$$d_n \widehat{KI}_{22} \simeq d_n \alpha h^{1/2} \widehat{KI}_{12} \longrightarrow^p \alpha C = E_g\left[\ln f_{\theta^*} - \ln g(x)\right], \tag{23}$$

where C is a known positive constant. Thus collecting all terms together:

$$\widehat{KI} \simeq \widehat{KI}_{11} - \frac{1}{2}\widehat{KI}_{12} - \left(\widehat{KI}_{21} + \widehat{KI}_{22}\right), \tag{24}$$

we have the next theorem:

THEOREM 3: *Given assumptions A1-A6, and given that* $nh^5 \longrightarrow 0$ *as* $n \longrightarrow \infty$, *then*

$$nh^{1/2}\left(\widehat{KI} + \frac{1}{2}\widehat{KI}_{12} + \widehat{KI}_{22}\right) \longrightarrow^d N(0, \sigma^2)$$

[15] In order to see this, it is just sufficient to rewrite $\widehat{KI}_{12}$ as $\int \left(\frac{\widehat{f_n} - E\widehat{f_n} + E\widehat{f_n} - g}{g}\right)^2 d\widehat{F}_n$.

where $\sigma^2 = 2\left\{\int K^2(u)du - \int \left[\int K(u)K(u+v)du\right]^2 dv\right\}$

Proof: See the Appendix.

To better understand the implication of A6 for the determination of the combination weights $p_j(\widehat{KI})$, it is helpful to rewrite the previous result as follows:

$$nh^{1/2}(\widehat{KI} + \frac{1}{2}\widehat{KI_{12}}) \sim^A N(m, \sigma^2) \tag{25}$$

where $m = KI(g, f_{\theta^*}) = \alpha C$, from (21) and (23). This implies that to estimate the mean of the distribution it is necessary to pin down the α, whose estimation is based on the 'plausibility' of the candidate models. Assumption A6 elicits the following definition of plausible model:

Def : $M_j = f_j(\theta, x)$ *is plausible, thus will be included in the set* $\mathcal{M}$*, if the expected value of its approximation error is equal to* αC

In other words, according to A6, all the competing models are on average expected to have the same distance from the true model g. Subsequently, as suggested by the definition of m, α could be estimated by a suitably normalized average of all models' misspecification:

$$\widehat{\alpha} = \frac{1}{J}\sum_j \widehat{KI}_j/C = \widehat{KI}(g, f_{\theta^*})/C \tag{26}$$

Therefore, to obtain $p_j(\widehat{KI})$ we have to employ the c.d.f. of a Normal with mean $E(\widehat{KI}_j)$ and variance σ^2. This entails that, if a model performs better than the average performance of all plausible models, that is $0 < ki_j < \widehat{m}_n$, then it receives an high weight in the models combination. On the other hand, if the model performs poorly relative to all other models, that is $ki_j > \widehat{m}_n$, then its probability of being correct ($p_j(\widehat{KI})$) will be low.

5 Finite sample performance of the NPQMLE estimator

In order to analyze the behavior of the parameter estimator in finite sample, we provide the results of a set of Montecarlo experiments, where we use the Kullback-Leibler distance between the true and the estimated model ($KI(g, f_j(\widehat{\theta}))$) to judge the goodness of the estimation methodology, and to compare it to QMLE.

We use 1000 iterations for each experiment. At each iteration, I first generate the data according to some distribution g that represents the true model; second, estimate the nonparametric density $\widehat{f}_n$ using a second order Gaussian kernel; third, determine the optimal value of the parameters minimizing the KI between $\widehat{f}_n$ and each candidate model $f_j(\theta, x)$; fourth, evaluate KI between g and $f_j(\widehat{\theta}, x)$ at $\widehat{\theta}_{NPQMLE}$ and $\widehat{\theta}_{QMLE}$ respectively. Finally, to evaluate the performance of NPQMLE and to compare it to QMLE we compute the average of $KI(g, f_j(\widehat{\theta}_{NPQMLE}))$ and of $KI(g, f_j(\widehat{\theta}_{QMLE}))$ out of 1000 stored values. The smoothing parameter h is chosen according to $h = 1.06\widehat{\sigma}_x n^{-\beta}$ where $0 < \beta < 1$. Further, in accordance with Theorem 3, β must satisfy $\frac{1}{5} < \beta < 1$.

The basic design is as follows:

$$X_i \sim G(9, 3)$$

that is, the true model is a family of univariate Gamma distributions $G(\varsigma, \lambda)$ with parameters $\varsigma = 9$ and $\lambda = 3$. We choose a set of three candidate models in which we also include the true ones. In particular, the

first model is $M_1 = G(\varsigma, \lambda)$, the second model $M_2 = N(\mu, \sigma^2)$ is a Normal and the third model $M_3 = W(\alpha, \beta)$ is a Weibull distribution.

We report the results for sample size $n = 400$, 800 and 1200, when $\beta = \frac{1}{4.5}, \frac{1}{4}$, in order to show the sensitivity of the parameter estimates to the choice of the smoothing parameter.

Table 1 indicates that the estimates of Gamma parameters using NPQMLE perform worse than QMLE, obviously because under the correct structure QMLE is MLE. In fact, in the case of NPQMLE $KI(g, f_j(\widehat{\theta})) = 0.0084$ for MLE $KI(g, f_j(\widehat{\theta})) = 0.0047$. This is due to estimation error introduced by the use of the nonparametric density as reference estimate. But, when the assumed model is not the correct ones, as in the case of Normal and Weibull, the NPQMLE method delivers an estimate of the parametric density that is closer to the true structure than those estimated by QMLE. As shown in the table, in the case of the Normal, using NPQMLE $KI(g, f_j(\widehat{\theta}))$ is equal to 0.0233, while for QMLE it equals 0.044. A very similar result is obtained for the Weibull distribution. Table 2 displays analogous results but, since the bandwidth h is smaller, NPQMLE provides parameters' estimates that are characterized by a lower bias and slightly higher variance. This overall causes a reduction of the distance between misspecified and true model, as it can be noticed by the lower values of $KI(g, f_j(\widehat{\theta}))$.

As the sample size increases, from Table 3 we can notice that NPQMLE delivers an estimate of the misspecified model that gets closer to the true one. This was expected, since as n increases the nonparametric density gets closer to the true model and this helps improving the estimation results. Further, in NPQMLE the distance between g and $f_j(\widehat{\theta})$ reduces approximately at the same rate as QMLE. This can be clearly seen observing for example, the reduction of the $KI(g, f_j(\widehat{\theta}))$ for both estimation methods in the case of Gamma. Nevertheless, under the misspecified models, NPQMLE still outperforms QMLE, in the sense that it still delivers a KI which is half of that obtained by QMLE.

6 Application to stock returns

6.1 A Set of simple models

I now apply the described prediction method to determine stock returns predictive density, that will subsequently be used by an investor to choose the optimal share to invest in the risky asset. Typically, due to the hypothesis of asset market efficiency, stock prices are assumed to follow a random walk, that is:

$$p_t = \mu + p_{t-1} + \epsilon_t,\ \epsilon_t IID, \text{ where } p_t = \log(P_t).$$

Further, since the most widespread assumption for the innovations ϵ_t is normality, stock returns are normally distributed with mean μ and variance σ^2. While contrasting evidence exists on the predictability of stock returns, there is substantial support against the normality assumption.

First, as reported by Campbell-Lo-Mackinay (1997)[16], the skewness for daily US stock returns tend to be negative for stock index and positive for individual stocks. Second, the excess Kurtosis for daily US stock returns is large and positive for both index and individual stocks. Both characteristics are further documented in Ullah-Pagan[17] (1999) using non-parametric estimation of monthly stock returns' density from 1834 to 1925. In their analysis is clearly shown that the density departs significantly from a normal,

[16] The Econometrics of Financial Markets, 1997, pag. 16 and 17.
[17] Nonparametric Econometrics, 1999, pag 71-74.

because of its asymmetry, the fat tails and the sharp peak around zero. Third, Diebold-Gunther and Tay (1998) in their application to density forecasting of daily S&P 500 returns indicate that the Normal forecasts are severely deficient. Finally, Knight-Satchell and Tran (1995) show that scale Gamma distributions are a very good model for UK FT100 index.

Given these facts, let assume that the set of candidate models for the risky asset's returns consists of three distributions: a Normal ($N(\mu, \sigma^2)$), a Fisher-Tippet[18] ($F(\alpha, \beta)$) and a mixture of general Gamma ($G(\varsigma, \lambda)$). The first model, as described above, derives from the 'convenient' version of random walk hypothesis. The second model is suggested by the empirical evidence reported in the first two points which advocates the use of extreme value distribution with more probability mass in the tail areas, and the third model is a direct consequence of the study by Finally, Knight-Satchell and Tran (1995).

Let X_t be the log of asset return for day t, it will be modelled using the following densities:

$$1)\ f(X_t; \mu, \sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(X_t - \mu)^2}{2\sigma^2},$$

$$2)\ f(X_t; \alpha, \beta) \equiv \frac{1}{\beta} \exp(\frac{X_t - \alpha}{\beta}) \exp(-\exp(\frac{X_t - \alpha}{\beta})).$$

The third model requires some more details since Gamma distribution is defined only for $0 \leq X_t \leq \infty$, as such the distribution for X_t will be a mixture of two Gammas. Following the authors, lets define the variable:

$$Z_t = \begin{matrix} 1 & \text{with probability } p \\ 0 & \text{with probability } \textit{1-p} \end{matrix}$$

where p is the proportion of returns that are less than a specified benchmark γ. It then follows that X_t is defined

$$X_t = \gamma + X_{1t}(1 - Z_t) - X_{2t} Z_t$$

where X_{jt} are independent random variables with density $f_j(\cdot)$. Hence if $Z_t = 1$, $X_t \leq \gamma$ and we sample from the X_2 distribution; if $Z_t = 0, X_t > \gamma$ and we sample from the X_1 distribution. $f_1(\cdot)$ and $f_2(\cdot)$ are defined as follow:

$$f_1(X_{1t}; \varsigma, \lambda) \equiv \frac{\lambda^\varsigma}{\Gamma(\varsigma)} (X_{1t} - \gamma)^{\varsigma - 1} \exp(-\lambda(X_{1t} - \gamma))$$

$$f_2(X_{2t}; \varsigma, \lambda) \equiv \frac{\lambda^\varsigma}{\Gamma(\varsigma)} (\gamma - X_{2t})^{\varsigma - 1} \exp(-\lambda(\gamma - X_{2t}))$$

6.2 The Data

To implement the empirical application I use daily closing price observations on the US S&P500 index over the period from December 1, 1969 to October 31, 2001, for a total of 7242 observations. The source of the data is DRI. Stock return X_t is computed as $\log(1 + R_t)$ where $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$. Descriptive statistics for the entire sample are provided in the following table.

[18] It is also known as double exponential distribution and a particular case of it is the Gumbel distribution.

	S&P500 index
Min. value	-0.08642
Max. value	0.087089
Mean	0.000319
Std. deviation	0.01005
Kurtosis	4.9333
Skewness	-0.10974

Table I

Furthermore, Ang and Bekaert (2001,2002) and Guidolin and Timmermann (2002) have stressed the importance of distinguishing between 'bear' and 'bull' regimes in modeling stock returns and indicate that these persistent regimes have important economic implications for investors' portfolio decisions. Based on these observations, I have chosen to divide the data in two groups. The first contains all samples relative to contraction (C) and the second includes all samples relative to expansion (E). These two phases of the business cycle typically coincide with 'bear' and 'bull' regimes of the stock market. This implies that the optimal model for asset returns is conditional on the specific regime, which for simplicity I assume to be known at the time of the empirical analysis. The next table indicates the periods of contraction and expansion included in the dataset.[19]

dates	1	2	3	4
C	12/69-11/70	11/73-3/75	7/81-11/82	3/01-10/01
E	4/75-1/80	7/80-7/81	12/82-6/90	4/91-2/01

Table II

Under the assumption that in each regime all subsamples are drawn from a fixed distribution, it is possible to create for each state a unique sample that include all contractions and all expansions respectively. Merging together all the recessions we obtain a sample of 1321 observations, while combining all expansions we obtain a sample of 5921 observations. The descriptive statistics for these two subsamples are reported in the following tables.

Expansion	S&P500 index
Min. value	-0.08642
Max. value	0.087089
Mean	0.00044
Std. deviation	0.009165
Kurtosis	7.1555
Skewness	-0.30326

Contraction	S&P500 index
Min. value	-0.05047
Max. value	0.05574
Mean	-0.00039
Std. deviation	0.0132
Kurtosis	1.05685
Skewness	0.26712

Table III

It is evident from Table I and III, that these data are not consistent with the common assumption that the true model for X_t is the Gaussian distribution. These values confirm previous studies where daily stock returns have been found to exhibit high excess Kurtosis and negative Skewness for index returns. Further,

[19] The contractions and expansions are those provided by NBER's Business Cycle Dating Committee for the US Economy, available at the website www.nber.org/cycles. I have excluded the recession of 1990-91 because of the limited number of observations that does not allow to estimte the nonparametric density with precision in that subsample.

it is very striking how these values differ across regimes. First, as found in other studies, contractions and in general bear regimes are characterized by high volatility and negative mean for stock return, which turns out to be a problem in determining the optimal share to invest in the risky asset. Second, while during expansions stock returns show a positive excess kurtosis (even bigger than that displayed in Table I for all data) and a negative Skewness (three times bigger than that for the entire sample), during contractions the excess Kurtosis is negative (lower than three) and the Skewness is positive. According to these simple descriptive statistics, it is reasonable to expect different optimal models for stock returns across these two regimes.

6.3 Empirical Results.

For each of these samples I estimate the univariate density of stock returns by Nadaraya-Watson kernel density estimators. For the Kernel function I employ the second-order Gaussian Kernel and the bandwidths are selected via least-squares cross-validation (Silverman, 1986, p48). Graphs of the nonparametric densities are reported in the appendix in Figures 1, 2 and 3.

I then use the Kullback-Leibler entropy to measure the distance between the estimated nonparametric density and each of the models belonging to the set $\mathcal{M}$. Minimizing this distance I obtain the parameter estimates for each candidate distribution and a value for $\widehat{KI}_j$, which allows me to achieve a ranking of all competing models and the subsequent weight for each of them in the final model combination. The estimated parameters for each distribution are reported below.

$N(\mu, \sigma^2)$	Entire sample	Expansion	Contraction
$\widehat{\mu}$	0.0004*	0.0005*	-0.0008*
$\widehat{\sigma}$	0.0082*	0.0075*	0.0123*
KLI	0.1897	0.1587	0.0513

*All estimates are significant at 1% level

$F(\alpha, \beta)$	Entire sample	Expansion	Contraction
$\widehat{\alpha}$	-0.00179*	-0.0014*	-0.00403*
$\widehat{\beta}$	0.008509*	0.00773*	0.01213*
KLI	0.9836	0.9209	0.3362

*All estimates are significant at 1% level

$G(\varsigma, \lambda)$	Entire sample	Expansion	Contraction
$\widehat{\varsigma}$	1.1104*	1.1212*	1.1237*
$\widehat{\lambda}$	146.3839*	160.6803*	97.4237*
$\widehat{\gamma}$	0.00031	0.00044	-0.00039
$\widehat{p}$	0.47878	0.465631	0.53637
$1-\widehat{p}$	0.52122	0.5343	0.46363
KLI	0.0468	0.0666	0.0776

*All estimates are significant at 1% level

Table IV

Examining the tables we see that all the estimates are intuitively reasonable and significantly different from zero. Comparing all the three models over the entire sample, we can notice that the model characterized

by the double Gamma outperforms the other two models. Its $\widehat{KI}$ assumes the lowest value (0.0468) which is four times smaller than that for the Normal and twenty time smaller than that of Fisher-Tippet. Also in the case of expansion, the double Gamma is clearly better than the other two models; its $\widehat{KI}$ equals 0.0666 which is half the value for the Normal. In contrast, for the sample including all contractions the Gaussian distribution performs slightly better than the double Gamma. The value of its $\widehat{KI}$ is equal to 0.0513 which is smaller than the respective value for the double Gamma (0.0776). Finally, both values are ten times smaller than the $\widehat{KI}$ for the Fisher-Tippet distribution. These results contradict the common assumption that the best unique model for the stock returns is the Gaussian distribution, and confirm that the optimal model changes across regimes. Further, since more than one model performs fairly well, and because each of them has properties that capture particular characteristics of return distribution, it seems reasonable to combine them.

It is important to stress some characteristics of the double Gamma, since it is overall the model that provides the best performance in terms of aggregate similarity to the data. First of all, it is worth mentioning that in all three samples the values of $\widehat{p}$ suggest that the sample proportions for negative returns are not very different from that of positive returns. Second, ς's estimates in all three samples are greater than unity, which entails that returns are well described by a bimodal density. All these features of the estimated model confirm the results that Knight-Satchell and Tran (1995) found in the case of UK stock returns.

The final step to compute the similarity-weighted predictive distribution $M(\widehat{\theta}_{M_j})$ consists in evaluating for each of the models under consideration the 'probability' $p_j(\widehat{KI})$ of being correct. It can be helpful to first provide the realizations of $\widehat{KI}_j$ for all models in each of the sample.

	All data	Expansion	Contraction
G	0.0468	0.0666	0.0776
N	0.1897	0.1587	0.0513
F	0.9836	0.9209	0.3362

Table V

The following table exhibits the value of $p(\widehat{KI}_j)$ for the three models under consideration.

	All data	Expansion	Contraction
G	0.8121	0.7811	0.5689
N	0.7033	0.7086	0.604
F	0.0779	0.0924	0.331

Table VI

As it can be noticed these values represent 'probabilities' before normalization since they do not sum up to unity. Results contained in table VI seem to confirm that this methodology in determining the "probability of being the correct model" works in the right direction. In fact, in each of the samples the p.d.f. with the lowest realization of the KI receives the highest $p_j(\widehat{KI})$, and hence it will receive the largest weight in the model combination. Further, the very poor performance of the Fisher-Tippet distribution with respect to the other two candidate models, suggests that it would be sensible to discard this model. Thus, in the next section I present the results obtained combining the Normal and the double Gamma according to the weights reported in the first two rows of Table VI.

6.4 In and Out-of-sample performance of model combination

Lets first consider the in-sample performance of model combination using the entire dataset from December 1, 1969 to October 31, 2001. In this case, after normalizing the $p(\widehat{KI}_j)$, the double Gamma $G(1.1104, 146.38)$ receives a weight of 0.5359 and the Normal $N(0.0004, (0.0082)^2)$ receives a weight of 0.4641. The Kullback-Leibler distance between the nonparametric density estimate and the model combination equals 0.0256, attaining a loss almost half of the best minimizer. If I consider the sample including all expansions, to the Gamma $G(1.1212, 160.68)$ it is assigned a weight equal to 0.5243 and to the Normal $N(0.0005, (0.0075)^2)$ a weight of 0.4757. This model combination delivers a distance from the nonparametric density equal to 0.0179 which is a third of that achieved by the best model. Finally, considering only contraction data, the Gamma $G(1.1237, 97.42)$ receives a weight of 0.4937, while the Normal $N(-0.0008, (0.0123)^2)$ attains a weight equal to 0.5063. In this case as well, the model combination outperforms the best model by achieving a KI equal to 0.0137, which is one fourth of the distance achieved by the best model.

Now, to verify the performance of NPQMLE and of the model combination out of sample, we analyze the previous results in the context of a different dataset, using the series of stock returns observed from November 1, 2001 to September 30, 2003, for a total number of observations of 479. This sample represents the most recent case of expansion, or more precisely recovery, according to the latest determination of the Business Cycle Committee of the NBER. The summary statistics are displayed below.

	S&P500 index
Min. value	-0.01842
Max. value	0.024204
Mean	-0.0000556
Std. deviation	0.00619
Kurtosis	0.932
Skewness	0.2804

Using this data, but the parameter estimates and the weights obtained from the expansion sample for the period December 1, 1969 to October 31, 2001, we evaluate the KI distance between the nonparametric density estimated in the new sample and the parametric model estimated in the previous sample. I obtain the following results: the KI between the model combination and $\widehat{f}_n$ is equal to 0.7649, between the Gamma distribution and $\widehat{f}_n$ is equal to 0.7749 and between the Normal and $\widehat{f}_n$ is 0.9235. That is, the model combination slightly outperforms both models, including the Gamma that in the case of expansion was the best minimizer. This result can be further corroborated using a larger out-of sample dataset and bootstrap methodology.

7 Investors' optimal asset allocation

7.1 The Optimization Problem

In this section, I first briefly describe the framework to derive the optimal portfolio choice under 'uncertainty', when an investor uses the similarity-weighted predictive distribution as the model on the basis of which to act. Second, I consider how the estimated model combination affects investor's optimal asset allocation.

Lets consider an individual making portfolio choice at time T , this choice involves two kind of assets: a risky asset which consists of a broad portfolio of stocks (S&P500 index), whose gross return at time t per unit invested at time $t-1$ is $1+R_t$, and a riskless asset whose gross return is $1+R_t^f$. The decision maker has access to the return histories over T periods, he knows in advance the future return of the riskless asset that in accordance with standard practice is assumed to be constant.

Lets define $r_t = \log(1+R_t)$ and $r_t^f = \log(1+R_t^f)$,then we can describe the investor's information set in the following way:

$$I_t \equiv \left[\{r_t\}_{t=0}^{T}, \{r_t^f\}_{t=0}^{T} and\ \{r^f\}_{t=T+1}^{H} \right].$$

He invests one unit of saving, divided between an amount $1-a$ in the safe asset and a in the risky asset, and then he holds on to the portfolio until date H.

Let $W(r_t, a)$ denote the value of the portfolio and suppose also that we are considering a self financing portfolio. Thus, the value of the portfolio at time $t = H$ is given by:

$$W(r_t, a) = (1-a) \prod_{t=T+1}^{H} (1+R_t^f) + a \prod_{t=T+1}^{H} (1+R_t) = (1-a) H r_t^f + a \sum_{t=T+1}^{H} r_t.$$

Lets also assume that utility depends only on the final value of the portfolio: $U(W(r_t, a))$.Then, the problem is to choose the best decision rule d that maps the observations contained in I_t into actions a, in other words: the optimal share to invest in the risky asset. This decision rule is obtained by maximizing the following expected utility:

$$E_{g_{r|s}}[U(W(r_t, a))]$$

$$s.t\ W_{t+H} = (1-a) H r_t^f + a \sum_{t=T+1}^{H} r_t \text{ and } a \in [0,1].$$

In order to simplify the analysis lets assume that H=1, such that the wealth form reduces to

$$W(\widetilde{r}_t, a) = (1-a) r_t^f + a r_t = r_t^f + a(r_t - r_t^f) = C + a\widetilde{r}_t,$$

where C is a constant and $\widetilde{r}_t$ is the excess return.

Example 1: *(CARA investor) Assume that $U(W(\widetilde{r}_t, a))$ is the utility function of an investor with negative exponential utility:*

$$U(W(\widetilde{r}_t, a)) = -\exp(-\delta W(\widetilde{r}_t, a))$$

$$E_{g_{r|s}}[U(W(\widetilde{r}_t, a))] = -K E_{g_{r|s}}[\exp(-\delta a \widetilde{r}_t)],$$

where δ is the risk aversion parameter, K is the expected utility relative to the riskless asset and $g_{r|s}$ is the distribution of the return r_t given the regime s, which is unknown. Typically, the return distribution $g_{r|s}$ is assumed to be Normal, consequently the expected utility results to be:

$$E_{g_{r|s}}[U(W(\widetilde{r}_t, a))] = -\exp(-a\delta\mu + \frac{1}{2}\delta^2 a^2 \sigma^2),$$

where $\mu = E(\widetilde{r}_t)$ and $\sigma^2 = Var(\widetilde{r}_t)$ are the mean and the variance of the Normal distribution. In this case the optimal share to invest in the risky asset is given by:

$$a(g_{r|s}) = \arg\max_a E_{g_{r|s}}[U(W(\widetilde{r}_t, a))]$$

and it is equal to:

$$a^* = \frac{\mu}{\delta\sigma^2}.$$

The economic agents use the personal model $g_{r|s}$ as if it represented the actual model. The problem is that assuming $g_{r|s}$ to be the density function of a Normal random variable, is likely that the model is misspecified.

What is described next is a characterization of expected utility maximization under 'risk' very similar to Gilboa-Schmilder(2001), that is the decision process is one in which the "decision maker first forms probabilistic beliefs and then uses them to reduce a decision problem under uncertainty to a decision problem under risk". In other words, the use of models selection and combination described in previous sections helps reducing the degree of model ambiguity, because it shrinks the set of candidate models into a unique distribution that characterizes the risk of the decision problem.

Example 2*: CARA investors with probabilistic belief $M(\widehat{\theta}_{M_j})$. In the context of example 1, lets suppose that the investor instead of assuming that the returns are Normally distributed, builds his probabilistic belief $g_{r/s}$ as described in section II and III. Moreover, lets assume that the current regime is known. Therefore, his model for the asset return is equal to $M(\widehat{\theta}_{M_j})$.*

$$E_{g_{r|s}}[U(W(\widetilde{r}_t, a))] = -KE_{M(\widehat{\theta}_{M_j})}[\exp(-\delta a\widetilde{r}_t)]$$

$$-E_{M(\widehat{\theta}_{M_j})}[\exp(-\delta ar)] = -\int \exp(-\delta ar)\cdot(\sum_j p_j(\widehat{KI})f_j(r,\widehat{\theta}))dr$$

$$= -\sum_j p_j(\widehat{KI})\int \exp(-\delta ar)f_j(r,\widehat{\theta}))dr = \sum_j p_j(\widehat{KI})(-E_{f_j}(\exp(-\delta ar).$$

If we define $t = \delta a$ we can rewrite the expected utility function as $-E[\exp(-tr)]$, which is the Moment Generating Function[20] *(MGF) after we account for the change in signs. This implies that the expected utility function, when the expectation is taken under the model $M(\widehat{\theta}_{M_j})$ is equal to*

$$= -\sum_j p_j(\widehat{KI})MGF_{f_j}(-t),$$

which is nothing more than the weighted average of the Moment Generating Function of each model included in $M(\widehat{\theta}_{M_j})$.

It is important to notice that the existence of a closed form solution for the optimal share a is still guaranteed. However, in this framework the optimal choice not only depends on the risk aversion and on the moments of the probability law of stock returns, but it also depends on the weights contained in the model combination. As such, it is affected by the measure of uncertainty about the true structure. Explicit formulas for each MGF and for the expected utility are provided in the Appendix.

[20] The Moment Generating Function $\mathrm{MGF}_r(t) = E_r(\exp tr)$.

7.2 Implications of model ambiguity for the optimal stock holdings

In this section, I compare the optimal shares selected using the model combination, with the same quantities obtained employing the best candidate models. Investors choose different shares to invest in the risky asset according to their level of risk aversion and investment horizon. I start reporting the results relative to each single model for three different level of risk aversion and for an investment horizon equal to one period. However, since returns are i.i.d, this particular choice of the time horizon does not have any impact.

Under the assumption that the Gamma distribution is the optimal model for stock returns, for all three samples under consideration I obtain the following results:

N=7242	a^*
R.A=2	1
R.A=6	0.41
R.A=10	0.27

E, N=5921	a^*
R.A=2	1
R.A=6	0.77
R.A=10	0.45

C, N=1321	a^*
R.A=2	0
R.A=6	0
R.A=10	0

Table VIII

In contrast, under the assumption that the Gaussian distribution is the optimal model, we have:

N=7242	a^*
R.A=2	1
R.A=6	0.99
R.A=10	0.594

E, N=5921	a^*
R.A=2	1
R.A=6	1
R.A=10	0.88

C, N=1321	a^*
R.A=2	0
R.A=6	0
R.A=10	0

Table VIV

In the case of the double Gamma distribution, the results relative to the entire sample, are very similar to those reported by Avramov (2000) for the i.i.d model (Figure 5, p63). The comparison is made even easier from the fact that I used his same values for the coefficient of the risk aversion. The values reported in table VIII seem very reasonable also when compared to the most recent evidence from the 2001 Survey of Consumer Finances[21]. Among the families holding stocks, corresponding to 21.3% of the interviewed population, on average the median value of wealth invested in stock is around 32.9% (which is in between 0.41 and 0.27). In contrast, the values obtained for the Gaussian distribution tend to overestimate the actual share, most likely because this model is not able to account for the fat tails of the distribution which can strongly affect the results.

In analyzing the results for the case of expansion (E), it can be noticed that for the double Gamma I obtain values very close to those of Guidolin and Timmermann under: no predictability, bull state probability equal to one, investment horizon equal one and by the same values for the risk aversion (Figure 5, Guidolin-Timmermann (2002)). On the contrary, these values are very different from those obtained using the Normal distribution.

Unfortunately the case of contraction does not provide very interesting results, due to the fact that with any model the estimated average of stock return is negative, causing the optimal share to be zero for any value of the risk aversion. Similar results are also reported by Guidolin and Timmermann in the case of

[21] All the values reported are obtained from Table B, pg 13 of "Recent changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances". Federal Reserve Bulletin.

bear regime, where values very close to zero are obtained for almost all considered values of the risk aversion ranging from 1 to 20.

Now let me show the values of the optimal share obtained using the combination of double Gamma and Gaussian distributions described in section 6.4. Since the expected utility is given by a linear combination of the MGF of each single distribution, then the optimal share is given by a linear combination of the shares found using each candidate models included in $M(\widehat{\theta}_{M_j})$.

N=7282	a^*
R.A=2	1
R.A=6	0.6792
R.A=10	0.4204

E, N=5921	a^*
R.A=2	1
R.A=6	0.8794
R.A=10	0.6546

C, N=1361	a^*
R.A=2	0
R.A=6	0
R.A=10	0

Table X

This implies that the investor, who fears misspecification and accounts explicitly for it through the model combination, invests more in the risky asset than what he would have invested using the double Gamma distribution as the unique probabilistic belief . This is due to the fact that now using a similarity-weighted distribution, the investor no longer assigns a probability of one to the mixture of Gamma, and hence does not overestimate the precision of the forecast provided by this model.

8 Conclusions

This paper proposes a method to estimate the probability density of a random variable of interest in the presence of model ambiguity. The first step consists in estimating and ranking the candidate parametric models minimizing the Kullback-Leibler 'distance' (KLD) between the nonparametric fit and the parametric fit. In the second step, the information content of the KLD is used to determine the weights in the model combination, even when the true structure does not necessarily belong to the set of candidate models.

This approach has the following features. First, it provides an explicit representation of model uncertainty exploiting models' misspecification. Second, it overcomes the necessity to have a specific prior over the set of models and about parameters belonging to each of the models under consideration. Finally, it is computationally extremely easy.

The NPQMLE estimator obtained in the first step is root-n consistent and asymptotically normally distributed. Thus, it preserves the same asymptotic properties of a full parametric estimator. Furthermore, when the misspecified model is used, it delivers 'better' finite sample performances than QMLE. However, it is important to bear in mind that such result is completely determined by the smoothing parameter.

To implement the model combination, using the technical machinery provided by previous studies on nonparametric entropy-based testing, I derive the asymptotic distribution of the Kullback-Leibler information between the nonparametric density and the candidate parametric model. Since the approximation error affects the asymptotic mean of the KLD's distribution, the latter varies with the underlying parametric model. Then, to determine the same distribution for all candidate models, employing an assumption technically equivalent to a Pitman alternative, I center the resulting Normal on the average performance of all plausible models. Consequently, the weights in the model combination are determined by the probability of obtaining a performance worse than that actually achieved, relatively to that attained on average by the other competing models.

The empirical application to daily stock returns indicates that, during the phases of expansion, the best model is the double Gamma distribution, while during the phases of recession is the Gaussian distribution. Moreover, the combination of the Normal and the double Gamma, according to the weights obtained with the described methodology, outperforms in- and out-of-sample all candidate models including the best one. Most likely, this result is due to the fact that none of the candidate models is the true structure, as such the models combination being a higher dimensional parametric alternative is able to approximate the data more closely.

This suggests that in decision contexts characterized by high uncertainty, such that it can be hard: to form specific priors, to conceive an exhaustive set of all possible models and/or to use the true complex structure, the proposed approach can provide a better hedge against the lack of knowledge of the correct model. Additionally, this methodology can also be used to form priors in training sample, before applying more sophisticated Bayesian averaging techniques.

This approach can be further extended to conditional distributions to address more challenging and complex prediction problems. I leave this problem to future research.

9 Appendix

9.1 Proof Theorem 1

The first step consists in showing that $KI_{jn}(\theta)$ converges at least in probability to the contrast function $KI_j(\theta)$.

$$KI_{jn}(\theta) - KI_j(\theta) = \sum_{i=1}^{n}(\ln \widehat{f_n}(x_i) - \ln f(\theta, x_i)\widehat{f_n}(x_i) - \int_x (\ln g(x) - \ln f(\theta, x)g(x)dx = \tag{27}$$

$$= \sum_{i=1}^{n}\left(\ln \widehat{f_n}(x_i)\right)\widehat{f_n}(x_i) - \int_x (\ln g(x))g(x)dx - \sum_{i=1}^{n}\ln f(\theta, x_i)\widehat{f_n}(x_i) + \int \ln f(\theta, x)g(x)dx = D1 - D2 \tag{28}$$

$$D1 = \sum_{i=1}^{n}\left(\ln \widehat{f_n}(x_i)\right)\widehat{f_n}(x_i) - \int_x (\ln g(x))g(x)dx \tag{29}$$

$$D2 = \sum_{i=1}^{n}\ln f(\theta, x_i)\widehat{f_n}(x_i) - \int \ln f(\theta, x)g(x)dx \tag{30}$$

$$D1 \longrightarrow^{p} 0 \text{ as } n \longrightarrow \infty, \text{ since } \sum_{i=1}^{n}\left(\ln \widehat{f_n}(x_i)\right)\widehat{f_n}(x_i) \longrightarrow^{p} \int_x (\ln g(x))g(x)dx \tag{31}$$

see Theorem 2, Dmitriev-Tarasenko (1972),

$$D2 = \sum_{i=1}^{n}\ln f(\theta, x_i)\left(\widehat{f_n}(x_i) - g(x_i)\right) + \left(\sum_{i=1}^{n}\ln f(\theta, x_i)g(x_i) - \int \ln f(\theta, x)g(x)dx\right) = D2_1 + D2_2 \tag{32}$$

$$\text{as } n \longrightarrow \infty,\ \widehat{f_n} - g \longrightarrow^p 0, \text{ then } D2_1 \longrightarrow 0 \tag{33}$$

and since $\sum_{i=1}^{\infty} \ln f(\theta, x_i) g(x_i) = \int \ln f(\theta, x) g(x) dx$, then $D2_2 \longrightarrow 0$.

We can conclude that $KI_{jn}(\theta) \longrightarrow^p KI_j(\theta)$, hence it is a *contrast* relative to the *contrast function* $KI_j(\theta)$ according to the Definitions 3 and 4 in Dhrymes (1998).

Further since $KI_{jn}(\theta)$ can be rewritten as $H_n(\theta) - H_n(\widehat{f_n})$, where

$$H_n(\widehat{f_n}) = -\sum_{i=1}^{n} \left(\ln \widehat{f_n}(x_i)\right) \widehat{f_n}(x_i) \text{ and } H_n(\theta) = -\sum_{i=1}^{n} \ln f(\theta, x_i) \widehat{f_n}(x_i) \tag{34}$$

then

$$H_n(\theta_1) - H_n(\theta_2) = [H_n(\theta_1) - H_n(\widehat{f_n})] - [H_n(\theta_2) - H_n(\widehat{f_n})]. \tag{35}$$

It follows that

$$H_n(\theta_1) - H_n(\theta_2) \rightarrow^p KI_j(\theta_1) - KI_j(\theta_2). \tag{36}$$

By the continuity of Kullback-Leibler Information and by A3, assumption (iii) of Theorem 1 in Dhrymes (1998) is justified. Then the consistency of the MC estimator $\widehat{\theta}_{M_j}$ follows immediately by this same theorem.

9.2 Proof Theorem 2:

By the mean value theorem around the parameter θ^*

$$0 = \nabla KI(\widehat{f}_n, f_{\widehat{\theta}}) \simeq \nabla KI(\widehat{f}_n, f_{\widehat{\theta}}) \mid_{\theta^*} + \nabla^2 KI(\widehat{f}_n, f_{\widehat{\theta}}) \mid_{\overline{\theta}} (\widehat{\theta}_n - \theta_n^*) \tag{37}$$

$$(\widehat{\theta}_n - \theta^*) \simeq -(\nabla^2 KI(\widehat{f}_n, f_\theta) \mid_{\overline{\theta}})^{-1} \cdot \nabla KI(\widehat{f}_n, f_\theta) \mid_{\theta^*} \tag{38}$$

$$(\widehat{\theta}_n - \theta^*) \simeq -\left(\sum_i \frac{\partial^2 \log f(\overline{\theta}, x_i)}{\partial \theta_i \partial \theta_j} \widehat{f}_n(x_i)\right)^{-1} \cdot \left(\sum_i \frac{\partial \log f(\theta^*, x_i)}{\partial \theta} \widehat{f}_n(x_i)\right)$$

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \simeq -\left(\frac{1}{n}\sum_i \frac{\partial^2 \log f(\overline{\theta}, x_i)}{\partial \theta_i \partial \theta_j} \widehat{f}_n(x_i)\right)^{-1} \cdot \left(\frac{1}{\sqrt{n}}\sum_i \frac{\partial \log f(\theta^*, x_i)}{\partial \theta} \widehat{f}_n(x_i)\right). \tag{39}$$

Let us define $s(\theta, x) = \frac{\partial \log f(\theta, x_i)}{\partial \theta} = \frac{\partial f(\theta, x)/\partial \theta}{f(\theta, x)}$

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \simeq -\left(\frac{1}{n}\sum_i \frac{\partial s(\overline{\theta}, x_i)}{\partial \theta} \widehat{f}_n(x_i)\right)^{-1} \cdot \left(\frac{1}{\sqrt{n}}\sum_i s(\theta^*, x_i) \widehat{f}_n(x_i)\right) = -(A_n(\overline{\theta}))^{-1} W_n(\theta^*). \tag{40}$$

Rewriting $W_n(\theta^*)$ as a second order U-statistic of the form

$$U_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{h} K(\frac{x_j - x_i}{h}) s(\theta^*, x_i) = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{h} K(\frac{x_j - x_i}{h}) [s(\theta^*, x_i) - s(\theta^*, x_j)] \tag{41}$$

where the last equality holds since $k(u) = -k(-u)$), we can notice that $W_n(\theta^*) = \sqrt{n}U_n$.

Applying first Lemma 3.1 and then Theorem 3.1 in Powell, Stock, and Stoker (1989), or similarly Lemma 3.3b. in Zheng (1996) we can show that $W_n(\theta^*)$ is asymptotically normally distributed and that it is $O_p\left(\frac{1}{\sqrt{n}}\right)$. Let define H_n in the following way:

$$H_n(x_i, x_j) = \frac{1}{h} K(\frac{x_j - x_i}{h}) \left[s(\theta^*, x_i) - s(\theta^*, x_j) \right]. \tag{42}$$

First, we need to verify that $E\left[\|H_n(x_i, x_j)\|^2 \right] = o(n)$. Let define $v^2(\theta^*, x) = E(s^2(\theta^*, x)/x)$ and $v(\theta^*, x) = E(s(\theta^*, x)/x)$,

$$E\left[\|H_n(x_i, x_j)\|^2 \right] = E\left[E\left(\|H_n(x_i, x_j)\|^2 / x_i, x_j \right) \right] = \tag{43}$$

$$= \frac{1}{h^2} \int \left\| K(\frac{x_j - x_i}{h}) \right\|^2 \left[v^2(\theta^*, x_i) + v^2(\theta^*, x_j) - 2v(\theta^*, x_i) v(\theta^*, x_j) \right] g(x_i) g(x_j) dx_i dx_j =$$

now using the change of variable from (x_i, x_j) to $(x_i, u = \frac{x_j - x_i}{h})$ we obtain

$$= \frac{1}{h^2} \int \|K(u)\|^2 \left[v^2(\theta^*, x_i) + v^2(\theta^*, x_i + hu) - 2v(\theta^*, x_i) v(\theta^*, x_i + hu) \right] g(x_i) g(x_i + hu) dx_i h du =$$

$$= O(\frac{1}{h}) = O(n(nh)^{-1}) = o(n) \text{ since } nh \to \infty. \tag{44}$$

This implies that $\sqrt{n}(U_n - \widehat{U}_n) = o_p(1)$. Thus, we need just to study the behavior of $\widehat{U}_n$ which is given by

$$\widehat{U}_n = E(r_n(x_i)) + \frac{2}{n} \sum_i r_n(x_i) - E(r_n(x_i)). \tag{45}$$

Let compute $r_n(x_i)$ which is defined in the following way:

$$r_n(x_i) = E(H_n(x_i, x_j)/x_i) = \int \frac{1}{h} K(\frac{x_j - x_i}{h}) \left[s(\theta^*, x_i) - v(\theta^*, x_j) \right] g(x_j) dx_j = \tag{46}$$

$$= \frac{1}{h} \int K(u) \left[s(\theta^*, x_i) - v(\theta^*, x_i + hu) \right] g(x_i + hu) h du = r(x_i) + t_n(x_i) \tag{47}$$

where

$$r(x_i) = \left[s(\theta^*, x_i) - v(\theta^*, x_i) \right] g(x_i) \tag{48}$$

and

$$t_n(x_i) = h^2 v'(\overline{x}_i, \theta^*) g'(\overline{x}_i, \theta^*) \int u^2 K(u) du = o(h^2). \tag{49}$$

This last expression has been obtained applying the mean value theorem to $v(\theta^*, x_i + hu)$ and $g(x_i + hu)$, which yields $v(\theta^*, x_i) + huv'(\overline{x}_i, \theta^*)$ and $g(\theta^*, x_i) + hug'(\overline{x}_i, \theta^*)$ where $\overline{x}_i$ lies in $[x_i, x_i + hu]$.

Further, we need to compute $E(r_n(x_i)) = E(H_n(x_i, x_j))$

$$E(H_n(x_i,x_j)) = \frac{1}{h}\int K(u)\left[v(\theta^*,x_i) - v(\theta^*,x_i+hu)\right] g(x_i)g(x_i+hu)dx_i h du =$$

$$= \int K(u)du \int \left[v(\theta^*,x_i) - v(\theta^*,x_i)\right] g^2(x_i)dx_i = 0. \tag{50}$$

So what we have to study is the asymptotic behavior of

$$\sqrt{n}\widehat{U}_n = \sqrt{n}E(r_n(x_i)) + \frac{2}{\sqrt{n}}\sum_i r_n(x_i) - E(r_n(x_i)) = \frac{2}{\sqrt{n}}\sum_i r(x_i) - E(r(x_i)) + \frac{2}{\sqrt{n}}\sum_i t_n(x_i) - E(t_n(x_i)) \tag{51}$$

where $r(x_i) = \left[s(\theta^*,x_i) - v(\theta^*,x_i)\right] g(x_i)$ and $E(r(z_i)) = E(E\left[\left(\left(s(\theta^*,x_i) - v(\theta^*,x_i)\right)g(x_i)\right)/x_i\right]) = 0$ and the last term of the above expression converges to zero in probability. Hence, the limiting distribution of $\sqrt{n}\widehat{U}_n$ is the same of $\frac{2}{\sqrt{n}}\sum_i r(x_i) = \frac{2}{\sqrt{n}}\sum_i \left[s(\theta^*,x_i) - v(\theta^*,x_i)\right] g(x_i)$.

By Lindeberg-Levy central limit theorem, we have that

$$W_n(\theta^*) = \sqrt{n}\widehat{U}_n \to^d N(0, B(\theta^*)) \text{ as } n \to \infty \tag{52}$$

$$B(\theta^*) = 4E(\left[s(\theta^*,x_i) - v(\theta^*,x_i)\right]^2 g(x_i)^2) = 4\int \left(s^2(\theta^*,x_i) + \left(v(\theta^*,x_i)\right)^2 - 2s(\theta^*,x_i)v(\theta^*,x_i)\right) g(x_i)^3 dx =$$

$$= \int \left(v^2(\theta^*,x_i) - \left(v(\theta^*,x_i)\right)^2\right) g(x_i)^3 dx_i = \int var(s(\theta^*,x_i))g(x_i)^3 dx_i = E(var(s(\theta^*,x_i))g(x_i)^2). \tag{53}$$

This implies that $W_n(\theta^*) = O_p\left(\frac{1}{\sqrt{n}}\right)$. It follows that

$$\sqrt{n}(\widehat{\theta}_n - \theta_n^*) = -(A_n(\overline{\theta}))^{-1}W_n(\theta^*) \to N(0, A(\theta^*))^{-1}B(\theta^*)A(\theta^*))^{-1}). \tag{54}$$

9.3 Proof Theorem 3:

KI can be rewritten in the following way:

$$KI = \int_x (\ln \widehat{f_n}(x) - \ln f_{\widehat{\theta}}(x))d\widehat{F}_n(x) = \int_x (\ln \widehat{f_n}(x) - \ln g(x))d\widehat{F}_n(x) - \int (\ln f_{\widehat{\theta}}(x) - \ln g(x))d\widehat{F}_n(x) = KI_1 - KI_2. \tag{55}$$

Similarly to Fan(1994), this representation is very helpful to examine the effect of estimating f_{θ_*} by $f_{\widehat{\theta}}$ on the limiting distribution of $\widehat{KI}$. From now on the index j for th model will be omitted.

I start examining the limiting distribution of $\widehat{KI}_1 = \frac{1}{n}\sum_i \ln\left(\frac{\widehat{f_n}(x_i)}{g(x_i)}\right)$ that by the Law of Large Numbers (LLN) can be considered a good approximation of $n^{-1}E((\ln \widehat{f_n}(x) - \ln g(x)) = n^{-1}KI_1$. This first part of the proof draws heavily upon Hall(1984) and Hong and White(2000).

Using this inequality $\left|\ln(1+u) - u + \frac{1}{2}u^2\right| \leq |u|^3$ for $|u| < 1$ and defining $u = \frac{\widehat{f_n} - g(x)}{g(x)} = \frac{\widehat{f_n}}{g(x)} - 1$ we obtain the following result:

$$\frac{1}{n}\sum_i \ln\left(\frac{\widehat{f_n}(x_i)}{g(x_i)}\right) - \frac{1}{n}\sum_i \left(\frac{\widehat{f_n}(x_i)-g(x_i)}{g(x_i)}\right) + \frac{1}{2n}\sum_i \left(\frac{\widehat{f_n}(x_i)-g(x_i)}{g(x_i)}\right)^2 \leq u^3. \tag{56}$$

Let define

$$\widehat{V}_{1n} = \frac{1}{n}\sum_i \left(\frac{\widehat{f_n}(x_i)-g(x_i)}{g(x_i)}\right)$$

and

$$\widehat{V}_{2n} = \frac{1}{n}\sum_i \left(\frac{\widehat{f_n}(x_i)-g(x_i)}{g(x_i)}\right)^2.$$

By Lemma 3.1 Hong-White (2000), under assumption A1 and A2, $nh^4/\ln n \to \infty$, $h \to 0$. Then:

$$\widehat{KI}_1 = \widehat{V}_{1n} - \frac{1}{2}\widehat{V}_{2n} + O_p(n^{-\frac{3}{2}}h^{-3}\ln n + h^6). \tag{57}$$

Now we have to analyze the terms $\widehat{V}_{1n}$ and $\widehat{V}_{2n}$. Let define $\overline{f}(x) = \int h^{-1}K(\frac{x_i-x}{h})g(x)dx$ and

$$a_n(x_i,x_j) = \frac{h^{-1}K_h(x_i-x_j) - \int h^{-1}K_h(x_i-x)g(x)dx}{g(x_i)}$$

$$b_n(x_i) = \frac{\int h^{-1}K_h(x_i-x)f(x)dx - g(x_i)}{g(x_i)}.$$

Then

$$\widehat{V}_{1n} = \frac{1}{n}\sum_i \left[\frac{\widehat{f_n}(x_i)-\overline{f}(x_i)}{g(x_i)} + \frac{\overline{f}(x_i)-g(x_i)}{g(x_i)}\right] = \frac{1}{n(n-1)h}\sum_i\sum_{j,i\neq j} a_n(x_i,x_j) + \frac{1}{n}\sum_i b_n(x_i) = \tag{58}$$

$$= \widehat{V}_{11n} + \widehat{B}_n,$$

where $\widehat{V}_{11n}$ is a second order U-statistic and it will affect the asymptotic distribution of $\widehat{KI}_1$. Similarly to Hall(1984) let rewrite $\widehat{V}_{11n}$ in the following way:

$$\widehat{V}_{11n} = \frac{1}{n(n-1)}\sum_i\sum_{j,i\neq j} H_{1n}(x_i,x_j)$$

$$H_{1n}(x_i,x_j) = \frac{1}{2h}\left(\frac{K_h(x_j-x_i) - \int K_h(x-x_i)g(x)dx}{g(x_i)} + \frac{K_h(x_i-x_j) - \int K_h(x_i-x)g(x)dx}{g(x_i)}\right) = J_n(x_i,x_j)+J_n(x_j,x_i) \tag{59}$$

$E(H_{1n}(x_i,x_j)/x_i) = 0$, then using Theorem 1 in Hall(1984) we can show that

$$\widehat{V}_{11n} = \left\{\frac{1}{n(n-1)}\sum_i\sum_{j,i\neq j} H_{11n}(x_i,x_j)\right\} \Big/ \left\{\frac{2E\left[H_{1n}^2(x_i,x_j)\right]}{n^2}\right\} \to^d N(0,1). \tag{60}$$

$$E\left[J_n^2(x_i,x_j)\right] = \frac{1}{4h^2}\int\int \frac{\left(K_h(x_j-x_i) - \int K_h(x_i-x)g(x)dx\right)^2}{g^2(x_i)} g(x_i)g(x_j)dx_i dx_j =$$

applying a change of variable from $(x_i, x_j) = (x_i, u)$ where $u = \frac{x_j - x_i}{h}$ we get the following expression

$$= \frac{1}{4h} \int \int \frac{K^2(u) + \left[h \int K(u) g(x_i + hu) du\right]^2 - 2K(u) \left[h \int K(u) g(x_i + hu) du\right]}{g^2(x_i)} g(x_i) g(x_i + hu) du \simeq$$

$$\simeq \frac{1}{4h} \int K^2(u) du + o(\frac{1}{h}) = O(\frac{1}{h}). \tag{61}$$

Similarly we can show that

$$E\left[J_n(x_i, x_j) J_n(x_j, x_i)\right] \simeq \frac{1}{4h} \int K^2(u) du + o(\frac{1}{h}) = O(\frac{1}{h}). \tag{62}$$

Then it follows that

$$E\left[H_{1n}^2(x_i, x_j)\right] = E\left[2J_n^2(x_i, x_j) + J_n(x_i, x_j) J_n(x_j, x_i)\right] = \frac{1}{h} \int K^2(u) du + o(\frac{1}{h}) = O(\frac{1}{h}), \tag{63}$$

and

$$\sigma_{1n}^2 = \frac{2}{n^2 h} \int K^2(u) du + o(\frac{1}{h}). \tag{64}$$

The second term in () is the expected value of a Bias term, that is

$$\widehat{B}_n = \frac{1}{n} \sum_i b_n(x_i) \simeq \frac{h^2}{2} \mu_2 \int g^{(2)}(x) dx + o(h^2), \tag{65}$$

where $g^{(2)}(x)$ is the second derivative of the p.d.f. Hence $\widehat{B}_n = O_p\left(n^{-1/2} h^2\right)$. Thus, what we obtain is

$$\widehat{V}_{1n} = \widehat{V}_{11n} + \widehat{B}_n \sim \sigma_{1n} N(0,1) + \frac{h^2}{2} \mu_2 \int g^{(2)}(x) dx + o(h^2). \tag{66}$$

$$\widehat{V}_{2n} = \frac{1}{n} \sum_i \left[\frac{\widehat{f_n}(x_i) - \overline{f}(x_i)}{g(x_i)} + \frac{\overline{f}(x_i) - g(x_i)}{g(x_i)}\right]^2 =$$

$$= \frac{1}{n} \sum_i \left[\frac{\widehat{f_n}(x_i) - \overline{f}(x_i)}{g(x_i)}\right]^2 + \frac{1}{n} \sum_i \left[\frac{\overline{f}(x_i) - g(x_i)}{g(x_i)}\right]^2 + \frac{2}{n} \sum_i \left(\frac{\widehat{f_n}(x_i) - \overline{f}(x_i)}{g(x_i)}\right) \left(\frac{\overline{f}(x_i) - g(x_i)}{g(x_i)}\right) = \tag{67}$$

$$= \widehat{V}_{21n} + \widehat{V}_{22n} + \widehat{V}_{23n}. \tag{68}$$

$$\widehat{V}_{21n} = \frac{1}{n} \sum_i \left(\frac{1}{n-1} \sum_{j, i \neq j} a_n(x_i, x_j)\right)^2 =$$

$$= \frac{1}{n(n-1)^2} \sum_i \sum_{j, i \neq j} a_n^2(x_i, x_j) + \frac{2}{n(n-1)} \sum_i \sum_{j \neq i} \sum_{z \neq j} a_n(x_i, x_j) a_n(x_i, x_z). \tag{69}$$

The first part is a variance term and it will affect the mean of the asymptotic distribution. The second term equals a twice centered degenerate U-statistic $\widehat{U}_{2n}$, which is of the same order of magnitude of $\widehat{V}_{11n}$ and it also affects the asymptotic distribution of KI.

As $n \to \infty$,by Lemma 2 Hall(1984) the first term of $\widehat{V}_{21n}$ is given by

$$\frac{1}{n(n-1)^2}\sum_i \sum_{j,i\neq j} a_n^2(x_i,x_j) = \sigma_n^2 + O_p(n^{-3/2}h^{-1}), \tag{70}$$

where $\sigma_n^2 = \frac{1}{2n}\sigma_{1n}^2$.

$$2\widehat{U}_{2n} = \frac{2}{n(n-1)}\sum_i \sum_{i\neq j} \int a_n(x_j,x)a_n(x_i,x)g(x)dx = \frac{2}{n(n-1)}\sum_i \sum_{i\neq j} H_{2n}(x_i,x_j), \tag{71}$$

$$H_{2n}(x_i,x_j) = \int \frac{1}{h^2}\left[\frac{K_h(x_i-x_j) - \int K_h(x_i-x_j)g(x_j)dx_j}{g(x_i)}\right]\left[\frac{K_h(x_i-x_z) - \int K_h(x_i-x_z)g(x_z)dx_z}{g(x_i)}\right]g(x_i)dx_i.$$

$$E\left[H_{2n}^2(x_i,x_j)\right] = \frac{1}{h^4}E\left[\int\left(\frac{K_h(x_i-x_j) - \int K_h(x_i-x_j)g(x_j)dx_j}{g(x_i)}\right)\left(\frac{K_h(x_i-x_z) - \int K_h(x_i-x_z)g(x_z)dx_z}{g(x_i)}\right)g(x_i)dx_i\right]^2 =$$

$$\frac{1}{h^4}\int\int\left[\int\left(\frac{K_h(x_i-x_j) - \int K_h(x_i-x_j)g(x_j)dx_j}{g(x_i)}\right)\left(\frac{K_h(x_i-x_z) - \int K_h(x_i-x_z)g(x_z)dx_z}{g(x_i)}\right)g(x_i)dx_i\right]^2 g(x_j)g(x_z)dx_jdx$$

$$= \frac{1}{h^4}\int\left[\int \frac{K_h(x_i-x_j)K_h(x_i-x_z)}{g^2(x_i)}g(x_i)dx_i\right]^2 g(x_j)g(x_z)dx_jdx_z + o(\frac{1}{h}) =$$

$$= \frac{1}{h^4}\int\left[h\int \frac{K(u)K(u+v)}{g(x_j+hu)}du\right]^2 g(x_j)g(x_j+hu-hz)dx_jhdv + o(\frac{1}{h}) = \frac{1}{h}\int \frac{1}{g^2(x_j)}\left[\int K(u)K(u+v)du\right]^2 g^2(x_j)dx_jdv =$$

$$\simeq h^{-1}\int\left[\int K(u)K(u+v)du\right]^2 dv + o(\frac{1}{h}). \tag{72}$$

By Lemma 3 in Hall(84), then $\widehat{U}_{2n}$ is asymptotically Normally distributed $N(0,\sigma_{2n}^2)$, where

$$\sigma_{2n}^2 \simeq 2n^{-2}h^{-1}\int\left[\int K(u)K(u+v)du\right]^2 dv. \tag{73}$$

Hence finally we have that

$$\widehat{V}_{21n} \sim \sigma_n^2 + O_p(n^{-3/2}h^{-1}) + 2\sigma_{2n}N(0,1). \tag{74}$$

$\widehat{V}_{22n} = \frac{1}{n}\sum_i\left[\frac{\overline{f}(x_i)-g(x_i)}{g(x_i)}\right]^2 = \frac{1}{n}\sum_i b_n^2(x_i)$, which is a purely deterministic Bias-squared term, and it will affect the mean of the asymptotic distribution. That is,

$$\frac{1}{n}\sum_i b_n^2 = \frac{h^4}{4}\mu_2^2\int \frac{\left(g^{(2)}(x)\right)^2}{g(x)}dx + o(h^4). \tag{75}$$

Finally we can analyze $\widehat{V}_{23n}$:

$$2\widehat{V}_{23n} = \frac{2}{n}\sum_i\left(\frac{\widehat{f_n}(x_i) - \overline{f}(x_i)}{g(x_i)}\right)\left(\frac{\overline{f}(x_i) - g(x_i)}{g(x_i)}\right) = \frac{2}{n(n-1)}\sum_i \widehat{H}_{3ni}, \tag{76}$$

similarly to Hall(1984) define

$$H_{3ni} = \sum_j a_n(x_i, x_j) b_n(x_i) = \frac{1}{h} \int \left[\frac{K_h(x_i - x) - \int K_h(x_i - x_j) g(x_j) dx_j}{g(x_i)} \right] \left(\frac{\overline{f}(x_i) - g(x_i)}{g(x_i)} \right) dx \qquad (77)$$

Under assumptions A1 and A2 and given that $EH_{23ni} = 0$, by Lemma1 in Hall(1984) we have that $2\widehat{V}_{23n}$ is asymptotically normally distributed with zero mean and variance given by:

$$\sigma_{3n}^2 \simeq 2n^{-1} h^4 \mu_2^2 \left[\int \frac{\left(g^{(2)}(x_i)\right)^2}{g(x_i)} dx_i - \left(\int \left(g^{(2)}(x_i) \right) dx_i \right)^2 \right], \qquad (78)$$

which can be easily seen if we consider that $\frac{\overline{f}(x_i) - g(x_i)}{g(x_i)} = \frac{h^2 \mu_2 g^{(2)}(x_i)}{g(x_i)}$and that

$$EH_{3ni}^2 = h^4 \mu_2^2 \left[\int \frac{\left(g^{(2)}(x_i)\right)^2}{g(x_i)} dx_i - \left(\int \left(g^{(2)}(x_i) \right) dx_i \right)^2 \right].$$

Also this term will affect the asymptotic distribution of $\widehat{KI}_1$.

To summarize all previous steps, we can rewrite the expansion of $\widehat{KI}_1$ in the following way:

$$\widehat{KI}_1 = \widehat{V}_{11n} + \widehat{B}_n - \frac{1}{2} \left(\widehat{V}_{21n} + \widehat{V}_{22n} + 2\widehat{V}_{23n} \right) \sim \qquad (79)$$

$$N(0, \sigma_{1n}^2) + \frac{h^2}{2} \mu_2 \int g^{(2)}(x) dx + o(h^2) - \frac{1}{2} \left(\sigma_n^2 + O_p(n^{-3/2} h^{-1}) + 2N(0, \sigma_{2n}^2) + \frac{h^4}{4} \mu_2^2 \int \frac{\left(g^{(2)}(x)\right)^2}{g(x)} dx + o(h^4) + 2N(0, \sigma_{3n}^2) \right).$$

Once more, following Hall(1984), from the definition of $\widehat{V}_{21n}$ and the fact that $nh \to \infty$, we have that the difference between $\frac{1}{n(n-1)} \sum_i \sum_{j \neq i} a_n^2(x_i, x_j)$ and σ_n^2 is negligible w.r.t. $2\widehat{U}_{2n}$, hence the previous expression can be rewritten as follows:

$$\widehat{KI}_1 \sim (nh^{1/2})^{-1} \sqrt{2} \sigma_1 N_1 - (nh^{1/2})^{-1} \sqrt{2} \sigma_2 N_2 - n^{-1/2} h^2 \sqrt{2} \sigma_3 N_3 + \widehat{B}_n - \frac{1}{2} c_n, \qquad (80)$$

where N_1, N_2 and N_3 are asymptotically normal N(0,1); and

$$\sigma_1 = \int K^2(u) du, \; \sigma_2 = \int \left[\int K(u) K(u+v) du \right]^2 dv \text{ and } \sigma_3 = \mu_2^2 \left[\int \frac{\left(g^{(2)}(x_i)\right)^2}{g(x_i)} dx_i - \left(\int \left(g^{(2)}(x_i) \right) dx_i \right)^2 \right],$$

$$\text{and } c_n = (nh)^{-1} \int K^2(u) du + \frac{h^4}{4} \mu_2^2 \int \left(\frac{g^{(2)}(x)}{g(x)} \right)^2 dx + o(n^{-1} h^{-1} + h^4). \qquad (81)$$

It is important to notice that $\widehat{B}_n$, which is $O_p(n^{-1/2} h^2)$, will asymptotically cancel out with $n^{-1/2} h^2 \sqrt{2} \sigma_3 N_3$, since they are of the same order of magnitude.

Thus, we have the following results: as $n \to \infty$, $h \to 0$, $nh \to \infty$ and $nh^5 \to 0$

$$nh^{1/2} (\widehat{KI}_1 + \frac{1}{2} c_n) \to^d \sqrt{2} \sigma_1 N_1 - \sqrt{2} \sigma_3 N_2. \qquad (82)$$

Since $aN(0,1)+bN(0,1)$ can be written as partial sum of martingale difference array, and it can be proved to be asymptotically normal $N(0,a^2+b^2)$ (see Hall(84) p.10), then we have that $\sqrt{2}\sigma_1 N_1 - \sqrt{2}\sigma_3 N_2 = nh^{1/2}(\widehat{KI}_1 + \frac{1}{2}c_n) \to \sqrt{2}(\sigma_1-\sigma_3)N(0,1)$.

Let now examine the term

$$KI_2 = \int (\ln f_{\widehat{\theta}}(x) - \ln g(x)) d\widehat{F}_n(x) = \int (\ln f_{\widehat{\theta}}(x_i) - \log f_{\theta^*}(x_i) + \log f_{\theta^*}(x_i) - \ln g(x_i)) d\widehat{F}_n(x_i).$$

We start examining the limiting distribution of

$$\widehat{KI}_2 = \frac{1}{n}\sum_{i=1}\left(\log f_{\widehat{\theta}}(x_i) - \log f_{\theta^*}(x_i)\right)\widehat{f}_n(x_i) + \frac{1}{n}\sum_{i=1}\left(\log f_{\theta^*}(x_i) - \log g(x_i)\right)\widehat{f}_n(x_i) = \widehat{KI}_{21} + \widehat{KI}_{22}, \quad (83)$$

that similarly of $\widehat{KI}_1$by the LLN, can be considered a good approximation of $n^{-1}E(\ln f_{\widehat{\theta}}(x) - \ln g(x))$. This part of the proof is based mainly on Zheng (1996).

Employing the same expansion used for $\widehat{KI}_1$, where now $u = \frac{f_{\widehat{\theta}}(x_i) - f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)}$:

$$\frac{1}{n}\sum_{i=1}\log\left(\frac{f_{\widehat{\theta}}(x_i)}{f_{\theta^*}(x_i)}\right) \simeq \frac{1}{n}\sum_{i=1}\frac{f_{\widehat{\theta}}(x_i) - f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)} - \frac{1}{2n}\sum_{i=1}\left(\frac{f_{\widehat{\theta}}(x_i) - f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)}\right)^2,$$

we can rewrite $\widehat{KI}_{21}$ in the following way:

$$\widehat{KI}_{21}(f_{\widehat{\theta}}, f_{\theta^*}) \simeq \frac{1}{n}\sum_{i=1}\left(\frac{f_{\widehat{\theta}}(x_i) - f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)}\right)\widehat{f}_n(x_i) - \frac{1}{2n}\sum_{i=1}\left(\frac{f_{\widehat{\theta}}(x_i) - f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)}\right)^2\widehat{f}_n(x_i) = I_{n1} - \frac{1}{2}I_{n2}. \quad (84)$$

Applying the mean value theorem to $f_{\widehat{\theta}}(x_i)$ we obtain:

$$f_{\widehat{\theta}}(x_i) - f_{\theta^*}(x_i) \cong \frac{\partial f_{\theta^*}(x_i)}{\partial\theta}(\widehat{\theta}-\theta^*) + \frac{1}{2}(\widehat{\theta}-\theta^*)'\frac{\partial^2 f_{\overline{\theta}}(x_i)}{\partial\theta\partial\theta'}(\widehat{\theta}-\theta^*);$$

thus,

$$I_{n1} = \frac{1}{n}\sum_{i=1}\frac{\widehat{f}_n(x_i)}{f_{\theta^*}(x_i)}\left(\widehat{f}_{\theta}(x_i) - f_{\theta^*}(x_i)\right) \simeq \quad (85)$$

$$\simeq \frac{1}{n}\sum_i \frac{\widehat{f}_n(x_i)}{f_{\theta^*}(x_i)}\frac{\partial f_{\theta^*}(x_i)}{\partial\theta}(\widehat{\theta}-\theta^*) + \frac{1}{2n}\sum_i(\widehat{\theta}-\theta^*)'\frac{\widehat{f}_n(x_i)}{f_{\theta^*}(x_i)}\frac{\partial^2 f_{\theta^*}(x_i)}{\partial\theta\partial\theta'}(\widehat{\theta}-\theta^*) =$$

$$= \frac{1}{n(n-1)}\sum_i\sum_j\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right)\frac{\partial f_{\theta^*}(x_i)/\partial\theta}{f_{\theta^*}(x_i)}(\widehat{\theta}-\theta^*) +$$

$$+(\widehat{\theta}-\theta^*)'\frac{1}{2n(n-1)}\sum_i\sum_j\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right)\frac{\partial^2 f_{\overline{\theta}}(x_i)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}(\widehat{\theta}-\theta^*) =$$

$$= S_{1n}(\widehat{\theta}-\theta^*) + (\widehat{\theta}-\theta^*)'S_{2n}(\widehat{\theta}-\theta^*). \quad (86)$$

It can be noticed that the U-statistic form of S_{1n} is the same as that of U_n defined in theorem 2. It follows that $S_{1n} = O_p(\frac{1}{\sqrt{n}})$.

$$E\left(S_{2n}\right)=\frac{1}{2n(n-1)}\sum_{i}\sum_{j}E\left[\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right)\frac{\partial^2 f_{\overline{\theta}}\left(x_i\right)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}\right], \tag{87}$$

$$E\left[\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right)\frac{\partial^2 f_{\overline{\theta}}\left(x_i\right)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}\right]=\frac{1}{h}\int\int K\left(\frac{x_j-x_i}{h}\right)\frac{\partial^2 f_{\overline{\theta}}\left(x_i\right)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}g(x_i)g(x_j)dx_idx_j=$$

$$=\int\int K\left(u\right)\frac{\partial^2 f_{\overline{\theta}}\left(x_i\right)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}g(x_i)g(x_i+hu)dx_idu. \tag{88}$$

Similarly to Dimitriev-Tarasenko(1973), applying the Cauchy-Schwartz inequality we obtain that

$$\lim\sup_{n\to\infty}E\left(S_{2n}\right)\leq\int\frac{\partial^2 f_{\overline{\theta}}\left(x_i\right)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}g^2(x)dx=E\left(\frac{\partial s(\overline{\theta},x)}{\partial\theta}g(x)\right); \tag{89}$$

further, since $E\widehat{f}_n(x)=g(x)$, applying Fatou-Lebesgue theorem we have that

$$\lim_{n\to\infty}E\left(S_{2n}\right)=E\left(\frac{\partial s(\overline{\theta},x)}{\partial\theta}g(x)\right). \tag{90}$$

Thus, we have that $S_{2n}=O_p(1)$. Taking into account that $\sqrt{n}(\widehat{\theta}-\theta^*)=O_p(1)$, which in turn implies that $(\widehat{\theta}-\theta^*)=O_p(\frac{1}{\sqrt{n}})$, it follows that $I_{n1}=S_{1n}(\widehat{\theta}-\theta^*)+(\widehat{\theta}-\theta^*)'S_{2n}(\widehat{\theta}-\theta^*)$ is equal to

$$I_{n1}=O_p(\frac{1}{\sqrt{n}})*O_p(\frac{1}{\sqrt{n}})+O_p(\frac{1}{\sqrt{n}})*O_p(1)*O_p(\frac{1}{\sqrt{n}})=O_p(\frac{1}{n}). \tag{91}$$

Now we have to consider I_{n2}:

$$I_{n2}=\frac{1}{n}\sum_{i=1}\left(\frac{\widehat{f}_\theta\left(x_i\right)-f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)}\right)^2\widehat{f}_n(x_i)\simeq(\widehat{\theta}-\theta^*)'\frac{1}{n(n-1)}\sum_{i}\sum_{j}\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right)\frac{\partial\ln f_{\overline{\theta}}\left(x_i\right)}{\partial\theta}\frac{\partial\ln f_{\overline{\theta}}\left(x_j\right)}{\partial\theta'}(\widehat{\theta}-\theta^*)\simeq \tag{92}$$

$$\simeq(\widehat{\theta}-\theta^*)'\left(\frac{1}{n(n-1)h}\sum_{i}\sum_{j}K\left(\frac{x_j-x_i}{h}\right)s(\overline{\theta},x_i)s(\overline{\theta},x_j)'\right)(\widehat{\theta}-\theta^*)=(\widehat{\theta}-\theta^*)'S_{3n}(\widehat{\theta}-\theta^*)'. \tag{93}$$

Similarly to S_{2n}, it can be shown that S_{3n} is $O_p\left(1\right)$. It follows that I_{n2}

$$I_{n2}=O_p\left(\frac{1}{\sqrt{n}}\right)*O_p\left(1\right)*O_p\left(\frac{1}{\sqrt{n}}\right)=O_p\left(\frac{1}{n}\right). \tag{94}$$

Finally, we get that:

$$\widehat{KI}_{21}(f_{\widehat{\theta}},f_{\theta^*})\simeq I_{n1}-\frac{1}{2}I_{n2}=O_p(\frac{1}{n})-\frac{1}{2}O_p\left(\frac{1}{n}\right)=O_p\left(\frac{1}{n}\right),$$

then it follows that

$$(nh^{1/2})\widehat{KI}_{21}(f_{\widehat{\theta}},f_{\theta^*})=(nh^{1/2})O_p\left(\frac{1}{n}\right)=O_p(h^{1/2})\to^p 0. \tag{95}$$

Now, the same expansion used for $\widehat{KI}_{21}$ can be applied to $\widehat{KI}_{22}(f_{\theta^*}, g)$:

$$\widehat{KI}_{22}(f_{\theta^*}, g) \cong \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f_{\theta^*}(x_i)-g(x_i)}{g(x_i)}\right)\widehat{f}_n(x_i) - \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{f_{\theta^*}(x_i)-g(x_i)}{g(x_i)}\right)^2\widehat{f}_n(x_i) = J_{n1} - \frac{1}{2}J_{n2}, \quad (96)$$

$$E\left(J_{1n}(f_{\theta^*}, g)\right) = E\left(\int\left(\frac{f_{\theta^*}(x_i)-g(x_i)}{g(x_i)}\right)\widehat{f}_n(x_i)g(x_i)dx_i\right) = \int\int K(u)\left(f_{\theta^*}(x)-g(x)\right)g(x+hu)dxdu. \quad (97)$$

Applying the same steps used for S_{2n} we can show that

$$\limsup_{n\to\infty} E\left(J_{1n}(f_{\theta^*}, g)\right) \leq \int\left(f_{\theta^*}(x)-g(x)\right)g(x)dx = E\left(f_{\theta^*}(x)-g(x)\right)$$

$$\lim_{n\to\infty} E\left(J_{1n}(f_{\theta^*}, g)\right) = E\left(f_{\theta^*}(x)-g(x)\right). \quad (98)$$

It follows that $J_{1n}(f_{\theta^*}, g) = O_p(1)$. Repeating the same steps once more for $J_{2n}(f_{\theta^*}, g)$ we obtain:

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{f_{\theta^*}(x_i)-g(x_i)}{g(x_i)}\right)^2\widehat{f}_n(x_i)\right) = E\left(\int\left(\frac{f_{\theta^*}(x_i)-g(x_i)}{g(x_i)}\right)^2\widehat{f}_n(x_i)g(x_i)dx_i\right) =$$

$$= E\left(\int\frac{\left(f_{\theta^*}(x_i)-g(x_i)\right)^2}{g(x_i)}\widehat{f}_n(x_i)dx_i\right) = \int\int K(u)\frac{\left(f_{\theta^*}(x)-g(x)\right)^2}{g(x)}g(x+hu)dxdu, \quad (99)$$

$$\limsup_{n\to\infty} E\left(J_{2n}(f_{\theta^*}, g)\right) \leq \int\left(f_{\theta^*}(x)-g(x)\right)^2dx \quad (100)$$

$$\lim_{n\to\infty} E\left(J_{2n}(f_{\theta^*}, g)\right) = \int\left(f_{\theta^*}(x)-g(x)\right)^2dx > 0. \quad (101)$$

Then also $J_{2n}(f_{\theta^*}, g) = O_p(1)$. This implies that $\widehat{KI}_{22}(f_{\theta^*}, g) = J_{n1} - \frac{1}{2}J_{n2} = O_p(1)$.

Then it is clear that given assumptions A1-A4, if $h \to 0$, $nh \to \infty$ then

$$\widehat{KI}_{22}(f_{\theta^*}, g) \to^p E\left(f_{\theta^*}(x)-g(x)\right) - \frac{1}{2}\int\left(f_{\theta^*}(x)-g(x)\right)^2dx \cong E\left[\ln f_{\theta^*} - \ln g\right], \quad (102)$$

this implies that $nh^{1/2}\widehat{KI}_{22} \to^p \infty$, hence we need to rescale it by $d_n = n^{-1}h^{-1/2}$ where $d_n \to 0$ as $n \to \infty$. This is embodied in assumption A6:

$$\widehat{KI}_{22} \simeq \alpha h^{1/2}c_n \quad (103)$$

Finally we can put all terms together:

$$KI = \int_x(\ln\widehat{f_n}(x) - \ln f_{\widehat{\theta}}(x))\widehat{f_n}(x)dx \cong \widehat{KI}_1 - \widehat{KI}_2 \sim$$

$$\sim \left[(nh^{1/2})^{-1}\sqrt{2}\sigma_1 N_1 - (nh^{1/2})^{-1}\sqrt{2}\sigma_2 N_2 - \frac{1}{2}c_n\right] - \left[\widehat{KI}_{21}(f_{\widehat{\theta}}, f_{\theta^*}) + \widehat{KI}_{22}(f_{\theta^*}, g)\right], \quad (104)$$

since we showed that

$$(nh^{1/2})\widehat{KI}_{21}(f_{\widehat{\theta}}, f_{\theta^*}) \rightarrow^p 0 \tag{105}$$

the entire expression for $(nh^{1/2})KI$ can be approximated in the following way

$$(nh^{1/2})\left[(nh^{1/2})^{-1}\sqrt{2}\sigma_1 N_1 - (nh^{1/2})^{-1}\sqrt{2}\sigma_2 N_2 - \frac{1}{2}c_n - \left(J_{n1} - \frac{1}{2}J_{n2}\right)\right]. \tag{106}$$

Thus, if $h \propto n^{-\beta}$ with $\beta \geqslant \frac{1}{5}$, $c_n \simeq C(nh)^{-1}$

$$(nh^{1/2})\left(KI + \frac{1}{2}c_n\right) \sim \sqrt{2}\sigma_1 N_1 - \sqrt{2}\sigma_2 N_2 + \alpha C \tag{107}$$

then,

$$(nh^{1/2})\left(KI + \frac{1}{2}c_n\right) \rightarrow^d N\left(\alpha C, 2\left(\sigma_1^2 - \sigma_2^2\right)\right) \tag{108}$$

9.4 Formula of MGF and expected utility

It can be shown that the moment generating function for the double Gamma distribution is:

$$M_R(t) = \exp(t\gamma)[pMGF(t) + (1-p)MGF(-t)] =$$
$$\exp(t\gamma)[p(1-\phi_1 t)^{-\zeta_1} + (1-p)(1+\phi_2 t)^{-\zeta_2}$$

hence $E(U(R))$ where $t = a\delta$ and $\phi_i = 1/\lambda_i$ is given by the following expression:

$$E_{g_{r|s}}[U(W(\widetilde{r}_t, a))] = -M_R(-t) = -\exp(-a\delta\gamma)[p(1-\phi_1 a\delta)^{-\zeta_1} + (1-p)(1+\phi_2 a\delta)^{-\zeta_2}]$$

For the Gumbel distribution we have the following expression:

$$M_R(t) = \exp(\alpha t)\Gamma(1-\beta t)$$

$$E_{g_{r|s}}[U(W(\widetilde{r}_t, a))] = -M_R(-a\delta) = -\exp(-\alpha a\delta)\Gamma(1+\beta a\delta)$$

For the Normal we have the well known result:

$$M_R(t) = \exp(t\mu - \frac{1}{2}t^2\sigma^2)$$

$$E_{g_{r|s}}[U(W(\widetilde{r}_t, a))] = -M_R(-a\delta) = \exp(-a\delta\mu + \frac{1}{2}a^2\delta^2\sigma^2).$$

References

[1] Ahmad, I.A. and P.E. Lin; A Nonparametric Estimation of the Entropy for Absolutely Continuous Distributions. *IEEE Transactions on Information Theory*, 22, pp.372-375, 1976.

[2] Aït-Sahalia Yacine; Nonparametric Pricing of Interest Rate Derivative Securities, *Econometrica*, Vol.64 No.3 (May 1996), 527-560.

[3] Ang A.,and G. Bekaert, International Asset Allocation with Regime Shifts, *Review of Financial Studies* 15, 4, pp.1137-87, 2002.

[4] Ang A.,and G. Bekaert, How Do Regimes Affect Asset Allocation, Columbia Business School, 2002.

[5] Avramov D., Stock Return Predictability and Model Uncertainty, *The Wharton School Working Paper*, May 2000.

[6] Avramov D., Stock Return Predictability and Model Uncertainty, *Journal of Financial Economics* 64, pp.423-458, 2002.

[7] Bedford T. and R.Cooke; Probabilistic Risk Analysis, Cambridge University Press 2001.

[8] Chen X. and Huang J.Z.; Semiparametric and Nonparametric Estimation via the Method of Sieves, Manuscript New York University, Noevmber 2002.

[9] Cremers K. J. M.; Stock Return Predictability: A Bayesian Model Selection Perspective, *The Review of Financial Studies* Vol.15, No.4, pp.1223-1249, Fall 2002.

[10] Cristobal Cristobal J.A., Roca P.F. and W.G. Manteiga, A Class of Linear Regression Parameter Estimators Constructed by Nonparametric Estimation, *The Annals of Statistics* 1987, Vol.15, No. 2, 603-09.

[11] Dhrymes P.J.; Topics in Advanced Econometrics Volume I and II, Springer-Verlag 1993.

[12] Dhrymes P.J; Identification and Kullback Information in the GLSEM, *Journal of Econometrics* 83, 163-184 (1998).

[13] Diebold F.X., T.A. Gunther and A.S. Tay, Evaluating Density Forecasts, *PIER Working Paper* 97-018.

[14] Diebold, F.X. and J.A.Lopez, Forecast Evaluation and Combination. In G.S.Maddala and C.R.Rao, *Handbook of Statistics,* Volume 14, pp.241-68. Amsterdam: North-Holland.

[15] Dmitriev, Yu G. and F.P. Tarasenko, On the Estimation of Functionals of the Probability Density and its Derivatives, *Theory of Probability and Its Application* 18, pp.628-33, 1973.

[16] Dmitriev, Yu G. and F.P. Tarasenko, On a Class of Nonparametric Estimates of Nonlinear Functionals, *Theory of Probability and Its Application* 19, pp.390-94, 1974.

[17] Dudewicz E.J. and Edward C. van der Mulen, The Empiric Entropy, A New Approach to Nonparametric Entropy Estimation, in Puri M., Vilaplana J.P. and Wertz W., New Perspectives in Theoretical and Applied Statistics, 1987.

[18] Ebrahimi N.,Maasoumi E. and Soofi E.S.; Ordering Univariate Distributions by Entropy and Variance, *Journal of Econometrics* 90, 1999 pag 317-336.

[19] Fan Y., Testing the goodness of fit of a parametric density function by kernel method, *Econometric Theory, 10,* 316-356, 1994.

[20] Giacomini R., Comparing Density Forecasts via Weighted Likelihood Ratio Tests: Asymptotic and Bootstrap Methods, working paper University of California San Diego, June 2002.

[21] Giacomini R. and H. White, Test of conditional predicitve ability, working paper University of California San Diego, April 2003.

[22] Gilboa I. and D. Schmeidler; Cognitive Foundations of Inductive Inference and Probability: An Axiomatic Approach, Mimeo March 2000.

[23] Gilboa I. and D. Schmeidler; Inductive Inference: An Axiomatic Approach, February 2001, forthcoming on Econometrica.

[24] Hall P.; Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators, *Journal of Multivariate Analysis* 14, 1-16 (1984).

[25] Hall P.; On Kullback-Leibler Loss and Density Estimation, *The Annals of Statistics*, Volume 15, Issue 4, 1491-1519 (Dec.,1987).

[26] Hansen L.P. and T.J. Sargent; Acknowledging Misspecification in Macroeconomic Theory, *Review of Economic Dynamics* 4, 519-535 2001.

[27] Härdle W.; Applied Nonparametric Regression, Econometric Society Monographs 1990.

[28] Hasminskii R.Z. and I.A. Ibragimov, On the Nonparametric Estimation of Functionals, in Mandl P. and M. Huskova, Proceedings of the Second Prague Symposium on Asymptotic Statistics, August 1978.

[29] Hendry D.F. and M.P.Clements; Pooling of Forecasts, *Econometrics Journal*, volume 5, pp.1-26, 2002.

[30] Henry M.; Estimating Ambiguity, Manuscript Columbia University 2001.

[31] Hong Y. and H. White, Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence, manuscript July 2000.

[32] Keuzenkamp H. A.; Probability, Econometrics and Truth, Cambridge University Press 2000.

[33] Knight J.L., Satchell S.E. and K.C. Tran, Statistical modelling of asymmetric risk in asset returns, *Applied Mathematical Finance* 2, 1995, 155-172.

[34] Maasoumi E. and Racine J.; Entropy and Predictability of Stock Market Returns, *Journal of Econometrics* 107, 2002 pags 291-312.

[35] Pagan A.and A. Ullah; Nonparametric Econometrics, Cambridge University Press 1999.

[36] Robinson, P.M., Consistent Nonparametric Entropy-Based Testing, *Review of Economic Studies* 1991, 58, 437-53.

[37] Roca P.F. and Manteiga W.G., Efficiency of a New Class of Linear Regression Estimates Obtained by Preliminary Nonparametric Estimation, in Puri M., Vilaplana J.P. and Wertz W., New Perspectives in Theoretical and Applied Statistics, 1987.

[38] Sawa T., Information Criteria for Discriminating Among Alternative Regression Models, *Econometrica* Vol.46, Nov.1978.

[39] Sims A.; Uncertainty Across Models, *The American Economic Review*, Volume 78, Issue 2, 163-67 (May, 1988).

[40] Sin C.Y. and H. White, Information Criteria for Selecting Possibly Misspecified Parametric Modles, *Journal of Econometrics* 71 (1996), pp207-225.

[41] Ullah A.; Entropy, Divergence and Distance Measures with Econometric Applications, *Journal of Statistical Planning and Inference* 49 (1996) 137-162.

[42] Uppal R. and T. Wang, Model Misspecification and Under-Diversification, mimeo January 2002.

[43] Vapnik V.N.; The Nature of Statistical Learning Theory, Springer-Verlag 2000 (math Q325.7.V37).

[44] White H., Maximum Likelihood Estimation of Misspecified Models, Econometrica, 1982, volume 50, number1.

[45] White H., Estimation, Inference and Specification Analysis, Cambridge University Press 1994.

[46] Zheng J.X., A Consistent Test of Functional Form Via Nonparametric Estimation Techniques, *Journal of Econometrics* 75 (1996) pp263-289.

[47] Zheng J.X, A Consistent Test of Conditional Parametric Distributions, *Econometric Theory*, 16, 2000, pp 667-691.

10 Tables

True:G(9,3)	$h \propto n^{-1/4.5}$	N=400		t=1000			
		Gamma		Normal		Weibull	
		par(1)	par(2)	par(1)	par(2)	par(1)	par(2)
NPQMLE	estim.θ	8.22	3.33	25.55	9.31	0.000031	3.11
	st.dev.	0.701	0.298	0.545	0.425	0.000015	0.136
	KI$(g, f(\widehat{\theta}))$		**0.0084**		**0.023**		**0.025**
QMLE	estim.θ	9.034	3.03	27.002	9.001	0.000022	3.173
	st.dev.	0.629	0.213	0.449	0.367	0.000013	0.131
	KI$(g, f(\widehat{\theta}))$		**0.0037**		**0.044**		**0.0445**

Table 1

True:G(9,3)	$h \propto n^{-1/4}$	N=400		t=1000			
		Gamma		Normal		Weibull	
		par(1)	par(2)	par(1)	par(2)	par(1)	par(2)
NPQMLE	estim.θ	8.586	3.194	25.536	9.103	0.000026	3.17
	st.dev.	0.756	0.291	0.543	0.425	0.000014	0.143
	KI$(g, f(\widehat{\theta}))$		**0.0062**		**0.0211**		**0.0236**

Table 2

True:G(9,3)	$h \propto n^{-1/4.5}$	N=800		t=1000			
		Gamma		Normal		Weibull	
		par(1)	par(2)	par(1)	par(2)	par(1)	par(2)
NPQMLE	estim.θ	8.36	3.27	25.53	9.21	0.000027	3.143
	st.dev.	0.521	0.214	0.379	0.315	0.000009	0.101
	KI$(g, f(\widehat{\theta}))$		**0.0049**		**0.019**		**0.0222**
QMLE	estim.θ	9.007	3.009	27.011	9.012	0.000021	3.166
	st.dev.	0.469	0.161	0.322	0.276	0.000073	0.098
	KI$(g, f(\widehat{\theta}))$		**0.002**		**0.042**		**0.0432**

Table 3

True:G(9,3)	$h \propto n^{-1/4}$	N=800		t=1000			
		Gamma		Normal		Weibull.	
		par(1)	par(2)	par(1)	par(2)	par(1)	par(2)
NPQMLE	estim.θ	8.65	3.15	25.511	9.05	0.000024	3.183
	st.dev.	0.555	0.212	0.384	0.317	0.000008	0.105
	KI$(g, f(\widehat{\theta}))$		**0.0035**		**0.0186**		**0.0211**

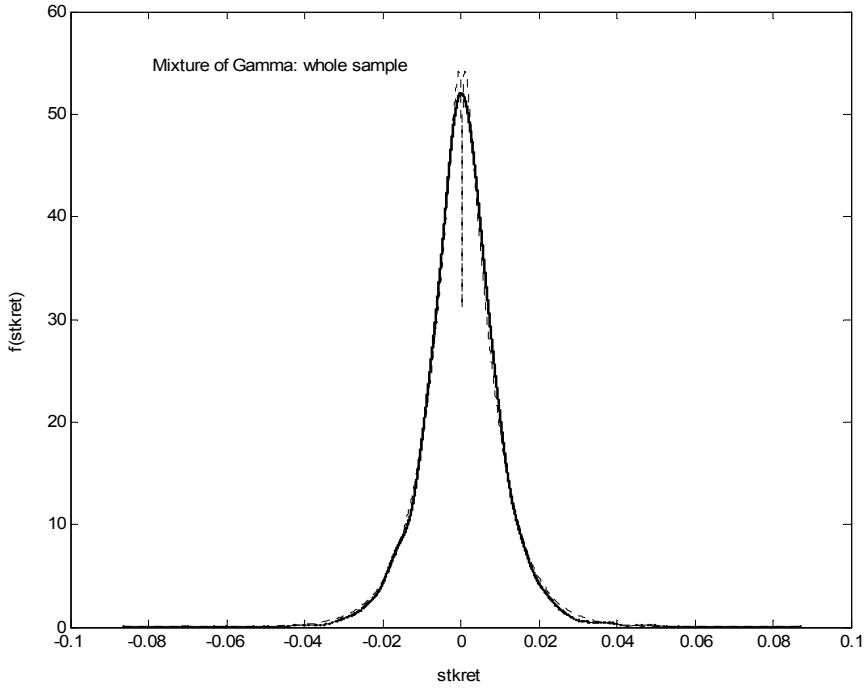
Table 4

True:G(9,3)	$h \propto n^{-1/4.5}$	N=1200		t=1000			
		Gamma		Normal		Weibull	
		par(1)	par(2)	par(1)	par(2)	par(1)	par(2)
NPQMLE	estim.θ	8.47	3.22	25.49	9.14	0.000025	3.157
	st.dev.	0.419	0.166	0.319	0.241	0.000007	0.081
	KI$(g, f(\widehat{\theta}))$		**0.0032**		**0.018**		**0.0207**
QMLE	estim.θ	9.01	2.999	26.992	8.998	0.000021	3.166
	st.dev.	0.353	0.121	0.262	0.208	0.000056	0.076
	KI$(g, f(\widehat{\theta}))$		**0.0012**		**0.041**		**0.042**

Table 5

True:G(9,3)	$h \propto n^{-1/4}$	N=1200		t=1000			
		Gamma		Normal		Weibull	
		par(1)	par(2)	par(1)	par(2)	par(1)	par(2)
NPQMLE	estim.θ	8.75	3.108	25.49	8.993	0.000023	3.196
	st.dev.	0.455	0.171	0.327	0.255	0.000006	0.085
	KI$(g, f(\widehat{\theta}))$		**0.0023**		**0.0174**		**0.0200**

Table 6



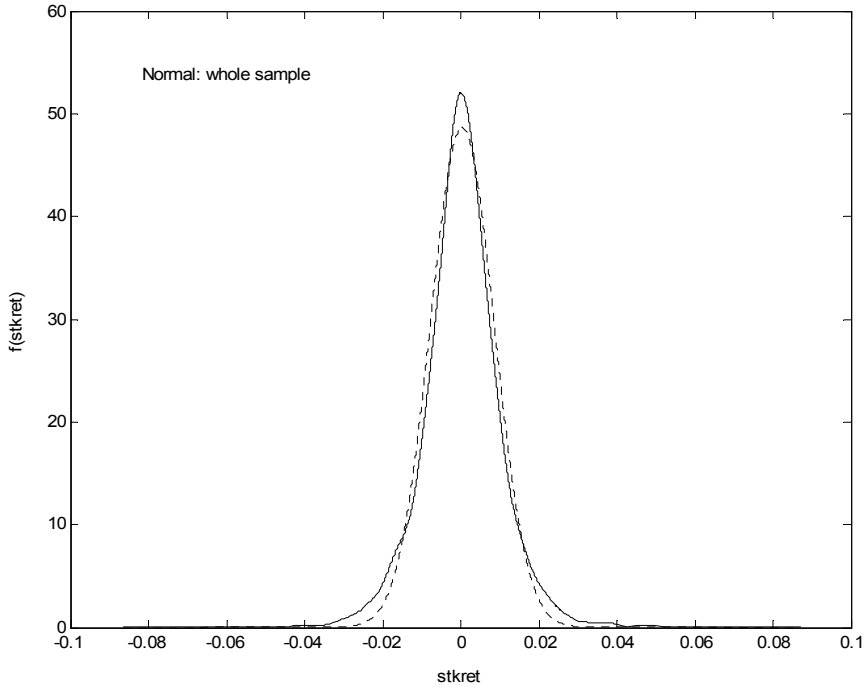
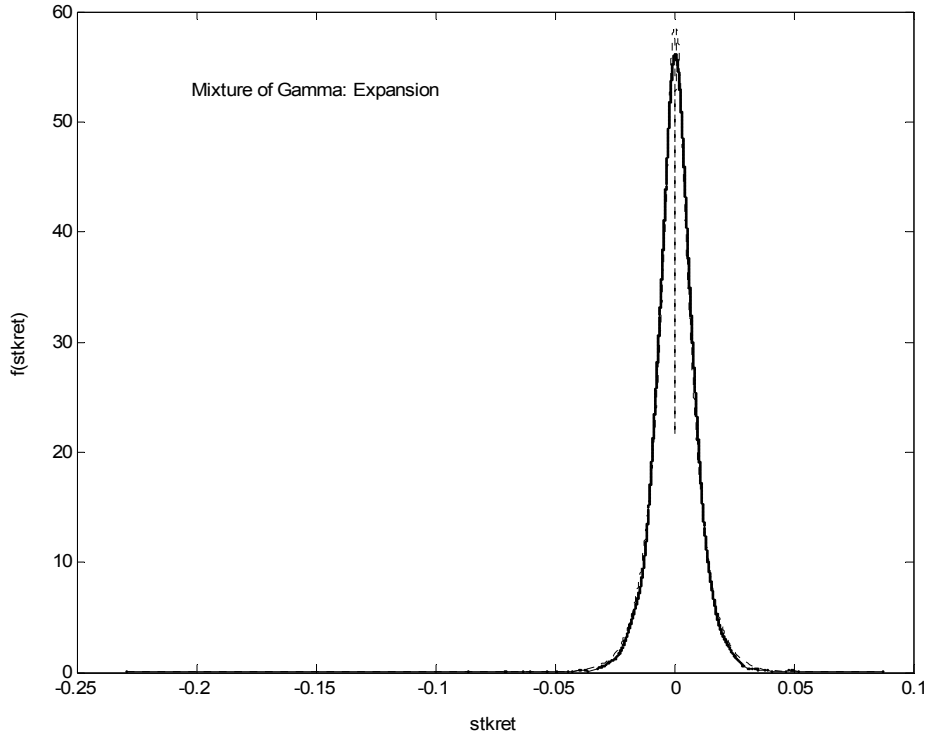


Figure 1 and 2 show the estimation result using the mixture of Gamma and the Normal Distributions, for the entire sample. The solid line represents the nonparametric density, the dotted line the parametric model.



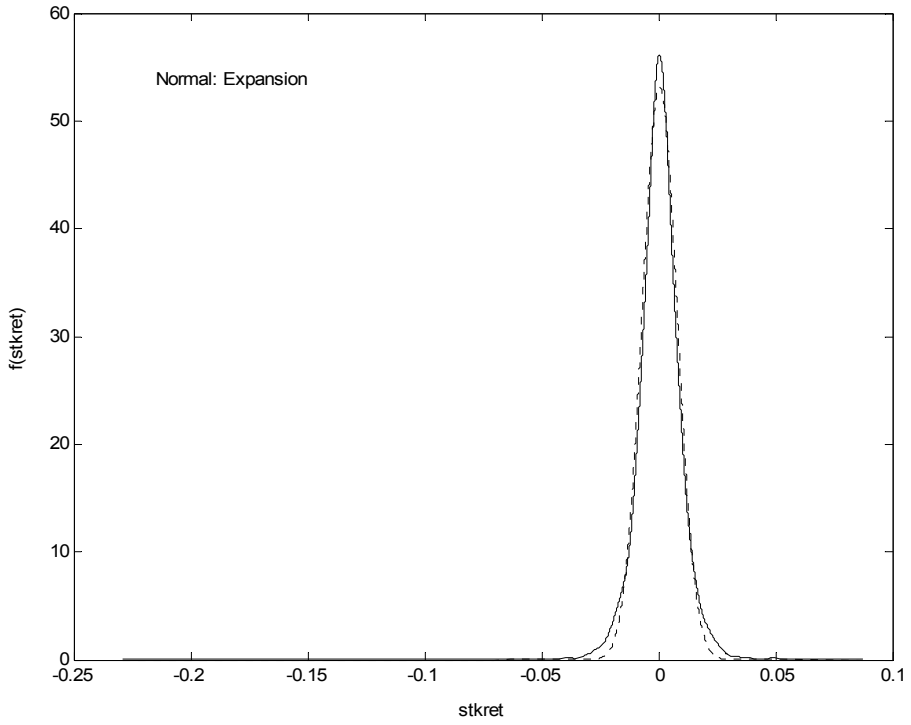
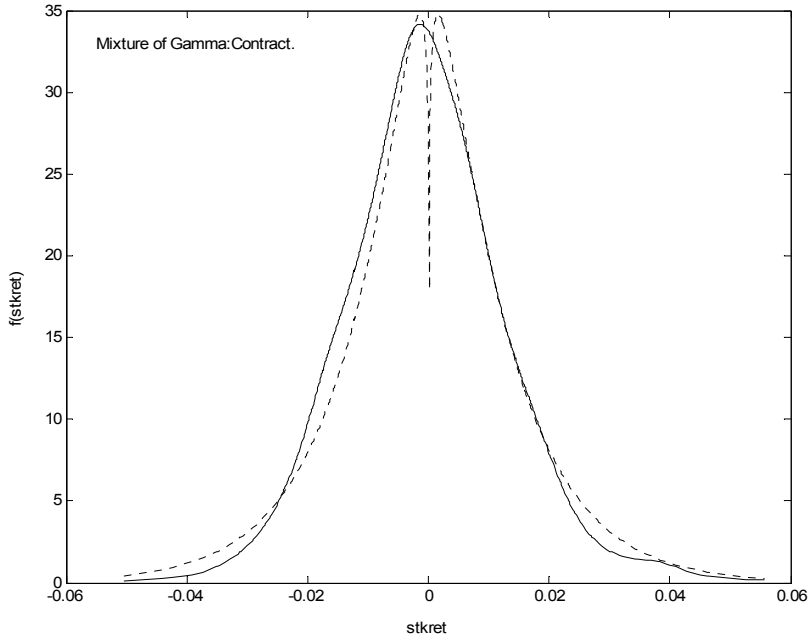


Figure 3 and 4 show the estimation result using the mixture of Gamma and the Normal Distributions, for expansion sample. The solid line represents the nonparametric density, the dotted line the parametric model.



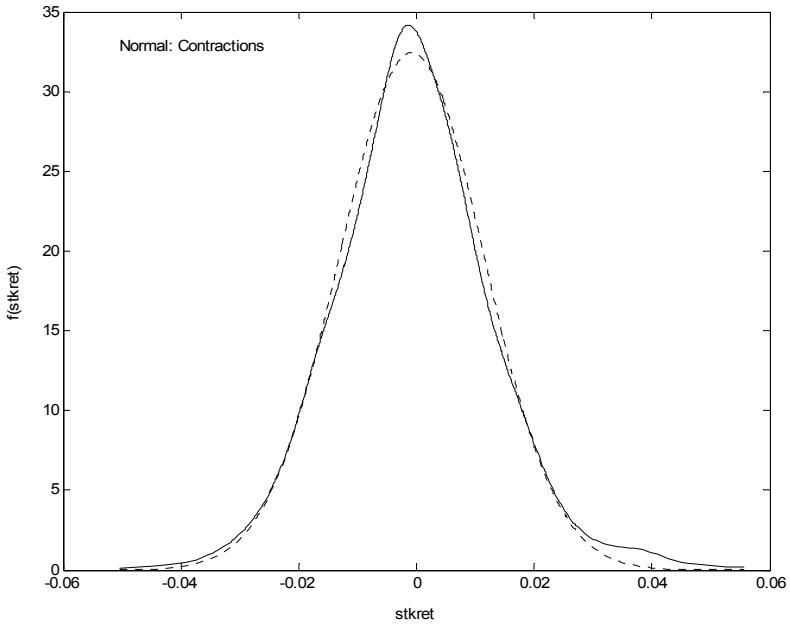


Figure 5 and 6 show the estimation result using the mixture of Gamma and the Normal Distributions, for contraction sample. The solid line represents the nonparametric density, the dotted line the parametric model.