Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Density Selection and Combination Under Model Ambiguity: An Application to Stock Returns

Stefania D'Amico

2005-09

Density Selection and Combination under Model Ambiguity: an Application to Stock Returns

Stefania D'Amico*

First Draft: May 2003, This Draft: January 2005

Abstract

This paper proposes a method for predicting 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 KLD between the nonparametric and the parametric density estimates is asymptotically normally distributed. This result leads to determining the weights in the model combination, using the distribution function of a Normal centered on the average performance of all plausible models. Consequently, the final weight is determined by the ability of a given model to perform better than the average. As such, this combination technique does not require the true structure to belong 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 superior to the Gaussian distribution in modeling stock returns, and that the combination outperforms each individual candidate model both in- and out-of-sample.

Keywords: Density forecast comparison, Kernel density estimation, Entropy, Model Combination.

*I would like to thank Phoebus Dhrymes for his guidance and many interesting discussions. I also wish to thank Jean Boivin, Xiaohong Chen, Mitali Das, Rajeev Dehejia, Stefano Eusepi, Mira Farka, Marc Henry, Alexei Onatski, Athanasios Orphanides and Jonathan Wright for valuable comments. I am solely responsible for the remaining errors and the opinions expressed in this paper do not necessarily reflect those of the Federal Reserve Board or the Federal Reserve System. Address: Division of Monetary Affairs, Board of Governors of the Federal Reserve System. E-mail: stefania.d'amico@frb.gov

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]

This paper proposes a method to quantify the plausibility of alternative probabilistic models and to combine them in a unique weighted predictive distribution, where the weights are function of the uncertainty about the correct model.

The following three basic observations motivate this analysis. First, even though econometric models are implemented in order to deal with uncertainty and guide decisions, very often they are developed without any reference to the "uncertainty about the model." Second, even when model uncertainty is acknowledged and a set of finely parameterized models is considered, a typical implicit assumption is that this set contains the true model. Third, although the approximating nature of a simple model is recognized, the information contained in the approximation error is rarely exploited.

In contrast, in this study, I investigate the problem of density prediction allowing for model ambiguity. Instead of specifying a unique statistical structure and treating it as the true model,

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

I consider a finite set of competing models not necessarily including the correct model. Thus, 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. This implies the presence of an approximation error whose information content can be exploited to combine models.

I develop a method of prediction that ranks different probabilistic models according to the sum of their similarities to past observations. The similarity is measured by the opposite of the distance, that is the Kullback-Leibler Information (KI), between the candidate model and the reference model that is approximated by a nonparametric density. The final weights used to combine models are a function of these distances which embody the uncertainty about the correct structure.

This modeling approach will permit one to study and exploit model misspecification which is defined as the discrepancy between the candidate and the actual model and is measured by the KI. Since the KI is given by the sum of the estimation and approximation errors and since the weights are function of the KI, through the models' weights, we are able to account for both errors and to extract information from a nonparametric estimate.

To implement this methodology, the paper shows that the Kullback-Leibler Information between the nonparametric fit and the parametric candidate model is asymptotically normally distributed with mean given by the model's approximation error.[2] This result leads to determining the weights in the model combination using the cumulative distribution function of a Normal centered on the average performance of all plausible models. As such, the final weight is determined by the ability

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

of a given model to provide a realization of misspecification that is lower than the average.

An important advantage of this method is that it increases the model's flexibility without compromising its parsimony. Because often, tightly parameterized models give better out-of-sample performance, parsimony is a desirable characteristic. As a result, the set of competing models consists of simple parametric alternatives, even when an infinite-dimensional approximation is available.[3] This increases the likelihood that the true model does not belong to the set of candidates and that more than one model can perform fairly well, such that it can be hard to distinguish among them. Under these circumstances, the model combination could provide a better hedge against the lack of knowledge of the correct structure and outperform both in-sample and out-of-sample each of the competing models. This is because the model combination, providing an explicit representation of uncertainty across models, gathers information from 'all' plausible ones. That is, model combination can be viewed as a device to increase the flexibility of the estimation procedure. Furthermore, if the weights in the model combination are not estimated as free parameters but are determined by the ignorance about the true structure, this extra flexibility does not imply the estimation of a higher number of parameters. This translates in a lower risk of overparameterization and in a potentially more robust out-of-sample performance.

I apply the proposed method to determine the predictive density of daily stock returns under different phases of the business cycle. This empirical application is motivated both by the difficulty in estimating the probability law of asset returns which usually are modelled with a misspecified

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

density function, and by the large availability of data for financial series which facilitates the use of nonparametric techniques. I find that the model combination outperforms in-sample and out-of-sample each candidate model including the single best minimizer. The results also indicate that in the small out-of-sample exercise, the model combination performs slightly better than the nonparametric density and than the mixture of models where the weights are estimated as free parameters. Furthermore, in the larger out-of-sample exercise its performance is only marginally worse than the last mentioned models that can be regarded as more complex alternatives.

This way of implementing probabilistic prediction is important 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). A subsequent work by Sin and White (1996) uses information criteria for selecting misspecified parametric models. Nevertheless, none of these studies makes use of a preliminary nonparametric estimation to distinguish among alternative models. Furthermore and more importantly, none of these papers focuses on model combination.

In the context of model combination, there are two main 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). Unlike the Bayesian approach, in this study it is not necessary to assume that the true structure belongs to the set of candidate models. Further, this selection and combination procedure

is based on the idea that 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 competing models. For this reason a pilot nonparametric density, summarizing all information contained in the data, is used to guide the estimation. Finally, this methodology, 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. This requires a limited amount of hypothetical reasoning since it relies directly on data that are available to any observer without ambiguity. The cognitive plausibility of my methodology is founded on case-based decision theory (CBDT). In particular the behavioral axioms of Inductive Inference developed by Gilboa and Schmeilder (2001) provide support for my prediction method[4].

The second vein, though characterized by a completely different approach, represents the studies about forecast evaluation and combination: Diebold and Lopez (1996), Hendry and Clements (2001) and Giacomini (2003) among others. Finally, there is a third strand partially related to this work. It 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 illustrates the model combination technique; Section III analyzes the asymptotic distribution of the uncertainty measure; Section IV contains

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

the empirical application to stock returns; and Section V concludes. Analytical proofs and technical issues are discussed in the Appendix.

2 Description of the selection and combination method

2.1 Model selection

I consider a prediction problem for which a finite set of parametric candidate models is given: $\mathcal{M} \equiv \{f_j(x,\theta), j = 1, ..., J\}_{\theta \in \Theta}$. The goal of the predictor is to rank these models and 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 $\Theta = \left\{\theta_{f_j} : f_j(x,\theta) \in \mathcal{M}\right\}$, and $\Theta \subset \mathcal{R}^d$.

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 $\widehat{f_n}\,(x)$ for each sample $q = 1, ...Q$. Subsequently, from the set Ω, I derive the set of past cases $\mathcal{C} = \left\{ \widehat{f_{nq}}(x) : x \in \Omega \right\}$, which is the final information that the predictor posses to judge the different models. 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 $\widehat{f_{nq}}(x)$ and parameter θ_{f_j}, representing the support that this case lends to the model $f_j(x,\theta)$ 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 nonparametric density $\widehat{f_{nq}}(x)$ and the model $f_j(x,\theta)$:

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

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

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

where i is the index for all observations contained in a sample q.

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:

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

or equivalently:

$$\sum_{q\in C} KI\left(\widehat{f_{nq}}(x), f_1(x,\theta)\right) < \sum_{q\in C} KI\left(\widehat{f_{nq}}(x), f_2(x,\theta)\right). \tag{4}$$

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 as described in D'Amico (2003a), that is:

$$\max_{\theta_{f_j}} \sum_{q \in C} w_{qj} = \min_{\theta_{f_j}} \sum_{q \in C} KI\left(\widehat{f_{nq}}(x), f_j(x,\theta)\right). \tag{5}$$

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

$$f_1 \succ f_2 \text{ IFF } \min_{\theta_{f_1} \in \Theta} \sum_{q \in C} KI\left(\widehat{f_{nq}}(x), f_1(x,\theta)\right) < \min_{\theta_{f_2} \in \Theta} \sum_{q \in C} KI\left(\widehat{f_{nq}}(x), f_2(x,\theta)\right), \tag{6}$$

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

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

2.2 Model Combination

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 $\widehat{KI}\left(\widehat{f_n}(x), f_j(x,\widehat{\theta})\right)$.

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 $\widehat{\theta}_{f_j}$, 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 by combining the models in the predictive distribution $M(\widehat{\theta}_{f_j}) = \sum_j p_j(\widehat{KI}) f_j(x, \widehat{\theta})$. The similarity-weight $p_j(\widehat{KI})$ can be loosely interpreted as the probability of model f_j being correct. In contrast, if the predictor selected a single distribution f_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 $M(\widehat{\theta}_{f_j})$, 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 to be the space of functions to which the true unknown model $g(x)$ belongs: by assumption $g(x)$ minimizes the KI over G. $F_{\Theta_{f_j}} \subseteq G$ 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 $f_j(x, \theta^*)$ in $F_{\Theta_{f_j}}$ to the function $g(x)$ is the p.d.f. that minimizes the KI over $F_{\Theta_{f_j}}$, while $f_j(x, \widehat{\theta}) \in F_{\Theta_{f_j}}$ minimizes the sample version of the KI. The distance between $f_j(x, \widehat{\theta})$ and $f_j(x, \theta^*)$ represents the estimation error that vanishes as $n \to \infty$. Instead, the approximation error[5] given by the distance between $f_j(x, \theta^*)$ and $g(x)$, can be reduced only if the dimension of $F_{\Theta_{f_j}}$ grows with the sample size. Model combination can therefore be considered as a method to increase the dimension of the parameter space accounting

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

for the approximation error.

Only if $F_{\Theta_{f_j}} \equiv G$, then $g(x) = f_j(x, \theta_0) = f_j(x, \theta^*)$ and $\widehat{\theta}$ is a consistent estimator of the true parameter θ_0. Typically, because of the advantages[6] offered by parsimonious models, $F_{\Theta_{f_j}}$ 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 f_j, and as such it can not be ignored.

Once it is decided to use the combinations of p.d.f. $M(\widehat{\theta}_{f_j})$ as predictive density, the main task consists in determining the probability $p_j(\widehat{KI})$. For this purpose, I show that (see the next section and the Appendix for more details) $\widehat{KI}_j$ minus a correction term ($m_n \cong dist(f_j(\theta^*), g)$), mainly due to the approximation error, is asymptotically distributed Normal $N(0, \sigma^2)$, 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 a misspecification $\widehat{KI}_j$ worse than the one actually obtained (ki). That is:

$$p_j(\widehat{KI}) = 1 - P(\widehat{KI}_j \leq ki). \tag{8}$$

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 $p_j(\widehat{KI})$

[6] Closed form solution, ease of interpretation and low computational costs.

of having a good model will be very high.

It is clear that to determine the weight it is just sufficient to compute the cumulative distribution function of a Normal with mean m_n and variance σ^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.

3 Asymptotic results

3.1 Assumptions

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

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: Given the set $\mathcal{M}$, 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^d, the family of distributions $F(\theta)$ has density $f(\theta, x)$ which are measurable in x for every $\theta \in \Theta$ and continuous in θ for every $x \in \Omega$; $E_g[\log g(x) - \log f(\theta, x)]$ exists and has a unique minimum at an interior point $\theta^*$of Θ; $\log f(\theta, x)$ is bounded by

a function $b(x)$ for all $\theta \in \Theta$, where $b(x)$ is integrable w.r.t. the true distribution G.

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

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 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)).

3.2 Asymptotic distribution of KI: heuristic approach

In order to obtain the weights in the models combination, as indicated by the formula (8), 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 $\widehat{KI}_j$ affect the mean and variance of the asymptotic distribution.

To simplify the notation I drop the index j and I 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:

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

$$= \int_x (\ln \widehat{f_n} - \ln g) d\widehat{F}_n - \int_x (\ln f_{\widehat{\theta}} - \ln g) d\widehat{F}_n = \widehat{KI}_1 - \widehat{KI}_2, \tag{9}$$

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[7]:

$$\widehat{KI}_1 \simeq \int_x \left(\frac{\widehat{f_n} - g}{g} \right) d\widehat{F}_n - \frac{1}{2} \int_x \left(\frac{\widehat{f_n} - g}{g} \right)^2 d\widehat{F}_n = \widehat{KI}_{11} - \frac{1}{2} \widehat{KI}_{12}, \tag{10}$$

where $\widehat{KI}_{11}$ is a stochastic element that will affect the asymptotic distribution of $\widehat{KI}$, while $\widehat{KI}_{12}$ is roughly[8] 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:

$$\widehat{KI}_2 = \int_x (\ln f_{\widehat{\theta}} - \ln f_{\theta^*}) d\widehat{F}_n + \int_x (\ln f_{\theta^*} - \ln g(x)) d\widehat{F}_n = \widehat{KI}_{21} + \widehat{KI}_{22}, \tag{11}$$

where $f_{\theta^*} = f_j(x/s, \theta^*)$.

[7] 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 $\ln(1+\gamma) \simeq \gamma - \frac{1}{2}\gamma^2$.

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

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:

$$d_n\widehat{KI}_{21} \longrightarrow^p 0. \tag{12}$$

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

$$\widehat{KI}_{22} \longrightarrow^p E_g\left[\ln f_{\theta^*} - \ln g(x)\right] = (-KI(g, f_{\theta^*})) \leq 0, \tag{13}$$

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:

$$\textbf{A6}\text{: } KI_{22} \sim \alpha h^{1/2} KI_{12}. \tag{14}$$

A6 requires that the mean of the approximation error is proportional to a quantity (KI_{12}) 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 $\beta > \frac{1}{5}$, $\widehat{KI}_{12} \sim C(nh)^{-1}$, then we obtain that:

$$d_n\widehat{KI}_{22} \longrightarrow^p \alpha C, \tag{15}$$

where C is a known positive constant. This assumption can be interpreted as a local misspecification, where the resulting local convergence rate is chosen such that it cancel out with the rate at which the misspecification would converge to infinity.

Thus collecting all terms together:

$$\widehat{KI} \simeq \widehat{KI}_{11} - \frac{1}{2}\widehat{KI}_{12} - \left(\widehat{KI}_{21} + \widehat{KI}_{22}\right), \tag{16}$$

we have the next theorem:

THEOREM 1: *Given assumptions A1-A6, and given that* $nh^5 \longrightarrow 0$ *as* $n \longrightarrow \infty$, *then*

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

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:

$$nh^{1/2}\left(\widehat{KI} + \frac{1}{2}\widehat{KI}_{12}\right) \sim^{A} N(m, \sigma^2) \tag{17}$$

where $m = \alpha C \simeq KI(g, f_{\theta^*})$, from (13) and (15). 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 : $f_j(\theta, x)$ *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:

$$\widehat{\alpha} = \frac{1}{J}\sum_{j} \widehat{KI}_j / C \simeq \overline{KI}(g, f_{\theta^*})/C, \tag{18}$$

where $E(\widehat{KI}_j)$ can be considered an approximation of the average specification error $\left(\overline{KI}(g, f_{\theta^*})\right)$ that can not be observed.

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 a large 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.

4 Application to stock returns

A common assumption to many models in finance, such as the capital asset pricing model (CAPM), the arbitrage pricing theory (APT) and the Black and Scholes option pricing theory, is that of normally distributed returns. The problem is that very often this assumption is not supported by empirical evidence. Financial asset returns posses distributions characterized by a sharp peak around zero, by tails heavier than those of the normal distribution and by a certain degree of asymmetry.

As early as 1963, Mandelbrot (1963) strongly rejected normality as a distributional model for asset returns and a subsequent work by Fama (1965) further corroborated such evidence. These studies give rise to a new probabilistic foundation for financial assets that was based on the Stable Paretian Distribution, which generalizes the Gaussian distribution and allows for heavy tails and

skewness. However, this kind of distributions had little success in practice, since they are characterized by infinite variance which is inappropriate for real data and further very often there is not a closed form expression for the density.

Given the importance of the subject, more recently many economists and statisticians have focused their attention on tests and models to describe the distribution of asset returns[9]. First, as reported by Campbell-Lo-Mackinlay (1997)[10], the skewness for daily US stock returns tend to be negative for stock indexes 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[11] (1999) using non-parametric estimation of monthly stock returns' density from 1834 to 1925. In their analysis it is clearly shown that the density departs significantly from a normal, 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.

4.1 A Set of simple models

I now apply the described prediction method to determine stock returns predictive density, that subsequently can be used to determine the optimal share to invest in the risky asset. Given the previous facts, let me assume that the set of candidate models for the risky asset's returns consists

[9] See for example, the Handbook of Heavy Tailed Distributions in Finance (2003), for a complete analysis of studies about modeling the distribution of several financial assets.

[10] The Econometrics of Financial Markets, 1997, pag. 16 and 17.

[11] Nonparametric Econometrics, 1999, pag 71-74.

of three distributions: a Normal ($N(\mu, \sigma^2)$), a Fisher-Tippet[12] ($F(\alpha, \beta)$) and a mixture of general Gamma ($G(\varsigma, \lambda)$).

The first model, derives from the 'convenient' version of random walk hypothesis. 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. The second model is suggested by the empirical evidence reported in the previous paragraph 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 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, let us define the variable:

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

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

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 $f_j(\cdot)$, $j = 1, 2$. 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:

$$3)\ f_1(X_{1t}; \varsigma, \lambda) \equiv \frac{\lambda^{\varsigma}}{\Gamma(\varsigma)}(X_{1t} - \gamma)^{\varsigma - 1} \exp(-\lambda(X_{1t} - \gamma))$$

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

4.2 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.

	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[13].

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 includes all contractions and all expansions respectively. Merging together all the recessions I obtain a sample of 1321 observations, while combining all expansions I 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 II

It is evident from Table I and II, 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

[13] 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.

returns. Further, 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.

4.3 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).

I then use the Kullback-Leibler entropy to measure the distance between the estimated non-parametric 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*
KI	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*
KI	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
KI	0.0468	0.0666	0.0776

*All estimates are significant at 1% level

Table III

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 $M(\widehat{\theta}_{f_j})$ 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 IV: Realized loss for each model

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 V:Optimal weight for each model

As it can be noticed these values represent 'probabilities' before normalization since they do not

sum up to unity. Results contained in table V 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 in order to conform the application to assumption A6. Thus, in the next section I present the results obtained combining only the Normal and the double Gamma distributions.

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

Lets first consider the in-sample performance of model combination. The results are summarized in the following table, where the values of KI for each single model are reported.

	All data	Expansion	Contraction
$w_g^{ki}G + w_n^{ki}N$	0.0256	0.0179	0.0137
G	0.0468	0.0666	0.0776
N	0.1897	0.1587	0.0513

Table VI: In-sample Results

Note: w_j^{ki} indicates the weight for model j obtained as function of KI

Using the entire dataset from December 1, 1969 to October 31, 2001- 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 the nonparametric KI and of the model combination out-of-sample, the previous results are analyzed 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

Table VII

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, I evaluate the KI distance between the nonparametric density estimated in the new sample ($\widehat{f}_{nOUT}$) and the parametric models estimated in the previous sample. I obtain the following results: the KI between the model combination

$(0.5243G_{IN} + 0.4757N_{IN})$ and $\widehat{f}_{nOUT}$ is equal to 0.7639, between the Gamma distribution and $\widehat{f}_{nOUT}$ is equal to 0.7749 and between the Normal and $\widehat{f}_{nOUT}$ 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.

Models	Expansion 2001-03
$w_g^{ki}G + w_n^{ki}N$	0.7639
G	0.7749
N	0.9235
$w_gG + w_nN$	0.8194
$\widehat{f}_{nIN}$	0.7927

Table VIII: Out-of-sample Results

Note: w_j^{ki} and w_j indicate the weight for model j obtained as function of KI and as free parameter respectively.

Another important comparison to carry out is the following. If the mixture of the Normal and Gamma distributions is estimated in-sample, where the weights are estimated as free parameters, how does this mixture perform with respect to the model combination, where the weights are a function of model misspecification? The mixture that minimizes the distance from the nonparametric density estimated in-sample relative to expansion is equal to $0.4863N(0.0006, 0.0067^2) + 0.5137G(1.0518, 127.996)$ and it delivers a KI equal to 0.0037, which is the smallest value obtained so far.

However, the out-of-sample fit of this mixture is worse than the fit obtained by model combination, since its distance from the nonparametric density estimated out-of-sample equals 0.8194. Hence, while increasing the number of parameters leads to better in-sample fit, it gives less good

out-of-sample results. On the contrary, when the weights are not unrestricted parameters, but are function of model misspecification, the out-of-sample fit seems to be more robust. This result regarding the not excellent out-of-sample performance of models that involve the estimation of a large number of unrestricted parameters is not uncommon (see for example Stock and Watson (1999), J.H. Wright (2003) and Cogley, Morozov and Sargent (2003)), even though there is not a definitive explanation for it.

To stress further this last point, I also control the out-of-sample performance of the nonparametric density estimated in-sample. The reason for this check should be clear if we think about the nonparametric density as an infinite-dimensional parametric alternative. As such, in-sample it represents the benchmark model, but what about its charcterization of the data out-of-sample? The answer is in line with the observation that highly parametrized models do not necessarily perform well out-of-sample. In fact, as shown in Table VIII the KI between the nonparametric fit obtained in-sample and the nonparametric fit out-of-sample is equal to 0.7927, which is somewhat worse than the model combination.

Are all these results further corroborated using a larger out-of sample dataset (i.e. 2506 observation rather than 479)? To verify the stability of the results I have redone the estimation using as in-sample data the stock return during all the expansions included from February 1961 to June 1990, and as out-of-sample data the stock returns from March 1991 to March 2001, which represents the longest expansion period available.

Models	Expansion 1961-90
$w_g^{ki}G + w_n^{ki}N$	0.0983
G	0.1995
N	0.5249
$w_gG + w_nN$	0.0359

Table IX: In-Sample results

In this case, the new double Gamma $G(1.1139, 436.15)$ achieves a KI equal to 0.1995 receiving a weight of 0.6555, while the Normal $N(0.0002, (0.0028)^2)$ obtains a KI that equals 0.5249, receiving a weight of 0.3445. The Kullback-Leibler distance between the nonparametric density estimate and the model combination equals 0.0983, attaining once more a loss half of the size of the best minimizer. On the other hand, the best mixture in-sample is given by $0.4355N(0.003, 0.0024^2) + 0.5645G(1.05, 349.68)$ and it delivers a KI equal to 0.0359, that is one third of the distance achieved by model combination.

Models	Expansion 1991-01
$w_g^{ki}G + w_n^{ki}N$	0.7786
G	0.8838
N	0.9714
$w_gG + w_nN$	0.7562
$\widehat{f}_{nIN}$	0.7637

Table X: Out-Sample results

The out-of-sample results, on the other hand, confirm only partially the previous findings. It still holds true that the model combination outperforms the best in-sample minimizer: its KI is equal to 0.7786 while the Gamma's KI equals 0.8838. However, the mixture delivers a distance from the nonparametric fit that equals 0.7562 that, in contrast to the previous out-of-sampel results, is marginally better than the KI achieved by the model combination. Further, even the nonparametric

fit attains a KI smaller than that of model combination: 0.7637 versus 0.7786.

These results are not surprising if we think about the large amount of observations in this out-of-sample exercise. Nevertheless, it is striking that a parsimonious model like the model combination does not perform much worse than these richer models. Based on both out-of-sample exercises, it is possible to conclude that the use of model combination, where the weight are function of the uncertainty about the true model, can provide a useful forecast tool.

5 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 information between the nonparametric fit and the parametric fit. In the second step, the information content of the KI 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.

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 KI'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 single model. This result can be 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. However, this explanation is not complete. The mixture of models where the weights are estimated as free parameters, even though is characterized by the same number of parameters does not perform like the model combination. Most likely, the information contained in model misspecification, when embodied in the weights of model combination, can improve the robustness of results to future mistakes.

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.

6 Appendix

6.1 Proof Theorem 1:

KI can be rewritten in the following way:

$$KI = \int_x (\ln \widehat{f_n}(x) - \ln f_{\widehat{\theta}}(x)) d\widehat{F}_n(x) = \int_x (\ln \widehat{f_n}(x) - \ln g(x)) d\widehat{F}_n(x) - \int (\ln f_{\widehat{\theta}}(x) - \ln g(x)) d\widehat{F}_n(x) = KI_1 - KI_2. \tag{19}$$

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 the single 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 $E((\ln \widehat{f_n}(x) - \ln g(x)) = KI_1$. 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 $u = \frac{\widehat{f_n}(x) - g(x)}{g(x)} = \frac{\widehat{f_n}(x)}{g(x)} - 1$ we obtain the following result:

$$\frac{1}{n}\sum_i \ln\left(\frac{\widehat{f_n}(x_i)}{g(x_i)}\right) - \frac{1}{n}\sum_i \left(\frac{\widehat{f_n}(x_i) - g(x_i)}{g(x_i)}\right) + \frac{1}{2n}\sum_i \left(\frac{\widehat{f_n}(x_i) - g(x_i)}{g(x_i)}\right)^2 \leq \sum_i u_i^3. \tag{20}$$

We can drop the absolute value because of Markov's inequality, see proof of Lemma 3.1 in Hong-White

(2000).

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:

$$\widehat{KI}_1 = \widehat{V}_{1n} - \frac{1}{2}\widehat{V}_{2n} + O_p(n^{-\frac{3}{2}}h^{-3}\ln n + h^6). \tag{21}$$

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

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

$$b_n(x_i) = \frac{h^{-1}\int K\left(\frac{x-x_i}{h}\right) g(x)dx - g(x_i)}{g(x_i)}.$$

Then

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

$$= \widehat{V}_{11n} + \widehat{B}_n, \tag{22}$$

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:

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

$$H_{1n}(x_i, x_j) = \frac{1}{2h}\left(\frac{K\left(\frac{x_j - x_i}{h}\right) - \int K\left(\frac{x-x_i}{h}\right) g(x)dx}{g(x_i)} + \frac{K\left(\frac{x_i - x_j}{h}\right) - \int K\left(\frac{x-x_i}{h}\right) g(x)dx}{g(x_i)}\right) \equiv J_n(x_i, x_j) + J_n(x_j, x_i) \tag{23}$$

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

$$\widehat{V}_{11n} = \left\{ \frac{1}{n(n-1)} \sum_i \sum_{j,i \neq j} H_{1n}(x_i, x_j) \right\} \Big/ \left\{ \frac{2E\left[H_{1n}^2(x_i, x_j)\right]}{n^2} \right\} \rightarrow^d N(0,1). \tag{24}$$

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

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

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

$$= \frac{1}{4h} \int K^2(u) du + o\left(\frac{1}{h}\right) = O\left(\frac{1}{h}\right). \tag{25}$$

Similarly we can show that

$$E\left[J_n(x_i, x_j) J_n(x_j, x_i)\right] = \frac{1}{4h} \int K^2(u) du + o\left(\frac{1}{h}\right) = O\left(\frac{1}{h}\right). \tag{26}$$

Then it follows that

$$E\left[H_{1n}^2(x_i, x_j)\right] = E\left[2J_n^2(x_i, x_j) + 2J_n(x_i, x_j) J_n(x_j, x_i)\right] = \frac{1}{h} \int K^2(u) du + o\left(\frac{1}{h}\right) = O\left(\frac{1}{h}\right), \tag{27}$$

and

$$\sigma_{1n}^2 = \frac{2}{n^2 h} \int K^2(u) du + o\left(\frac{1}{h}\right). \tag{28}$$

The second term in (22) is the expected value of a Bias term, that is

$$\widehat{B}_n = \frac{1}{n} \sum_i b_n(x_i) \simeq \frac{h^2}{2} \mu_2 \int g^{(2)}(x) dx + o(h^2), \tag{29}$$

where $g^{(2)}(x)$ is the second derivative of the p.d.f. and $\mu_2 = \int u^2 k(u) du$. Hence $\widehat{B}_n = O_p\left(n^{-1/2} h^2\right)$. Thus, what we obtain is

$$\widehat{V}_{1n} = \widehat{V}_{11n} + \widehat{B}_n \sim \sigma_{1n} N(0,1) + \frac{h^2}{2}\mu_2 \int g^{(2)}(x)dx + o(h^2). \tag{30}$$

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

$$\frac{1}{n}\sum_i \left[\frac{\widehat{f_n}(x_i) - \overline{f}(x_i)}{g(x_i)}\right]^2 + \frac{1}{n}\sum_i \left[\frac{\overline{f}(x_i) - g(x_i)}{g(x_i)}\right]^2 + \frac{2}{n}\sum_i \left(\frac{\widehat{f_n}(x_i) - \overline{f}(x_i)}{g(x_i)}\right)\left(\frac{\overline{f}(x_i) - g(x_i)}{g(x_i)}\right) \tag{31}$$

$$= \widehat{V}_{21n} + \widehat{V}_{22n} + \widehat{V}_{23n}. \tag{32}$$

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

$$= \frac{1}{n(n-1)^2}\sum_i \sum_{j,i\neq j} a_n^2(x_i, x_j) + \frac{2}{n(n-1)}\sum_i \sum_{j\neq i}\sum_{z\neq j} a_n(x_i, x_j)a_n(x_i, x_z). \tag{33}$$

The first term is a variance term and it will affect the mean of the asymptotic distribution. As $n \to \infty$, by Lemma 2 Hall(1984) the first term of $\widehat{V}_{21n}$ is given by:

$$\frac{1}{n(n-1)^2}\sum_i \sum_{j,i\neq j} a_n^2(x_i, x_j) = \sigma_n^2 + O_p(n^{-3/2}h^{-1}), \tag{34}$$

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

The second term equals a twice centered degenerate U-statistic $\widehat{U}_n$, which is of the same order of magnitude of $\widehat{V}_{11n}$ and it also affects the asymptotic distribution of $\widehat{KI}_1$.

$$2\widehat{U}_n = \frac{2}{n(n-1)}\sum_i \sum_{i\neq j} \int a_n(x_j, x)a_n(x_i, x)g(x)dx = \frac{2}{n(n-1)}\sum_i \sum_{i\neq j} H_{2n}(x_i, x_j), \tag{35}$$

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

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

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

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

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

$$= h^{-1}\int\left[\int K(u)K(u+v)du\right]^2 dv + o\left(\frac{1}{h}\right). \tag{36}$$

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

$$\sigma_{2n}^2 \simeq 2n^{-2}h^{-1}\int\left[\int K(u)K(u+v)du\right]^2 dv. \tag{37}$$

Finally we have that

$$\widehat{V}_{21n} \sim \sigma_n^2 + O_p(n^{-3/2}h^{-1}) + \sqrt{2}\sigma_{2n}N(0,1). \tag{38}$$

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

$$\frac{1}{n}\sum_i b_n^2 = \frac{h^4}{4}\mu_2^2\int\frac{\left(g^{(2)}(x)\right)^2}{g(x)}dx + o(h^4). \tag{39}$$

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

$$2\widehat{V}_{23n} = \frac{2}{n}\sum_i\left(\frac{\widehat{f_n}(x_i)-\overline{f}(x_i)}{g(x_i)}\right)\left(\frac{\overline{f}(x_i)-g(x_i)}{g(x_i)}\right) = \frac{2}{n(n-1)}\sum_i H_{3n}(x_i,x_j), \tag{40}$$

similarly to Hall(1984) define

$$H_{3n}(x_i,x_j) = \sum_j a_n(x_i,x_j)b_n(x_i) = \frac{1}{h}\int\left[\frac{K\left(\frac{x-x_i}{h}\right)-\int K\left(\frac{x_j-x_i}{h}\right)g(x_j)dx_j}{g(x_i)}\right]\left(\frac{\overline{f}(x_i)-g(x_i)}{g(x_i)}\right)dx_i\ . \tag{41}$$

Under assumptions A1 and A2 and given that $EH_{3n} = 0$, by Lemma 1 in Hall(1984) we have that $2\widehat{V}_{23n}$ is asymptotically normally distributed with zero mean and variance given by:

$$\sigma_{3n}^2 \simeq 2n^{-1}h^4\mu_2^2\left[\int \frac{\left(g^{(2)}(x_i)\right)^2}{g(x_i)}dx_i - \left(\int \left(g^{(2)}(x_i)\right)dx_i\right)^2\right], \tag{42}$$

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

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

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:

$$\widehat{KI}_1 = \widehat{V}_{11n} + \widehat{B}_n - \frac{1}{2}\left(\widehat{V}_{21n} + \widehat{V}_{22n} + 2\widehat{V}_{23n}\right) \sim \tag{43}$$

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

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. $2\widehat{U}_n$, hence the previous expression can be rewritten as follows:

$$\widehat{KI}_1 \sim (nh^{1/2})^{-1}\sqrt{2}\sigma_1 N_1 - (nh^{1/2})^{-1}\sqrt{2}\sigma_2 N_2 - n^{-1/2}h^2\sqrt{2}\sigma_3 N_3 + \widehat{B}_n - \frac{1}{2}c_n, \tag{44}$$

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

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

$$\text{and } c_n = (nh)^{-1}\int K^2(u)du + \frac{h^4}{4}\mu_2^2\int\left(\frac{g^{(2)}(x)}{g(x)}\right)^2 dx + o(n^{-1}h^{-1} + h^4). \tag{45}$$

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 $nh^5 \to 0$

$$nh^{1/2}(\widehat{KI}_1 + \frac{1}{2}c_n) \to^d \sqrt{2}\sigma_1 N_1 - \sqrt{2}\sigma_2 N_2.$$

Since $aN(0,1) + bN(0,1)$ can be proved to be asymptotically normal $N(0, a^2 + b^2)$, then we have that $nh^{1/2}(\widehat{KI}_1 + \frac{1}{2}c_n) \to \sqrt{2}(\sigma_1 - \sigma_2)N(0,1)$.

Let us 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

$$\widehat{KI}_2 = \frac{1}{n}\sum_{i=1}\left(\log f_{\widehat{\theta}}(x_i) - \log f_{\theta^*}(x_i)\right)\widehat{f}_n(x_i) + \frac{1}{n}\sum_{i=1}\left(\log f_{\theta^*}(x_i) - \log g(x_i)\right)\widehat{f}_n(x_i) = \widehat{KI}_{21} + \widehat{KI}_{22}, \quad (46)$$

that similarly of $\widehat{KI}_1$by the LLN, can be considered a good approximation of $E(\ln f_{\widehat{\theta}}(x) - \ln g(x))$. 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:

$$\widehat{KI}_{21}(f_{\widehat{\theta}}, f_{\theta^*}) \simeq \frac{1}{n}\sum_{i=1}\left(\frac{f_{\widehat{\theta}}(x_i) - f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)}\right)\widehat{f}_n(x_i) - \frac{1}{2n}\sum_{i=1}\left(\frac{f_{\widehat{\theta}}(x_i) - f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)}\right)^2\widehat{f}_n(x_i) = I_{n1} - \frac{1}{2}I_{n2}. \quad (47)$$

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

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

where $\overline{\theta}$ lies between $\widehat{\theta}$ and θ^*.

Thus,

$$I_{n1} = \frac{1}{n}\sum_{i=1}\frac{\widehat{f}_n(x_i)}{f_{\theta^*}(x_i)}\left(\widehat{f}_\theta(x_i) - f_{\theta^*}(x_i)\right) \simeq \tag{48}$$

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

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

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

$$S_{1n}(\widehat{\theta}-\theta^*) + (\widehat{\theta}-\theta^*)'S_{2n}(\widehat{\theta}-\theta^*). \tag{49}$$

It can be noticed that the U-statistic form of S_{1n} is the same as that of U_n defined in theorem 2 D'Amico (2003a)[14]. It follows that $S_{1n} = O_p(\frac{1}{\sqrt{n}})$.

$$E(S_{2n}) = \frac{1}{2n(n-1)}\sum_i\sum_j E\left[\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right)\frac{\partial^2 f_{\overline{\theta}}(x_i)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}\right], \tag{50}$$

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

$$\int\int K(u)\frac{\partial^2 f_{\overline{\theta}}(x_i)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}g(x_i)g(x_i+hu)dx_i du. \tag{51}$$

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

$$\limsup_{n\to\infty} E(S_{2n}) \leq \int \frac{\partial^2 f_{\overline{\theta}}(x_i)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}g^2(x)dx; \tag{52}$$

then

$$E(\|S_{2n}\|) \leq \int\int K(u)\left\|\frac{\partial^2 f_{\overline{\theta}}(x_i)/\partial\theta\partial\theta'}{f_{\theta^*}(x_i)}\right\| g(x_i)g(x_i+hu)dx_i du = O(1)$$

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

[14]The appendix of this paper is available upon request.

$$I_{n1} = O_p(\frac{1}{\sqrt{n}}) * O_p(\frac{1}{\sqrt{n}}) + O_p(\frac{1}{\sqrt{n}}) * O_p(1) * O_p(\frac{1}{\sqrt{n}}) = O_p(\frac{1}{n}). \tag{53}$$

Now we have to consider I_{n2}:

$$I_{n2} = \frac{1}{n}\sum_{i=1}\left(\frac{\widehat{f}_\theta(x_i) - f_{\theta^*}(x_i)}{f_{\theta^*}(x_i)}\right)^2 \widehat{f}_n(x_i) \simeq (\widehat{\theta}-\theta^*)'\frac{1}{n(n-1)}\sum_i\sum_j \frac{1}{h}K\left(\frac{x_j - x_i}{h}\right)\frac{\partial \ln f_{\overline{\theta}}(x_i)}{\partial\theta}\frac{\partial \ln f_{\overline{\theta}}(x_j)}{\partial\theta'}(\widehat{\theta}-\theta^*) \tag{54}$$

$$= (\widehat{\theta} - \theta^*)' S_{3n} (\widehat{\theta} - \theta^*)'. \tag{55}$$

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

$$I_{n2} = O_p\left(\frac{1}{\sqrt{n}}\right) * O_p(1) * O_p\left(\frac{1}{\sqrt{n}}\right) = O_p\left(\frac{1}{n}\right). \tag{56}$$

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

$$(nh^{1/2})\widehat{KI}_{21}(f_{\widehat{\theta}}, f_{\theta^*}) = (nh^{1/2})O_p\left(\frac{1}{n}\right) = O_p(h^{1/2}) \to^p 0. \tag{57}$$

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

$$\widehat{KI}_{22}(f_{\theta^*}, g) \cong \frac{1}{n}\sum_{i=1}^n\left(\frac{f_{\theta^*}(x_i) - g(x_i)}{g(x_i)}\right)\widehat{f}_n(x_i) - \frac{1}{2n}\sum_{i=1}^n\left(\frac{f_{\theta^*}(x_i) - g(x_i)}{g(x_i)}\right)^2\widehat{f}_n(x_i) = J_{n1} - \frac{1}{2}J_{n2}, \tag{58}$$

$$E\left(J_{1n}(f_{\theta^*}, g)\right) = E\left(\int\left(\frac{f_{\theta^*}(x_i) - g(x_i)}{g(x_i)}\right)\widehat{f}_n(x_i)g(x_i)dx_i\right) = \int\int K(u)\left(f_{\theta^*}(x) - g(x)\right)g(x+hu)dxdu. \tag{59}$$

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

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

$$E\left(\|J_{1n}\|\right) \leq \int\int K(u)\left\|f_{\theta^*}(x) - g(x)\right\| g(x+hu)dxdu = O(1)$$

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:

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

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

$$\limsup_{n\to\infty} E\left(J_{2n}(f_{\theta^*}, g)\right) \leq \int \left(f_{\theta^*}(x) - g(x)\right)^2 dx \tag{60}$$

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-A5, if $h \to 0$, $nh \to \infty$, then

$$\widehat{KI}_{22}(f_{\theta^*}, g) \to^p E\left(f_{\theta^*}(x) - g(x)\right) - \frac{1}{2}\int \left(f_{\theta^*}(x) - g(x)\right)^2 dx, \tag{61}$$

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, which implies:

$$\widehat{KI}_{22} \simeq \alpha h^{1/2} c_n \tag{62}$$

Finally we can put all terms together:

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

$$\left[(nh^{1/2})^{-1}\sqrt{2}\sigma_1 N_1 - (nh^{1/2})^{-1}\sqrt{2}\sigma_2 N_2 - \frac{1}{2}c_n\right] - \left[\widehat{KI}_{21}(f_{\hat{\theta}}, f_{\theta^*}) + \widehat{KI}_{22}(f_{\theta^*}, g)\right], \tag{63}$$

since we showed that

$$(nh^{1/2})\widehat{KI}_{21}(f_{\widehat{\theta}}, f_{\theta^*}) \rightarrow^p 0 \tag{64}$$

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

$$(nh^{1/2})\left[(nh^{1/2})^{-1}\sqrt{2}\sigma_1 N_1 - (nh^{1/2})^{-1}\sqrt{2}\sigma_2 N_2 - \frac{1}{2}c_n - \left(J_{n1} - \frac{1}{2}J_{n2}\right)\right]. \tag{65}$$

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

$$(nh^{1/2})\left(\widehat{KI} + \frac{1}{2}c_n\right) \sim \sqrt{2}\sigma_1 N_1 - \sqrt{2}\sigma_2 N_2 + \alpha C \tag{66}$$

then,

$$(nh^{1/2})\left(\widehat{KI} + \frac{1}{2}c_n\right) \rightarrow^d N\left(\alpha C, 2\left(\sigma_1^2 - \sigma_2^2\right)\right). \tag{67}$$

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] Cogley T., S. Morozov and T.J. Sargent, Bayesian Fan Charts for U.K. Inflation: Forecasting and Sources of Uncertainty in an Evolving Monetary System, *Working paper No. 2003/44,* Center for Financial Studies, 2003.

[10] 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.

[11] D'Amico S., Quasi Maximum Likelihood Estimation via a Pilot Nonparametric Estimate, mimeo 2003.

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

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

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

[15] 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.

[16] 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.

[17] Dimitriev, 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.

[18] 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.

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

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

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

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

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

[24] Gilboa I. and D. Schmeidler; Inductive Inference: An Axiomatic Approach, *Econometrica*, January 2003, v. 71, iss. 1, pp. 1-26.

[25] Gilboa I. and D. Schmeidler, A Theory of Case-Based Decisions, Cambridge University Press 2001.

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

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

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

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

[30] 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.

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

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

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

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

[35] 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.

[36] Knox T.A., Analytical Methods for Learning How to Invest when Returns are Uncertain; Manuscript University of Chicago Graduate School of Business, August 2003.

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

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

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

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

[41] Skouras S., Decisionmetrics: A decision-based approach to econometric modelling; Manuscript Santa Fe Institute, November 2001.

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

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

[44] Stock H.J. and M.W. Watson, Forecasting Inlfation, *Journal of Monetary Economics 44 (1999) 293-335.*

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

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

[47] Vapnik V.N., The Nature of Statistical Learning Theory, Springer-Verlag 2000.

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

[49] Wright J.H., Forecasting U.S. Inflation by Bayesian Model Averaging, International Finance Discussion Papers, Number 780, Board of Governors of the Federal Reserve System, September 2003.

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

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