FEDERAL RESERVE BANK *of* ATLANTA **WORKING PAPER SERIES**

General Aggregation of Misspecified Asset Pricing Models

Nikolay Gospodinov and Esfandiar Maasoumi

Working Paper 2017-10
November 2017

Abstract: This paper proposes an entropy-based approach for aggregating information from misspecified asset pricing models. The statistical paradigm is shifted away from parameter estimation of an optimally selected model to stochastic optimization based on a risk function of aggregation across models. The proposed method relaxes the perfect substitutability of the candidate models, which is implicitly embedded in the linear pooling procedures, and ensures that the aggregation weights are selected with a proper (Hellinger) distance measure that satisfies the triangle inequality. The empirical results illustrate the robustness and the pricing ability of the aggregation approach to stochastic discount factor models.

JEL classification: C13, C52, G12

Key words: entropy, model aggregation, asset pricing, misspecified models, oracle inequality, Hellinger distance

---

The authors thank Mark Fisher, Ruixuan Liu, and the participants at the Max King Conference (Monash University) and the Econometrics Workshop (University of Kansas) for useful discussions and suggestions. The views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Please address questions regarding content to Nikolay Gospodinov, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street NE, Atlanta, GA 30309-4470, 404-498-7892, nikolay.gospodinov@atl.frb.org, or Esfandiar Maasoumi, Emory University, Department of Economics, Rich Memorial Building 324, 1602 Fishburne Drive, Atlanta, GA 30322-2240, esfandiar.maasoumi@emory.edu.

1 Introduction

Stochastic discount factor (SDF) models are routinely rejected when confronted with data. We examine certain aggregations of these models when all are assumed to be misspecified and the true SDF process is not included in the choice set. To be sure, all models are misspecified by design as they are constructed to be simple approximations to a complex 'Data Generation Process' (DGP). The DGP is a 'latent object', and models of it are simplified/directed/partial maps. This is especially true when these models are incompletely specified and are estimated by moment matching. Despite the obvious nature of the statement above, its accommodation in practice remains inconsistent and even contradictory in many instances. We argue that the traditional inference objectives require a more careful consideration when all models are expressly allowed to be misspecified.

The analysis of misspecified moment condition models is still in its infancy. This is a fertile ground for important future research; see Lars Hansen's Nobel lecture, Hansen (2013). When there are several candidate models, their respective 'pseudo-true' objects that may allow a misspecification-consistent analysis, are relative objects, specific to each model and even to the estimation criteria that quantify them (GMM, Kullback-Leibler, Likelihood). Model selection and model averaging, and certainly policy analysis, do not have clearly defined objectives in this setting.

Partial effects, for instance, would refer to different conditional distributions and parameters, as provided by each model. This problem is only partially mitigated in some situations, as in the context of comparing misspecified asset pricing models using the Hansen-Jagannathan (Hansen and Jagannathan, 1991, 1997) distance that uses the inverse of the same second moment matrix of the test assets to weigh the pricing errors for all candidate models. But there is a larger problem here that is inherent to 'model selection' which is designed to choose only one of the candidate models and ignores the information in the remaining models. Model selection may be meaningful only if the 'true' DGP model were in the set of candidate models (*the dictionary*) and the procedure is consistent. This is a highly unrealistic situation as all models are misspecified. Indeed, 'consistency' in selection seems dubious when the true DGP is not included. A better alternative that has been favored for informal reasons, and has recently received further theoretical justification is "aggregation" which includes averaging and pooling.

Bernando and Smith (1994) offer a characterization and a taxonomy of the different views regarding model comparison and selection. The first perspective, that includes Bayesian model

averaging and frequentist model selection, is conditioned on one of the models being 'true'. In this approach, the ambiguity about the true model is resolved asymptotically and in the limit, the mixture that summarizes the beliefs about the individual models assigns a weight of one to one of the models. Diebold (1991) provides an illuminating example of this in the context of Bayesian forecast combination. Another possibility is also to assume that a 'true model/DGP' exists but is too complicated or cumbersome to implement, and all of the candidate models are viewed as approximations and hence misspecified. The third view dispenses completely with the self-contradictory notion of a 'true model' and treats the candidate models as genuinely misspecified either because they are believed to represent different aspects of the underlying DGP or because the underlying structure is completely unknown. "If models are misspecified in an indeterminate manner, then we should not be aiming at the discovery of the 'true data generating process'" (Maasoumi, 1993). Reasonable models may be statistically consistent with aspects of the data emanating from the latent DGP.

Earlier attempts to accommodate misspecified models in econometrics date back to the mid and late 70s.[1] These attempts stayed with the dominant statistical paradigm, then and now, of inference on parameters and the risk of decision making and forecasting, driven by parameter estimation uncertainty. A very important recent strand of the literature in mathematical sciences and engineering places risk of model choice at the center of statistical inquiry. This proves to be much more appropriate and productive when all candidate models are misspecified and we seek to aggregate over them. We give a brief overview of this approach in the next section as it is equally adept at handling searches for best aggregative densities, regression functions, and other similar objects.[2]

In this paper, we take the view that the DGP/'true model' is not known to be among the competing models. This is similar in spirit to Geweke and Amisano (2011, 2012) for prediction pooling of misspecified models. We develop a generalized entropy-based approach to mixing information from different models. The minimum Shannon entropy or Kullback-Leibler (KL) information criterion used by Geweke and Amisano (2011, 2012) and Hall and Mitchell (2007) is a special case of this framework. In this paper, our generalization is facilitated by the fact that we are not mixing densities so that the combination does not need to commute with any possible marginalization of the distributions involved (McConway, 1981; Genest, Weerahandi and Zidek, 1984). More importantly,

[1] To exemplify, see Maasoumi (1977, 1978, 1990).

[2] An early example of thinking of unknown functions as an aggregation problem is Maasoumi (1987).

unlike Geweke and Amisano (2011), we choose a divergence measure for selecting the mixture weights which is a proper measure of distance since it is symmetric and it satisfies the triangle inequality. Generalized entropy also allows us to relax the perfect substitutability of the candidate models which is implicitly embedded in the linear pooling procedures.[3] Our closeness measure is also useful for clustering subsets of models which might be particularly useful and informative if the set of candidate models is large. The model clustering will identify similar attributes across models and act effectively as a dimension reduction device by reducing the set of information-enhancing models. This is a 'big data' problem and we will briefly allude to penalization methods that are similar in spirit.

The SDF framework for asset pricing provides an arguably perfect laboratory studying the problem of model aggregation. It is widely documented that most, if not all, asset pricing models of equity returns are strongly rejected by the data.[4] Despite this evidence of misspecification, these asset pricing models can still collectively provide a useful guide for investment decisions or measuring investment performance. Gospodinov, Kan and Robotti (2013) propose a general methodology for model comparison and ranking of competing, possibly misspecified, asset pricing models that are estimated and evaluated using the Hansen-Jagannathan distance. Stutzer (1995) considers an information-theoretic approach to diagnosing asset pricing models. In a recent paper, Ghosh, Julliard and Taylor (2017) develop an entropy-based modification of the SDF that may price assets correctly. Unlike these papers, we use the generalized entropy measures of divergence to combine information from a set of misspecified models and elicit some features of the SDF. The latter is our 'latent' object or process.

Our contributions can be summarized as follows. On methodological side, we propose an information-theoretic approach to aggregating information in misspecified asset pricing models. The optimal aggregator takes a harmonic mean form with geometric and linear weighting schemes as special cases. The generalized entropy criterion that underlies our approach allows us to circumvent two serious drawbacks of the standard linear pooling. First, it ensures that the divergence measure between the densities of the pricing errors of candidate models is a proper distance measure that is positive, symmetric and satisfies the triangular inequality (Maasoumi, 1993). Second,

[3] Linear aggregation is dominant in the stochastic optimization literature and elsewhere.

[4] It is possible that the null of correct specification is not rejected even when the model is misspecified due to a failure of the rank condition. Gospodinov, Kan and Robotti (2016) show that the power of invariant tests for overidentifying restrictions in linear asset pricing models does not exceed the nominal size when the rank condition is violated.

the use of the harmonic mean as an aggregator relaxes the infinite substitutability assumption between models which is implicit in linear aggregation. On the practical side, our mixing procedure employs information from all models by assigning weights depending on the model's contribution to the overall reduction of the pricing errors. The weighted stochastic discount factor preserves the integrity of each structural model and pools the relevant information from each model in a bounded risk sense. This stands in sharp contrast with the existing methods in the literature that either select factors from a set of candidate factors or choose a single ('least misspecified') model from a set of candidate models. Both of these cases result in loss of information from omitting factors or models. Our empirical analysis reports non-trivial improvements (in terms of pricing error reduction) from aggregation.

Ultimately, the reason why so many studies find that almost all kinds of pooling and mixing methods 'perform well' can be readily gleaned from the classical results in a standard linear regression. Constraints (such as omitted components), even false constraints, are variance (uncertainty) reducing, with a cost on correct centering (bias). But the latter has an uncertain characterization when the true DGP/model is not known. Stochastic optimization techniques, paired with information criteria as suitable risk measures, reflect more deeply this phenomenon.

The rest of the paper proceeds as follows. Section 2 introduces the stochastic optimization paradigm. Section 3 discusses the main setup for evaluating asset pricing models/SDFs and introduces our ideal aggregate functions as well as the stochastic, risk-based approach to model aggregation. Section 4 describes the candidate consumption-based asset pricing models and presents the empirical results. Section 5 concludes.

2 Stochastic Optimization as a General Paradigm

2.1 Some Preliminaries

Consider the case where one is interested in estimating a functional $f(\cdot)$. If the true form of this functional is 'unknowable', estimation and inference would appear infeasible. However, one could infuse information from a set of auxiliary (partially specified) models that could elicit some aspects of the functional $f(\cdot)$ with bounded risk involving oracle inequalities. Examples of $f(\cdot)$ include conditional mean functions in regression models, densities, and other latent objects such as stochastic discount factors (SDFs). It becomes convenient, possibly inevitable, to shift the statistical paradigm away from optimal (parameter) estimation to a 'stochastic optimization' paradigm that

is detailed below.

Suppose there is a finite list (dictionary) $\mathcal{F}$ of candidate auxiliary models that embed certain theoretical or empirical features of the underlying DGP. The aggregation/stochastic optimization approach we adopt does not require a fully articulated structural model and does not assume that this dictionary contains a 'true' model. All models contained in the dictionary are statistical approximations. The proposed method will construct an aggregator that mimics (in terms of some data-dependent metric) the performance of the best (or least misspecified) model in the class. The aggregation estimator minimizes an empirical risk function that satisfies certain oracle inequalities (Rigollet, 2012; Rigollet and Tsybakov, 2012). Model selection, that assigns weights of one or zero to individual models, proves to be suboptimal. When the dictionary contains a mixture of linear and nonlinear, possibly non-nested, auxiliary models, this aggregation scheme arrives at a 'comprehensive' model. The aggregation provides an approximate mapping between the comprehensive and auxiliary models but this mapping, unlike in the standard case of a fully specified structural model, is perturbed by a component that reflects uncertainty about the underlying object $f(\cdot)$.

For simplicity, we introduce the main ideas and notation in the context of probability density functions but they can be easily adapted to more general functions of fixed mass. Let $Z_1, ..., Z_T$ denote observations of the random variable Z with an unknown density $f \in \mathcal{F}$, and $L : Z \times \mathcal{F} \rightarrow \mathbb{R}$ be a measurable loss function with a corresponding risk function $\mathcal{R} : \mathcal{F} \rightarrow \mathbb{R}$ defined as

$$\mathcal{R}(f_Z, f) = E[L(f_Z, f)], \ f \in \mathcal{F}, \tag{1}$$

where f_Z denotes any candidate distribution for Z. The oracle f^* is defined as

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathcal{R}(f_Z, f) \tag{2}$$

or $\mathcal{R}(f_Z, f^*) \leq \mathcal{R}(f_Z, f)$ for all $f \in \mathcal{F}$. Let

$$\mathcal{R}_T(f_T, f) = \frac{1}{T} \sum\nolimits_{t=1}^{T} L(f_t, f) \tag{3}$$

be the empirical version of the risk function $\mathcal{R}(f_Z, f)$, where f_T is the sample analog of f_Z. In the case of quadratic risk, for example, it takes the form $\mathcal{R}_T(f_T, f) = \|f_T - f\|_2 = \frac{1}{T} \sum\nolimits_{t=1}^{T} (f_t - f)^2$, where $\|\cdot\|_2$ denotes the L_2 norm.

When the interest lies in density or model aggregation, the stochastic optimization problem constructs a sample aggregator $\tilde{f}_T$ of available functions $f_1, ..., f_M$ in the $\mathcal{F}$ dictionary by mimicking

the oracle $\inf_{f\in\mathcal{F}}\mathcal{R}(f_Z,f)$. The functions $f_1,...,f_M$ are either given or obtained from a prior training sample (by sample splitting, for example). In the construction of the aggregator, these functions are evaluated at the sample values $Z_1,...,Z_T$. Then, for a constant $C\geq 1$, we have some version of the following "expectations" oracle inequality (Rigollet, 2015)

$$E[\mathcal{R}_T(\tilde{f}_T,f)]\leq C\inf_{f\in\mathcal{F}}\mathcal{R}(f_Z,f)+\triangle_{T,M}, \tag{4}$$

where $\triangle_{T,M}>0$ is a remainder term that characterizes the performance of the aggregator $\tilde{f}_T$.[5] Furthermore, for every $\delta>0$, the following error probability bound is established:

$$\Pr\left\{\mathcal{R}_T(\tilde{f}_T,f)\leq C\inf_{f\in\mathcal{F}}\mathcal{R}(f_Z,f)+\triangle_{T,M,\delta}\right\}\geq 1-\delta. \tag{5}$$

More generally, a balanced oracle inequality takes the form

$$E[\mathcal{R}_T(\tilde{f}_T,f)]\leq C\left[\inf_{f\in\mathcal{F}}\mathcal{R}(f_Z,f)+\tilde{\triangle}_{T,M}(f)\right], \tag{6}$$

with $\triangle_{T,M}=C\sup_{f\in\mathcal{F}}\tilde{\triangle}_{T,M}(f)$. An exact or sharp oracle inequality is obtained when $C=1$.

One popular example focuses on quadratic risk and the regression model

$$Y_t=g(X_t)+\epsilon_t, \tag{7}$$

where ϵ_t is $N(0,\sigma^2)$. $g(\cdot)$ is unknown and is modelled by functions in the dictionary $\mathcal{F}=\{f_1,...,f_M\}$.[6] Consider the linear aggregator

$$\tilde{f}_T^{(w)}=\sum_{i=1}^{M}w_if_i. \tag{8}$$

The bound for the risk $E[\mathcal{R}_T(\tilde{f}_T^{(w)},f)]$, where $\hat{w}$ denotes the least squares estimator of $w=(w_1,...,w_M)'$

$$\hat{w}=\underset{w\in\mathbb{R}^M}{\operatorname{argmin}}\frac{1}{T}\sum_{t=1}^{T}(Y_t-\tilde{f}_T^{(w)}(X_t))^2, \tag{9}$$

is provided in Rigollet and Tsybakov (2012) and Rigollet (2015). A few remarks are warranted here. First, $\inf_w\mathcal{R}(f^{(w)},f)>0$ when the candidate models are misspecified and a 'true' model is not part of the dictionary. Obtaining a sharp oracle inequality ($C=1$) in this case is important since it minimizes the impact of this systematic bias term (Rigollet and Tsybakov, 2012). Alternatively, one could construct adaptive weights by judiciously parameterizing the parameter space of w as

[5] $\triangle_{T,M}$ is free of f and $\tilde{f}_T$ and varies depending on the process and underlying assumptions; often for *iid* samples.
[6] The functions f_i, $i=1,...,M$, are either given or estimated with prior data samples.

a function of the sample size in such a way that this bias vanishes asymptotically. Finally, to minimize the magnitude of the remainder term in bounding the empirical risk, one could resort to penalized convex aggregation as discussed below (see also Rigollet and Tsybakov, 2012). Birgé (2013) shows that in the case of quadratic risk, the remaining term can be quite large and suggests a different way of aggregation based on T-estimators (Birgé, 2006).

Another interesting example is the density function of a variable $Y_t : f(Y_t)$. Suppose we have M density forecast models for the conditional density of $Y_t|X_{it}$ for $i = 1, ..., M$, $f(Y_t|X_{it}) = f_i$. We would like to aggregate the information in the M candidate models to form a density forecast for Y_t. Since we are interested in the unconditional density of Y_t, the aggregation weights should be based on the divergence between $f(Y_t)$ and the unconditional version of $f(Y_t|X_{it})$:

$$g_{it} = E_{X_i}\left\{f(Y_t|X_{it})\right\} = \int f(Y_t|x)dP_{it}(x), \tag{10}$$

where P_{it} is the marginal distribution of X_{it}. If $\hat{P}_{it}$ denotes an estimate of P_{it}, then

$$g^*_{it} = \int f(Y_t|x)d\hat{P}_{it}(x). \tag{11}$$

This can be performed by resampling only the predictors X_{it}, and g^*_{it} is an empirical average of $f(Y_t|X_{it})$ over the X_{it}.

2.2 Convex Aggregation

The distinction between 'model selection' and 'model aggregation' is important. The former has a zero-one weighting scheme that picks the model with smallest risk. This is known to be suboptimal relative to 'model aggregation' in which weights and aggregation penalties are obtained over a set of models in order to optimize a risk measure (Yang, 2000; Rigollet and Tsybakov, 2012).

The approach outlined below offers generality with respect to the risk function $\mathcal{R}(\tilde{f}, f)$. As before, the arguments in this section are developed for probability density functions but can be extended to more general functions. Assumption 1 below states some regularity conditions on the data.

ASSUMPTION 1. *Let (Z, A) be a measurable space and v be a σ-finite measure on (Z, A). Let $(Z_1, ..., Z_T)$ denote a sample of T iid observations from an unknown density f on Z with respect to v. Finally, let $\mathcal{F}$ be a finite dictionary of cardinality M of density functions $\{f_1, ..., f_M\}$ such that $\max_{f_i \in \mathcal{F}} \|f/f_i\|_\infty < \infty$.*

Further, consider the flat simplex for a set of model weights $w = (w_1, ..., w_M)$:

$$\mathcal{W}^M = \left\{ w \in \mathbb{R}^M : w_i \geq 0, \sum\nolimits_{i=1}^{M} w_i = 1 \right\}. \tag{12}$$

Then, the convex (weighted average) aggregator of the candidates $\{f_1, ..., f_M\}$ is given by

$$f^{(w)} = \sum_{i=1}^{M} w_i f_i, \; w \in \mathcal{W}^M, \tag{13}$$

with its estimator denoted by $\tilde{f}_T^{(w)}$. Model selection is a special case with $w \equiv e_i = (0, 0, ..., 1, 0, ..., 0)$ with $i = 1, ..., M$.

Let the pseudo-true density aggregator be defined as

$$f_w^* = \underset{w \in \mathcal{W}^M}{\text{argmin}} E[L(f^{(w)}, f)]. \tag{14}$$

Oracle inequalities are established relative to $\mathcal{R}(f_w^*, f) = E[L(f_w^*, f)]$ both in terms of expectations and probability. The following lemma summarizes these results.

LEMMA 1. *Suppose that Assumption 1 holds. Then, for some $C \geq 1$,*

$$E[\mathcal{R}_T(\tilde{f}_T^{(w)}, f)] \leq C \min_{w \in \mathcal{W}^M} \mathcal{R}(f^{(w)}, f) + \triangle_{T,M} \tag{15}$$

and for every $\delta > 0$,

$$\Pr\left\{ \mathcal{R}_T(\tilde{f}_T^{(w)}, f) \leq C \min_{w \in \mathcal{W}^M} \mathcal{R}(f^{(w)}, f) + \triangle_{T,M,\delta} \right\} \geq 1 - \delta, \tag{16}$$

where $\triangle_{T,M}$ and $\triangle_{T,M,\delta}$ are remainder terms that do not depend on f or f_i, $i = 1, ..., M$.

When the density properties of the w are recognized, one may incorporate penalties for departures of the distribution of weights (w) from a priori distributions or desired distributions of weights (π) that may reflect an ordering of the models. For example, consider the linear aggregator $\tilde{f}_w = \sum_{i=1}^{M} w_i f_i$ of an unknown regression function f. Then, the aggregation weights may solve the following penalized optimization problem

$$\min_{w \in \mathcal{W}^M} \left[\sum_{i=1}^{M} w_i \mathcal{R}_T(\tilde{f}_T^{(w)}, f) + \frac{\beta}{T} \mathcal{KL}(w, \pi) \right], \tag{17}$$

where $\beta > 0$ is a penalty parameter, $\mathcal{KL}(w, \pi) = \sum_{i=1}^{M} w_i \ln\left(\frac{w_i}{\pi_i}\right)$ is the Kullback-Leibler divergence between w and π, and $\pi \in \mathcal{W}^M$ is a prior probability density. This could also be a convenient device

when M is large relative to T, as in variable selection problems with 'big data' attributes. The solution for the above penalized optimization problem is driven by the form of the entropy divergence function. With the Kullback-Leibler divergence, the aggregation weights take an exponential form

$$w_i^* = \frac{\exp(-T\mathcal{R}_T(\tilde{f}_T^{(w)}, f)/\beta)\pi_i}{\sum_{j=1}^M \exp(-T\mathcal{R}_T(\tilde{f}_T^{(w)}, f)/\beta)\pi_j}. \tag{18}$$

Note that this is the quasi-Bayesian approach of Chernozhukov and Hong (2003) where the estimates of w can be obtained using MCMC methods.

Rigollet and Tsybakov (2012) show that the aggregator $\tilde{f}_T^{(w)} = \sum_{i=1}^M w_i^* f_i$ in the regression setup above with $\beta \geq 4\sigma^2$ satisfies the following balanced oracle inequality

$$E[\mathcal{R}_T(\tilde{f}_T^{(w)}, f)] \leq \min_{w \in \mathcal{W}^M} \left[\sum_{i=1}^M w_i \mathcal{R}(f_i, f) + \frac{\beta}{T} \mathcal{KL}(w, \pi) \right]. \tag{19}$$

Furthermore, by restricting $\mathcal{R}(f_i, f)$ to the vertices of the simplex $\mathcal{W}^M$ with the choice of π to be the uniform distribution on $\{1, ..., M\}$ we have the oracle inequality[7]

$$E[\mathcal{R}_T(\tilde{f}_T^{(w)}, f)] \leq \min_{1 \leq i \leq M} \mathcal{R}(f_i) + \frac{\beta \ln(M)}{T}. \tag{20}$$

By contrast, a model selection procedure that selects only one function in the dictionary is suboptimal as its remainder term is of higher order $\sqrt{\ln(M)/T}$ (see Rigollet and Tsybakov, 2012) whereas $\ln(M)/T$ is the desired minimax rate.

2.3 General Aggregation

To infer the form of the aggregator, we follow a general entropy-based approach proposed by Maasoumi (1986) for characterizing the solution for $\tilde{f}$ by selecting a distribution which is as close as possible to the multivariate distribution of f_i's. Maasoumi (1986) shows that generalizing the pairwise criteria of divergence to a general multivariate context results in the following measure of divergence:

$$\tilde{D}_\rho(\tilde{f}, f; w) = \sum_{i=1}^M w_i \mathcal{R}_{T,\rho}(\tilde{f}, f_i), \tag{21}$$

where

$$\mathcal{R}_{T,\rho}(\tilde{f}, f_i) = \frac{1}{\rho(\rho+1)} \sum_{t=1}^T \tilde{f}_t \left[\left(\frac{\tilde{f}_t}{f_{i,t}} \right)^\rho - 1 \right]. \tag{22}$$

[7] Note that the vertices are the selector vectors e_i, $i = 1, ..., M$, introduced above and $\sum_{i=1}^M w_i \mathcal{R}(f_i, \tilde{f}) = \sum_{i=1}^M e_i \mathcal{R}(f_i, f) = \mathcal{R}(f_i)$.

$\mathcal{R}_{T,\rho}(\tilde{f}, f_i)$ is the generalized entropy divergence between the aggregator $\tilde{f}$ and each of the prospective models f_i. The aggregator that minimizes $\tilde{D}_\rho(\tilde{f}, f; w)$ subject to $\sum_{i=1}^{M} w_i = 1$ is given by

$$\tilde{f}_t^* \propto \left[\sum_{i=1}^{M} w_i f_{i,t}^{-\rho}\right]^{-1/\rho}. \tag{23}$$

Note that the linear and convex pooling of models are obtained as special cases. For example, the dominant (convex) aggregator $\tilde{f}_t^{(w)} = \sum_{i=1}^{M} w_i f_{i,t}$ is an ideal aggregator function by the Kullback-Leibler divergence ($\rho = -1$).

What emerges from the literature is quite compelling. First, risk of aggregator functions dominates the model selection approach in terms of oracle bounds on expected losses. Second, the commonly used L_2 risk function has bounds that depend on dominating measure, and risk may be unbounded (see Birgé, 2006, 2013). Finally, quadratic risk is not a distance between distributions as it depends on the particular dominating measure. Hellinger distance is invariant to this and is a measure of distance between distributions and suitable regression functions.

This aspect of distance functions for distributions is emphasized in Maasoumi's (1993) survey of entropy functions and relative entropy functions. Granger, Maasoumi and Racine (2004) advocate a member of the generalized entropy divergence measures (see also Cressie and Read, 1984) which is a scaled normalization of the Hellinger distance. More specifically, let P and Q be probability measures with densities p and q with respect to a dominating measure ν. The generalized entropy or Cressie-Read divergence from Q to P is given by

$$D_\eta(P, Q) = \int \phi_\eta \left(dQ/dP\right) dQ, \tag{24}$$

where

$$\phi_\eta(x) = \frac{1}{\eta(\eta+1)}\left(x^{\eta+1} - 1\right) \tag{25}$$

is the Cressie-Read power divergence family of functions. More specifically,

$$D_\eta(P, Q) = \int \left(1 - \left(\frac{p}{q}\right)^\eta\right) q d\nu \text{ for } \eta \in \mathbb{R}. \tag{26}$$

When $\eta \to 0$, we obtain the Kullback-Leibler divergence measure

$$D_0(P, Q) = \int \ln\left(\frac{p}{q}\right) q d\nu = \mathcal{KL}(P, Q). \tag{27}$$

Similarly, the case $\eta = -1/2$ corresponds to the Hellinger distance measure

$$D_{-1/2}(P, Q) = \int \left(p^{1/2} - q^{1/2}\right)^2 d\nu = \mathcal{H}(P, Q). \tag{28}$$

Unlike the other measures in the Cressie-Read divergence family, the Hellinger distance is a proper measure of distance since it is positive, symmetric and it satisfies the triangle inequality. Kitamura, Otsu, and Evdokimov (2013) show the robustness of the Hellinger distance to perturbations of probability measures.

To fix the notation for what follows, let $\tilde{f}^{(w)} = \left[\sum_{i=1}^{M} w_i f_i^{1/2}\right]^2$ be the aggregator based on the Hellinger distance for the dictionary $\{f_1, ..., f_M\}$ with $\tilde{f}_T^{(w)}$ being its sample analog. Furthermore, $\mathcal{H}(\tilde{f}^{(w)}, f)$ is the corresponding risk function, where $\mathcal{H}$ denotes the Hellinger distance. Finally, let $\inf_{\tilde{f}^{(w)}} \sup_{f \in \mathcal{F}} \mathcal{H}(\tilde{f}^{(w)}, f)$ denote the minimax risk over $\mathcal{F}$. The following result is adapted from Birgé (2006) and provides a justification for our proposed aggregation approach in the rest of the paper.

LEMMA 2. *Suppose that Assumption 1 holds. Then,*

$$E[\mathcal{H}_T(\tilde{f}_T^{(w)}, f)] \leq C \left[\min_{w \in \mathcal{W}^M} \mathcal{H}(\tilde{f}^{(w)}, f) + \triangle_{T,M}\right], \tag{29}$$

where $C \geq 1$ and $\triangle_{T,M}$ is a remainder term. Moreover, the minimax risk over $\mathcal{F}$ is bounded by $C\triangle_{T,M}$.

As mentioned above, $\mathcal{H}(\tilde{f}^{(w)}, f) > 0$ under model misspecification. But with Hellinger distance and minmaxity, the risk remains under control even if the models are misspecified.

The bounds so far are established under the assumption that the data are *iid*. The extensions to the time series context are more involved and can be implemented using the conditional predictive density approach of Yang (2000) or the composite marginal likelihood approach (see Varin, 2008; Varin, Reid and Firth, 2011; among others). While our empirical application uses time series data, the returns and the risk factors are largely serially uncorrelated. Some of the bound results may continue to hold if the independence is replaced by a martingale difference assumption. However, a rigorous treatment of the time series case is left for future research.

3 Aggregation of Misspecified Asset Pricing Models

In the SDF setup considered below, the distance minimization is performed subject to restrictions imposed by the asset pricing model. The primal problem which targets the unknown functional of interest can be conveniently transformed to a dual problem. The immutable part (unknown functional) of the risk function falls out of the dual problem. It is important to stress that while this approach explicitly recognizes that the auxiliary models are misspecified, the "oracle SDF" is

still guided and proscribed by economic theory. An alternative would be a data-driven (model-free) approach to approximating the unknown function using (semi) non-parametric methods (see, for example, Donoho and Johnstone, 1994; Cai, Ren, and Sun, 2015). This approach is better tailored for model fit or prediction (as in machine learning) and will not be considered in this paper. In contrast, our aggregation method can be regarded as formal information nesting (information-theoretic) of various theory-based factor models that would inform policy makers and investors of data based support. Our data dependent model weights, w_i, will rank competing models, if so desired.

3.1 SDF and Hansen-Jagannathan Distance

Let R denote the returns on N test assets and $m \in \mathcal{M}$ be an admissible stochastic discount factor (SDF) that prices the test assets correctly,

$$E[Rm] = q, \tag{30}$$

where q denotes an $N \times 1$ vector of payoffs (a vector of ones if R are gross returns). Furthermore, let $y(\gamma)$ be a candidate stochastic discount factor that depends on a k-vector of unknown parameters $\gamma \in \Gamma$, where Γ is the parameter space of γ. If $y(\gamma)$ prices the N test assets correctly, then the vector of pricing errors, $e(\gamma)$, of the test assets is exactly zero:

$$e(\gamma) = E[Ry(\gamma)] - q = 0_N. \tag{31}$$

However, the pricing errors are nonzero when the asset-pricing model is misspecified. The squared Hansen-Jagannathan (Hansen and Jagannathan, 1991, 1997) distance

$$\delta^2 = \min_{\gamma \in \Gamma} \min_{m \in \mathcal{M}} E[(y(\gamma) - m)^2] \tag{32}$$

provides a misspecification measure of $y(\gamma)$ and can be used for estimating the unknown parameters γ. This is the standard L_2 norm between the functionals $y(\gamma)$ and m. It is sometimes more convenient to solve the following dual problem:

$$\delta^2 = \min_{\gamma \in \Gamma} \max_{\lambda \in \Re^N} E[y(\gamma)^2 - (y(\gamma) - \lambda' R)^2] - 2\lambda' q, \tag{33}$$

where λ is an $N \times 1$ vector of Lagrange multipliers. Note that $\lambda' R$ provides the smallest correction, in mean squared sense, to $y(\gamma)$ in order to make it an admissible SDF. Note that for a given SDF

$y(\gamma)$ and γ, the vector of Lagrange multipliers and the squared Hansen-Jagannathan distance can be expressed as

$$\lambda = U^{-1}e(\gamma), \tag{34}$$

and

$$\delta^2(\gamma) = e(\gamma)'U^{-1}e(\gamma), \tag{35}$$

where $U = E[RR']$.

Importantly, Hansen and Jagannathan (1991) provide a maximum pricing error interpretation of the distance $\delta(\gamma)$. Consider a portfolio a with unit second moment, i.e., $a'Ua = 1$. By the Cauchy-Schwartz inequality, the squared pricing error of this portfolio is

$$(a'e(\gamma))^2 = (a'U^{\frac{1}{2}}U^{-\frac{1}{2}}e(\gamma))^2 \leq (a'Ua)[e(\gamma)'U^{-1}e(\gamma)] = \delta^2(\gamma). \tag{36}$$

Specifically, the portfolio $a = U^{-1}e(\gamma)/\delta(\gamma)$ has a pricing error $\delta(\gamma)$. Then,

$$\max_{a:\, a'Ua=1} |a'e(\gamma)| = \delta(\gamma), \tag{37}$$

and $\delta(\gamma)$ can be interpreted as the maximum pricing error that one can obtain from using $y(\gamma)$ to price the test assets.

The Hansen-Jagannathan distance has an information-theoretic interpretation too. Let P be the data generating measure and Φ denote a family of probability measures that satisfy the asset pricing restrictions ($m \in \mathcal{M}$). The goal is to find a probability measure Q with minimal entropy divergence from the empirical measure P, defined as the solution to the following inverse problem

$$\min_{Q\in\Phi} D_\eta(P,Q) = \int \phi_\eta\left(dQ/dP\right) dQ \tag{38}$$

$$\text{subject to } \int e(\gamma) dQ = 0_N, \tag{39}$$

where $\phi_\eta(\cdot)$ denotes again the Cressie-Read divergence family. A candidate SDF $y(\gamma)$ defines a measure Q^y with density $dQ^y = \frac{y(\gamma)}{E[y(\gamma)]}dP$ and a relative entropy (with respect to P) given by $E\left[\frac{y(\gamma)}{E[y(\gamma)]}\phi_\eta\left(\frac{y(\gamma)}{E[y(\gamma)]}\right)\right]$. The model (SDF) $y(\gamma)$ is misspecified if $y(\gamma) \notin \mathcal{M}$.

Almeida and Garcia (2012) show that for a fixed vector of parameters γ, the primal and dual problems in the SDF framework can be written as

$$\delta_\eta(\gamma) = \min_{m\in\mathcal{M}} E\left[\frac{(1+m-y(\gamma))^{\eta+1}-1}{\eta(\eta+1)}\right] \tag{40}$$

and

$$\delta_{\eta}(\gamma) = \max_{\lambda \in \Re^N} \lambda' q - E\left[\frac{(\eta \lambda' R)^{\frac{\eta+1}{\eta}}}{\eta+1} + (y(\gamma)-1)\lambda' R + \frac{1}{\eta(\eta+1)}\right], \tag{41}$$

respectively. The dual problem for the Hansen-Jagannathan distance is obtained for $\eta = 1$ (see Almeida and Garcia, 2012; Ghosh, Julliard and Taylor, 2017).

There is a small but growing literature on evaluating asset pricing models using entropy measures (Stutzer, 1995; Kitamura and Stutzer, 2002; Almeida and Garcia, 2012; Backus, Chernov and Zin, 2014; Bakshi and Chabi-Yo, 2014; Ghosh, Julliard and Taylor, 2016; among others). Several of these papers derive optimal lower bounds on the SDFs and develop diagnostics that measure how far a model deviates from these entropy bounds. However, this analysis does not fully embrace the inherent misspecification of all asset pricing models and is still conducted in a "model selection" mode. Also, while some of the used entropy divergence measures nicely help to demonstrate how higher-order moments of the distribution can account for much of the entropy of the SDFs, they are not "distance" measures (metricness). Our point of departure from the existing literature is two-fold. First, we adopt an entropy-driven approach to model aggregation that explicitly recognizes the misspecification of the candidate SDFs. Second, we employ the Hellinger distance, due to its metricness and other theoretical and robustness properties, in estimating and aggregating the individual models.

3.2 SDF Aggregator

Suppose there are M proposed misspecified models, $\hat{y}_{i,t} = y_{i,t}(\hat{\gamma}_i)$, $i = 1, ..., M$ and $t = 1, ..., T$, for the unknowable true model m. While $\hat{y}_{i,t}$ is evaluated at $t = 1, ..., T$, they are based on estimates $\hat{\gamma}_i$ from a prior training sample of size N. In this respect, the effective number of sample observations is $N+T$, where the candidate models are estimated using observations $1, ..., N$ and the aggregation weights are estimated using observations $N+1, ..., N+T$. We allow for both linear and nonlinear SDF specifications as well as nested and non-nested SDFs. For the sake of argument, we assume that the model parameters for each model are estimated by minimizing the Hansen-Jagannathan distance. Our approach in this paper is to treat each model as an incomplete 'indicator' of the latent DGP. Then, a model averaging rule would aggregate information from all of these models and construct a pseudo-true model $\tilde{y}$.

Here, we follow Maasoumi (1986) in characterizing the solution for $\tilde{y}$. Let $y_t = (\hat{y}_{1,t}, ..., \hat{y}_{M,t})'$ be the i-th row of the $T \times M$ matrix Y and $\tilde{y} = h(\hat{y}_1, ..., \hat{y}_M)$, where h is an aggregator or index

function. Note that it might be more convenient to work with the estimated pricing errors $e_i(\hat{\gamma}_i)$, $i = 1, ..., M$, instead of $\hat{y}_i$'s. We are interested in finding the aggregator $\tilde{y}_t$ with a distribution that is as close as possible to the multivariate distribution of $\hat{y}_i$'s. Maasoumi (1986) generalized the pairwise criteria of divergence to a general multivariate context, as follows:

$$D_\rho(\tilde{y}, Y; w) = \sum_{i=1}^{M} w_i \left\{ \sum_{t=1}^{T} \tilde{y}_t \left[\left(\frac{\tilde{y}_t}{y_{i,t}} \right)^\rho - 1 \right] \Big/ \rho(\rho+1) \right\}, \tag{42}$$

The aggregator that minimizes $D_\rho(\tilde{y}, Y; w)$ subject to $\sum_{i=1}^{M} w_i = 1$ is given by

$$\tilde{y}_t^* \propto \left[\sum_{i=1}^{M} w_i y_{i,t}^{-\rho} \right]^{-1/\rho}. \tag{43}$$

Note that the linear pooling of models is obtained as a special case when $\rho = -1$ and the Hellinger distance aggregator is obtained for $\rho = -1/2$.

In order to implement the above aggregation scheme, we need to estimate the unknown parameters $w = (w_1, ..., w_M)'$ and ρ. We propose two methods for estimating these parameters.

The first method is, for given $(\hat{y}_{1,t}, ..., \hat{y}_{M,t})'$ obtained in a preliminary step by minimizing the Hansen-Jagannathan distance for each model, set $\rho = -1$ and construct the pricing errors of the aggregator

$$\tilde{e}_T(w) = \frac{1}{T} \sum_{t=1}^{T} R_t \left[\sum_{i=1}^{M} w_i \hat{y}_{i,t} \right] - q. \tag{44}$$

Then, the unknown aggregation weights w are obtained as

$$\hat{w} = \arg\min \tilde{e}_T(w)' \left(\frac{1}{N} \sum_{t=1}^{N} R_t R_t' \right)^{-1} \tilde{e}_T(w) \tag{45}$$

subject to the restrictions $w_i \geq 0$ for $i = 1, .., M$ and $\sum_{i=1}^{M} w_i = 1$. Note also that these parameters can be estimated by any member of the Cressie-Read divergence family. We use the Hansen-Jagannathan distance estimator due to its computational simplicity and maximum pricing error interpretation.

The other possibility is to estimate w by minimizing the distance of the aggregator's distribution from a desired distribution. Let P be a probability measure associated with some benchmark model with density p, and q denote the density of the Hellinger distance aggregator $\tilde{y}_t(w) = \left[\sum_{i=1}^{M} w_i \hat{y}_{i,t}^{1/2} \right]^2$. Using the generalized entropy (Cressie-Read) divergence from Q to P defined in (24)-(25) and

imposing $\eta = -1/2$, we obtain the scaled Hellinger distance $\mathcal{H} \propto D_{-1/2}(P, Q)$

$$\mathcal{H} = \frac{1}{2}\int \left(p^{1/2}(x) - q^{1/2}(x)\right)^2 dx. \tag{46}$$

Estimate of x is obtained by minimizing $\mathcal{H}$ with respect to w, subject to the relevant restrictions. In practical implementation, we estimate p and q by a kernel density estimator and the integral in (46) is evaluated numerically. The choice of a benchmark model is discussed in the next section.

4 Empirical Analysis

4.1 Data and Asset-Pricing Models

We analyze five popular nonlinear asset-pricing models. The SDF for these models is log-linear in the factors and takes the form $y_t(\gamma) = \exp(\gamma' \tilde{f}_t)$.

1. CAPM of Brown and Gibbons (1985):

$$y_t^{CAPM}(\alpha, \beta) = \beta(1-k)^{-\alpha} R_{m,t}^{-\alpha} \tag{47}$$

   or

$$\ln(y_t^{CAPM}(\gamma)) = \gamma_0 + \gamma_1 \ln(R_{m,t}), \tag{48}$$

   where R_m is the gross market return, β is the discount rate, $\alpha > 0$ is the coefficient of relative risk aversion, k is the proportion of wealth consumed in every period, $\gamma_0 = -\alpha \ln(\beta(1-k))$ and $\gamma_1 = -\alpha$.

2. Consumption CAPM (CCAPM):

$$y_t^{CCAPM}(\alpha, \beta) = \beta \left(\frac{C_t}{C_{t-1}}\right)^{-\alpha} \tag{49}$$

   or

$$\ln(y_t^{CCAPM}(\gamma)) = \gamma_0 + \gamma_1 c_t, \tag{50}$$

   where C denotes real per capita consumption of non-durable goods (seasonally adjusted), $c_t = \ln(C_t) - \ln(C_{t-1})$ is the growth rate in nondurable consumption, $\gamma_0 = \ln(\beta)$ and $\gamma_1 = -\alpha$.

3. Ultimate consumption (UC) model of Parker and Julliard (2005):

$$y_t^{UC}(\alpha, \beta) = \beta \left(\frac{C_{t+s}}{C_{t-1}}\right)^{-\alpha} \tag{51}$$

or

$$\ln(y_t^{UC}(\gamma)) = \gamma_0 + \gamma_1 c_t^s, \tag{52}$$

where $c_t^s = \ln(C_{t+s}) - \ln(C_{t-1})$ and $s > 0$.

4. Non-expected utility (EZ) model of Epstein and Zin (1989, 1991) and Weil (1989):

$$y_t^{EZ}(\alpha, \beta, \sigma) = \beta^{\frac{1-\alpha}{1-\sigma}} \left(\frac{C_t}{C_{t-1}}\right)^{-\sigma\left(\frac{1-\alpha}{1-\sigma}\right)} R_{m,t}^{\frac{\sigma-\alpha}{1-\sigma}}, \tag{53}$$

where $1/\sigma \geq 0$ is the elasticity of intertemporal substitution. Note that the restriction $\alpha = \sigma$ reduces the model to the standard expected utility model (nonlinear CCAPM). The logarithm of the SDF is given by

$$\ln(y_t^{EZ}(\gamma)) = \gamma_0 + \gamma_1 c_t + \gamma_2 \ln(R_{m,t}), \tag{54}$$

where $\gamma_0 = 1 - \ln(\beta)$, $\gamma_1 = -\frac{(1-\alpha)(\sigma(1-\phi)+\phi)}{1-\sigma}$, and $\gamma_2 = \frac{\sigma-\alpha}{1-\sigma}$.

5. Durable consumption CAPM (D-CCAPM) of Yogo (2006):

$$y_t^{D-CAPM}(\alpha, \beta, \sigma, \phi) = \beta^{\frac{1-\alpha}{1-\sigma}} \left(\frac{C_t}{C_{t-1}}\right)^{-\sigma\left(\frac{1-\alpha}{1-\sigma}\right)} \left(\frac{C_{d,t}/C_t}{C_{d,t-1}/C_{t-1}}\right)^{\phi(1-\alpha)} R_{m,t}^{\frac{\sigma-\alpha}{1-\sigma}}, \tag{55}$$

where C_d is consumption of durable goods and $\phi \in [0, 1]$ is the budget share of durable consumption. When $\phi = 0$, we have the classical non-expected (Epstein-Zin) utility model. By imposing the additional restriction $\alpha = \sigma$, we obtain the standard expected utility model (nonlinear CCAPM). After taking logarithms, we have

$$\ln(y_t^{D-CAPM}(\gamma)) = \gamma_0 + \gamma_1 c_t + \gamma_2 c_{d,t} + \gamma_3 \ln(R_{m,t}), \tag{56}$$

where $\gamma_0 = 1 - \ln(\beta)$, $\gamma_1 = -\frac{(1-\alpha)(\sigma(1-\phi)+\phi)}{1-\sigma}$, $\gamma_2 = \phi(1-\alpha)$, and $\gamma_3 = \frac{\sigma-\alpha}{1-\sigma}$.

In summary, the traditional CCAPM is nested within the EZ model when $\alpha = \sigma$ while D-CCAPM nests EZ ($\phi = 0$) and CCAPM ($\phi = 0$ and $\alpha = \sigma$). The UC model is strictly non-nested with all the other models.

As a benchmark model ('pivot') for computing the Hellinger distance, we use the three-factor (FF3) model of Fama and French (1993)

$$y_t^{FF3}(\gamma) = \gamma_0 + \gamma_1 r_{m,t} + \gamma_2 smb_t + \gamma_3 hml_t, \tag{57}$$

where r_m denotes the excess return on the market portfolio, smb is the return difference between portfolios of stocks with small and large market capitalizations, and hml is the return difference between portfolios of stocks with high and low book-to-market ratios ("value" and "growth"

stocks, respectively). The FF3 model is one of the most successful empirical models and the information contained in the *smb* and *hml* factors is somewhat orthogonal to the information in the consumption-based CAPM models considered above.[8]

The test asset returns are the monthly gross returns on the value-weighted 25 Fama-French size and book-to-market ranked portfolios, and the 17 industry portfolios from Kenneth French's website. The sample period is February 1959 to December 2015. The consumption data that is used to construct the growth rates c_t, c_t^s and $c_{d,t}$, is real per capita, seasonally adjusted consumption of non-durable and durable goods from the Bureau of Economic Analysis. The excess return $r_{m,t}$ on the value-weighted stock market index (NYSE-AMEX-NASDAQ) is obtained from Kenneth French's website. The gross market return is constructed by adding the one-month T-bill rate to the excess return. The data for the *smb* and *hml* factors is also collected from Kenneth French's website. For the UC model of Parker and Julliard (2005), we use $s = 23$. This reduces the effective sample period to February 1959 – December 2013 or 659 observations. The only persistent factor is the accumulated consumption growth c_t^s in the UC model. All other factors, as well as the returns on the test assets, do not exhibit serial correlation and the their statistical properties provide a reasonable approximation to the regularity conditions in our theoretical framework.

All models are estimated using a fixed 'training' sample of size $N = 419$. This is kept fixed in this paper. The remaining (20 years) part of the sample $N + 1, ..., N + T$, where $T = 240$, is used for estimation of w_i and evaluation of the aggregate model. This allows us to treat the estimated functionals (SDFs) as given in estimating and evaluating the aggregator. An additional advantage of this setup is that misspecification of the model reflects not only its pricing ability but also its parameter stability over time. In the first set of results, we leave the last 20 years of the sample (January 1994 to December 2013) for model aggregation and evaluation. In a second experimental design, the training sample is the most recent part of the sample period and the evaluation is over the initial 20 years (February 1959 to January 1979).

Unknown parameters are estimated by minimizing the Hansen-Jagannathan distance in (33) over the training sample which is equivalent to

$$\hat{\gamma} = \underset{\gamma \in \Gamma}{\arg\min}\ e_T(\gamma)' \left(\frac{1}{N} \sum\nolimits_{t=1}^{N} R_t R_t' \right)^{-1} e_T(\gamma), \tag{58}$$

[8] Another candidate for a benchmark model would be the non-parametric estimate of a comprehensive model. Such a model is examplified in Cai, Ren, and Sun (2015). On the other hand, a robust pivot can be provided by a constant SDF model which is the least favorable specification for pricing the test assets.

where $e_T(\gamma)$ denotes the sample pricing errors of the model. Plugging in the estimated parameters but using data for the second part of the sample $N+1, ..., N+T$, the sample Hansen-Jagannathan distance is given by

$$\hat{\delta} = \sqrt{e_T(\hat{\gamma})' \left(\frac{1}{T} \sum\nolimits_{t=N+1}^{N+T} R_t R_t' \right)^{-1} e_T(\hat{\gamma})}. \tag{59}$$

4.2 Results

As described in the previous section, the parameters, and hence the SDF, for each individual model are estimated by minimizing the Hansen-Jagannathan distance over N observations in a training sample. We consider two aggregators. The first aggregator is a linear aggregator $\left[\sum_{i=1}^{M} w_i \hat{y}_{i,t}\right]$ coupled with a vector of weight w that are obtained by minimizing the Hansen-Jagannathan distance as in (45). The second aggregator is the aggregator $\left[\sum_{i=1}^{M} w_i y_{i,t}^{1/2}\right]^2$ with a weight vector that minimizes the Hellinger distance between the densities of the aggregator and the pivot (3-factor Fama-French model-FF3) as in (46). Both aggregators estimate the aggregation weights over the remaining T observations. Starting values for weights are the inverse of the Hansen-Jannathan distances, i.e., $\hat{w}_i = (1/\hat{\delta}_i)/\sum_{i=1}^{M}(1/\hat{\delta}_i)$ for $i = 1, ..., M$.

Regardless of the form of the aggregators, all models are evaluated in terms of the HJ distance, computed over observations $N+1$ to $N+T$ ($T = 240$). It should be emphasized that the Hellinger distance aggregator is put at disadvantage since its risk function used for aggregation and estimation of weights is different than the one used for evaluation.[9] Nevertheless, it is useful to document the robustness properties of this aggregator even though we expect its performance to be inferior to the performance of the HJD aggregator.

Tables 1 and 2 report the values of the Hansen-Jagannathan (HJ) distances of the five consumption-based asset pricing models, the benchmark (FF3) model and the aggregator. The HJ distances for the individual models are computed with data from the evaluation sample but using the parameter estimates from the training sample. For the aggregator, the candidate SDFs are estimated from the training sample and treated as fixed. The aggregation weights, which are also reported in the Tables 1 and 2, are then estimated over the evaluation sample either by minimizing the Hansen-Jagannathan distance ($\hat{w}_{-1}$) or the Hellinger distance ($\hat{w}_{-1/2}$) and the resulting aggregator SDF is

[9] The HJ distance is, in fact, a non-optimal GMM estimator with a fixed weighting matrix. The fixed weighting matrix, set to the inverse of the second moment matrix of the test asset returns, provides an objective criteria for comparing pricing errors across competing asset pricing models. While maximum-entropy estimation, including the Hellinger distance estimator, can also be interpreted as a GMM-type estimator, it results in an implicit weighting matrix that is model-specific and makes the comparison of pricing performance across models difficult.

used for computing the corresponding HJ distance. Note that the risk functions for model evaluation and estimation are different and our choice of HJ distance for model evaluation is dictated by our desire to ensure consistency across the different models and the appealing economic interpretation of this risk function. Specification test (HJ distance test) comfortably rejects the null of correct specification for all models. Thus, aggregation is over misspecified models.

In order to assess the robustness of the aggregation procedure across different portfolios of test assets, we consider the following portfolios: (1) 25 Fama-French and 17 industry portfolios, (2) only 25 Fama-French portfolios, and (3) only 17 industry portfolios. As documented in the literature, the 3-factor Fama-French model performs best for pricing the 25 Fama-French portfolios. This should present a challenge for our aggregation since none of the consumption-based models provide proxies of the *smb* and *hml* factors in the FF3 model.

Table 1 presents the results for the evaluation sample January 1994 – December 2013 and Table 2 reports the results for the evaluation sample February 1959 – January 1979. These results clearly illustrate the advantages of our aggregation method. Aggregation reduces the pricing errors relative to the candidate models. It also fares very well relative to the empirical best, here the Fama-French, model when the 25 Fama-French portfolios are used as test assets. This is reassuring since in general practice, the latter may be unknown or indeterminate. Another interesting observation is that CAPM dominates FF3 model in Table 1 even for the 25 Fama-French test assets. This may appear surprising since the Fama-French factors are constructed by sorting the underlying portfolio returns. But since our model evaluation is performed "out-of-sample", the higher pricing errors of the FF3 model reflect its larger parameter instability over time.

The aggregator functions (linear and Hellinger) do about equally well in both evaluation periods. But they suggest widely different weights to different models fitted to the same or different sample periods. Performance of both average functions is better for the most recent evaluation sample including the Great recession. Candidate model's performance is erratic, but aggregate model's performance, whatever the aggregator, is stable and reliable. Aggregation would seem to be robust, and adapt to what is commonly regarded as "regime change" in econometric jargon.

The HJD aggregator largely dominates across models, assets and evaluation periods. It is interesting to note that the HJD weight estimation is coupled with linear aggregation. That is the case of infinite substitution between models. The model with the smallest HJD will ultimately get the highest weight. In this sense, the HJD aggregation is acting like model selection as the shrinkage is done towards the model that minimizes the HJ distance. In the case of the Hellinger

distance, the models are "finitely substitutable." This implies more "hedging" by the Hellinger distance aggregation as it takes away weight from CAPM and assigns it to the EZ model.

Figure 1 plots the SDFs for each model and the Hellinger weighted SDF that uses information from all models for the combined 25 Fama-French and 17 industry portfolios. The aggregator SDF strikes a balance between volatility of the different models. Although the aggregation method shrinks the SDF towards the SDF of the FF3, it cannot match fully the performance of this pivot.[10] But it is still interesting to see that the aggregator closely resembles the dynamics and performance (in terms of pricing errors) of the benchmark model despite of the different information sets.[11] It should be emphasized again that the aggregator based on the Hellinger distance uses different estimation (Hellinger) and evaluation (Hansen-Jagannathan) risk functions. This makes its performance even more impressive as in most cases its pricing ability exceeds or is very close to the best performing candidate model.

The aggregation methods are also quite robust to different sets of test assets as they adapt and recalibrate the weights across the different models. It is interesting to note that for the evaluation sample January 1994 – December 2013, the aggregators load largely on the CAPM and EZ models with the weights on the other models being near zero. This sparsity of the aggregation scheme may prove to be particularly beneficial when the set of candidate models is large. Overall, the robust performance of the proposed aggregation method suggests that combining information from different, possibly misspecified models, may offer substantial advantages. Even if the aggregator is dominated by an individual model, we can not know, a priori, which model will do well over a particular sample for a particular set of test assets. Therefore, in the risk sense, the model aggregation is ideal.

4.3 Simulations

We conduct a small Monte Carlo simulation experiment to assess the properties of the proposed model aggregators. The time series sample size is $N + T = 600$ with $N = 360$ and $T = 240$, and the number of Monte Carlo replications is 1,000. Let $Y_t = [f_t', \ r_t']'$, where $r_t = \ln(R_t)$, with

$$\mu = E[Y_t] = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{60}$$

[10] This phenomenon is expected from the oracle inequalities mentioned earlier. While the model corresponding to $\inf_{i=1,\ldots,M} \mathcal{R}_T(f_i, f)$ is not discoverable here, the "risk" of the Fama-French model is the smallest.

[11] In unreported results, we relax the positivity constraint on w which allows some poorly behaved models to receive a negative weight in the aggregation procedure. Interestingly, this provides further, and often substantial, reduction of the pricing errors which is accompanied by a much higher volatility of the pricing kernel.

and

$$V = \text{Var}[Y_t] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}. \tag{61}$$

We use two sets of test asset returns: (i) the 25 Fama-French portfolios, and (ii) the 17 industry portfolios. We consider four consumption-based models – CAPM, CCAPM, EZ and D-CCAPM – with factors $\ln(R_{m,t})$, c_t, and $c_{d,t}$. As in the empirical application, the benchmark model is the Fama-French 3 factor model with factors $r_{m,t}$, smb_t, and hml_t. We assume that

$$\begin{bmatrix} f_t \\ r_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \right). \tag{62}$$

The covariance matrix of the simulated factors and returns, V, is is set equal to the sample covariance matrix from the data.

We investigate two scenarios: first, when all of the models are misspecified and second, when one of the models (CAPM, in particular) is correctly specified. In the first case (misspecified models), the means of the simulated returns are set equal to the sample means of the actual returns since all of the estimated models are rejected by the data. For generating data from a correctly specified model, we use the properties of the log-normal distribution and write the pricing errors for a log-linear SDF as

$$\begin{aligned} e(\gamma) &= E[R_t y_t(\gamma)] - 1_N = E[\exp(r_t + \gamma_0 + \gamma_1' f_t)] - 1_N \\ &= \exp\left(\gamma_0 + \mu_2 + 0.5\gamma_1' V_{11}\gamma_1 + V_{21}\gamma_1 + 0.5\text{Diag}(V_{22})\right) - 1_N. \end{aligned} \tag{63}$$

It then follows that a model is correctly specified if and only if

$$\mu_2 = -0.5\text{Diag}(V_{22}) - \left(\gamma_0 + 0.5\gamma_1' V_{11}\gamma_1\right) 1_N - V_{21}\gamma_1. \tag{64}$$

Thus, we can set the mean of the simulated returns μ_2 as in (64) to ensure that one of the models is correctly specified.

Note that, by construction, the statistical nature of the underlying data generating mechanism is the same in the training and evaluation samples. This was not the case in the empirical example which spans several business cycles, crisis periods and possible parameter shifts. This lack of regime-switching in the data generating process allows the aggregators to assign weights based purely on pricing performance and not on the statistical stability of the models. This is expected to induce more mixing across models.

Tables 3 and 4 report the simulation results for the individual asset pricing models and the two aggregators based on the Hansen-Jagannathan distance (HJD) and Hellinger distance (HEL). The

aggregators use information in the four consumption-based models – CAPM, CCAPM, EZ and D-CCAPM – while FF3 is used as a pivot for the HEL aggregation. The estimation of the parameters and the construction of the aggregators is exactly the same as described in the previous sections. Tables 3 and 4 report the mean, median, 10% and 90% quantiles of the empirical distribution of the Hansen-Jagannathan distance as a metric for evaluating the pricing performance of all models and aggregators. The tables also present the median of the Monte Carlo distribution of estimated weights that the aggregators assign to each model.

For the case when all models are misspecified (Table 3), SDF aggregation offers a substantial improvement in pricing performance. The HJD aggregator dominates uniformly the HJD measures of individual models used for aggregation and fares favorably even to the FF3 model which is the most successful empirical model for pricing the 25 Fama-French portfolios. For the 17 industry portfolios, for example, the HJD aggregator readily outperforms the Fama-French 3 factor model. The Hellinger distance aggregator again appears to robustify away from the best performing individual model and distribute the weights more evenly across models. Despite the mismatch between the risk functions for aggregation and pricing performance evaluation, the HEL aggregator achieves some of the smallest pricing errors.

When one of the models is true (Table 4), it is not surprising to see that this model (CAPM) dominates the other individual models although it is probably somewhat surprising that the aggregation weights are still fairly equally distributed over competing models. This is partly due to the fact that CAPM is nested within some of the other consumption-based models. But, more importantly, this also illustrates the "insurance" value of mixing by attaching a "premium" to the possibility of choosing catastrophically false individual models.

5 Conclusions

Economic models are misspecified by design as they try to approximate a complex and often an unknown (and possibly unknowable) true data generating process. Instead of selecting a single model for pricing assets, decision making or forecasting, aggregating information from all these models may adapt better to the underlying uncertainty and result in a more robust approximation. Information theory and generalized entropy provide the natural theoretical foundation for dealing with these types of uncertainty and partial specification. We capitalize on some insights from the information-theoretic approach and propose a mixture method for aggregating information

from different misspecified asset pricing models. The optimal aggregator takes a harmonic mean form with geometric and linear weighting schemes as special cases. In addition, the generalized entropy criterion that underlies our approach allows us to circumvent some serious drawbacks of the standard linear pooling. The application of the aggregator to combining consumption-based asset pricing models demonstrates the advantages of our approach. Density forecasting using a large set of diverse, partially specified models is another natural application of the proposed method. Extending the oracle inequality approach, which is used to bound the risk of the aggregation method, to time series data and more general entropy measures is a promising venue for future research.

References

Almeida, C., and R. Garcia, 2012, Assessing misspecified asset pricing models with empirical likelihood estimators, *Journal of Econometrics* 170, 519–537.

Backus, D., M. Chernov, and S. Zin, 2014, Sources of entropy in representative agent models, *Journal of Finance* 69, 51–99.

Bakshi, G., and F. Chabi-Yo, 2014, New entropy restrictions and the quest for better specified asset pricing models, Dice Center WP 2014-07, Ohio State University.

Bernardo, J., and A. Smith, 1994, *Bayesian Theory*, Wiley.

Birgé, L., 2006, Model selection via testing : An alternative to (penalized) maximum likelihood estimators, *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques* 42, 273–325.

Birgé, L., 2013, Model selection for density estimation with L_2-loss, unpublished manuscript.

Brown, D. P., and M. Gibbons, 1985, A simple econometric approach for utility-based asset pricing models, *Journal of Finance* 40, 359–381.

Cai, Z., Y. Ren, and L. Sun, 2015, Pricing kernel estimation: A local estimating equation approach, *Econometric Theory* 31, 560–580.

Chen, X., and S. C. Ludvigson, 2009, Land of addicts? An empirical investigation of habit-based asset pricing models, *Journal of Applied Econometrics* 24, 1057–1093.

Chernozhukov, V., and H. Hong, 2003, An MCMC approach to classical estimation, *Journal of Econometrics* 115, 293–346.

Cressie, N., and T. Read, 1984, Multinomial goodness of fit tests, *Journal of the Royal Statistical Society B* 46, 440–464.

Diebold, F. X., 1991, A note on Bayesian forecast combination procedures, in *Economic Structural Change: Analysis and Forecasting* (P. Hackl and A. H. Westlund, eds.), 225–232.

Donoho, D. L., and I. M. Johnstone, 1994, Ideal spatial adaptation by wavelet shrinkage, *Biometrika* 81, 425–455.

Epstein, L. G., and S. E. Zin, 1989, Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework, *Econometrica* 57, 937–968.

Epstein, L. G., and S. E. Zin, 1991, Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical investigation, *Journal of Political Economy* 99, 555–576 .

Fama, E. F., and K. R. French, 1993, Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 3–56.

Genest, C., S. Weerahandi, and J. V. Zidek, Aggregating opinions through logarithmic pooling, *Theory and Decision* 17, 61–70.

Geweke, J., and G. Amisano, 2011, Optimal prediction pools, *Journal of Econometrics* 164, 130–141.

Geweke, J., and G. Amisano, 2012, Prediction with misspecified models, *American Economic Review: Papers & Proceedings* 102, 482–486.

Ghosh, A., C. Julliard, and A. P. Taylor, 2017, What is the consumption-CAPM missing? An information-theoretic framework for the analysis of asset pricing models, *Review of Financial Studies* 30, 442–504.

Gospodinov, N., R. Kan, and C. Robotti, 2013, Chi-squared tests for evaluation and comparison of asset pricing models, *Journal of Econometrics* 173, 108–125.

Gospodinov, N., R. Kan, and C. Robotti, 2016, Spurious inference in reduced-rank asset-pricing models, unpublished manuscript.

Granger, C. W., E. Maasoumi, and J. C. Racine, 2004, A dependence metric for possibly nonlinear processes, *Journal of Time Series Analysis* 25, 649–669.

Hall, S. G., and J. Mitchell, 2007, Combining density forecasts, *International Journal of Forecasting* 23, 1–13.

Hansen, L. P., 2013, Uncertainty outside and inside economic models, *Nobel Prize Lecture*.

Hansen, L. P., and R. Jagannathan, 1991, Implications of security market data for models of dynamic economies, *Journal of Political Economy* 99, 225–262.

Hansen, L. P., and R. Jagannathan, 1997, Assessing specification errors in stochastic discount factor models, *Journal of Finance* 52, 557–590.

Kitamura, Y., T. Otsu, and K. Evdokimov, Robustness, infinitesimal neighborhoods, and moment restrictions, *Econometrica* 81, 1185–1201.

Kitamura, Y., and M. Stutzer, 2002, Connections between entropic and linear projections in asset pricing estimation, *Journal of Econometrics* 107, 159–174.

Maasoumi, E, 1977, *A Study of Improved Methods for the Estimation of the Reduced Forms of Simultaneous Equations based on 3SLS Estimators*, Ph.D. Thesis, London School of Economics.

Maasoumi, E, 1978, A modified Stein-like estimator for the reduced form coefficients of simultaneous equations, *Econometrica* 46, 695–703.

Maasoumi, E., 1986, The measurement and decomposition of multi-dimensional inequality, *Econometrica* 54, 991–997.

Maasoumi, E., 1987, Unknown regression functions and information efficient functional forms: An interpretation, *Advances in Econometrics* 5, 301–309.

Maasoumi, E., 1990, How to live with misspecification if you must, *Journal of Econometrics* 44, 67–86.

Maasoumi, E., 1993, A compendium to information theory in economics and econometrics, *Econometric Reviews* 12, 137–181.

Maasoumi, E., and J. S. Racine, 2002, Entropy and predictability of stock market returns, *Journal of Econometrics* 107, 291–312.

McConway, K. J., 1981, Marginalization and linear opinion pools, *Journal of the American Statistical Association* 76, 410–414.

Parker, J. A., and C. Julliard, 2005, Consumption risk and the cross section of expected returns, *Journal of Political Economy* 113, 185–222.

Rigollet, P., 2012, Kullback–Leibler aggregation and misspecified generalized linear models, *Annals of Statistic*s 40, 639–665.

Rigollet, P., 2015, *High Dimensional Statistics*, Lecture Notes, MIT.

Rigollet, P., and A. B. Tsybakov, 2012, Sparse estimation by exponential weighting, *Statistical Science* 27, 558–575.

Stutzer, M., 1995, A Bayesian approach to diagnosis of asset pricing models, *Journal of Econometrics* 68, 367–397.

Varin, C., 2008, On composite marginal likelihoods, *Advances in Statistical Analysis* 92, 1–28.

Varin, C., N. Reid, and D. Firth, 2011, An overview of composite likelihood methods, *Statistica Sinica* 21, 5–42.

Weil, P., 1989, The equity premium puzzle and the risk-free rate puzzle, *Journal of Monetary Economics* 24, 401–421.

Yang, Y., 2000, Mixing strategies for density estimation, *Annals of Statistics* 28, 75–87.

Yogo, M., 2006, A consumption-based explanation of expected stock returns, *Journal of Finance* 61, 539–580.

Table 1: Empirical results for individual models and SDF aggregators.

Evaluation period: 1994:1–2013:12.

	CAPM	CCAPM	UC	EZ	D-CCAPM	FF3	HJD agg.	HEL agg.
25 Fama-French + 17 industry portfolios								
$\hat{\delta}$	0.5237	0.5663	0.5874	0.5409	0.5405	0.5268	0.5212	0.5294
$\hat{w}_{-1}$	0.6831	0.0000	0.0928	0.2240	0.0000			
$\hat{w}_{-1/2}$	0.2479	0.0000	0.0000	0.7182	0.0338			
25 Fama-French portfolios								
$\hat{\delta}$	0.4481	0.4913	0.4717	0.4620	0.4657	0.4527	0.4478	0.4554
$\hat{w}_{-1}$	0.8647	0.0000	0.0000	0.1352	0.0001			
$\hat{w}_{-1/2}$	0.1950	0.0000	0.0000	0.8048	0.0001			
17 industry portfolios								
$\hat{\delta}$	0.2035	0.2460	0.2374	0.2456	0.2283	0.2424	0.2035	0.2110
$\hat{w}_{-1}$	0.9944	0.0006	0.0001	0.0007	0.0042			
$\hat{w}_{-1/2}$	0.4628	0.0036	0.0000	0.4714	0.0623			

Notes: This table reports the estimates for the Hansen-Jagannathan distance $\hat{\delta}$, the aggregation weights $\hat{w}_{-1}$ obtained by minimizing the Hansen-Jagannathan distance (HJD agg.), and the aggregation weights $\hat{w}_{-1/2}$ for the method based on minimizing the Hellinger distance (HEL agg.) between the densities of the aggregator and the FF3 model.

Table 2: Empirical results for individual models and SDF aggregators.

Evaluation period: 1959:2–1979:1.

	CAPM	CCAPM	UC	EZ	D-CCAPM	FF3	HJD agg.	HEL agg.
25 Fama-French + 17 industry portfolios								
$\hat{\delta}$	0.6908	0.6621	0.6656	0.6963	0.6933	0.6556	0.6622	0.6651
$\hat{w}_{-1}$	0.0000	0.9996	0.0003	0.0000	0.0000			
$\hat{w}_{-1/2}$	0.0000	0.8012	0.0000	0.0051	0.1936			
25 Fama-French portfolios								
$\hat{\delta}$	0.5125	0.4613	0.4710	0.5073	0.4889	0.4499	0.4607	0.5113
$\hat{w}_{-1}$	0.0000	0.8806	0.0001	0.0001	0.1192			
$\hat{w}_{-1/2}$	0.9866	0.0133	0.0000	0.0000	0.0000			
17 industry portfolios								
$\hat{\delta}$	0.2004	0.1971	0.2188	0.1965	0.1992	0.2373	0.1965	0.1981
$\hat{w}_{-1}$	0.0020	0.0007	0.0000	0.9973	0.0000			
$\hat{w}_{-1/2}$	0.0419	0.0009	0.0003	0.3345	0.6224			

Notes: This table reports the estimates for the Hansen-Jagannathan distance $\hat{\delta}$, the aggregation weights $\hat{w}_{-1}$ obtained by minimizing the Hansen-Jagannathan distance (HJD agg.), and the aggregation weights $\hat{w}_{-1/2}$ for the method based on minimizing the Hellinger distance (HEL agg.) between the densities of the aggregator and the FF3 model.

Table 3: Simulation results for individual models and SDF aggregators.

Case (i): all models are misspecified.

	CAPM	CCAPM	EZ	D-CCAPM	FF3	HJD agg.	HEL agg.
25 Fama-French portfolios							
mean $\hat{\delta}$	0.4713	0.4831	0.4780	0.4834	0.4533	0.4577	0.4708
median $\hat{\delta}$	0.4683	0.4786	0.4737	0.4794	0.4501	0.4545	0.4680
10% quant. $\hat{\delta}$	0.3944	0.4038	0.3975	0.4035	0.3722	0.3787	0.3920
90% quant. $\hat{\delta}$	0.5535	0.5670	0.5626	0.5717	0.5338	0.5396	0.5540
mean $\hat{w}_{-1}$	0.3512	0.1775	0.1422	0.3291			
mean $\hat{w}_{-1/2}$	0.1766	0.1420	0.2586	0.4228			
17 industry portfolios							
mean $\hat{\delta}$	0.3000	0.3036	0.3101	0.3213	0.3081	0.2908	0.3010
median $\hat{\delta}$	0.2985	0.3008	0.3070	0.3162	0.3077	0.2889	0.3013
10% quant. $\hat{\delta}$	0.2285	0.2280	0.2341	0.2407	0.2360	0.2166	0.2276
90% quant. $\hat{\delta}$	0.3717	0.3781	0.3888	0.4020	0.3839	0.3615	0.3730
mean $\hat{w}_{-1}$	0.4047	0.3347	0.1030	0.1575			
mean $\hat{w}_{-1/2}$	0.3230	0.2174	0.1718	0.2878			

Notes: This table reports the Monte Carlo estimates for the Hansen-Jagannathan distance $\hat{\delta}$ (mean, median, 10% quantile, and 90% quantile), the mean aggregation weights $\hat{w}_{-1}$ obtained by minimizing the Hansen-Jagannathan distance (HJD agg.), and the mean aggregation weights $\hat{w}_{-1/2}$ for the method based on minimizing the Hellinger distance (HEL agg.) between the densities of the aggregator and the FF3 model. The sample size is 600 and the number of Monte Carlo simulations is 1,000.

Table 4: Simulation results for individual models and SDF aggregators.

Case (ii): CAPM is correctly specified and all other models are misspecified.

	CAPM	CCAPM	EZ	D-CCAPM	FF3	HJD agg.	HEL agg.
25 Fama-French portfolios							
mean $\hat{\delta}$	0.3370	0.3490	0.3433	0.3507	0.3459	0.3286	0.3387
median $\hat{\delta}$	0.3339	0.3477	0.3426	0.3498	0.3414	0.3262	0.3369
10% quant. $\hat{\delta}$	0.2728	0.2820	0.2763	0.2771	0.2789	0.2648	0.2737
90% quant. $\hat{\delta}$	0.4066	0.4212	0.4127	0.4239	0.4149	0.3977	0.4082
mean $\hat{w}_{-1}$	0.4344	0.2353	0.1523	0.1781			
mean $\hat{w}_{-1/2}$	0.3360	0.1402	0.2218	0.3020			
17 industry portfolios							
mean $\hat{\delta}$	0.2657	0.2680	0.2744	0.2833	0.2770	0.2563	0.2666
median $\hat{\delta}$	0.2633	0.2654	0.2696	0.2784	0.2746	0.2548	0.2644
10% quant. $\hat{\delta}$	0.2042	0.2065	0.2119	0.2142	0.2108	0.1947	0.2040
90% quant. $\hat{\delta}$	0.3300	0.3318	0.3442	0.3591	0.3439	0.3198	0.3309
mean $\hat{w}_{-1}$	0.4003	0.3490	0.0908	0.1599			
mean $\hat{w}_{-1/2}$	0.3241	0.2010	0.1983	0.2766			

Notes: This table reports the Monte Carlo estimates for the Hansen-Jagannathan distance $\hat{\delta}$ (mean, median, 10% quantile, and 90% quantile), the mean aggregation weights $\hat{w}_{-1}$ obtained by minimizing the Hansen-Jagannathan distance (HJD agg.), and the mean aggregation weights $\hat{w}_{-1/2}$ for the method based on minimizing the Hellinger distance (HEL agg.) between the densities of the aggregator and the FF3 model. The sample size is 600 and the number of Monte Carlo simulations is 1,000.

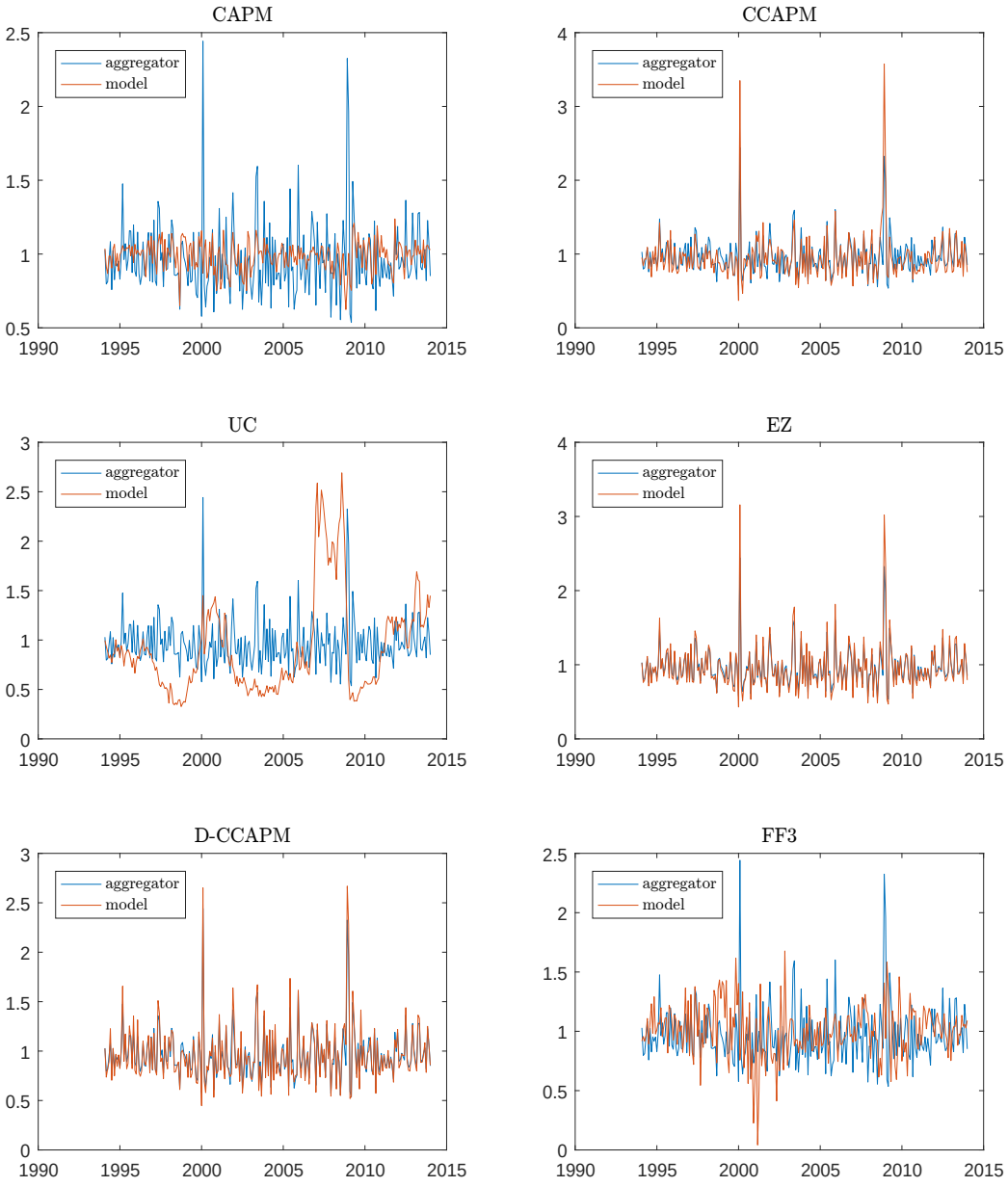

Figure 1: SDFs for individual models and aggregator based on the Hellinger distance for the January 1994 – December 2013 evaluation sample.