Generalized Aggregation of Misspecified Models: With An Application to Asset Pricing*

Nikolay Gospodinov† and Esfandiar Maasoumi‡

December 2018

Abstract

We propose a generalized aggregation approach for model averaging. The entropy-based optimal criterion is a natural choice for aggregating information from many possibly misspecified models as it adapts better to the underlying model uncertainty and results in a more robust approximation. Unlike almost all other approaches in the existing literature, we do not require a true data generation process/model to be contained in the set of models - neither implicitly nor in otherwise popular limiting forms. A shift in paradigm is suggested which prioritizes stochastic optimization and aggregation of information about outcomes over parameter estimation of an optimally selected model. Stochastic optimization is based on a risk function of aggregators across models that satisfies oracle inequalities. Generalized aggregators relax the common perfect substitutability of the candidate models, implicit in linear pooling techniques. The aggregation weights are data-driven and obtained from a proper (Hellinger) distance measure. The empirical results illustrate the economic significance of the aggregation approach in the context of stochastic discount factor models.

JEL Classification: C13, C52, G12.

Keywords: Entropy; Model aggregation; Asset pricing; Misspecified models; Oracle bounds; Hellinger distance.

*We would like to thank Zhungwu Cai, Ruixuan Liu, Richard Luger and seminar and conference participants at Emory University, McGill University, Indiana University, University of Kansas, City University of Hong Kong, University of Southern California, International Symposium on Financial Engineering and Risk Management (Fudan University), Conference in Honor of Max King (Monash University), and CIREQ Econometrics Conference on "Recent Advances in the Method of Moments" 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.

†Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street, N.E., Atlanta, GA 30309-4470, Email: nikolay.gospodinov@atl.frb.org

‡Emory University, Department of Economics, Rich Memorial Building 324, 1602 Fishburne Drive, Atlanta, GA 30322-2240, Email: esfandiar.maasoumi@emory.edu

1 Introduction

Asset pricing models are routinely rejected when confronted with data. This state of affairs is not unusual in many fields. All practitioners profess to accept the truism that "all models are misspecified." Yet, with notable few exceptions, inference methods and decision making proceed as though the data generating process (DGP) is either known or the model uncertainty around it is asymptotically inconsequential. DGP is a latent, possibly unknowable, object and models of it are – by construction – only simplified, incomplete or directed maps. This is especially true when models are partially specified and are estimated by moment matching. While it seems natural that model uncertainty should be explicitly recognized and adequately incorporated in statistical inference, this is rarely done in data analysis of moment condition models. Standard practice typically acknowledges only sampling variability and parameter uncertainty but not model uncertainty. Model averaging or aggregation, discussed below, provides an intriguing alternative for dealing with irreducible model uncertainty in order to elicit some features of the latent object of interest.

Misspecified models can still be useful for informing policy makers and investors in their decision making. While earlier attempts to accommodate misspecified models in econometrics date back to the late 70s and early 80s (Maasoumi, 1977, 1978, 1990; White, 1982; Gouroeroux, Monfort and Trognon, 1984; among others), the analysis of uncertain moment condition models is still in its infancy. This is a fertile ground for future research (see Lars Hansen's (2013) Nobel lecture and advances in misspecification-robust inference in moment condition models that include Maasoumi and Phillips, 1982; Inoue and Hall, 2003; Gospodinov, Kan and Robotti, 2013; among others). These attempts generally focus on inference on model parameters. There are some conceptual and implementational hurdles, however, that arise in the analysis of multiple misspecified models. With several candidate models, each is characterized by its own 'pseudo-true' objects and parameters that are specific to each model and even to estimation criteria, instrument sets, smoothing parameters etc. (Antoine, Proulx and Renault, 2018). Analytical and policy objectives may not coincide with these pseudo-true objects.

An important recent strand of literature in mathematical sciences and engineering shifts the statistical paradigm away from parameter estimation and places the risk of model choice at the center of statistical inquiry. This is appropriate and productive when an aggregate object serves as an optimal inference for the output of models. Some traditional objects of inference, especially

partial effects, require far more careful reconsideration when all models are expressly misspecified. For instance, different partial effects are referenced by different conditional distributions and parameters in each model. Instead, an empirical counterpart of an optimal aggregate/average is a properly defined object that represents desired outcomes, such as forecast density, functionals, stochastic discount factor etc.[1]

Most model averaging methods assume that the set of candidate models contains the true model. For example, the Bayesian approach assumes that the true model is contained under the support of the prior. Diebold (1991) provides an illuminating example of this in the context of Bayesian forecast combination. In this approach (Bernando and Smith, 1994), the ambiguity about the true model is resolved asymptotically and in the limit, the mixture that summarizes the beliefs about the individual models would assign a weight of one to a single model. In the limit, this is akin to model selection since it is designed to choose only one of the candidate models and ignores the information in the remaining models. This model selection procedure loses its consistency and robustness properties when the true DGP is not in the set of candidate models. As Maasoumi (1993) argues, "if models are misspecified in an indeterminate manner, we should not be aiming at the discovery of the 'true data generating process'." Similarly, Monfort (1996) remarks that "the search for a well-specified model is something like the quest for the Holy Grail." Our approach dispenses completely with the self-contradictory notion of a true model and treats the candidate models as genuinely misspecified.

In econometrics, our approach is similar in spirit to Geweke and Amisano (2011, 2012) for prediction pooling from misspecified models. In contrast, we develop a generalized entropy-based approach to mixing information from different models. The minimum Shannon entropy or Kullback-Leibler information criterion used by Geweke and Amisano (2011, 2012) and Hall and Mitchell (2007) is a special case of our framework.[2] Importantly, unlike Geweke and Amisano (2011), we choose a proper metric for selecting the mixture weights which is a "distance" since it is symmetric and satisfies the triangle inequality. Our closeness metric is also useful for clustering subsets of models which might be ranked as more informative in a large set of candidate models. The model clustering will identify similar attributes across models and act effectively as a "dimension" reduction device. This is a 'big data' problem and we will briefly allude to penalization methods

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

[2] In this paper, our generalization is facilitated by the fact that we are not mixing densities, necessarily, so that the aggregator does not need to commute with any possible marginalization of the distributions involved (McConway, 1981; Genest, Weerahandi and Zidek, 1984).

that are similar in spirit. The data-driven model weights can also identify situations in which linear aggregation is optimal.[3]

The stochastic discount factor (SDF) framework for asset pricing provides an evidently suitable setting for assessing the benefits of model aggregation. It is widely documented that most, if not all, asset pricing models of equity returns are strongly rejected by the data, and finding a robust set of factors that adequately span the space of SDFs remains elusive.[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 and Jagannathan (1991, 1997) 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. Gagliardini and Ronchetti (2016) and Antoine, Proulx and Renault (2018) characterize the properties of pseudo-true SDFs in a conditional framework. 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] It is also important to contrast the model aggregation approach with the model ambiguity literature. While both approaches are trying to accommodate model misspecification, model ambiguity framework chooses a single reference model a priori based on certain considerations such as analytical tractability, for instance. The aim there is to express the costs of such a model being misspecified when it is the center of an "epsilon ball" of other undefined models/laws (see the seminal work of Hansen and Sargent, 2001, 2008, as well as some recent extensions such as Bonhomme and Weidner, 2018, and Christensen and Connault, 2018). The cost of this model ambiguity is represented in terms of wider asymptotic confidence intervals for inferential objects (such as parameters and partial effects) from the reference model, when the "epsilon" deviation is of a certain small order of magnitude in the sample size. Unlike model aggregation, there are no expressed models competing unambiguously with the focus model in this approach.

[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 (2017) 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 (pooling). On the practical side, our mixing procedure employs information from all models by assigning data-driven 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.

It is instructive to preview the form and the empirical performance of our aggregator using a simple example of 12-month ahead forecasting of U.S. core inflation (CPI less food and energy). The models considered are the Phillips curve model, integrated moving average (1,1) model (Stock and Watson, 2007), commodity-based (convenience yield) model (Gospodinov, 2016), historical average and the Blue Chip survey of expected CPI inflation. The individual model forecasts at time t are denoted by $f_{i,t}$, $i = 1, ..., M$. Our general aggregator takes the form

$$\tilde{f}_t = \left[\sum\nolimits_{i=1}^{M} w_i f_{i,t}^{-\rho} \right]^{-1/\rho}.$$

We set $\rho = -1/2$ and the aggregation weights w_i are estimated by minimizing the "distance" between the aggregator and a pivot/desired density. The data is at monthly frequency for the period 1988:01–2018:09 with 2002:01-2018:09 for out-of-sample evaluation. The out-of-sample forecasts for the "pivot" (Blue Chip survey) and the aggregator are plotted in Figure 1.

Figure 1 about here

Forecast performance is evaluated using a wide range of Bregman loss functions (Patton, 2018). The forecast improvements of the aggregate over the individual model forecasts was substantial across all loss functions (in excess of 60% over the bests performing (convenience yield) model). The aggregate forecast is unbiased and the forecast weights exhibit interesting dynamics over time as the relative performance of the individual models changes. This should be viewed against the backdrop of the well-documented challenges in forecasting core inflation.

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 Program for Misspecified Models

2.1 Some Preliminaries

Suppose one is interested in estimating an unknown functional $f(\cdot)$. Information from a set of auxiliary (partially specified) models is available about this latent $f(\cdot)$. Examples of $f(\cdot)$ include conditional mean functions in regression models, densities, and other latent objects such as stochastic discount factors (SDFs). We consider a shift the statistical paradigm from (parameter) estimation to a "stochastic optimization" paradigm that is detailed below.

Let there be a finite list (dictionary) $\mathcal{F}$ of candidate models that intend to embed certain theoretical or empirical features of the underlying DGP. The stochastic optimization approach does not require a fully articulated structural model and does not assume that this dictionary contains a 'true' model. It will construct an aggregator that minimizes an empirical risk relative to a pseudo-best aggregate. Because it is data driven, it has the potential to adapt to a least misspecified, or the true DGP, were it to be in the class. The aggregation estimator satisfies certain oracle inequalities on expected error, or probability of errors (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, models, the 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 $L : Z \times \mathcal{F} \to \mathbb{R}$ be a measurable loss function with a corresponding risk function $\mathcal{R} : \mathcal{F} \to \mathbb{R}$ 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 risk function $\mathcal{R}(f_Z, f)$, where f_T is the sample analog of f_Z. In the case of quadratic risk it is $\mathcal{R}_T(f_T, f) = \|f_T - f\|_2 = \frac{1}{T} \sum\nolimits_{t=1}^{T} (f_t - f)^2$, where $\|\cdot\|_2$ denotes the L_2 norm.

When interest lies in density or model aggregation, one 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 prior training samples (by sample splitting, for example). These functions are evaluated at the sample values $Z_1, ..., Z_T$. Then, for a constant $C \geq 1$, the following is an "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 under certain conditions:

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

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 $C = 1$.

It is instructive to illustrate some of the main ideas with the popular example of quadratic risk and regression model

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

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

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

[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.

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 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 $f(Y_t|X_{it})$:

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

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 many weights (and aggregation penalties) are obtained 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 $\mathcal{R}(\tilde{f}, f)$.[7] 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

[7] As before, the arguments in this section are developed for probability density functions but can be extended to more general functions.

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[8]

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

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

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

as possible to the multivariate distribution of f_i's. We assume that $f_{i,t} \geq 0$ for all i and t. This is automatically satisfied for density functions and no-arbitrage stochastic discount factors as well as positively-valued variables. In other situations, this condition will require a more judicious choice of variables; gross returns instead of net returns, for example. 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

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

$\mathcal{R}_{T,\rho}(\tilde{f}, f_i)$ is the generalized entropy divergence between the aggregator $\tilde{f}$ and each of the prospective models f_i. The aggregator that minimizes $\tilde{D}_\rho(\tilde{f}, f; w)$ 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 minimaxity, 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,

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

where z denotes a non-zero $N \times 1$ vector of payoffs (a vector of ones if R are gross returns). In order to make this pricing equation consistent with the absence of arbitrage opportunities, $\mathcal{M}$ may need to be replaced by the set of nonnegative admissible SDFs $\mathcal{M}^+$. 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:

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

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:

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

where λ is an $N \times 1$ vector of Lagrange multipliers.[9] The term $\lambda' R$ 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

[9] To ensure non-negativity of the SDF, it may be necessary to replace $(y(\gamma) - \lambda' R)$ with $(y(\gamma) - \lambda' R)^+$, where $(a)^+ \equiv \max[a, 0]$. See Gospodinov, Kan and Robotti (2016) for the analysis in this case. The nonlinear SDFs that we consider below satisfy automatically the non-negativity constraint and this modification is superfluous.

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

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

and

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

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 Cross-Validation Inference

In-sample evaluation and model comparison of asset pricing models is the dominant approach for assessing their pricing performance. But this in-sample framework is plagued by econometric problems that can have serious adverse effects on inference in and across models. For example, presence of weak or spurious factors, rank failure, parameter instability, non-invariance to normalizations etc. have been shown to result in highly misleading inference procedures. Also, the asymptotic framework for model evaluation and comparison changes abruptly depending on whether the models are correctly specified or misspecified, nested, non-nested or overlapping etc.

While out-of-sample evaluation is immune to some of these problems, the structure of the traditional asset pricing models is not well-suited for that without fully specifying the dynamics of the factors. For this reason, we will resort to cross validation that adapts better to the model uncertainty in constructing pseudo out-of-sample pricing errors for model evaluation and estimation of aggregation weights.

Let $\hat{e}_{i,(-t)}(\gamma_i) = \frac{1}{T-1}\sum_{j\neq t} R_j y_{i,j}(\gamma_i) - 1_N$ and $\hat{U}_{(-t)} = \frac{1}{T-1}\sum_{j\neq t} R_j R_j'$ be the leave-one-out estimators of $e_i(\gamma_i)$ and U for model $i = 1, ..., M$. These estimators are obtained by removing the t-th observation from R and $y_i(\gamma)$ and computing the sample quantities with the remaining $T-1$ observations. Parameters for model i are then estimated as

$$\hat{\gamma}_{i,(-t)} = \underset{\gamma_i\in\Gamma}{\arg\min}\ \hat{e}_{i,(-t)}(\gamma_i)'\hat{U}^{-1}_{(-t)}\hat{e}_{i,(-t)}(\gamma_i). \tag{42}$$

This, in turn, is used for constructing the leave-one-out estimator of the SDF $\hat{y}_{i,(-t)} = y_{i,(-t)}(\hat{\gamma}_{i,(-t)})$ and the cross-validation version of the Hansen-Jagannathan distance

$$\hat{\delta}_{i,(-t)} = \sqrt{\hat{e}_{i,(-t)}(\hat{\gamma}_{i,(-t)})'\hat{U}^{-1}_{(-t)}\hat{e}_{i,(-t)}(\hat{\gamma}_{i,(-t)})}. \tag{43}$$

If R_t or $y_{i,t}(\gamma_i)$ are serially correlated or h-dependent for some $h > 1$, the leave-one-out procedure should be replaced with a leave-h-out cross-validation which removes $h-1$ data points on both sides of the t-th observation. The cross-validation distance is expected to reflect better the underlying model uncertainty and to provide a more reliable statistical measure of pricing performance. The next section uses the cross-validation approach for estimating aggregation weights.

Under some regularity conditions (see Gospodinov, Kan and Robotti, 2013), if $\delta > 0$,

$$\sqrt{T}(\hat{\delta}^2_{i,(-t)} - \delta_i^2) \xrightarrow{d} N(0, V_{\delta_i}), \tag{44}$$

where $V_{\delta_i} = \sum_{k=-\infty}^{\infty} E[v_{i,t}(\gamma_i^*)v_{i,t+k}(\gamma_i^*)]$ and $v_{i,t}(\gamma_i^*) = y_{i,t}(\gamma_i^*)^2 - [y_{i,t}(\gamma_i^*) - \lambda_i^{*\prime}R_{i,t}]^2 - 2\lambda_i^{*\prime}z - \delta_i^2$.

A similar result can be used for pairwise model selection between models i and j. Under the null $H_0 : \delta_i^2 = \delta_j^2$, we have (Hansen, Heaton and Luttmer, 1995; Gospodinov, Kan and Robotti, 2013)

$$\sqrt{T}(\hat{\delta}_{i,(-t)}^2 - \hat{\delta}_{j,(-t)}^2) \xrightarrow{d} N(0, V_{\delta_i j}), \tag{45}$$

where $V_{\delta_i j} = \sum_{k=-\infty}^{\infty} E[v_{ij,t} v_{ij,t+k}] > 0$ and $v_{ij,t} = v_{i,t}(\gamma_i^*) - v_{j,t}(\gamma_j^*)$. It is important to emphasize that the result in (45) holds only if $\sigma_d^2 \neq 0$.

Unfortunately, the implementation of the model selection test will depend on whether the models are nested, non-nested or overlapping. This makes its implementation quite cumbersome with further complications for multiple model comparison. As a related issue, since all candidate models are believed to be misspecified, choosing only one will result in loss of information and inflated risk. In a simplified context, Yang (2003) shows that the ratio of the risks associated with a test-based model (density) selection and model (density) averaging, respectively is strictly greater than one.

3.3 SDF Aggregator

Suppose there are M proposed misspecified models for the unknowable true model m with $\hat{y}_{i,(-t)} = y_{i,(-t)}(\hat{\gamma}_{i,(-t)})$, $i = 1, ..., M$ and $t = 1, ..., T$, denoting their corresponding leave-one-out SDF estimates. 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.[10] 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 SDF $\tilde{y}$.

Here, we follow Maasoumi (1986) in characterizing the solution for $\tilde{y}$. Let $\hat{y}_{(-t)} = (\hat{y}_{1,(-t)}, ..., \hat{y}_{M,(-t)})'$ be the t-th row of the $T \times M$ matrix Y and $\tilde{y} = h(\hat{y}_{(-1)}, ..., \hat{y}_{(-T)})$, where h is an aggregator or index function. Similarly, let $\hat{e}_{(-t)} = (\hat{e}_{1,(-t)}, ..., \hat{e}_{M,(-t)})'$ denote the corresponding vector of leave-one-out pricing errors. We are interested in constructing the aggregator $\tilde{y}$ with a distribution that is as close as possible to the multivariate distribution of $\hat{y}_i$'s. Maasoumi (1986) generalized the pairwise

[10] Alternatively, one could estimate the parameters using other entropy-based estimators (Almeida and Garcia, 2012), including the Hellinger distance. Furthermore, one could estimate the parameters and the aggregation weights by sample splitting instead of cross-validation. An earlier version of the paper explored this possibilities but the results are not reported here to conserve space.

criteria of divergence to a general multivariate context as follows

$$D_\rho(\tilde{y}, Y; w) = \sum_{i=1}^{M} w_i \left\{ \sum_{t=1}^{T-1} \tilde{y}_{(-t)} \left[\left(\frac{\tilde{y}_{(-t)}}{y_{i,(-t)}} \right)^\rho - 1 \right] \Big/ \rho(\rho+1) \right\}. \tag{46}$$

The aggregator that minimizes $D_\rho(\tilde{y}, Y; w)$ takes the form

$$\tilde{y}_t^* \propto \left[\sum_{i=1}^{M} w_i y_{i,(-t)}^{-\rho} \right]^{-1/\rho}. \tag{47}$$

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 what follows, we set $\rho = -1/2$ that renders the Hellinger distance aggregator $\tilde{y}_{(-t)}(w) = \left[\sum_{i=1}^{M} w_i \hat{y}_{i,(-t)}^{1/2} \right]^2$. In order to implement the this aggregation scheme, we need to estimate the unknown parameters $w = (w_1, ..., w_M)'$ The estimation of w is performed by minimizing the distance of the aggregator's distribution from a desired distribution. Let P be a probability measure associated with pricing errors from some benchmark model (pivot) with density p, and q denote the density of (leave-one-out) pricing errors of the Hellinger distance aggregator $\tilde{y}_{(-t)}(w)$. 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 $\mathcal{H} \propto D_{-1/2}(P, Q)$

$$\mathcal{H} = \frac{1}{2} \int \left(p^{1/2}(x) - q^{1/2}(x) \right)^2 dx. \tag{48}$$

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 (48) 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 four 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):

$$y_t^{CAPM}(\alpha, \beta) = \beta(1-k)^{-\alpha} R_{m,t}^{-\alpha} \tag{49}$$

or

$$\ln(y_t^{CAPM}(\gamma)) = \gamma_0 + \gamma_1 \ln(R_{m,t}), \tag{50}$$

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):

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

or

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

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. 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 $\gamma_2 = \frac{\sigma-\alpha}{1-\sigma}$.

4. 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 $\alpha = \sigma$).

As a 'pivot' for computing the Hellinger distance, we use a model with a constant as a single factor. This choice is intended to robustify the aggregator with respect to the least favorable model specification. This is an important point that should be taken into account in the performance evaluation of the Hellinger aggregator. We also report results for 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.[11] It should be noted that this model is not used in the estimation of aggregation weights and only as a benchmark for single model performance.

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 January 1969 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. The factors, as well as the returns on the test assets, do not exhibit serial correlation and their statistical properties provide a reasonable approximation to our regularity conditions and the leave-one-out cross validation framework.

[11] 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.

4.2 Results

The parameters for each model and the aggregation weights for the Hellinger aggregator are estimated as described in Sections 3.2 and 3.3. For the sake of comparison, we also report another aggregator with weights that are the inverse of the Hansen-Jagannathan distances, i.e., $\hat{w}_i = (1/\hat{\delta}_i)/\sum_{i=1}^{M}(1/\hat{\delta}_i)$ for $i = 1, ..., M$. All models are then evaluated using the cross-validation version of the Hansen-Jagannathan distance. 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.[12]

Table 1 reports the values of the Hansen-Jagannathan (HJ) distances of the four consumption-based asset pricing models, the external benchmark (FF3) model and the two aggregators. The table also presents the aggregation weights for the Hellinger distance ($\hat{w}_{-1/2}$). The resulting aggregator SDF is used for computing the corresponding HJ distance. Again, the choice of different risk functions for model evaluation (HJ distance) and estimation (Hellinger distance) 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.

The results in Table 1 clearly illustrate the advantages of our aggregation method. Aggregation reduces the pricing errors relative to the candidate models. In all cases, the Hellinger aggregation (HEL agg.) approach dominates the individual models with sometimes significant pricing improvements.[13] The documented improvements are against the backdrop that the HJ distance of asset pricing models are usually fairly small (given their high degree of misspecification) and the use of

[12] The Hansen-Jagannathan 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.

[13] 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 a further reduction of the pricing errors.

very naive model (constant SDF) as a pivot for estimating the aggregation weights. The Hellinger aggregator also fares very well relative to the empirically best performing (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 EZ and D-CCAPM dominate FF3 model for combined (25 Fama-French and 17 industry) set of portfolio returns. But since the cross-validation has the flavor of "out-of-sample" evaluation, the higher pricing errors of the FF3 model may reflect its larger parameter instability over time.

Table 1 about here

The HJ-distance aggregator (HJD agg.) also delivers improvements over individual models. It is interesting to note that the HJD involves linear aggregation of models. That is the case of infinite substitution between models. The model with the smallest HJD will ultimately get the highest weight. 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 EZ, for example, and assigns it to the D-CCAPM model. Even when candidate model's performance is erratic, aggregate model's performance, whatever the aggregator, is stable and reliable. Overall, the aggregation approaches appear robust and adapt well to "regime changes" in the data.

Figure 2 plots the SDFs for four models and the Hellinger-weighted SDF that uses information from all models for the combined 25 Fama-French and 17 industry portfolios. Since, for these test assets, the aggregator SDF assigns most of the weight to the EZ model, it adapts to the volatility of the SDF for this model.

Figure 1 about here

In summary, the aggregation method appears to be quite robust to different sets of test assets as it recalibrates the weights across the different models. It is interesting to note that the Hellinger aggregator loads largely on the EZ and D-CCAPM 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. Our experiments reveal that this sparsity is even more pronounced when the aggregation is achieved via linear pooling where the shrinkage is done towards the best performing model. By contrast, the Hellinger aggregator appears to robustify away from the best performing individual model and distribute weights more evenly across models. 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 preferable.

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 $T = 200$ and the number of Monte Carlo replications is 1,000. Let $Y_t = [f_t', \ r_t']'$, where $r_t = \ln(R_t)$, with

$$\mu = E[Y_t] = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{58}$$

and

$$V = \text{Var}[Y_t] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}. \tag{59}$$

For test asset returns, we use 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 Hellinger aggregator assigns weights to these models with a pivot given by the constant SDF. Also, the external benchmark model is the Fama-French 3 factor model with factors $r_{m,t}$, smb_t, and hml_t. We assume that

$$\begin{bmatrix} f_t \\ r_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \right). \tag{60}$$

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 (D-CCAPM, 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

$$\begin{aligned} e(\gamma) &= E[R_t y_t(\gamma)] - 1_N = E[\exp(r_t + \gamma_0 + \gamma_1' f_t)] - 1_N \\ &= \exp\left(\gamma_0 + \mu_2 + 0.5\gamma_1' V_{11}\gamma_1 + V_{21}\gamma_1 + 0.5\text{Diag}(V_{22})\right) - 1_N. \end{aligned} \tag{61}$$

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

$$\mu_2 = -0.5\text{Diag}(V_{22}) - \left(\gamma_0 + 0.5\gamma_1' V_{11}\gamma_1\right) 1_N - V_{21}\gamma_1. \tag{62}$$

Thus, we can set the mean of the simulated returns μ_2 as in (62) to ensure that one of the models is correctly specified.

Unlike the empirical example, which spans several business cycles, crisis periods and structural changes, the 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 2 and 3 report the simulation results for the individual asset pricing models and the Hellinger distance aggregator (HEL agg.). The estimation of the parameters and the construction of the aggregator is exactly the same as described in the previous sections. Tables 2 and 3 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. The tables also present the mean of the Monte Carlo distribution of estimated weights that the aggregator assigns to each model.

Table 2 about here

For the case when all models are misspecified (Table 2), SDF aggregation offers a substantial improvement in pricing performance. The aggregator dominates uniformly the HJ-distance measures of individual models used for aggregation. Despite the mismatch between the risk functions for aggregation and pricing performance evaluation, the Hellinger aggregator achieves the smallest pricing errors. Interestingly, the aggregator does not assign the largest weight to the model with smallest HJ distance. Instead, it distributes the weights across models with the largest loading being assigned to the most comprehensive (D-CCAPM) model.

Table 3 about here

The results are similar when one of the models is true (Table 3). Since D-CCAPM nests the other model specifications, it is not surprising that the more parsimonious parameterizations have smaller HJ distances. The aggregator again dominates all individual models. It continues to load mostly on the D-CCAPM model but the aggregation weights are still fairly equally distributed over competing models even if the true model is in the candidate set. This 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 new generalized 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.

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 value when the true DGP/model is not known. Stochastic optimization techniques, paired with information criteria as suitable risk measures, reflect more deeply this phenomenon.

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 dependent 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.

Andrews, D. W. K, 1994, Empirical process methods in econometrics, in: R. F. Engle and D. L. McFadden, (Eds.), *Handbook of Econometrics* vol.4, 2247–2294, North-Holland, Amsterdam.

Antoine, B., K. Proulx, and E. Renault, 2018, Pseudo-true SDFs in conditional asset pricing models, *Journal of Financial Econometrics*, forthcoming.

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.

Bonhomme, S., and M. Weidner, Minimizing sensitivity to model misspecification, Working Paper, University of Chicago.

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.

Christensen, T., and B. Connault, 2018, Counterfactual sensitivity and robustness, Working Paper.

Chudik, A., G. Kapetanios, and M. H. Pesaran, 2018, A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models, *Econometrica*, forthcoming.

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.

Gagliardini, P., and D. Ronchetti, 2016, Comparing asset pricing models by the conditional Hansen-Jagannathan distance, Working Paper.

Genest, C., S. Weerahandi, and J. V. Zidek, 1984, 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., 2016, The role of commodity prices in forecasting U.S. core inflation, Atlanta Fed Working Paper 2016-5.

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, On the properties of the constrained Hansen-Jagannathan distance, *Journal of Empirical Finance* 36, 121–150.

Gospodinov, N., R. Kan, and C. Robotti, 2017, Spurious inference in reduced-rank asset-pricing models, *Econometrica* 85, 1613–1628.

Gourieroux, C., A. Monfort, and A. Trognon, 1984, Pseudo maximum likelihood methods: Theory, *Econometrica* 52, 681–700.

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., J. C. Heaton, and E. G. J. Luttmer, 1995, Econometric evaluation of asset pricing models, *Review of Financial Studies* 8, 237–274.

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.

Hansen, L. P., and T. J. Sargent, 2001, Robust control and model uncertainty, *American Economic Review* 91, 60–66.

Hansen, L. P., and T. J. Sargent, 2008, *Robustness*, Princeton University Press, Princeton, NJ.

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 P. C. B. Phillips, 1982, On the behavior of inconsistent instrumental variable estimators, *Journal of Econometrics* 19, 183–201.

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.

Monfort, A., 1996, A reappraisal of misspecified econometric models, *Econometric Theory* 12, 597–619.

Patton, A., 2018, Comparing possibly misspecified forecasts, Working Paper.

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.

Stock, J. H., and M. W. Watson, 2007, Why has inflation become harder to forecast?, *Journal of Money, Credit and Banking* 39, 3–33.

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.

White, H., 1982, Maximum likelihood estimation of misspecified models, *Econometrica* 50, 1–25.

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

Yang, Y., 2001, Adaptive regression by mixing, *Journal of the American Statistical Association* 96, 574–588.

Yang, Y., 2003, Regression with multiple candidate models: Selecting or mixing?, *Statistica Sinica* 13, 783–809.

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.

	CAPM	CCAPM	EZ	D-CCAPM	FF3	HJD agg.	HEL agg.
25 Fama-French + 17 industry portfolios							
$\hat{\delta}$	0.4605	0.4678	0.4401	0.4436	0.4478	0.4405	0.4367
$\hat{w}_{-1/2}$	0.0253	0.1788	0.7943	0.0017			
25 Fama-French portfolios							
$\hat{\delta}$	0.3571	0.3991	0.3653	0.3682	0.3383	0.3468	0.3428
$\hat{w}_{-1/2}$	0.1432	0.0132	0.0928	0.7508			
17 industry portfolios							
$\hat{\delta}$	0.1450	0.1436	0.1421	0.1466	0.1322	0.1385	0.1373
$\hat{w}_{-1/2}$	0.0749	0.0025	0.0020	0.9206			

Notes: This table reports the estimates for the Hansen-Jagannathan distance $\hat{\delta}$ for individual models and two aggregators: HJD agg. with aggregation weights that are inverses of the $\hat{\delta}_i$'s (scaled to sum up to one) and HEL agg. with aggregation weights $\hat{w}_{-1/2}$ obtained by minimizing the Hellinger distance between the densities of the aggregator and a benchmark (constant SDF) model.

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

Case (i): all models are misspecified.

	CAPM	CCAPM	EZ	D-CCAPM	HEL agg.
mean $\hat{\delta}$	0.3203	0.3315	0.3295	0.3393	0.3026
median $\hat{\delta}$	0.3217	0.3337	0.3333	0.3504	0.3042
10% quant. $\hat{\delta}$	0.2462	0.2540	0.2510	0.2560	0.2247
90% quant. $\hat{\delta}$	0.4002	0.4224	0.4258	0.4483	0.3849
mean $\hat{w}_{-1/2}$	0.0929	0.2657	0.0991	0.5423	

Notes: This table reports the Monte Carlo estimates for the Hansen-Jagannathan distance $\hat{\delta}$ (mean, median, 10% quantile, and 90% quantile), 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 pivot (constant SDF model). The sample size is 200 and the number of Monte Carlo simulations is 1,000.

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

Case (ii): D-CCAPM is correctly specified.

	CAPM	CCAPM	EZ	D-CCAPM	HEL agg.
mean $\hat{\delta}$	0.2826	0.2898	0.2893	0.2981	0.2666
median $\hat{\delta}$	0.2834	0.3000	0.2983	0.3172	0.2680
10% quant. $\hat{\delta}$	0.2165	0.2203	0.2169	0.2200	0.1991
90% quant. $\hat{\delta}$	0.3557	0.3735	0.3725	0.3912	0.3421
mean $\hat{w}_{-1/2}$	0.1002	0.2787	0.1007	0.5205	

Notes: This table reports the Monte Carlo estimates for the Hansen-Jagannathan distance $\hat{\delta}$ (mean, median, 10% quantile, and 90% quantile), 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 pivot (constant SDF model). The sample size is 200 and the number of Monte Carlo simulations is 1,000.

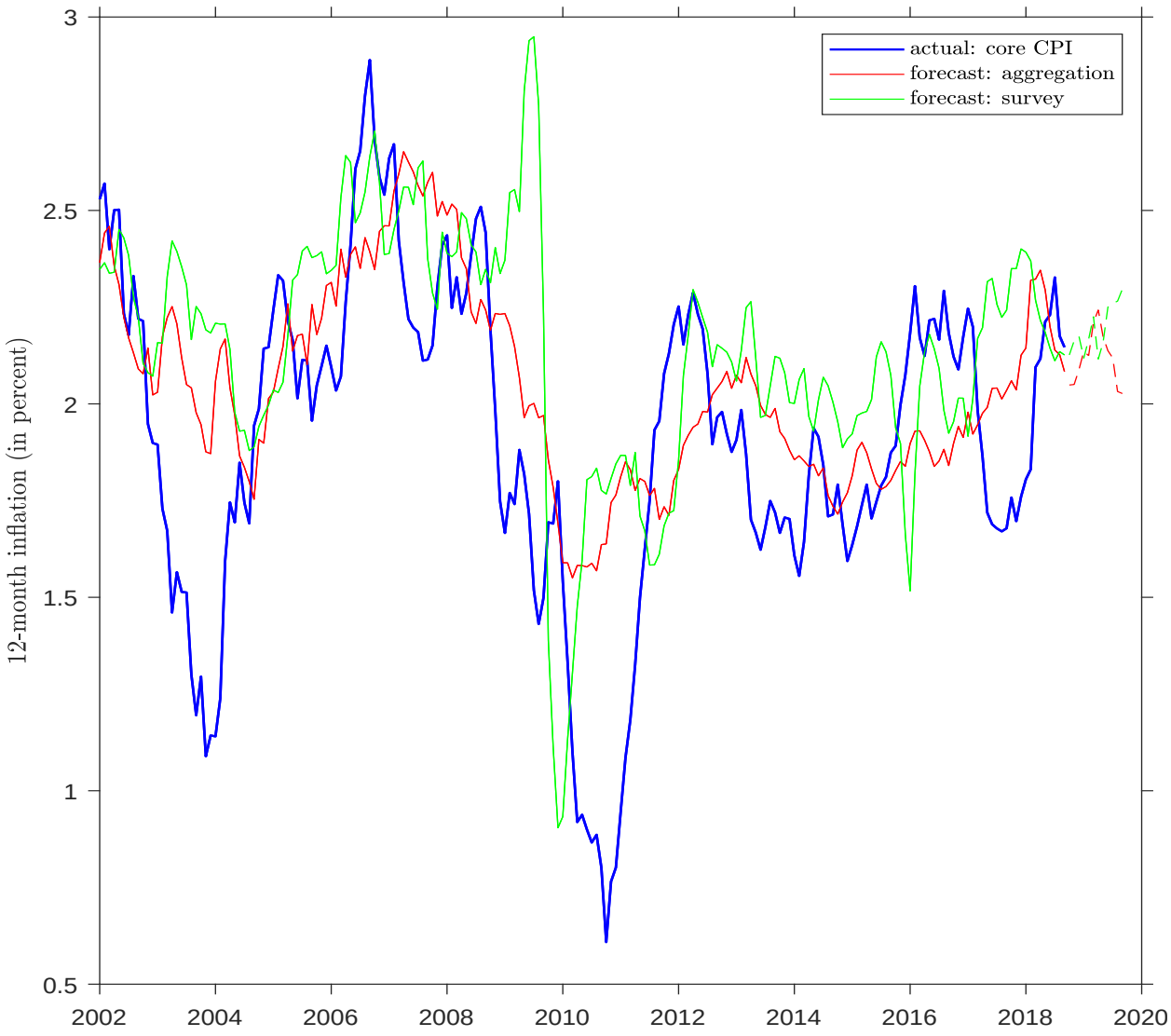

Figure 1: 12-month ahead, out-of-sample, forecasts of core inflation from the Blue Chip survey and the proposed aggregator.

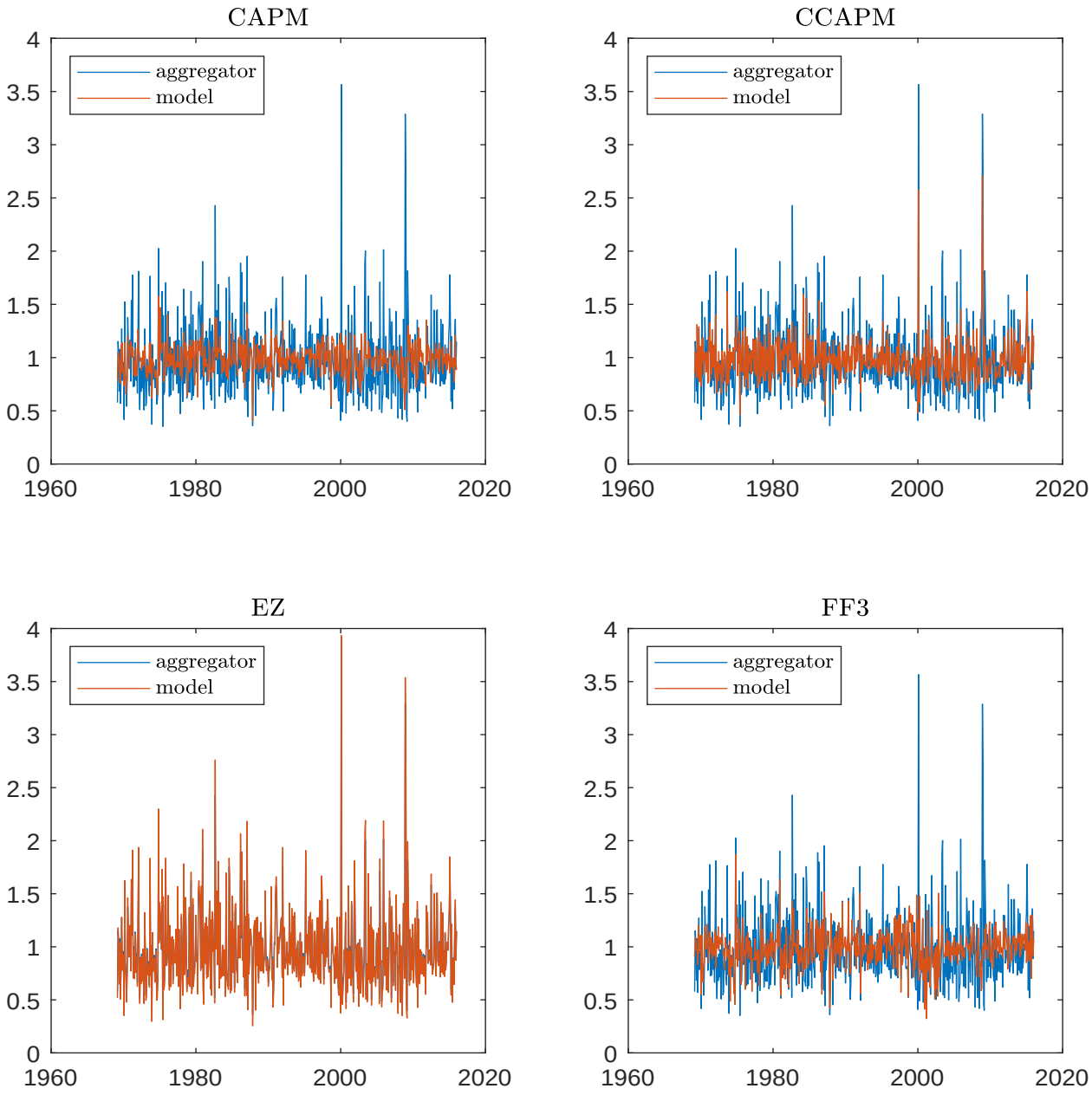

Figure 2: SDFs for individual models and aggregator based on the Hellinger distance for the January 1994 – December 2013 evaluation sample.