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

Nikolay Gospodinov† and Esfandiar Maasoumi‡

October 2019

Abstract

We propose a generalized aggregation approach for model averaging. The entropy-based optimal criterion is a natural choice for aggregating information from many "globally" misspecified models as it adapts better to the underlying model uncertainty and obtains more robust approximations. Unlike almost all other approaches in the existing literature, we do not require a "reference model", or a true data generation process/model contained in the set of models - neither implicitly nor in otherwise popular limiting forms. This shift in paradigm 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. Our generalized aggregators relax the common perfect substitutability of the candidate models, implicit in linear averaging and pooling. The aggregation weights are data-driven and obtained from a proper (Hellinger) distance measure. The empirical results illustrate the performance and the economic significance of the aggregation approach in the context of stochastic discount factor models and inflation forecasting.

JEL Classification: C13, C52, G12.

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

*We would like to thank the Editor (Viktor Todorov), two referees, 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), CIREQ Econometrics Conference on "Recent Advances in the Method of Moments", and the 2019 Winter Meeting of the Econometric Society 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 "global" model misspecification 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; Gourieroux, 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; Hall and Inoue, 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. The approach in the literature on model ambiguity is also different as it assumes "local" misspecification around a given "reference model".

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. This is appropriate and productive when an aggregate object serves for optimal inference

about outcomes. Some traditional objects of inference, especially partial effects, require far more careful reconsideration when all models are globally 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 a recent comprehensive review of model averaging methods, see Steel, 2018). 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. 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 globally 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] 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 acts effectively as a "dimension" reduction device. This is not possible with other non-metric measures of "divergence" from information theory.

The stochastic discount factor (SDF) framework for asset pricing provides an evidently suitable

[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).

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.[3] 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, 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 hyperbolic mean form with generalized geometric and linear aggregation schemes as special cases. The generalized entropy criterion that underlies our approach allows us to circumvent two serious drawbacks of the standard weighted averages. 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, the use of a hyperbolic 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

[3] 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.

improvements 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

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

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 1988:01–2019:07 with the subsample 2002:01-2019:07 used 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) and can be summarized as follows. The forecast improvements of the aggregate over the individual model forecasts was substantial across all loss functions (in excess of 60% over the best performing (convenience yield) model). The aggregate forecast is unbiased (in a Mincer-Zernowitz regression) 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 has the following structure. 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 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 of the statistical paradigm from parameter estimation to a "stochastic optimization" paradigm that is detailed below. In what follows, we assume that $f(\cdot) > 0$.

Suppose that there exists a finite list (dictionary) $\mathcal{F}$ of candidate models that is aimed to embed certain theoretical or empirical features of the underlying DGP.[4] 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 even a true DGP, were it to be in the class. 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 $f(\cdot)$.

To fix notation and main ideas, suppose that the unknown $f(\cdot)$ is approximated by M functions in the dictionary $\mathcal{F} = \{f_1, ..., f_M\}$ from a sample $Z_1, ..., Z_T$ such that $\max_{f_i \in \mathcal{F}} \|f/f_i\|_\infty < \infty$.[5] Consider the flat simplex for a set of weights $w = (w_1, ..., w_M)$:

$$\mathcal{W}^M = \left\{ w \in \mathbb{R}^M : w_i \geq 0, \sum\nolimits_{i=1}^{M} w_i = 1 \right\}. \tag{2}$$

For a given risk function $\mathcal{R} : \mathcal{F} \to \mathbb{R}$, the pseudo-true aggregator of the candidates $\{f_1, ..., f_M\}$ is defined as

$$f_w^* = \underset{w \in \mathcal{W}^M}{\operatorname{argmin}} \mathcal{R}(f^{(w)}, f). \tag{3}$$

[4] The assumption that the number of candidate models is finite is made for convenience. In fact, it may be desirable – for constructing a better approximation of the comprehensive model – to allow for the number of models in this class to expand as a function of the sample size. This possible extension is beyond the scope of this paper.

[5] The functions f_i $(i = 1, ..., M)$, evaluated at the sample values $Z_1, ..., Z_T$, are either given or estimated with prior data samples.

The approach outlined below offers generality with respect to the risk function $\mathcal{R}(f^{(w)}, f)$. The sample aggregator, denoted by $\tilde{f}^{(w)}$, is constructed by mimicking the pseudo-true aggregator. Its performance is evaluated using the empirical risk function $\mathcal{R}_T(\tilde{f}^{(w)}, f)$ and "oracle" inequalities are established relative to $\mathcal{R}(f^{(w)}, f)$ both in terms of expectations and probability.

More specifically, for some constant $C \geq 1$,

$$E[\mathcal{R}_T(\tilde{f}^{(w)}, f)] \leq C \min_{w \in \mathcal{W}^M} \mathcal{R}(f^{(w)}, f) + \triangle_{T,M} \tag{4}$$

and for every $\delta > 0$,

$$\Pr\left\{\mathcal{R}_T(\tilde{f}^{(w)}, f) \leq C \min_{w \in \mathcal{W}^M} \mathcal{R}(f^{(w)}, f) + \triangle_{T,M,\delta}\right\} \geq 1 - \delta, \tag{5}$$

where $\triangle_{T,M}$ and $\triangle_{T,M,\delta}$ are remainder terms that do not depend on f or f_i, $i = 1, ..., M$. More generally, a balanced oracle inequality takes the form

$$E[\mathcal{R}_T(\tilde{f}^{(w)}, f)] \leq C \left[\min_{w \in \mathcal{W}^M} \mathcal{R}(f^{(w)}, f) + \triangle_{T,M,w}\right], \tag{6}$$

with $\triangle_{T,M} = C \max_{w \in \mathcal{W}^M} \triangle_{T,M,w}$. An exact or sharp oracle inequality is obtained when $C = 1$.

A few remarks are warranted here. First, $\min_{w \in \mathcal{W}^M} \mathcal{R}(f^{(w)}, f) > 0$ 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; see also Section 2.3). 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. Furthermore, to minimize the magnitude of the remainder term in bounding the empirical risk, one could resort to penalized convex aggregation (Rigollet and Tsybakov, 2012) that imposes penalties for departures from a priori or desired distributions of weights reflecting, for example, an ordering of the models. Finally, 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' with an oracle inequality remainder term for model selection being of higher order than the desired minimax rate (Rigollet and Tsybakov, 2012).

2.2 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. As before, 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) proposed generalizing the pairwise criteria of divergence to a general multivariate context, as follows:

$$\tilde{D}_\rho(\tilde{f}, f; w) = \sum_{i=1}^{M} w_i \mathcal{R}_{T,\rho}(\tilde{f}, f_i), \tag{7}$$

where

$$\mathcal{R}_{T,\rho}(\tilde{f}, f_i) = \frac{1}{\rho(\rho+1)} \sum_{t=1}^{T} \tilde{f}_t \left[\left(\frac{\tilde{f}_t}{f_{i,t}} \right)^\rho - 1 \right]. \tag{8}$$

$\mathcal{R}_{T,\rho}(\tilde{f}, f_i)$ 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

$$\tilde{f}_t^{(w)} \propto \left[\sum_{i=1}^{M} w_i f_{i,t}^{-\rho} \right]^{-1/\rho}. \tag{9}$$

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 ($\rho = -1$). The case $\rho = -1/2$ corresponds to the Hellinger distance aggregator.

The literature on stochastic optimization reveals several insights. First, the risk of an aggregator function dominates that of model selection in terms of oracle bounds on expected loss. Second, the commonly used L_2 risk function has bounds that depend on the 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 the dominating measure, thus is a proper measure of distance between distributions, as well as suitable regression 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, and $P \ll Q$ denote absolute continuity of P relative to Q. The generalized entropy or Cressie-Read divergence from Q to P is given by

$$D_\eta(P, Q) = \int \phi_\eta \left(dP/dQ \right) dQ, \tag{10}$$

if $P \ll Q$ and $+\infty$ otherwise, where

$$\phi_\eta(x) = \frac{1}{\eta(\eta+1)}\left(x^{\eta+1} - 1\right) \tag{11}$$

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

$$D_\eta(P,Q) = \int \left(1 - \left(\frac{p}{q}\right)^\eta\right) q d\nu \text{ for } \eta \in \mathbb{R}. \tag{12}$$

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

$$D_0(P,Q) = \int \ln\left(\frac{p}{q}\right) q d\nu = \mathcal{KL}(P,Q). \tag{13}$$

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

$$D_{-1/2}(P,Q) = \int \left(p^{1/2} - q^{1/2}\right)^2 d\nu = \mathcal{H}(P,Q). \tag{14}$$

It is easy to show that the Hellinger distance[6] is bounded from above by the Kullback-Leibler divergence measure, i.e., $\mathcal{H}(P,Q) \leq \mathcal{KL}(P,Q)$.

Given a dictionary $\{f_1, ..., f_M\}$, the aggregator based on the Hellinger distance is given by $\tilde{f}^{(w)} = \left[\sum_{i=1}^M w_i f_i^{1/2}\right]^2$ 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 with $\mathcal{H}(\tilde{f}^{(w)}, f) > 0$ under model misspecification. Under certain conditions, the minimax risk, defined as $\inf_{\tilde{f}^{(w)}} \sup_{f \in \mathcal{F}} \mathcal{H}(\tilde{f}^{(w)}, f)$ over $\mathcal{F}$, is bounded (Birgé, 2006) and remains under control even if the models are misspecified.

2.3 Alternative Approaches

Convex aggregation of the candidates $\{f_1, ..., f_M\}$, given by

$$\tilde{f}^{(w)} = \sum_{i=1}^M w_i f_i,\ w \in \mathcal{W}^M, \tag{15}$$

has been the dominant framework for combining information from different models. As pointed out above, the convex aggregator is obtained from our general aggregator for a particular choice

[6] In this paper we use the Hellinger distance as an aggregator and not as an estimation criterion. Antoine and Dovonon (2018) show that while the Hellinger distance is robust to perturbations of probability measures under local misspecification (Kitamura, Otsu, and Evdokimov, 2013), it is not robust under global misspecification. For estimation purposes, we use the GMM estimator which is robust to global misspecification (Hall and Inoue, 2003; Gospodinov, Kan and Robotti, 2013) and invoke the appealing properties of the Hellinger distance for model aggregation only.

of the parameter ρ. Furthermore, model selection is a special case of convex aggregation with $w \equiv e_i = (0, 0, ..., 1, 0, ..., 0)$ with $i = 1, ..., M$.

For example, when the density properties of the w are recognized, one may incorporate penalties for departures from a priori or desired distributions of weights that may reflect an ordering of the models. With the Kullback-Leibler divergence as a penalty function, the aggregation weights take an exponential form (Claeskens and Hjort, 2008; Rigollet and Tsybakov, 2012)

$$w_i^* = \frac{\exp(-T\mathcal{R}_T(\tilde{f}^{(w)}, f)/\beta)\pi_i}{\sum_{j=1}^{M} \exp(-T\mathcal{R}_T(\tilde{f}^{(w)}, f)/\beta)\pi_j}, \tag{16}$$

where $\beta > 0$ is a penalty parameter, and $\pi \in \mathcal{W}^M$ is a prior probability density. Given the closed form of the above expression, this proves to be a convenient device when M is large relative to T, as in variable selection problems with "big data" attributes.

While still a special case of our aggregator, this approach can be applied to general risk functions and possibly non-linear models. One could obtain sharper results for aggregating linear models under more specific choices of a risk function. A large literature has now emerged for model averaging procedures and estimators with optimal weights determined by minimizing Mallows' criterion (Hansen, 2007), mean squared error (Liang, Zou, Wan and Zhang, 2011), jackknife or cross-validation criterion (Hansen and Racine, 2012), Kullback-Leibler distance (Zhang, Zou and Carroll, 2015), etc. We would like to stress that these model averaging procedures impose strong conditions on the structure and degree of potential misspecification of the models. Furthermore, sharp optimality results under these criteria are often established only within the context of linear regression models.

It is helpful to emphasize the contrast between our model averaging approach and the "model ambiguity" framework. In the model ambiguity framework, a reference model plays a central role and the unknown true model belongs to a neighborhood of the approximate reference model. For example, suppose that the true structural model contains unobservables with a distribution P which is allowed to vary over a neighborhood of size ϵ of a reference distribution P^* while other structural features of the model structure are maintained. This small neighborhood is of small order of magnitude in the sample size, defining a "local" misspecification of the reference model.[7] More formally, consider a divergence measure $\phi_\eta(x)$ and an ϵ-neighborhood defined as

$$\mathcal{N}_\epsilon = \{P^* \in \mathcal{P} : D_\eta(P, P^*) \leq \epsilon\}, \tag{17}$$

[7] For a comprehensive discussion, 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 (2019).

where $D_\eta(P, P^*) = \int \phi_\eta\left(dP/dP^*\right) dP^*$ if $P \ll P^*$ and $+\infty$ otherwise, and $\phi_\eta(x)$ is given in eq. (11). While Kullback-Leibler $\phi_0\left(dP/dP^*\right) = \ln\left(dP/dP^*\right)$ is a popular choice of defining this neighborhood, other choices such as Pearson's χ^2 divergence also provide analytical tractability in expressing the degree of model misspecification or ambiguity in an "ϵ-ball" of other undefined models/laws. 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 – say $h_L(\mathcal{N}_\epsilon)$ and $h_U(\mathcal{N}_\epsilon)$ over $P \in \mathcal{N}_\epsilon$ – when the ϵ-deviation is of a certain small order of magnitude in the sample size. Unlike our model aggregation, there are no expressedly defined models competing unambiguously with the reference model.

We close this section by noting that the Bayesian approach to model averaging commonly assumes that the "true model" is contained under the support of the prior.[8] Similarly, the frequentist approach typically allows only for a shrinking (with the sample size) deviation from the true model; i.e., the model is "locally" misspecified. While appropriate in some contexts, this dominant mode of analysis is driven largely by analytical convenience and may be an actual impediment in dealing with the case when all models are "globally" misspecified.[9]

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 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 of various theory-based models that would inform policy makers and investors.

[8] If one interprets the prior as providing probabilistic weights of various parametric states, the Bayesian model averaging fits naturally into the conceptual structure of the model aggregation framework.

[9] In the context of asset pricing models below, we define a model to be globally misspecified if there exists no value for the parameter vector, indexing this model, that sets all pricing errors to zero.

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,

$$E[Rm] = a, \tag{18}$$

where a denotes a non-zero $N \times 1$ vector of payoffs. In the case when R are gross returns, $a = 1_N$, where 1_N is a vector of ones. 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:

$$e(\gamma) = E[Ry(\gamma)] - a = 0_N. \tag{19}$$

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

$$\delta = \min_{\gamma \in \Gamma} \min_{m \in \mathcal{M}} E[(y(\gamma) - m)^2] \tag{20}$$

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:

$$\delta = \min_{\gamma \in \Gamma} \max_{\lambda \in \Re^N} E[y(\gamma)^2 - (y(\gamma) - \lambda' R)^2] - 2\lambda' a, \tag{21}$$

where λ is an $N \times 1$ vector of Lagrange multipliers.[10] The term $\lambda' R$ provides the smallest correction, in mean squared sense, to $y(\gamma)$ in order to make it an admissible SDF. Importantly, Hansen and Jagannathan (1991) demonstrate that δ 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 as well. Let P be the data generating measure and Φ denote a family of probability measures that satisfy the asset

[10] 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

$$\min_{Q \in \Phi} D_\eta(P, Q) = \int \phi_\eta \left(dQ/dP \right) dQ \tag{22}$$

$$\text{subject to } \int e(\gamma) dQ = 0_N, \tag{23}$$

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

$$\delta_\eta(\gamma) = \min_{\gamma \in \Gamma} \min_{m \in \mathcal{M}} E\left[\frac{(1 + m - y(\gamma))^{\eta+1} - 1}{\eta(\eta+1)} \right] \tag{24}$$

and

$$\delta_\eta(\gamma) = \max_{\lambda \in \Re^N} \lambda' a - E\left[\frac{(\eta \lambda' R)^{\frac{\eta+1}{\eta}}}{\eta+1} + (y(\gamma) - 1) \lambda' R + \frac{1}{\eta(\eta+1)} \right], \tag{25}$$

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 entropy divergence measures 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. 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 properties, in estimating and aggregating the individual models.

It is useful to highlight an important aspect of the aggregation approach. The asset pricing restrictions typically hold in a conditioning setup. This requires the estimation and evaluation to be performed either by operating directly on the conditional moment restrictions implied by

economic theory or by augmenting the unconditional moment restrictions with conditioning information via scaled factors and returns. The aggregation approach attempts essentially to provide an approximate comprehensive model that encompasses, in a flexible way, the conditional asset-pricing restrictions. An additional appealing feature of the aggregation approach is that it can accommodate situations where different asset pricing models hold over nonoverlapping periods in the sample. Time-varying aggregation weights and cross-validation methods, discussed below, prove to be convenient tools for dealing with various forms of regime switching and structural instability.

3.2 Cross-Validation Inference and Aggregation

Let $U = E[RR']$ and assume that U is nonsingular so that none of the test assets is redundant. 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

$$\lambda = U^{-1}e(\gamma), \tag{26}$$

and

$$\delta(\gamma) = e(\gamma)'U^{-1}e(\gamma). \tag{27}$$

The estimators of γ and λ are then defined as

$$\hat{\gamma} = \arg\min_{\gamma\in\Gamma} \hat{e}(\gamma)'\hat{U}^{-1}\hat{e}(\gamma), \tag{28}$$

where $\hat{e}(\gamma) = \frac{1}{T}\sum_{t=1}^{T} R_t y_t(\gamma) - z$ and $\hat{U} = \frac{1}{T}\sum_{t=1}^{T} R_t R_t'$, and $\hat{\lambda} = \hat{U}^{-1}\hat{e}(\hat{\gamma})$. The conditions for the consistency of these estimates are outlined in Appendix A.1. Note that this framework is general enough to accommodate both linear and nonlinear SDF specifications.

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. 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.[11]

[11] Alternatively, one could estimate the aggregation weights by sample splitting instead of cross-validation. An earlier version of the paper explored this possibility but the results are not reported here to conserve space.

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

$$\hat{\gamma}_{i,(-t)} = \underset{\gamma_i\in\Gamma}{\arg\min}\ \hat{e}_{i,(-t)}(\gamma_i)'\hat{U}_{(-t)}^{-1}\hat{e}_{i,(-t)}(\gamma_i). \tag{29}$$

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

$$\hat{\delta}_{i,(-t)} = \hat{e}_{i,(-t)}(\hat{\gamma}_{i,(-t)})'\hat{U}_{(-t)}^{-1}\hat{e}_{i,(-t)}(\hat{\gamma}_{i,(-t)}). \tag{30}$$

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 $\delta > 0$,

$$\sqrt{T}(\hat{\delta}_{i,(-t)} - \delta_i) \xrightarrow{d} N(0, V_{\delta_i}), \tag{31}$$

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

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

$$\sqrt{T}(\hat{\delta}_{i,(-t)} - \hat{\delta}_{j,(-t)}) \xrightarrow{d} N(0, V_{\delta_i j}), \tag{32}$$

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 (32) holds only if $V_{\delta_i j} \neq 0$.

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.[12] As a related issue, since all candidate models are treated as incomplete 'indicators' of the latent DGP, choosing only one will result in

[12] In a quasi-likelihood framework, some recent papers (Schennach and Wilhelm, 2017; Liao and Shi, 2019) propose procedures that control asymptotic size uniformly without explicit knowledge of the structure of the competing models.

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. Instead, we resort to a model averaging rule that would aggregate information from all of these models and construct a pseudo-true SDF $\tilde{y}$.[13]

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 with $\hat{e}_{(-t)} = (\hat{e}_{1,(-t)}, ..., \hat{e}_{M,(-t)})'$ being the vector of leave-one-out pricing errors. The generalized aggregator that minimizes

$$D_\rho(\tilde{y}, Y; w) = \sum_{i=1}^{M} w_i \left\{ \sum_{t=1}^{T-1} \tilde{y}_{(-t)} \left[\left(\frac{\tilde{y}_{(-t)}}{y_{i,(-t)}} \right)^\rho - 1 \right] \Big/ \rho(\rho+1) \right\}. \tag{33}$$

again takes the form

$$\tilde{y}^{(w)}_{(-t)} \propto \left[\sum_{i=1}^{M} w_i y_{i,(-t)}^{-\rho} \right]^{-1/\rho}. \tag{34}$$

Appendix A.2 provides a bound for the linear aggregator ($\rho = -1$). While deriving an explicit bound for the general case is much more involved, the result above suggests that, for a given set of weights w, the bound for the linear aggregator also serves as a bound for the optimal aggregator.

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 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 $\tilde{e}_{(-t)}$ of the Hellinger distance aggregator $\tilde{y}^{(w)}_{(-t)}$. Using the generalized entropy (Cressie-Read) divergence from Q to P defined in (10)-(11) and imposing $\eta = -1/2$, we obtain the scaled Hellinger distance $\mathcal{H} \propto D_{-1/2}(P, Q)$

$$\mathcal{H} = \frac{1}{2} \int \left(p^{1/2}(x) - q^{1/2}(x) \right)^2 dx. \tag{35}$$

In practical implementation, we estimate p and q using a kernel density estimator and the integral in (35) is evaluated numerically. The model weights w are then obtained by minimizing $\mathcal{H}$ with respect to w, subject to the relevant restrictions. The choice of a benchmark model is discussed in the next section.

[13] For earlier applications of model averaging in asset pricing, see Pastor and Stambaugh (2000) and Avramov (2002), among others.

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):

$$y_t^{CAPM}(\alpha, \beta) = \beta(1-k)^{-\alpha} R_{m,t}^{-\alpha} \tag{36}$$

or

$$\ln(y_t^{CAPM}(\gamma)) = \gamma_0 + \gamma_1 \ln(R_{m,t}), \tag{37}$$

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):

$$y_t^{CCAPM}(\alpha, \beta) = \beta \left(\frac{C_t}{C_{t-1}}\right)^{-\alpha} \tag{38}$$

or

$$\ln(y_t^{CCAPM}(\gamma)) = \gamma_0 + \gamma_1 c_t, \tag{39}$$

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):

$$y_t^{EZ}(\alpha, \beta, \sigma) = \beta^{\frac{1-\alpha}{1-\sigma}} \left(\frac{C_t}{C_{t-1}}\right)^{-\sigma\left(\frac{1-\alpha}{1-\sigma}\right)} R_{m,t}^{\frac{\sigma-\alpha}{1-\sigma}}, \tag{40}$$

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

$$\ln(y_t^{EZ}(\gamma)) = \gamma_0 + \gamma_1 c_t + \gamma_2 \ln(R_{m,t}), \tag{41}$$

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):

$$y_t^{D-CAPM}(\alpha,\beta,\sigma,\phi) = \beta^{\frac{1-\alpha}{1-\sigma}} \left(\frac{C_t}{C_{t-1}}\right)^{-\sigma\left(\frac{1-\alpha}{1-\sigma}\right)} \left(\frac{C_{d,t}/C_t}{C_{d,t-1}/C_{t-1}}\right)^{\phi(1-\alpha)} R_{m,t}^{\frac{\sigma-\alpha}{1-\sigma}}, \tag{42}$$

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

$$\ln(y_t^{D-CAPM}(\gamma)) = \gamma_0 + \gamma_1 c_t + \gamma_2 c_{d,t} + \gamma_3 \ln(R_{m,t}), \tag{43}$$

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 $\alpha = \sigma$).[14]

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. To assess the robustness of the aggregator to the choice of a pivot, we consider as an alternative pivot the three-factor (FF3) model of Fama and French (1993)

$$y_t^{FF3}(\gamma) = \gamma_0 + \gamma_1 r_{m,t} + \gamma_2 smb_t + \gamma_3 hml_t, \tag{44}$$

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).[15] The FF3 model is one of the most successful empirical models, especially for some of the test assets that we consider. It would be interesting to see how the substantial differences in the individual pricing performance of the FF3 model and constant SDF would affect the properties of the aggregator.

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

[14] As suggested by a referee, we can further enlarge the set of models with reduced-form factor models with a large number of factors and characteristics (see, for example, Kozak, Nagel and Santosh, 2018, 2019).

[15] Another candidate for a benchmark model would be the non-parametric estimate of a comprehensive model. Such a model is exemplified in Cai, Ren, and Sun (2015).

website.[16] 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.

4.2 Results

The parameters for each model and the aggregation weights for the Hellinger aggregator are estimated as described in Section 3.2. We consider two aggregators based on the Hellinger distance depending on whether the choice of a pivot is the constant SDF ("HEL ag1") or the FF3 model ("HEL ag2"). For the sake of comparison, we also report an equal-weight aggregator ("EW ag") with weights $1/M$ assign to each model. All models are then evaluated using the cross-validation version of the Hansen-Jagannathan distance. The choice of the Hansen-Jagannathan distance for model evaluation and comparison is dictated by its appealing economic interpretation and our desire to ensure consistency with the existing literature. Also, because the cross-validation has the flavor of "out-of-sample" evaluation, this criterion tends to penalize models that are exhibit structural instability. 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.[17]

Table 1 reports the values of the Hansen-Jagannathan (HJ) distances of the four consumption-based asset pricing models, the two benchmarks (constant SDF, denoted by "Ben 1", and FF3, denoted by "Ben 2") model and the three aggregators: "EW ag", "HEL ag1" and "HEL ag2". The table also presents the aggregation weights for the two Hellinger distance aggregators ($\hat{w}^{(1)}_{-1/2}$ and $\hat{w}^{(2)}_{-1/2}$). The resulting SDF aggregators are used for computing the corresponding HJ distance.

[16] Additional results for bond portfolios, equity options and currency returns as test assets are available from the authors upon request.

[17] 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.

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.

Table 1 about here

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 approach offers pricing improvements and is close to the best-performing individual model.[18] Non-trivial reduction of pricing errors is observed even with the constant SDF model as a pivot for estimating the aggregation weights. The Hellinger aggregator with the FF3 model as a pivot further reduces the pricing errors, especially for the 25 Fama-French portfolios, although these reductions are relatively modest. The improvements are achieved by assigning a larger weight to the best-performing model.[19] This, of course, may not be desirable if the dominant model or the pivot are characterized by parameter instability over time. By contrast, the less informative pivot implies more "hedging" by robustifying away from the best model and distributing the weights more evenly across "finitely substitutable" models. The result that the performance of the aggregator is fairly robust to the choice of a pivot is reassuring since in general practice, the properties of the models and test assets may be unknown or indeterminate. The equal-weight aggregator (EW ag) also delivers improvements over individual models but it is dominated by the Hellinger aggregators. 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 aggregator (with the constant SDF pivot) 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 and D-CCAPM models, it adapts to the volatility of the SDF for these models.

Figure 2 about here

[18] 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.

[19] This sparsity of the aggregation scheme, where the shrinkage is done towards the best-performing model, may prove to be beneficial when the set of candidate models is large.

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. 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, 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

$$\mu = E[Y_t] = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{45}$$

and

$$V = \text{Var}[Y_t] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}. \tag{46}$$

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

$$\begin{bmatrix} f_t \\ r_t \end{bmatrix} \sim t_8 \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \right), \tag{47}$$

where t_8 denotes a multivariate t-distribution with 8 degrees of freedom. 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

$$\begin{aligned} e(\gamma) &= E[R_t y_t(\gamma)] - 1_N = E[\exp(r_t + \gamma_0 + \gamma_1' f_t)] - 1_N \\ &= \exp\left(\gamma_0 + \mu_2 + 0.5\gamma_1' V_{11}\gamma_1 + V_{21}\gamma_1 + 0.5\text{Diag}(V_{22})\right) - 1_N. \end{aligned} \tag{48}$$

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

$$\mu_2 = -0.5\text{Diag}(V_{22}) - \left(\gamma_0 + 0.5\gamma_1' V_{11}\gamma_1\right) 1_N - V_{21}\gamma_1. \tag{49}$$

Thus, we can set the mean of the simulated returns μ_2 as in (49) 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 two Hellinger distance aggregators (HEL ag1 and HEL ag2) based on the two pivots, constant SDF and FF3 model, respectively. 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 results in pricing improvements and robust performance. Despite the mismatch between the risk functions for aggregation and pricing performance evaluation, the Hellinger aggregator produces pricing errors that are close to the best-performing and most comprehensive model (D-CCAPM). While both versions of the aggregator loads mostly on the model with smallest HJ distance, it also assigns non-trivial weights to the more parsimoniously parameterized models. As in the empirical example, the use of a substantially better pricing model as a pivot (FF3) offers only minimal improvements which confirms the relative insensitivity of the aggregator to the choice of a benchmark model.

Table 3 about here

The results are similar when one of the models is true (Table 3). The aggregator 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 more general entropy measures and data structures is a promising venue for future research.

Appendix

A.1 Parameter Estimation

In this section, we outline some regularity conditions to ensure the stochastic equicontinuity of the sample HJ-distance and the consistency of $\hat{\theta} = [\hat{\gamma}', \hat{\lambda}']'$ for its corresponding pseudo-true value θ^*. Let

$$\varphi_t(\theta) \equiv y_t(\gamma)^2 - (y_t(\gamma) - \lambda' R_t)^2 - 2\lambda' a. \tag{A.1}$$

Assumption A. *Assume that*

(i) $\varphi_t(\theta)$ *is* m*-dependent,*

(ii) *the parameter space* Θ *is compact,*

(iii) $\varphi_t(\theta)$ *is continuous in* $\theta \in \Theta$ *almost surely,*

(iv) $|\varphi_t(\theta_1) - \varphi_t(\theta_2)| \leq A_t |\theta_1 - \theta_2| \ \forall \ \theta_1, \theta_2 \in \Theta$*, where* A_t *is a bounded random variable that satisfies* $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E[|A_t|^{2+\omega}] < \infty$ *for some* $\omega > 0$,

(v) $\sup_{\theta\in\Theta} E[|\varphi_t(\theta)|^{2+\omega}] < \infty$ *for some* $\omega > 0$,

(vi) *the population dual problem* $\arg\min_{\gamma\in\Gamma} \max_{\lambda\in\Re^N} E[\varphi_t(\theta)]$ *has a unique solution* θ^* *which is in the interior of* Θ.

Assumptions A(i)–A(v) ensure the stochastic equicontinuity of $\varphi_t(\theta)$ (see Andrews, 1994) and imply that

$$\sup_{\theta\in\Theta} \left| \frac{1}{T} \sum_{t=1}^{T} \varphi_t(\theta) - E[\varphi_t(\theta)] \right| \xrightarrow{p} 0. \tag{A.2}$$

The uniform convergence in (A.2) along with the identification condition in Assumption A(vi) deliver the consistency of $\hat{\theta} = [\hat{\gamma}', \hat{\lambda}']'$ for θ^* in an individual SDF model.[20]

A.2 Aggregation Bounds

This section considers the convex SDF aggregator and shows that its Hellinger distance risk is bounded. Let $U^{-1/2}a = x_j$ for $j = 1, ..., N$, and $U^{-1/2}E[Ry(\gamma)] = \mu_j(\gamma)$, and note that γ can be

[20] Alternatively, one could estimate the parameters using other entropy-based estimators (Almeida and Garcia, 2012), including the Hellinger distance.

obtained equivalently from the nonlinear cross-asset regression

$$X_j = \mu_j(\gamma, Z_j) + \epsilon_j, \tag{A.3}$$

where ϵ_j are the scaled average pricing errors with variance $\sigma^2 = \text{Var}(\epsilon_j)$.[21] Under correct model specification and conditional on the data, ϵ_j has mean zero and known density $h\left(\frac{x-\mu(\cdot)}{\sigma}\right)$. If there is uncertainty about the distribution h, multiple candidate distributions can also be accommodated as in Yang (2001). The joint density of the data is denoted by

$$p_{\mu,\sigma} = \frac{1}{\sigma} h\left(\frac{x-\mu(\cdot)}{\sigma}\right) \tag{A.5}$$

with respect to the product measure of v and P_Z.

Suppose that there is a finite dictionary of M models that provide the SDF estimates $\hat{y}_{i,t} = \hat{y}_{i,t}(\hat{\gamma}_i)$, and hence $\hat{\mu}_{i,j} = \hat{\mu}_{i,j}(\hat{\gamma}_i, Z_j)$ for $i = 1, ..., M$, $t = 1, ..., T$ and $j = 1, ..., N$. Define the weights

$$w_i = \frac{\pi_i \prod_{j=1}^N h\left((X_j - \hat{\mu}_{i,j})/\hat{\sigma}_i\right)}{\sum_{k=1}^M \pi_k \prod_{j=1}^N h\left((X_j - \hat{\mu}_{k,j})/\hat{\sigma}_k\right)} \tag{A.6}$$

with π_i denoting a set of prior nonnegative weights that sum up to one ($1/M$, for instance). Then, the aggregator is assumed to take the form

$$\tilde{\mu}_j = \sum_{i=1}^M w_i \hat{\mu}_{i,j}. \tag{A.7}$$

For bounding the risk of the SDF aggregator, we need the following conditions (see Yang, 2001, 2004).

Assumption B. *Assume that*

(i) *there exists a constant* $B_1 > 0$ *such that* $\sup_i \left\|\hat{\mu}_{i,j} - \mu\right\|_\infty \leq B_1\sigma$ *for all* $i \geq 1$,

(ii) *there exist constants* $0 < \xi_1 \leq 1 \leq \xi_2 < \infty$ *such that* $\xi_1 \leq \hat{\sigma}_{i,j}^2/\sigma^2 \leq \xi_2$ *for all* $i \geq 1$ *and* $j \geq 1$,

(iii) *there exists a constant* $B_2 > 0$ *such that, for the pair* $0 < s_0 < 1$ *and* $C > 0$,

$$\int h(x) \ln \frac{h(x)}{(1/s)h((x-c)/s)} \leq B_2\left((1-s)^2 + c^2\right) \tag{A.8}$$

for $s_0 \leq s \leq s_0^{-1}$ *and* $-C < c < C$.

[21] For linear SDFs, $y_t(\gamma) = \gamma' \tilde{f}_t$, $\hat{\gamma}$ is the OLS estimator

$$\hat{\gamma} = (\hat{D}'\hat{U}^{-1}\hat{D})^{-1}\hat{D}'\hat{U}^{-1}1_N, \tag{A.4}$$

where $\hat{U} = \frac{1}{T}\sum_{t=1}^T R_t R_t'$ and $\hat{D} = \frac{1}{T}\sum_{t=1}^T R_t \tilde{f}_t'$.

Conditions (i) and (ii) put some restrictions on the behavior of the scaled pricing errors. Assumption B(iii) allows for certain types of non-Gaussianity of the density of pricing errors.

LEMMA 1. *Let $H(\tilde{\mu}_j, p_{\mu,\sigma})$ denote the risk under the Hellinger distance. Then, under Assumptions A and B, we have*

$$E[\mathcal{H}(p_{\mu,\sigma}, \tilde{\mu}_j)] \leq \inf_i \left\{ \ln(1/\pi_i) + \mathcal{C}(B_1, B_2, \xi_1, \xi_2) \left(E(\hat{\sigma}_i^2 - \sigma^2)^2 + \sum\nolimits_{j=1}^{N} E \left\| \hat{\mu}_{i,j} - \mu \right\|^2 \right) \right\}, \quad \text{(A.9)}$$

where the term $\mathcal{C}(B_1, B_2, \xi_1, \xi_2)$ is a function of the constants in Assumption B.

The proof of Lemma 1 follows similar arguments as in Yang (2001, 2003) and using that $\mathcal{H}(P, Q) \leq \mathcal{KL}(P, Q)$. While this bound is not operational, it can be used to inform the choice of candidate models, all of which are allowed to be misspecified.

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., and P. Dovonon, 2018, Robust estimation with exponentially tilted Hellinger distance, Working Paper, Simon Fraser University.

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

Avramov, D., 2002, Stock return predictability and model uncertainty, *Journal of Financial Economics* 64, 423–458.

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, arXiv:1807.02161.

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.

Christensen, T., and B. Connault, 2019, Counterfactual sensitivity and robustness, Working Paper, arXiv:1904.00989.

Claeskens, G., and N. L. Hjort, 2008, *Model Selection and Model Averaging*, Cambridge University Press, Cambridge, UK.

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, A. R., and A. Inoue, 2003, The large sample behaviour of the generalized method of moments estimator in misspecified models, *Journal of Econometrics* 114, 361–394.

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

Hansen, B. E., 2008, Least squares model averaging, *Econometrica* 75, 1175–1189.

Hansen, B. E., and J. S. Racine, 2012, Jackknife model averaging, *Journal of Econometrics* 167, 38–46.

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.

Kozak, S., S. Nagel, and S. Santosh, 2018, Interpreting factor models, *Journal of Finance* 73, 1183–1223.

Kozak, S., S. Nagel, and S. Santosh, 2019, Shrinking the cross-section, *Journal of Financial Economics*, forthcoming.

Liang, H., G. Zou, A. T. K. Wan, and X. Zhang, 2011, Optimal weight choice for frequentist model average estimators, *Journal of the American Statistical Association* 106, 1053–1066.

Liao, Z., and X. Shi, 2019, A nondegenerate Vuong test and a post selection confidence interval for semi/nonparametric models, Working Paper, UCLA.

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.

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.

Pastor, L., and R. Stambaugh, 2000, Comparing asset pricing models: an investment perspective, *Journal of Financial Economics* 56, 335–381.

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.

Schennach, S. M., and D. Wilhelm, 2017, A simple parametric model selection test, *Journal of the American Statistical Association* 112, 1663–1674.

Steel, M. F. J., 2018, Model averaging and its use in economics, Working Paper, arXiv:1709.08221.

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.

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.

Zhang, X., G. Zou, and R. J. Carroll, 2015, Model averaging based on Kullback-Leibler distance, *Statistica Sinica* 25, 1583–1598.

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

	CAPM	CCAPM	EZ	D-CCAPM	Ben 1	Ben 2	EW ag	HEL ag1	HEL ag2
25 Fama-French + 17 industry portfolios									
$\sqrt{\hat{\delta}}$	0.4569	0.4612	0.4344	0.4344	0.4690	0.4431	0.4388	0.4352	0.4343
$\hat{w}^{(1)}_{-1/2}$	0.1445	0.0077	0.6295	0.2182					
$\hat{w}^{(2)}_{-1/2}$	0.0049	0.0285	0.1982	0.7684					
25 Fama-French portfolios									
$\sqrt{\hat{\delta}}$	0.3540	0.3768	0.3434	0.3410	0.3773	0.3351	0.3456	0.3437	0.3410
$\hat{w}^{(1)}_{-1/2}$	0.1264	0.0002	0.8731	0.0004					
$\hat{w}^{(2)}_{-1/2}$	0.0003	0.0229	0.0007	0.9761					
17 industry portfolios									
$\sqrt{\hat{\delta}}$	0.1416	0.1374	0.1354	0.1348	0.1417	0.1267	0.1362	0.1356	0.1350
$\hat{w}^{(1)}_{-1/2}$	0.0202	0.3049	0.6013	0.0735					
$\hat{w}^{(2)}_{-1/2}$	0.0281	0.0672	0.1904	0.7143					

Notes: This table reports the estimates for the (square root of the) Hansen-Jagannathan distance $\sqrt{\hat{\delta}}$ for individual models and three aggregators: "EW ag" with equal $(1/M)$ aggregation weights, and "HEL ag1" and "HEL ag2" with aggregation weights $\hat{w}^{(1)}_{-1/2}$ and $\hat{w}^{(2)}_{-1/2}$ obtained by minimizing the Hellinger distance between the densities of the aggregator and two benchmark models, constant SDF and FF3 model, respectively.

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

Case (i): all models are misspecified.

	CAPM	CCAPM	EZ	D-CCAPM	HEL ag1	HEL ag2
mean $\sqrt{\hat{\delta}}$	0.3184	0.3176	0.3091	0.2987	0.3031	0.3016
median $\sqrt{\hat{\delta}}$	0.3169	0.3146	0.3071	0.2944	0.3009	0.2987
10% quant. $\sqrt{\hat{\delta}}$	0.2443	0.2433	0.2344	0.2226	0.2247	0.2242
90% quant. $\sqrt{\hat{\delta}}$	0.3965	0.3980	0.3918	0.3782	0.3826	0.3811
mean $\hat{w}^{(1)}_{-1/2}$	0.0636	0.2821	0.0961	0.5582		
mean $\hat{w}^{(2)}_{-1/2}$	0.0824	0.1127	0.1475	0.6574		

Notes: This table reports the Monte Carlo estimates for the (square root of the) Hansen-Jagannathan distance $\sqrt{\hat{\delta}}$ (mean, median, 10% quantile, and 90% quantile), and the mean aggregation weights $\hat{w}^{(1)}_{-1/2}$ and $\hat{w}^{(1)}_{-1/2}$ for the methods based on minimizing the Hellinger distance (HEL ag1 and HEL ag2) between the densities of the aggregator and the corresponding pivot (constant SDF or FF3), respectively. 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 ag1	HEL ag2
mean $\sqrt{\hat{\delta}}$	0.2859	0.2859	0.2770	0.2673	0.2715	0.2700
median $\sqrt{\hat{\delta}}$	0.2832	0.2827	0.2735	0.2644	0.2689	0.2668
10% quant. $\sqrt{\hat{\delta}}$	0.2143	0.2148	0.2060	0.1982	0.2024	0.1993
90% quant. $\sqrt{\hat{\delta}}$	0.3584	0.3600	0.3516	0.3378	0.3455	0.3431
mean $\hat{w}^{(1)}_{-1/2}$	0.0738	0.2610	0.0866	0.5786		
mean $\hat{w}^{(2)}_{-1/2}$	0.1051	0.1137	0.1574	0.6238		

Notes: This table reports the Monte Carlo estimates for the (square root of the) Hansen-Jagannathan distance $\sqrt{\hat{\delta}}$ (mean, median, 10% quantile, and 90% quantile), and the mean aggregation weights $\hat{w}^{(1)}_{-1/2}$ and $\hat{w}^{(1)}_{-1/2}$ for the methods based on minimizing the Hellinger distance (HEL ag1 and HEL ag2) between the densities of the aggregator and the corresponding pivot (constant SDF or FF3), respectively. The sample size is 200 and the number of Monte Carlo simulations is 1,000.

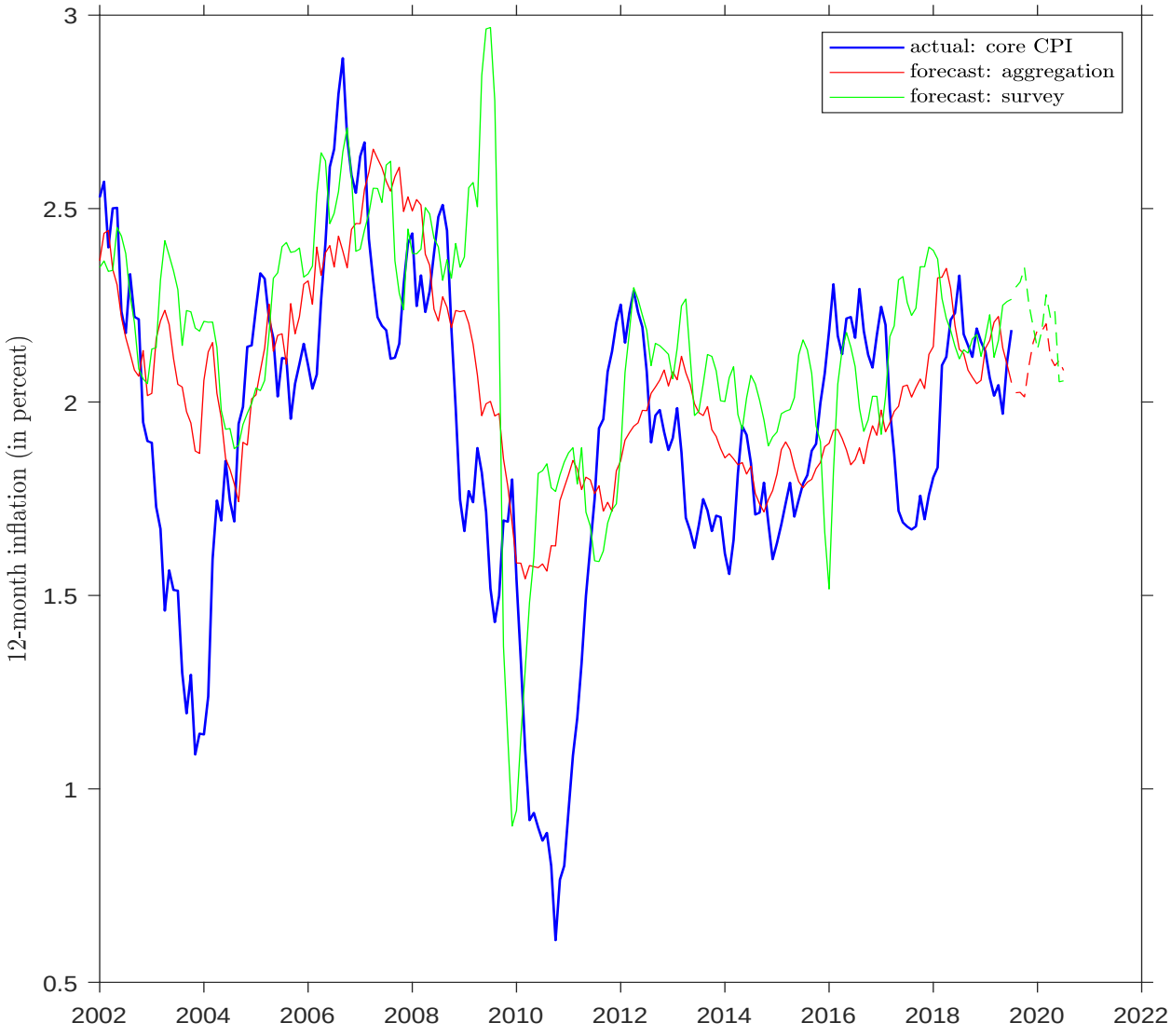

Figure 1: 12-month ahead, out-of-sample, forecasts of core inflation from the Blue Chip survey and the proposed aggregator.

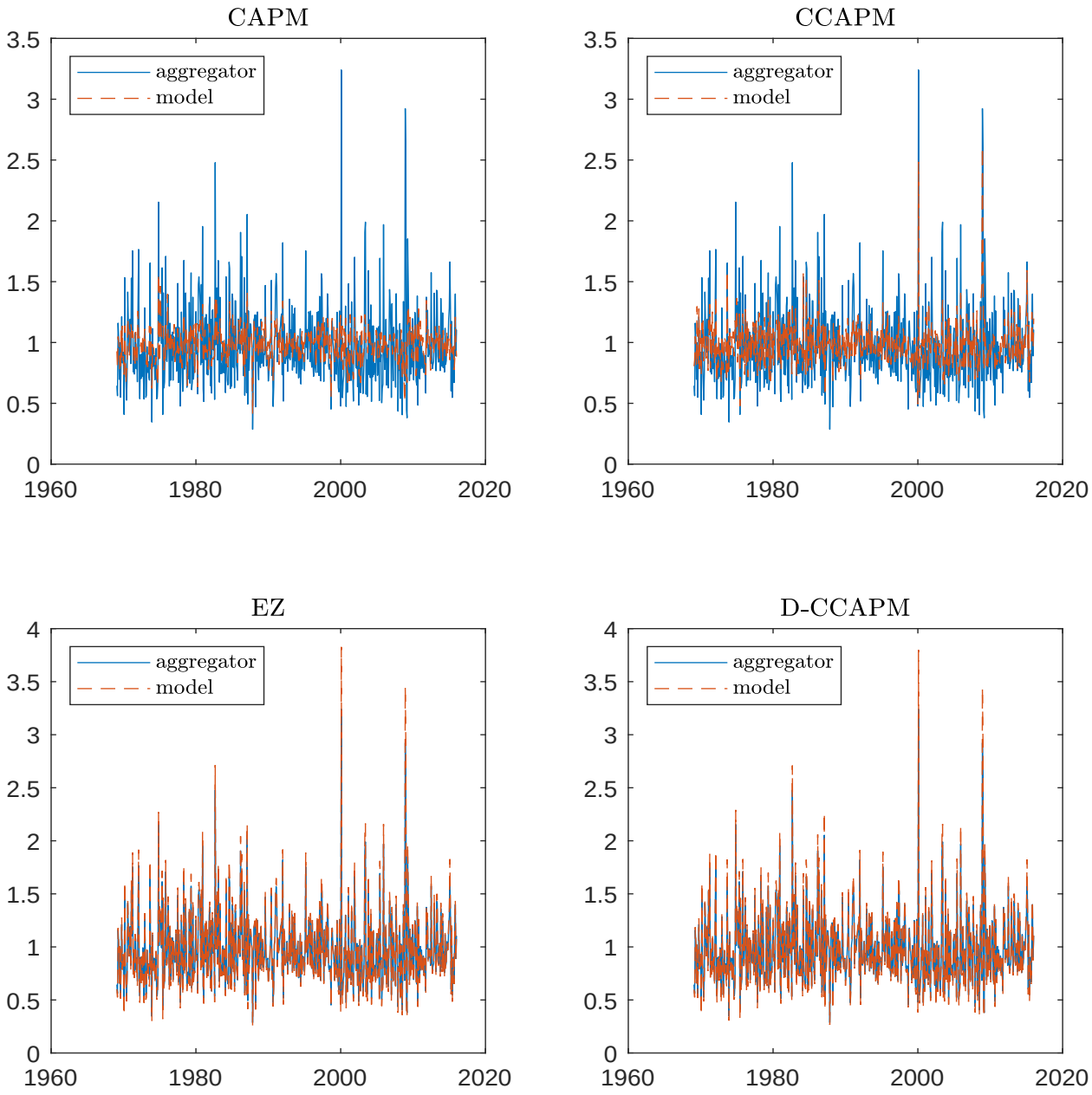

Figure 2: SDFs for individual models and the aggregator based on the Hellinger distance for the January 1969 – December 2015 period (test assets: 25 Fama-French and 17 industry portfolio returns).