General Aggregation of Misspecified Asset Pricing Models

Nikolay Gospodinov
Federal Reserve Bank of Atlanta

Esfandiar Maasoumi
Emory University

August 1, 2018

Model Misspecification and Main Idea

- Overwhelming evidence that most, if not all, economic models are misspecified.
- We adopt the view that dispenses completely with the notion of a true model and treats the candidate models as genuinely misspecified:
  1. because they approximate or represent different aspects of latent DGP;
  2. or because the underlying structure is completely unknown.
- Misspecified models can still be useful for informing policy makers and investors in their decision making...but one needs to proceed carefully
  1. Perform a model selection procedure ("least misspecified" model)
     - statistical inference (on the pseudo-true values of the model) needs to adequately incorporate model uncertainty
  2. Combine information from all models by model aggregation to elicit some features of the latent object of interest
     - the statistical paradigm is shifted away from parameter estimation of an optimally selected model
     - interest lies in some unknown functional (conditional mean, forecast density, stochastic discount factor etc.)

Motivation and Context

- Accounting for model uncertainty:
  - Model $m = (S, \gamma) \in \mathcal{M}$, where S is the model structure (functional form, distributional assumptions, heteroskedasticity, time dependence) and γ are parameters specific to the model structure S.
  - There is both parameter and model uncertainty.
  - What must be done is integrating over both S and γ

    $$p(y|x,\mathcal{M}) = \int_{\mathcal{M}} p(y|x,m)p(m|x)dm = \int\int p(y|x,S,\gamma)p(S,\gamma|x)dSd\gamma.$$

  - Instead, we often condition on a specific model structure S^*

    $$p(y|x,\mathcal{M}) = p(y|x,S^*) = \int\int p(y|x,S^*,\gamma^*)p(\gamma|x,S^*)d\gamma$$

    ignoring model uncertainty.
- Model averaging is a way of dealing with model uncertainty. But most model averaging methods assume that $\mathcal{M}$ contains the true model.
- We are usually interested in some functional f given the data. But for model averaging to make sense, f needs to be the same for all models.

Entropy-Based Aggregation

- Information-theoretic approach to aggregation:
  - adapts better to the underlying uncertainty surrounding DGP.
- M proposed misspecified models $\{f_1, ..., f_M\}$; $\tilde{f}$ is the aggregator.
  - each model is an incomplete 'indicator' of the latent object of interest.
- Consider the flat simplex $\mathcal{W}^M = \left\{ w \in \mathbb{R}^M : w_i \geq 0, \sum_{i=1}^M w_i = 1 \right\}$.
- The empirical risk function $\mathcal{R}_{T,\rho}(\tilde{f}, f_i)$ is the generalized entropy divergence between the aggregator $\tilde{f}$ and each prospective models f_i:

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

- The aggregator that minimizes $\sum_{i=1}^M w_i \mathcal{R}_{T,\rho}(\tilde{f}, f_i)$, $w \in \mathcal{W}^M$, is

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

  - linear ($\rho = -1$), geometric ($\rho \to 0$) and Hellinger ($\rho = -1/2$) pooling are special cases.

Example: Forecasting U.S. Core Inflation

- Monthly data for 1988:01–2018:02.
- 12-month forecasts of U.S. core (CPI less food and energy) inflation.
- Models:
  - BC: Blue Chip survey of expected CPI inflation
  - PC: Phillips curve model
  - HA: Historical average
  - MA: IMA(1,1) model (Stock and Watson, 2007)
  - CY: Simplified commodity-based (convenience yield) model (Gospodinov and Ng, 2013; Gospodinov, 2017)
  - AG: Hellinger distance ($\rho = -1/2$) aggregator of PC, HA, MA and CY (BC is used as pivot)
- Recursive model estimation (initial sample 1988:01–1996:12)
- Aggregation weights are estimated by minimizing the Hellinger distance between the aggregator and pivot densities over a training sample (initial sample 1997:01–2001:12).
- Out-of-sample evaluation: 2002:01-2018:02.

Example: Forecasting U.S. Core Inflation

Bregman loss functions (Patton, 2017) for different forecasting models

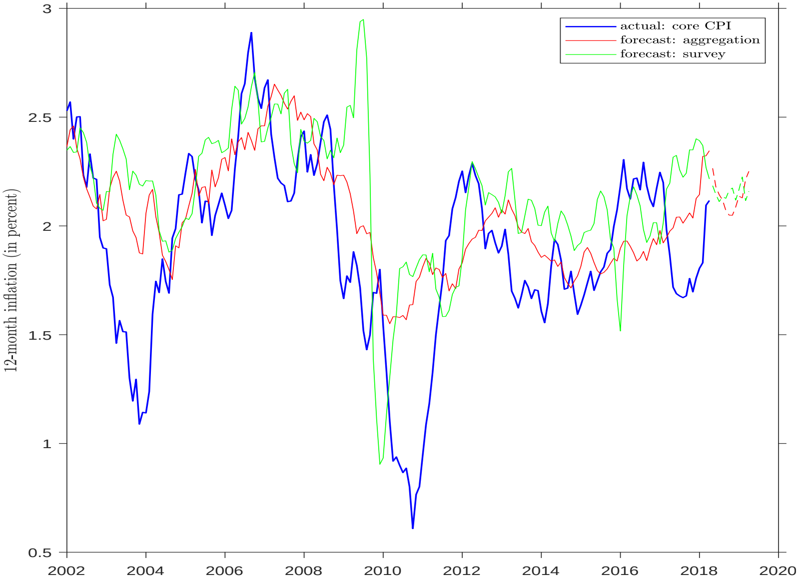
	PC	HA	MA	BC	CY	AG
Homogeneous Bregman Loss ($k > 1$)						
$k = 1.1$	2.4408	2.0272	2.2855	1.7521	1.4074	1.0000
$k = 2$ (MSE)	1.9695	2.051	2.0913	1.7726	1.5628	1.0000
$k = 3$	1.7382	2.0339	1.8989	1.7948	1.7676	1.0000
$k = 3.5$	1.6745	2.0101	1.8125	1.8065	1.8835	1.0000
$k = 4$	1.6297	1.9781	1.7323	1.8188	2.0092	1.0000
Non-homogeneous (exponential) Bregman Loss ($a \neq 0$)						
$a = -1$	2.6709	2.0111	2.4709	1.7349	1.2609	1.0000
$a = -0.5$	2.2476	2.0495	2.2796	1.7528	1.3907	1.0000
$a \to 0$ (MSE)	1.9695	2.0515	2.0913	1.7726	1.5628	1.0000
$a = 0.5$	1.7912	2.0157	1.9117	1.7959	1.7875	1.0000
$a = 1$	1.6796	1.9434	1.7433	1.8235	2.0788	1.0000

All losses are expressed as ratios to that of the aggregator (AG) model.

Example: Forecasting U.S. Core Inflation

- Dominance of forecast aggregation across ALL loss functions
  - the forecast improvements are quite large
  - improvements are largest when over-predictions are penalized more heavily than under-predictions
  - unbiased forecast: Mincer-Zarnowitz regression (intercept=-0.0192, slope=0.9221)
- For the individual models, BC and CY work best except when over-predictions are very costly.
- Largest weights are assigned to the CY model.
- Interesting dynamics of forecast weights over time.
- Some evidence against perfect substitutability of candidate models, which is implicitly embedded in the linear pooling ($\rho = -1$).
- The aggregator can be adapted to some other model instead of BC (we prefer BC because it's model-free).
- "Intercept corrections" à la Klein/Theil lead to further improvements.
- Reminder: forecasting core inflation is really challenging.

Example: Forecasting U.S. Core Inflation



Oracle Inequalities and Bounds

- Model aggregation as a stochastic optimization approach.
- Let functional $f(\cdot)$ be the unknown object to be inferred.
- Suppose that a finite list (*dictionary*) $\mathcal{F}$ of candidate auxiliary models is available.
- Stochastic optimization minimizes an empirical risk function that satisfies oracle inequalities (Rigollet, 2012; Rigollet and Tsybakov, 2012).
  - model aggregation with aggregation weights obtained from the stochastic optimization problem;
  - model selection assigns weights of one or zero to individual models: it proves to be suboptimal.
- Let $Z_1, ..., Z_T$ denote observations of the random variable Z with an unknown distribution.
- Let $L : Z \times \mathcal{F} \to \mathbb{R}$ be a measurable loss function with a corresponding risk function $\mathcal{R} : \mathcal{F} \to \mathbb{R}$ defined as

$$\mathcal{R}(f) = \mathbb{E}[L(Z, f)], \ f \in \mathcal{F}.$$

Oracle Inequalities and Bounds

- The *oracle* f^* is defined as $f^* = \inf_{f \in \mathcal{F}} \mathcal{R}(f)$.
  - "oracle" because it cannot be constructed without knowledge of the true functional.
- The goal is to construct an aggregator $\tilde{f}$ of $f_1, ..., f_M$ in the $\mathcal{F}$ dictionary by mimicking the oracle $\inf_{f \in \mathcal{F}} \mathcal{R}(f)$.
- *Oracle bound* (in expectation): there exists a constant $C \geq 1$ such that

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

  - the remainder term $\triangle_{T,M}(\mathcal{F}) > 0$ characterizes the performance of the aggregator: explicit function of M and sample size T;
  - the goal is to find an optimal (smallest possible) $\triangle_{T,M}(\mathcal{F})$: a difficult problem especially with dependent data and general functional forms;
  - if the model is misspecified, $\inf_{f \in \mathcal{F}} \mathcal{R}(f) > 0$;
  - it is therefore desirable to obtain a bound with a leading constant $C = 1$ (sharp inequality);
  - again, this is a challenging task.

Entropy-Based Aggregators

- Let P and Q be probability measures with densities p and q with respect to a dominating measure ν.
- Generalized entropy divergence from Q to P is given by

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

where $\phi_\eta(x) = \frac{1}{\eta(\eta+1)}\left(x^{\eta+1} - 1\right)$ is the Cressie-Read family, or

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

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

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

  - the case $\eta = -1/2$ corresponds to the Hellinger distance measure (the only proper measure of distance in the class)

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

Hellinger-Distance Aggregator

- 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 with $\tilde{f}_T^{(w)}$ being its sample analog;
  - $\mathcal{H}(\tilde{f}^{(w)}, f)$ be the risk function based on the Hellinger distance.
- Then (see also Birgé, 2006, 2013),

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

  where $C \geq 1$ and $\triangle_{T,M}$ is a remainder term.
- Moreover, the minmax risk over $\mathcal{F}$ is bounded by $C\triangle_{T,M}$.
- Note that $\mathcal{H}(\tilde{f}^{(w)}, f) > 0$ under model misspecification.
- But with Hellinger distance and minmaxity, the risk remains under control even if the models are misspecified.
  - Kitamura, Otsu, and Evdokimov (2013); Antoine and Dovonon (2017) for the robustness properties of the Hellinger distance.

HJ-Distance

- Let m_t represent an admissible SDF at time t and let $\mathcal{M}$ be the set of all admissible SDFs.
- An SDF m_t is admissible if it prices the test assets correctly, i.e.,

$$\mathbb{E}[R_t m_t] = 1_N.$$

- Suppose that $y_t(\gamma)$ is a candidate SDF at time t that depends on the vector of unknown parameters $\gamma \in \Gamma$
  - linear SDF $y_t(\gamma) = x_t'\gamma$, where x_t are K ($K < N$) risk factors.
- Model is correctly specified if $\exists$ a $\gamma \in \Gamma$ such that $y_t(\gamma) \in \mathcal{M}$.
- Model is misspecified if $y_t(\gamma) \notin \mathcal{M}$ for all $\gamma \in \Gamma$.
- Hansen and Jagannathan (1991, 1997) suggested using

$$\delta = \min_{\gamma \in \Gamma} \min_{m_t \in \mathcal{M}} \left(\mathbb{E}[(y_t(\gamma) - m_t)^2]\right)^{\frac{1}{2}}$$

  as a misspecification measure for $y_t(\gamma)$.
- We refer to δ as the Hansen-Jagannathan distance (HJD).

HJ-Distance

- To preview what's coming, HJD can be interpreted as a quadratic risk for *stochastic optimization* with misspecified models
  - Almeida and Garcia (2012) show that for a fixed vector of parameters γ, the primal problem in the SDF framework can be written as

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

  - The primal problem for the HJD is obtained for $\eta = 1$. The normalized Hellinger distance follows for $\eta = -1/2$.
- HJD is "oracle" since m_t is an unknown/unknowable latent object.
- It is often more convenient to solve the following dual problem:

$$\delta^2 = \min_{\gamma \in \Gamma} \max_{\lambda \in \Re^N} \mathbb{E}[y_t(\gamma)^2 - (y_t(\gamma) - \lambda' R_t)^2 - 2\lambda' 1_N],$$

where λ is an N-vector of Lagrange multipliers.

- m_t no longer plays a role!!!

HJ-Distance

- Let $\theta = [\gamma', \lambda']'$ and $\theta^* = [\gamma^{*\prime}, \lambda^{*\prime}]'$ be defined as

$$\theta^* = \arg\min_{\gamma\in\Gamma} \max_{\lambda\in\Re^N} \mathbb{E}[L_t(\theta)],$$

  where $L_t(\theta) \equiv y_t(\gamma)^2 - (y_t(\gamma) - \lambda' R_t)^2 - 2\lambda' 1_N$.
- By rearranging the dual problem, it is easy to show that

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

  where $U = \mathbb{E}[R_t R_t']$ and $e(\gamma^*) = \mathbb{E}[R_t y_t(\gamma^*) - 1_N]$, and

$$\delta^2 = e(\gamma^*)' U^{-1} e(\gamma^*).$$

- Then, the estimator $\hat{\theta} = [\hat{\gamma}', \widehat{\lambda}']'$ can be obtained sequentially as

$$\hat{\gamma} = \arg\min_{\gamma\in\Gamma} e_T(\gamma)' U_T^{-1} e_T(\gamma),$$

  and $\widehat{\lambda} = \hat{U}^{-1} e_T(\hat{\gamma})$, where U_T is the sample analog of U.
    - a non-optimal GMM estimator with a fixed weighting matrix U_T^{-1}.

Consumption-Based Models and SDF Aggregation

- Dictionary of SDF models:
  - CAPM (Brown and Gibbons, 1985)
  - Consumption CAPM
  - Non-expected utility model (Epstein and Zin, 1989, 1991; Weil, 1989)
  - Durable consumption CAPM (Yogo, 2006)
  - External habit model (Abel, 1990)
- Auxiliary models are misspecified, but economic theory still provides guidance to mimicking the oracle SDF.
- The primal problem targets unknown functional of interest, but is transformed to the dual.
  - The immutable part of risk drops out.
- Our aggregation method is information nesting.
- Data dependent model weights, w_i, will rank competing models.
- An alternative is a data-driven (model-free) approach to approximating the unknown function using non-parametric methods.
  - This is better suited to a 'machine learning' approach.

Evidence of Misspecification: Asset Pricing Models

HJ-distance estimation of SDF models (t-stats and p-vals)							
Model	*market*	cg_t	cd_t	cg_{t-1}	smb_t	hml_t	*Spec. Test*
CAPM	2.70 [2.35]						0.00
CCAPM		−1.41 [−1.29]					0.00
Epstein-Zin	3.31 [2.76]	−2.14 [−2.08]					0.00
D-CCAPM	3.14 [2.60]	−1.94 [−1.84]	−0.79 [−0.79]				0.00
External habit		−1.81 [−1.57]		−1.14 [−1.14]			0.00
Fama-French	1.92 [1.66]				−2.29 [−1.92]	−2.70 [−2.48]	0.00

Notes: Test assets: 25 Fama-French + 17 industry portfolios. Sample period: 1959:02 - 2012:12. Rank test is testing the null of a reduced rank of D. Misspecification-robust t-stats in square brackets.

- All models are rejected!
  - Still, it is common practice to use GMM standard errors for correctly specified models even when the model is rejected by the data.
  - Allowing for model uncertainty reduces the statistical significance (especially for non-traded factors).

SDF Aggregation: Some Specifics

- M proposed misspecified models, $\hat{y}_{i,t} = y_{i,t}(\hat{\gamma}_i)$, $i = 1, ..., M$, for the unknowable true SDF m.
- The estimates $\hat{\gamma}_i$ of the pseudo-true values γ_i^* are obtained from a prior training sample of size N by minimizing the HJD for each model.
- The effective number of sample observations is $N + T$
  - candidate models are estimated using observations $1, ..., N$
  - aggregation weights are estimated using observations $N + 1, ..., N + T$.
- Then, a model averaging rule would aggregate information from all of these models and construct a pseudo-true SDF $\tilde{y}$.
- We are interested in finding the aggregator $\tilde{y}_t$ with a distribution that is as close as possible to the distributions of $\hat{y}_i$'s.
- The *risk* of the aggregator $\tilde{y}_t$ has an oracle component relative to m. This is common to all empirical decisions.
- All decisions are "stochastically optimizing" (empirical) risk of $\tilde{y}_t$.

SDF Aggregation: Some Specifics

- Parameters for model i are estimated over the training sample ($t = 1, ..., N$) as

$$\hat{\gamma}_i = \underset{\gamma_i \in \Gamma}{\arg\min}\ e_T(\gamma_i)' \left(\frac{1}{N} \sum_{t=1}^{N} R_t R_t' \right)^{-1} e_T(\gamma_i),$$

  where $e_T(\gamma_i)$ denotes the sample pricing errors of model $_i$.
- The SDFs $\hat{y}_{i,t} = y_{i,t}(\hat{\gamma}_i)$, $i = 1, ..., M$, are constructed by plugging in the estimated parameters but using data for the second part of the sample $N + 1, ..., N + T$.
- Recall that the aggregator that minimizes GE risk takes the form

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

  - under quadratic risk ($\rho = -1$), we obtain linear pooling.
  - under Hellinger-distance risk ($\rho = -1/2$), $\tilde{y}_t \propto \left[\sum_{i=1}^{M} w_i y_{i,t}^{1/2} \right]^2$.
- **Two methods** for estimating w.

SDF Aggregation: Some Specifics

- **Method 1**: HJ-distance approach.
- For given $(\hat{y}_{1,t}, ..., \hat{y}_{M,t})'$, construct the pricing errors of the aggregator

$$\tilde{e}_T(w) = \frac{1}{T}\sum\nolimits_{t=N+1}^{N+T} R_t \left[\sum\nolimits_{i=1}^{M} w_i \hat{y}_{i,t}\right] - 1_N.$$

- The unknown weights w are obtained by minimizing the HJ-distance of $\tilde{e}_T(\theta)$

$$\tilde{\delta} = \sqrt{\tilde{e}_T(w)' \left(\frac{1}{T}\sum\nolimits_{t=N+1}^{N+T} R_t R_t'\right)^{-1} \tilde{e}_T(w)},$$

subject to $w_i \geq 0$ and $\sum_{i=1}^{M} w_i = 1$.

SDF Aggregation: Some Specifics

- **Method 2**: minimizing the Hellinger distance (consistent risk function).
- Let p be the density of some favored benchmark model ("pivot"), and q the density of the aggregator $\tilde{y}_t(\theta) = \left[\sum_{i=1}^{M} w_i y_{i,t}^{1/2}\right]^2$.
- Minimize the Hellinger distance (with respect to w)

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

  subject to $w_i \geq 0$ and $\sum_{i=1}^{M} w_i = 1$.
- Starting values for weights 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 $i = 1, ..., M$.
- Densities p and q are estimated by a kernel density estimator and the integral in $\mathcal{H}$ is evaluated numerically.
- The choice of a benchmark model: Fama-French 3-factor model.

Simulations

- Factors and returns are simulated from a multivariate normal distribution with parameters calibrated to the data.
- Sample size is $N + T = 600$ with $N = 360$ and $T = 240$.
- Two scenarios: (i) all models are misspecified and (ii) CAPM is "true" but all other models are misspecified.
- Two sets of test asset returns: (i) the 25 Fama-French portfolios, and (ii) the 17 industry portfolios.
- Models for aggregation: CAPM, CCAPM, EZ and D-CCAPM.
- Benchmark model: FF3.
- Aggregators: HJ distance and Hellinger distance.
- Evaluation metric for pricing performance: HJ distance.
- HJD aggregator is expected to work the best: But how does it compare to individual models?
- HEL aggregator is expected to show robustness: But how does it assign weights compared to HJD aggregator?

Simulations: All Models are Misspecified

	CAPM	CCAPM	EZ	D-CCAPM	FF3	HJD agg.	HEL agg.
				25 Fama-French portfolios			
mean $\hat{\delta}$	0.4713	0.4831	0.4780	0.4834	0.4533	0.4577	0.4708
median $\hat{\delta}$	0.4683	0.4786	0.4737	0.4794	0.4501	0.4545	0.4680
mean $\hat{w}_{-1}$	0.3512	0.1775	0.1422	0.3291			
mean $\hat{w}_{-1/2}$	0.1766	0.1420	0.2586	0.4228			
				17 industry portfolios			
mean $\hat{\delta}$	0.3000	0.3036	0.3101	0.3213	0.3081	0.2908	0.3010
median $\hat{\delta}$	0.2985	0.3008	0.3070	0.3162	0.3077	0.2889	0.3013
mean $\hat{w}_{-1}$	0.4047	0.3347	0.1030	0.1575			
mean $\hat{w}_{-1/2}$	0.3230	0.2174	0.1718	0.2878			

- SDF aggregation offers a substantial improvement in pricing performance.
- HJD aggregator dominates uniformly the individual models used for aggregation.
- HEL aggregator appears to robustify away from the best performing individual model and distribute weights more evenly across models.

Simulations: CAPM is Correctly Specified

	CAPM	CCAPM	EZ	D-CCAPM	FF3	HJD agg.	HEL agg.
				25 Fama-French portfolios			
mean $\hat{\delta}$	0.3370	0.3490	0.3433	0.3507	0.3459	0.3286	0.3387
median $\hat{\delta}$	0.3339	0.3477	0.3426	0.3498	0.3414	0.3262	0.3369
mean $\hat{w}_{-1}$	0.4344	0.2353	0.1523	0.1781			
mean $\hat{w}_{-1/2}$	0.3360	0.1402	0.2218	0.3020			
				17 industry portfolios			
mean $\hat{\delta}$	0.2657	0.2680	0.2744	0.2833	0.2770	0.2563	0.2666
median $\hat{\delta}$	0.2633	0.2654	0.2696	0.2784	0.2746	0.2548	0.2644
mean $\hat{w}_{-1}$	0.4003	0.3490	0.0908	0.1599			
mean $\hat{w}_{-1/2}$	0.3241	0.2010	0.1983	0.2766			

- Even when one of the models is true, HJD aggregation dominates.
- Somewhat surprising that aggregation weights are still fairly equally distributed over competing models.
  - partly due to the fact that CAPM is nested within other models.
  - it also illustrates the "insurance" value of mixing by penalizing the possibility of choosing catastrophically false individual models.

Concluding Remarks

- Economic models are misspecified by design as they try to approximate a complex/unknown/unknowable DGP.
- Instead of selecting a single model for policy analysis or decision making, aggregating information from all models may adapt better to the underlying uncertainty and result in a more robust approximation.
- Information theory provides the natural theoretical foundation for dealing with these types of uncertainty and partial specification.
- We capitalize on some insights from the information-theoretic approach and propose a mixture method for aggregating information from different misspecified asset pricing models.
- The generalized entropy criterion that underlies our approach allows us to circumvent some drawbacks of the standard linear pooling.
- Potentially wide applicability in (micro, macro, labor) economics using a large set of diverse, partially specified models.