Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models

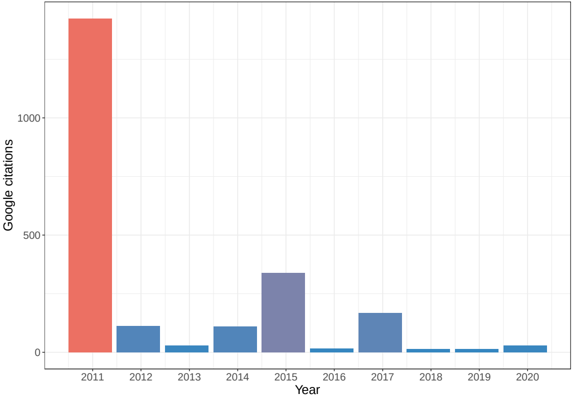
Svetlana Bryzgalova [1] Jiantao Huang [2] Christian Julliard [2]

[1]London Business School [2]London School of Economics

July 6, 2020

AFA Presidential Addresses

Leaders of the profession try to encourage deep thinking and promote new ideas:



2011: John Cochrane coins the term “factor zoo” ($\sim$ 1430)

2017: Campbell Harvey on p-hacking and factor zoo ($\sim$ 170)

2015: Luigi Zingales “Does Finance Benefit Society?” ($\sim$ 340)

Factors everywhere

Linear factor models are widely in different areas of asset management:

- Designing a trading strategy: market, size, value, momentum, etc...
- Understanding the main sources of risk premia on the market
- Measuring performance: which alpha?
- Understanding price impact and limits to arbitrage
- Reflecting strategy capacity and crowded trades

Factors are often based on stock characteristics or economy-wide features

Starting with Fama and French, asset pricing factors has been at the center of academic research

Empirical asset pricing: The big picture

Two facts:

I. Even simple linear factor models are often hard to evaluate
- Uncertainty in R^2 and standard errors: Lewellen et al (2010)
- Identification: useless factors seem relevant, e.g. Kleibergen (2009)

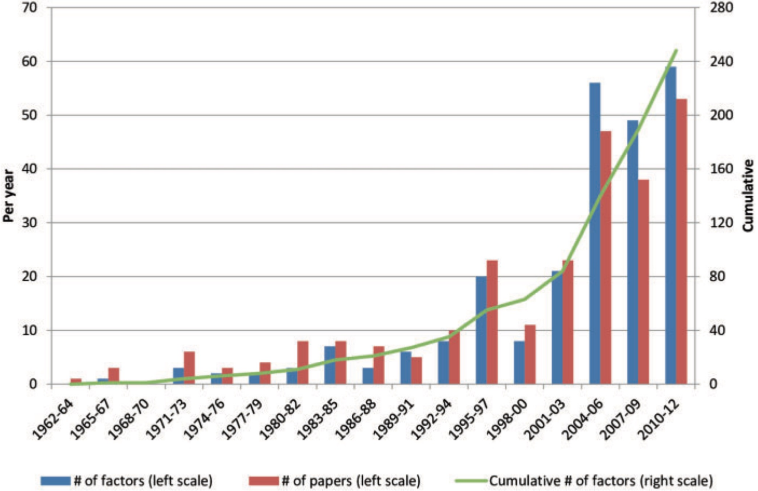
II. A zoo of asset pricing factors: ≥ 400
- literally quadrillions of possible models
- model uncertainty, factor and model selection/aggregation, etc.

Current steady state:

$\Rightarrow$ An active re-evaluation of many empirical findings

$\Rightarrow$ A call for new empirical designs, and estimators.

Factor production mill: Harvey et al (2015)



Big data, computing power, and machine learning increase factor production further

Walking through the factor zoo

Cochrane (2011) called for solutions and dimension reduction:

- Spanning tests: Barillas and Shanken (2016), Chib et al (2020)
- Lasso: Giglio and Xiu (2019), Neuhierl et al (2019)
- Reduced rank regression: Huang and Zhou (2018)
- PCA in anomalies space: Kozak et al (2019)
- Latent factors: Lettau and Pelger (2019, 2020)
- Characteristics + latent factors: Fan et al (2015), Kelly et al (2018)
- Weak factors: Kleibergen (2009), Gospodinov et al (2013, 2019)

Key observation: so far, no general method that

1. can handle quadrillion models: both selection and aggregation
2. works for both tradable and non tradable factors
3. remains statistically valid: misspecification and identification

This paper, in a nutshell

Novel framework for estimating linear factor models: simple, robust, and applicable to high dimensional problems

Single model estimation: Bayesian Fama-MacBeth

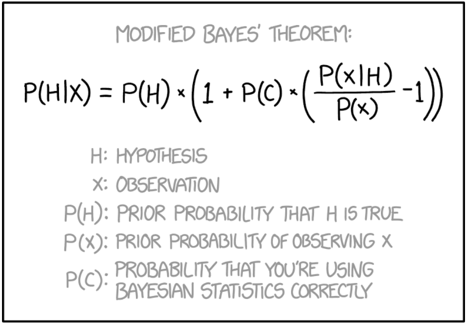
- Estimating and testing risk premia for small and large cross-sections
- Confidence intervals for risk premia and any other measure, e.g. R^2
- Automatically robust to spurious and level factors
- As fast as doing an OLS/GLS regression, and gets everything at once

Model analysis, comparison and aggregation: BMA

- factor selection
- designed to be robust to identification problems
- measures model uncertainty: selection and averaging
- misspecification: inference on risk premia across possible models

Roadmap

1. Define the model: assets and factors
2. Estimating a specific factor model
   ⇒ introduce likelihood
3. Factor and model selection
   ⇒ introduce integrated likelihood
4. How to deal with godzillions of models
   ⇒ sampling


MODIFIED BAYES' THEOREM:
$P(H|X) = P(H) \times \left(1 + P(C) \times \left(\frac{P(x|H)}{P(x)} - 1\right)\right)$
H: HYPOTHESIS
X: OBSERVATION
P(H): PRIOR PROBABILITY THAT H IS TRUE
P(X): PRIOR PROBABILITY OF OBSERVING X
P(C): PROBABILITY THAT YOU'RE USING BAYESIAN STATISTICS CORRECTLY

Identification pops up in Steps 2-3.

Standalone Models

What makes something a risk factor?

All the risk premia (or most) is captured by a given set of factors

- Their optimal combination (or its proxy) has the highest Sharpe ratio possible
- All the other strategies performance can be attributed to our factors

To be the true sources of risk, our factors should explain average returns on different strategies

- not just being correlated with those returns, but
- higher/lower comovement with the factor (β) should lead to higher/lower average return of the strategy

Cross-sectional test: higher beta on the factor(s) should align with different cross-sectional returns on trading strategies.

Classic Fama-MacBeth regressions

Take a set of widespread, economically important investment strategies

Step 1. Measure comovement between strategies/portfolios and factors

Step 2. Check if this comovement translates into higher/lower expected returns of the strategies

⇒ if this cross-sectional relationship exists, the factor is priced
⇒ it's a source of risk premium on the market

Linear Factor Models and Fama-MacBeth

Factor exposures of returns, $\boldsymbol{\beta_f} \in \mathbb{R}^{N \times K}$, are recovered from time series regressions:

$$\boldsymbol{R}_t = \boldsymbol{a} + \boldsymbol{\beta_f} \boldsymbol{f}_t + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{R}_t \in \mathbb{R}^N$ denotes excess returns, and $\boldsymbol{f}$ are demeaned factors.

Factor risk premia, $\boldsymbol{\lambda_f} \in \mathbb{R}^K$, are then estimated from the cross-sectional regression:

$$\bar{\boldsymbol{R}} = \lambda_c 1_{\boldsymbol{N}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{f}} \boldsymbol{\lambda_f} + \boldsymbol{\alpha}$$

$$\widehat{\boldsymbol{\lambda}}_{ols} = (\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\beta}})^{-1} \widehat{\boldsymbol{\beta}}^\top \bar{\boldsymbol{R}},$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{f}}$ denotes the time series estimate, $\widehat{\boldsymbol{\beta}} = (1_{\boldsymbol{N}} \ \widehat{\boldsymbol{\beta}}_{\boldsymbol{f}})$ and $\boldsymbol{\lambda}^\top = (\lambda_c \ \boldsymbol{\lambda}_{\boldsymbol{f}}^\top)$

Note: risk premia estimates crucially depend on whether $\boldsymbol{\beta}$ has full rank.

A factor example: Consumption CAPM

Test assets:

25 portfolios, created from stocks sorted by market size and book-to-market ratio
→ good spread in small/large, value/growth strategies

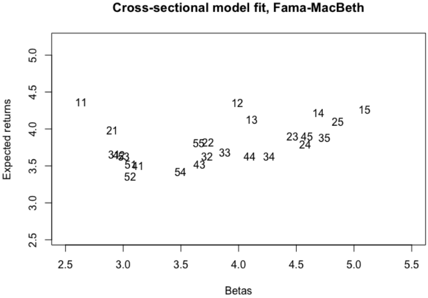
Consumption price of risk: $\lambda = 0.2\ [t = 1.3]$. Is this identified?

Spurious factors: all the betas are close to 0

Level factors: not enough spread in betas

Consequences:

- spurious factors are deemed significant
- often drive out the true sources of risk
- R^2 are inflated
- hard to evaluate model validity, etc.



Big problem for OLS/GLS with rank deficiency generated by useless (and level) factors.

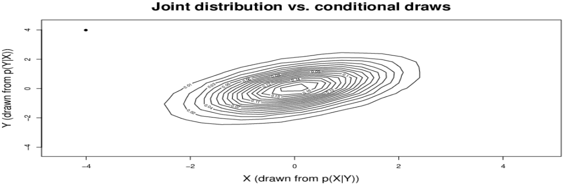
Bayesian basics I

Parameter as random variables: ⇒ focus on their "posterior" distribution

- start with a prior about parameter value/range
- update it, given the data (posterior distribution)
- learning view of working with data

Posterior via conditionals (aka "Gibbs"):

- Characterize (and evaluate) the joint distribution of e.g. parameters X and Y via $p(X|Y)$ and $p(Y|X)$.



Bayesian Fama-MacBeth

Time series properties ⇒ betas ⇒ risk premia

Step 1:

Time series stage is always valid, and under a diffuse prior yield the posterior of betas:

$$\boldsymbol{B}|\boldsymbol{\Sigma}, data \sim \mathcal{MVN}\left(\widehat{\boldsymbol{B}}_{ols}, \boldsymbol{\Sigma} \otimes (\boldsymbol{F}^\top \boldsymbol{F})^{-1}\right), \quad \text{where} \quad \boldsymbol{B} = \begin{pmatrix} \boldsymbol{a}^\top \\ \beta_{\boldsymbol{f}}^\top \end{pmatrix}$$

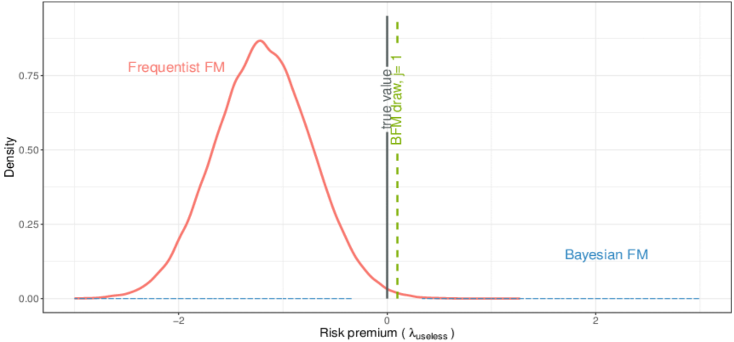
$$\boldsymbol{\Sigma}|data \sim \mathcal{W}^{\text{-}1}\left(T - K - 1, T\widehat{\boldsymbol{\Sigma}}_{ols}\right),$$

Step 2:

If the model is correctly specified $\mathbb{E}[\boldsymbol{R}_t] = \boldsymbol{a} = \boldsymbol{\beta}\boldsymbol{\lambda}$, therefore:

The posterior distribution of $\boldsymbol{\lambda}$ conditional on $\boldsymbol{B}$, $\boldsymbol{\Sigma}$ and the data, is a constant at $(\boldsymbol{\beta}^\top\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top\boldsymbol{a}$.

Risk premia estimates of a useless factor



Posterior distribution of risk premia (blue dashed line) from BFM-OLS estimation of a misspecified one-factor model based on a single simulation with $T = 1000$, and asymptotic distribution of the frequentist Fama-MacBeth estimate (red solid line). Dotted line corresponds to the pseudo-true value of the parameter (defined to be 0 for of a useless factor).

Intuition: the draws $\boldsymbol{\lambda}_{(j)} = (\boldsymbol{\beta}_{(j)}^{\top}\boldsymbol{\beta}_{(j)})^{-1}\boldsymbol{\beta}_{(j)}^{\top}\boldsymbol{a}_{(j)}$ will often flip sign as $\boldsymbol{\beta} \to 0$.

$\Rightarrow$ such posterior is centered around 0.

Estimation if risk premia

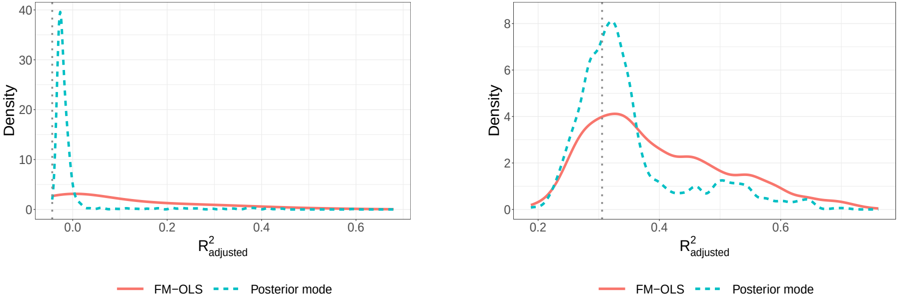
Numerous simulations:

- for strong factors, the prior has no impact on risk premia estimates
- valid confidence intervals, etc.
- weak factors are reliably recognized
- works well in small samples
- could be used with a small or large number of test assets
- very fast: 1-2 seconds to compute everything on a standard PC

Along with risk premia, the method automatically produces any other quantity that depends on it, and their confidence intervals:

- R-squared (cross-sectional fit),
- Sharpe ratio,
- alphas of portfolios, tail risk, etc.

Distribution of R^2_{adj} in models with strong and useless factors

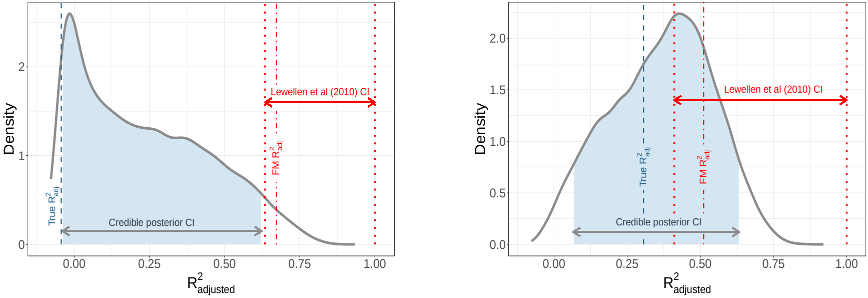


(a) Model with a useless factor

(b) Model with strong and useless factors

Asymptotic distribution of cross-sectional R^2 under different model specifications across 1,000 simulations of sample size $T = 20,000$. Blue dash lines correspond to the distribution of the posterior mode for R^2_{adj}, while red solid lines depict the pointwise sample distribution of R^2_{adj} evaluated at the frequentist Fama-MacBeth estimates. Grey dotted line stands for the true value of R^2_{adj}.

Estimation uncertainty of cross-sectional R^2



(a) Model with a useless factor

(b) Model with strong and useless factors

Posterior densities of cross-sectional R^2_{adj} in one representative simulation with centered 90% confidence interval (shaded area). Blue dashed line denotes the true R^2_{adj}. Red dash-dot line depicts Fama-MacBeth R^2_{adj} estimate with 90% Lewellen et al (2010) confidence intervals (red dotted lines).

Notable examples and 25 size-and-value Fama-French portfolios

		FM		BFM		
Model	Factors	$\hat{\lambda}_j$	R^2_{adj}	$\bar{\lambda}_j$	$R^2_{adj,mode}$	$R^2_{adj,median}$
		Panel A: OLS				
Liquidity-CAPM	Intercept	0.973^{*} $[-0.084, 2.030]$	36.24 $[-9.09, 100.00]$	1.162^{**} $[0.175, 2.120]$	34.09 $[-2.39, 61.46]$	30.27
Pastor and Stambaugh (2000)	LIQ	3.057^{**} $[0.727, 5.388]$		1.785 $[-1.237, 4.150]$		
	MKT	-0.281 $[-1.350, 0.788]$		-0.449 $[-1.371, 0.509]$		
Scaled CCAPM	Intercept	1.046 $[-0.848, 2.940]$	25.67 $[-14.29, 100.00]$	1.791^{**} $[0.001, 3.723]$	34.36 $[-4.76, 62.07]$	29.19
Lettau and Ludvigson (2001)	*cay*	1.817 $[-0.653, 4.288]$		0.791 $[-1.347, 2.686]$		
	ΔC_{nd}	0.713^{*} $[-0.030, 1.456]$		0.303 $[-0.462, 0.951]$		
	$\Delta C_{nd} \times cay$	0.804 $[-1.645, 3.253]$		0.301 $[-1.911, 2.270]$		
Durable CCAPM	Intercept	2.214 $[-1.037, 5.465]$	52.38 $[28.00, 100.00]$	2.780^{*} $[-0.184, 5.751]$	47.1 $[1.20, 69.91]$	40.78
Yogo (2006)	ΔC_{nd}	0.743^{*} $[-0.025, 1.511]$		0.357 $[-0.207, 0.832]$		
	ΔC_d	-0.057 $[-0.719, 0.605]$		0.014 $[-0.668, 0.693]$		
	MKT	0.083 $[-3.322, 3.489]$		-0.495 $[-3.395, 2.555]$		
		Panel B: GLS				
q-factor model	Intercept	1.305^{***} $[0.779, 1.831]$	55.03 $[24.40, 96.40]$	1.277^{***} $[0.702, 1.879]$	47.28 $[32.45, 64.19]$	48.54
Hou, Xue, and Zhang (2014)	ROE	0.295^{*} $[-0.026, 0.615]$		0.266 $[-0.087, 0.640]$		
	IA	0.270^{***} $[0.104, 0.437]$		0.265^{***} $[0.093, 0.450]$		
	ME	0.251^{***} $[0.161, 0.341]$		0.246^{***} $[0.144, 0.345]$		
	MKT	-0.749^{***} $[-1.268, -0.229]$		-0.720^{**} $[-1.292, -0.156]$		
HC-CAPM	Intercept	2.730^{***} $[1.458, 4.002]$	56.36 $[30.18, 83.64]$	2.759^{***} $[1.379, 4.095]$	58.24 $[9.67, 75.07]$	49.26
Jagannathan and Wang (1996)	ΔY	-0.421^{**} $[-0.742, -0.099]$		-0.241 $[-0.598, 0.114]$		
	MKT	-0.717 $[-1.979, 0.545]$		-0.740 $[-2.073, 0.622]$		

Risk premia estimates and cross-sectional fit for a selection of models on a cross-section of 25 Fama-French monthly excess returns. Each model is estimated via OLS and GLS. We report point estimates and 95% confidence intervals for risk premia, which are constructed based on the asymptotic normal distribution, and cross-sectional R^2 and its (5%, 95%) confidence level constructed as in Lewellen et al(2010) for FM estimation. For Bayesian Fama-MacBeth estimation we report: posterior mean of λ ($\bar{\lambda}_j$), its (2.5%, 97.5%) credible interval, posterior mode and median of cross-sectional R^2_{adj} and centered 90% credible intervals. *, ** and *** denote, respectively, 90%, 95% and 99% level significance.

Model Uncertainty

Bayesian basics II

"All models are wrong, but some are useful."
Box (1976)

Model posterior probabilities:

E.g. suppose there are two possible models:

$$\begin{aligned}\Pr(\text{model } 0|\text{data}) &= \frac{\Pr(\text{data}|\text{model } 0)\Pr(\text{model } 0)}{\Pr(\text{data})} \\ &= \frac{\Pr(\text{data}|\text{model } 0)\Pr(\text{model } 0)}{\Pr(\text{data}|\text{model } 0)\Pr(\text{model } 0) + \Pr(\text{data}|\text{model } 1)(1-\Pr(\text{model } 0))}\end{aligned}$$

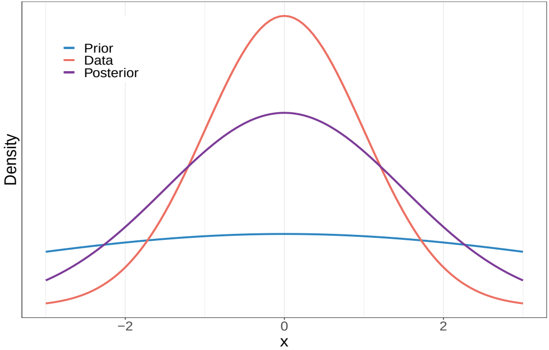
where $\Pr(\text{data}|\text{model } 0) = \int \text{likelihood}(\text{data}|\text{Model } 0, \theta^{(0)})\pi(\theta^{(0)})d\theta^{(0)}$, $\theta^{(0)}$ are the model 0 parameters and π their prior distribution.

Useful for:

- factor and model selection
- testing hypothesis and computing their *p*-values
- averaging/combine models (more on this later)

From prior to posterior probabilities

$$\textbf{Prob of model} \propto \textbf{Prob(data} \mid \textbf{model)} = \int \underbrace{\textbf{Likelihood} \times \textbf{prior}}_{\textbf{posterior}}$$



- If both prior and likelihood are (almost) flat, is the posterior a proper distribution?
- Can we integrate the area under it?
- With a flat prior for risk premia and weak factors present, posterior is not integrable
- Useless/level factors will be selected with probability approaching 1

Simulation: Probability of retaining risk factors with flat prior

	T	55%	57%	59%	61%	63%	65%
Panel A: strong factors							
f_{strong}	200	0.860	0.845	0.830	0.812	0.792	0.771
	600	0.987	0.985	0.985	0.983	0.981	0.979
	1000	0.998	0.998	0.997	0.996	0.996	0.995
Panel B: useless factors							
$f_{useless}$	200	1.000	0.995	0.982	0.940	0.862	0.726
	600	1.000	0.999	0.998	0.988	0.971	0.920
	1000	1.000	1.000	1.000	0.999	0.990	0.971
Panel C: strong and useless factors							
f_{strong}	200	0.928	0.912	0.891	0.878	0.860	0.838
	600	0.994	0.994	0.992	0.991	0.991	0.989
	1000	0.999	0.999	0.999	0.999	0.999	0.999
$f_{useless}$	200	0.955	0.894	0.788	0.642	0.489	0.360
	600	0.957	0.895	0.764	0.618	0.461	0.354
	1000	0.957	0.893	0.787	0.645	0.483	0.357

The table shows the frequency of retaining risk factors for different choice sets across 1,000 simulations of different size (T=200, 600, and 1,000). In Panel A, the candidate risk factor is truly cross-sectionally priced and strongly identified, while in Panel B they are not. Panel C summarizes the case of using both strong and useless candidate factors in the model. A candidate factor is retained in the model, if its posterior probability, $\Pr(\gamma_j = 1|data)$, is greater than a certain threshold, i.e. 55%, 57%, 59%, 61%, 63% and 65%.

“Spike-and-slab” prior as regularisation

A natural fix: regularisation to restore integrability.

We introduce a mixture prior for the risk premium of the *j*-th factor.

- When $\gamma_j = 1 =$ factor should be included → $\lambda_j | \sigma^2, \gamma_j = 1 \sim \mathcal{N}(0, \sigma^2 \psi_j)$.
- When $\gamma_j = 0 =$ factor should be excluded → prior is a Dirac at zero.

Heterogeneous shrinkage: we rely on the partial correlation between factors and test assets

$$\psi_j = \psi \times \boldsymbol{\rho_j}^\top \boldsymbol{\rho_j}$$

Note: it’s a prior on the *SR* of the factor:

$\psi \in [10, 20] \Rightarrow$ SRs as large as 0.92–2.62 are in 95% prior coverage for a strong factor.

Level factors: use demeaned correlations.

Does not affect strong factors, solves the problem of weak ones, and yields all the closed-form solutions.

Bayesian basics III

"Ex pluribus unum"
Augustine of Hippo (circa 397-400)

Bayesian Model Averaging:

If we are interested in some Δ well-defined for every model (risk premia, SR, or VaR):

$$\mathbb{E}\left[\Delta|\text{data}\right] = \sum_{j=0}^{M} \mathbb{E}\left[\Delta|\text{data}, \text{model} = j\right] \Pr\left(\text{model} = j|\text{data}\right)$$

where $\mathbb{E}\left[\Delta|\text{data}, \text{model} = j\right] = \lim_{L\to\infty} \frac{1}{L}\sum_{i=1}^{L} \Delta(\theta_i^{(j)})$ and $\theta_i^{(j)}$ are the j-th model posterior parameter draws.

Continuous spike-and-slab & how to handle quadrillion models

Given the number of factors in the literature, we might actually want to compare quadrillions, rather than millions, of models.

This can be done in our framework with two modifications:

1. Specify the λ prior as: $\lambda_j|\gamma_j, \sigma^2 \sim \mathcal{N}(0, r(\gamma_j)\psi_j\sigma^2)$
   where $r(1) = 1$ and $r(0) = r \lll 1$, i.e. we have a continuous Spike-and-slab.

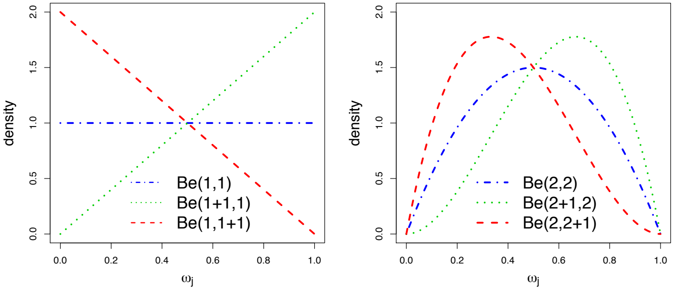
$\Rightarrow$ similar, integrable, posterior as before

2. Treat γ_j as a parameter to be sampled, so that factor posterior probabilities can be estimated as $\mathbb{E}[\gamma_j|\text{data}]$

$\Rightarrow$ Markov Chain ("Gibbs") over the space of possible models. Large computational efficiency gains since low probability models get endogenously under sampled.

Sampling γ_j and sparsity

Factor inclusion prior describes model sparsity: $\pi(\gamma_j = 1|\omega_j) = \omega_j, \quad \omega_j \sim Beta(a_\omega, b_\omega)$.



Note: beliefs about

1. Sharpe ratio achievable with one factor,
2. Sharpe ratio achievable in the economy,
3. sparsity of the true model

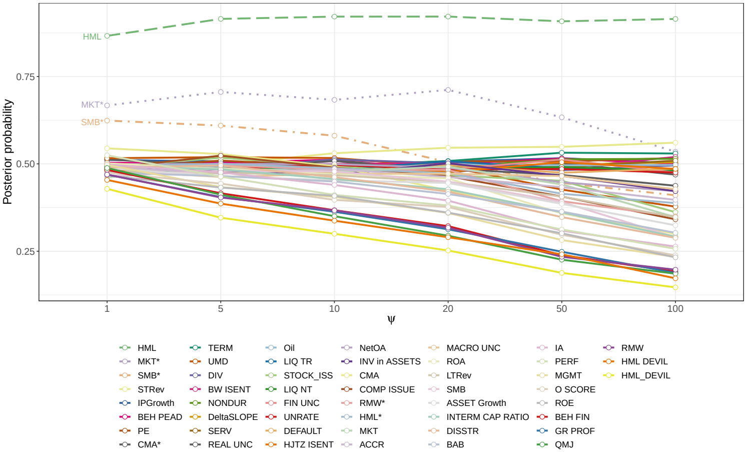
fully determine prior parameters.

Quadrillions of Models

Evaluating 2.25 quadrillion (10^{15}) models

- Test assets: 25 Fama-French size and book-to-market + 30 Industry portfolios
- We consider 51 notable factors proposed in the previous literature, yielding $2^{51} \approx 2.25$ quadrillion models,
  - $\approx$ 126 quadrillion (regressions)
  - $\approx$ 25,000 galaxies (in stars)
  - $\approx$ 15 brains (in synapses)
  - $\approx$ 63,000 Sala-i-Martin (1997) (at today's cpu speed)
- Extensions:
  - Out-of-sample exercise
  - Cross-sectional uncertainty: test assets
- Implemented in R, 10-18 cores (Python code is in progress!)

Posterior probabilities of the factors, $\mathbb{E}\left[\gamma_j|\text{data}\right]$



Posterior probabilities of factors, $\mathbb{E}\left[\gamma_j|\text{data}\right]$, estimated over the 1973:10-2016:12 sample using a cross-section of 25 Fama-French size and book-to-market and 30 Industry test asset portfolios. The prior distribution for the j-th factor inclusion is a $Beta(1,1)$, yielding a prior expectation for γ_j equal to 50%. Posterior probabilities are plotted for $\psi \in [1, \ 100]$.

Note: MKT* and SMB* = Daniel et al. (2020) MKT and SMB with hedged unpriced components.

Posterior factor inclusion probabilities and risk premia

	$\mathbb{E}[\gamma_j \mid \text{data}]$						$\mathbb{E}[\lambda_j \mid \text{data}]$						
	ψ:						ψ:						
Factors:	1	5	10	20	50	100	1	5	10	20	50	100	$\bar{F}$
HML	0.866	0.915	0.921	0.921	0.908	0.915	0.173	0.263	0.281	0.290	0.292	0.300	0.377
MKT⋆	0.667	0.706	0.683	0.712	0.633	0.535	0.074	0.170	0.207	0.259	0.268	0.229	0.514
SMB⋆	0.624	0.609	0.581	0.505	0.446	0.410	0.057	0.105	0.115	0.105	0.104	0.108	0.215
STRev	0.513	0.498	0.530	0.546	0.549	0.561	0.003	0.013	0.025	0.047	0.095	0.149	0.438
IPGrowth	0.511	0.507	0.488	0.506	0.516	0.502	0.000	-0.001	-0.001	-0.002	-0.005	-0.008	0.097*
BEH_PEAD	0.503	0.503	0.512	0.499	0.515	0.500	0.003	0.010	0.016	0.025	0.048	0.070	0.619
PE	0.486	0.509	0.494	0.508	0.507	0.517	-0.001	-0.003	-0.004	-0.005	-0.011	-0.019	6.770*
CMA⋆	0.513	0.495	0.509	0.486	0.468	0.437	0.001	0.000	-0.002	-0.004	-0.008	-0.010	0.242
TERM	0.477	0.478	0.494	0.508	0.532	0.530	0.001	0.003	0.006	0.011	0.024	0.038	0.962*
UMD	0.516	0.519	0.517	0.492	0.426	0.378	0.019	0.050	0.067	0.082	0.091	0.098	0.646
DIV	0.491	0.484	0.513	0.502	0.482	0.496	0.000	0.000	-0.001	-0.001	-0.003	-0.005	0.926*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
BAB	0.475	0.465	0.450	0.414	0.363	0.303	-0.011	-0.023	-0.024	-0.021	-0.013	-0.009	0.921
IA	0.498	0.476	0.440	0.395	0.309	0.263	-0.011	-0.014	-0.012	-0.011	-0.008	-0.008	0.409
PERF	0.525	0.463	0.411	0.381	0.312	0.257	-0.038	-0.055	-0.053	-0.053	-0.047	-0.047	0.651
MGMT	0.497	0.434	0.405	0.360	0.282	0.234	0.017	0.035	0.041	0.041	0.037	0.034	0.631
O_SCORE	0.475	0.443	0.397	0.377	0.297	0.238	-0.010	-0.015	-0.015	-0.015	-0.012	-0.004	0.020
ROE	0.467	0.432	0.408	0.360	0.301	0.232	0.005	0.015	0.020	0.023	0.022	0.017	0.555
BEH_FIN	0.483	0.416	0.367	0.322	0.239	0.191	-0.005	0.003	0.006	0.010	0.008	0.010	0.760
GR_PROF	0.469	0.405	0.363	0.313	0.248	0.188	0.012	0.011	0.011	0.011	0.011	0.008	0.199
QMJ	0.488	0.412	0.350	0.294	0.226	0.186	0.020	0.021	0.019	0.015	0.013	0.013	0.405
RMW	0.470	0.404	0.367	0.317	0.234	0.197	0.012	0.022	0.025	0.026	0.022	0.019	0.292
SKEW	0.454	0.386	0.337	0.290	0.241	0.173	0.000	0.000	0.000	0.000	0.000	0.000	0.438
HML_DEVIL	0.429	0.346	0.300	0.252	0.188	0.147	0.009	0.013	0.010	0.007	0.004	0.003	0.356

Posterior probabilities of factors, $\mathbb{E}\left[\gamma_j \mid \text{data}\right]$, and posterior mean of factor risk premia, $\mathbb{E}\left[\lambda_j \mid \text{data}\right]$. The last column reports sample average returns for the tradable factors. The data is monthly, 1973:10 to 2016:12. Test assets: cross-section of 25 Fama-French size and book-to-market and 30 Industry portfolios. Numbers denoted with the asterisk in the last column correspond to the return on the factor-mimicking portfolio of the nontradable factor, constructed by a linear projection of its values on the set of 51 test assets, and scaled to have the same volatility as the original nontradable factor. Factors have a prior probability of inclusion of 50%.

Post. factor probs and risk premia: no more than 5 factors (2.6M models)

	$\mathbb{E}[\gamma_j \mid \text{data}]$						$\mathbb{E}[\lambda_j \mid \text{data}]$						
	ψ:						ψ:						
Factors:	1	5	10	20	50	100	1	5	10	20	50	100	$\bar{F}$
HML	0.501	0.727	0.768	0.776	0.759	0.739	0.104	0.213	0.240	0.252	0.253	0.249	0.377
MKT*	0.212	0.362	0.386	0.393	0.381	0.354	0.030	0.114	0.149	0.178	0.199	0.197	0.514
SMB*	0.202	0.312	0.315	0.300	0.275	0.260	0.025	0.083	0.099	0.105	0.105	0.103	0.215
PERF	0.118	0.135	0.130	0.119	0.103	0.092	-0.013	-0.033	-0.038	-0.039	-0.037	-0.035	0.651
CMA	0.116	0.132	0.128	0.120	0.107	0.099	0.008	0.019	0.022	0.023	0.023	0.022	0.351
STOCK_ISS	0.095	0.112	0.123	0.129	0.125	0.117	-0.006	-0.022	-0.034	-0.045	-0.053	-0.055	0.515
COMP_ISSUE	0.096	0.108	0.115	0.119	0.117	0.110	0.007	0.026	0.040	0.054	0.065	0.068	0.497
MKT	0.074	0.069	0.085	0.110	0.132	0.133	0.003	0.012	0.025	0.045	0.068	0.075	0.563
UMD	0.087	0.089	0.093	0.093	0.089	0.085	0.004	0.013	0.019	0.025	0.029	0.032	0.646
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
QMJ	0.098	0.076	0.064	0.052	0.040	0.034	0.007	0.010	0.009	0.008	0.007	0.006	0.405
BAB	0.082	0.069	0.065	0.060	0.053	0.046	-0.003	-0.007	-0.008	-0.010	-0.011	-0.010	0.921
IA	0.087	0.071	0.064	0.056	0.048	0.042	-0.003	-0.005	-0.006	-0.006	-0.006	-0.005	0.409
O_SCORE	0.080	0.062	0.055	0.047	0.039	0.032	-0.002	-0.004	-0.005	-0.005	-0.004	-0.003	0.020
ROE	0.078	0.059	0.053	0.047	0.038	0.032	0.001	0.003	0.004	0.005	0.005	0.005	0.555
GR_PROF	0.081	0.056	0.045	0.037	0.028	0.022	0.004	0.005	0.004	0.003	0.002	0.002	0.199
BEH_FIN	0.077	0.055	0.046	0.038	0.030	0.024	-0.002	-0.001	0.000	0.001	0.001	0.001	0.760
SKEW	0.078	0.052	0.042	0.034	0.026	0.021	0.000	0.000	0.000	0.000	0.000	0.000	0.438
RMW	0.073	0.047	0.039	0.033	0.025	0.021	0.002	0.002	0.002	0.003	0.002	0.002	0.292
HML_DEVIL	0.065	0.040	0.032	0.025	0.019	0.015	0.001	0.002	0.002	0.001	0.001	0.001	0.356

Posterior probabilities of factors with Dirac spike-and-slab, and posterior mean of factor risk premia, $\mathbb{E}\left[\lambda_j \mid \text{data}\right]$. The data is monthly, 1973:10 to 2016:12. Test assets: cross-section of 25 Fama-French size and book-to-market and 30 Industry portfolios. Numbers denoted with the asterisk in the last column correspond to the return on the factor-mimicking portfolio of the nontradable factor, constructed by a linear projection of its values on the set of 51 test assets, and scaled to have the same volatility as the original nontradable factor. Factors have a prior probability of inclusion of 10.38%.

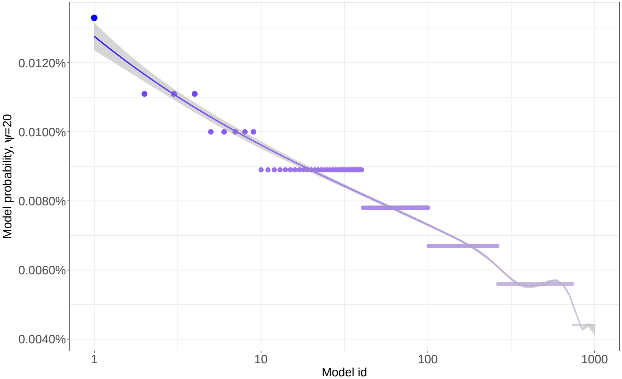
The most likely models

Factor models with highest posterior probability (continuous spike-and-slab, $\psi = 20$)

	model:									
factor:	1	2	3	4	5	6	7	8	9	10
HML	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MKT*	✓	✓		✓	✓	✓	✓	✓		✓
SMB*	✓		✓	✓	✓	✓		✓	✓	✓
STRev		✓		✓	✓	✓		✓	✓	
IPGrowth	✓	✓			✓	✓				
BEH_PEAD		✓	✓			✓	✓		✓	✓
PE	✓		✓		✓	✓			✓	
CMA*	✓		✓				✓	✓	✓	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
O_SCORE		✓	✓		✓		✓		✓	✓
ROE	✓			✓						
BEH_FIN		✓	✓							
GR_PROF					✓	✓			✓	✓
QMJ									✓	✓
RMW				✓	✓				✓	
SKEW	✓	✓								
HML_DEVIL										
Probability (%)	0.0133	0.0111	0.0111	0.0111	0.0100	0.0100	0.0100	0.0100	0.0100	0.0089

Factors and posterior model probabilities of ten most likely specifications computed using the continuous spike and slab approach, $\psi = 20$, 51 factors, and all possible models with up to 5 factors, yielding about 2.25 quadrillion models and a model prior probability of the order of 10^{-16}. Specifications organised by columns with the symbol ✓ indicating that the factor in the corresponding row is included. The data is monthly, 1973:10 to 2016:12. Test assets: cross-section of 25 Fama-French size and book-to-market and 30 Industry portfolios.

Model probabilities: massive model uncertainty



Posterior model probabilities, estimated over the 1973:10-2016:12 sample using a cross-section of 25 Fama-French size and book-to-market and 30 Industry test asset portfolios and $\psi = 20$.

Note: cumulative probability of best 1000 $\approx$ 6%

Posterior probabilities of notable models and “robust” factors model

We have found a lot of uncertainty about the “true” model...

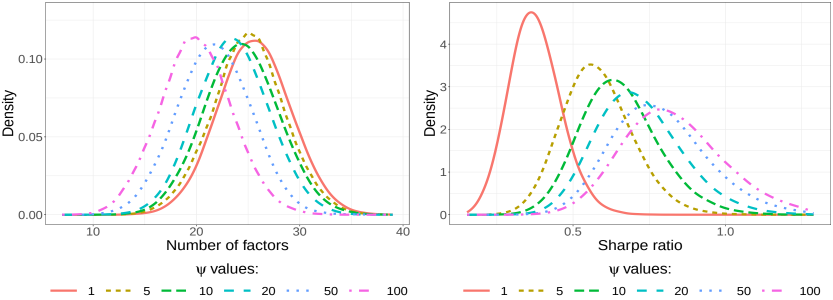
	ψ:					
model:	1	5	10	20	50	100
CAPM	0.02	0.01	0.01	0.01	0.03	0.07
Fama and French (1992)	0.05	0.01	0.01	0.01	0.01	0.00
Fama and French (2016)	0.09	0.06	0.05	0.04	0.03	0.02
Carhart (1997)	0.05	0.01	0.01	0.01	0.01	0.01
Hou, Xue, Zhang (2015)	0.02	0.01	0.00	0.00	0.00	0.00
Pastor and Stambaugh (2000)	0.02	0.01	0.01	0.01	0.02	0.02
Asness, Frazzini and Pedersen (2014)	0.10	0.06	0.04	0.04	0.03	0.02
Robust Factors Model (HML, MKT*, SMB*)	0.64	0.85	0.87	0.88	0.88	0.86

Posterior model probabilities for the specifications in the first column, for different values of ψ, computed using the Dirac spike-and-slab prior. The model in the last row uses the HML, MKT* and SBM* factors described. The data is monthly, 1970:01 to 2017:12. Test assets: cross-section of 25 Fama-French size and book-to-market and 30 Industry portfolios.

... but there is little uncertainty that the models we have are not good enough.

Note: none of the above models is among the 1000 most likely ones.

The posterior number of factors (sparsity in observables) and SR



⇒ evidence of commonality in the risks spanned by the various factors.

- no single (and small) set of factors
- many factors reflecting the same underlying risks
- some factors only help to pin down betas better, but are not sources of risk premia

OOS performance

Are we just overfitting?

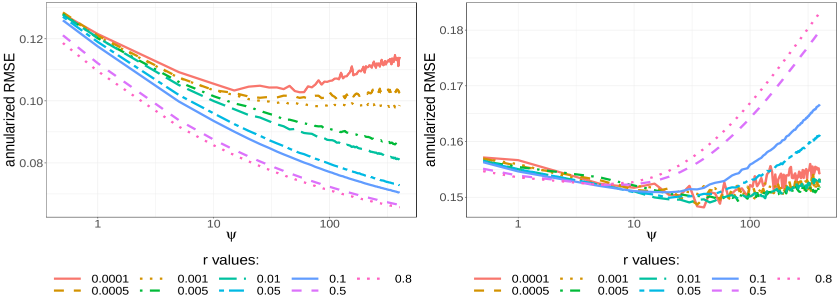
Many machine learning methods suffer from overfitting.

An out-of-sample exercise:

- split the sample in two
- estimate model parameters in one sample
- fit cross-sectionally on the other sample

Evaluate model fit in- and out-of-sample.

Are we just overfitting? Out-of-Sample performance



(a) In-sample pricing errors

(b) Out-of-sample pricing errors

Cross-sectional model fit RMSE obtained with a two-fold cross-validation for different values of shrinkage parameters, r and ψ. Units are annualized Sharpe ratios.

Economically-motivated priors help out-of-sample!

BMA captures about 85% of the total SR available.

Cross-Sectional Uncertainty

Cross-section sample

A systematic review of the recent papers with cross-sectional asset pricing:

- Everything published in top journals, 2007-2017
  JF, JFE, RFS, AER, Ecta, REStud, QJE, JME, MS, JFQA, RoF, AoF, JEF, JBF
- Everything presented at top conferences, 2007-2017
  AFA, WFA, NBER AP, NBER MF, NBER SI, EFA, MFA, NFA, SFS Cavalcade, Macro-Finance Society, Adam Smith, Gerzensee, Duke-UNC AP
- SSRN working papers with the upload date 2005-2017
  Keywords: "cross-section factor", "factor asset pricing", "cross-section of returns", "cross-sectional asset pricing"

$>$ 100 cross-sectional models

25 Potential cross-sections

ID	Cross-section	Weight	ID	Cross-section	Weight
1	25 size-and-b/m + 30 industry	32.09%	14	25 size-and-b/m + 25 size-and-mom + 5 b/m	0.61%
2	25 size-and-b/m + 49 industry	32.09%	15	49 industry + 5 b/m	1.50%
3	25 size-and-b/m + 10 mom + 30 industry	3.54%	16	30 industry + 25 size-and-mom	1.50%
4	25 size-and-b/m + 10 mom+ 25 size-and-mom	2.65%	17	49 industry + 25 size-and-mom	1.50%
5	10 mom + 49 industry	6.52%	18	25 size-and-b/m + 10 mom + 10 b/m + 10 size	0.36%
6	25 size-and-b/m+ 10 b/m + 30 industry	2.18%	19	25 size-and-b/m + 10 mom + 10 b/m + 25 size-and-mom	0.18%
7	25 size-and-b/m + 10 momentum + 49 industry	2.65%	20	10 momentum + 10 b/m + 49 industry	0.44%
8	25 size-and-b/m + 10 size + 30 industry	1.63%	21	25 size and b/m + 10 mom + 10 size + 25 size-and-mom	0.13%
9	10 b/m + 49 industry	4.01%	22	10 mom + 10 size + 49 industry	0.33%
10	25 size-and-b/m + 10 mom + 25 size-and-mom	1.22%	23	25 size-and-b/m + 10 mom + 30 industry + 5 bm	0.09%
11	10 size + 40 industry	3.01%	24	25 size and b/m + 10 b/m + 10 size + 30 industry	0.11%
12	25 size-and-b/m + 30 industry + 5 b/m	0.82%	25	10 b/m + 10 size + 49 industry	0.20%
13	25 size-and-b/m + 25 size-and-mom + 5 industry	0.61%			

Composite cross-sections built from the base test assets and their weight, based on the revealed preferences, which is calculated as the product of the empirical use frequency for each of the corresponding base assets.

Sample 2^{51} factor models for each of the cross-sections ≈ jointly evaluate 3.78 quintillion (10^{18}) regressions

Aggregation across cross-sections: revealed preferences i.e. proportional to the empirical probability of using each cross-section

Aggregating everything: Revealed Preferences

	$\mathbb{E}[\gamma_j\|\text{data}]$						$\mathbb{E}[\lambda_j\|\text{data}]$						
	ψ:						ψ:						
Factors:	1	5	10	20	50	100	1	5	10	20	50	100	$\bar{F}$
HML	76.44%	80.52%	79.93%	78.33%	75.92%	72.51%	0.130	0.204	0.218	0.224	0.226	0.221	0.377
MKT*	67.50%	69.97%	67.92%	65.50%	58.05%	51.27%	0.092	0.187	0.222	0.247	0.245	0.226	0.514
STRev	51.82%	52.61%	53.69%	55.36%	55.04%	52.57%	0.008	0.031	0.052	0.083	0.133	0.164	0.438
UMD	52.66%	52.73%	51.53%	49.04%	43.26%	37.21%	0.026	0.066	0.086	0.103	0.117	0.116	0.646
BEH_PEAD	50.93%	50.95%	51.42%	52.53%	52.87%	51.21%	0.004	0.012	0.020	0.033	0.058	0.078	0.619
ACCR	49.42%	50.18%	51.15%	49.97%	46.10%	41.42%	-0.010	-0.025	-0.036	-0.046	-0.060	-0.068	0.343
SMB*	57.64%	54.77%	51.10%	46.09%	40.80%	36.37%	0.040	0.068	0.073	0.071	0.073	0.074	0.215
DIV	49.41%	49.11%	50.78%	50.11%	49.00%	49.10%	0.000	0.000	0.000	0.000	0.000	0.000	0.933*
BW_ISENT	49.19%	50.32%	50.35%	50.83%	49.77%	49.19%	0.000	0.001	0.002	0.004	0.007	0.011	0.055*
CMA*	49.95%	50.26%	50.27%	48.73%	46.94%	43.51%	0.000	0.000	-0.001	-0.003	-0.005	-0.006	0.242
LIQ_TR	49.39%	49.92%	50.16%	50.92%	50.64%	49.78%	0.000	0.004	0.010	0.022	0.051	0.081	0.438
REAL_UNC	49.56%	49.06%	50.05%	48.51%	48.26%	45.37%	0.000	0.000	0.000	0.000	0.000	0.000	0.046*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
DISSTR	50.23%	46.43%	42.97%	37.00%	30.13%	24.49%	0.064	0.092	0.094	0.088	0.078	0.069	0.475
O_SCORE	47.33%	44.47%	40.64%	36.89%	29.23%	23.52%	-0.012	-0.019	-0.020	-0.020	-0.015	-0.010	0.020
SMB	43.72%	41.54%	39.74%	36.44%	29.82%	23.45%	0.021	0.036	0.041	0.042	0.038	0.031	0.257
ROE	47.18%	43.32%	39.19%	34.14%	27.31%	21.84%	0.004	0.010	0.014	0.016	0.015	0.013	0.555
PERF	51.69%	44.62%	38.96%	34.21%	27.05%	22.11%	-0.041	-0.053	-0.049	-0.045	-0.039	-0.036	0.651
MGMT	49.13%	42.81%	38.78%	33.65%	26.03%	20.70%	0.019	0.036	0.039	0.039	0.034	0.029	0.631
GR_PROF	47.85%	41.41%	37.31%	32.07%	25.35%	19.74%	0.023	0.032	0.035	0.035	0.031	0.026	0.199
BEH_FIN	47.84%	40.62%	36.23%	31.04%	23.24%	18.48%	-0.001	0.010	0.014	0.017	0.015	0.013	0.760
SKEW	46.66%	40.57%	36.12%	30.93%	24.75%	19.08%	0.000	0.000	0.000	0.000	0.000	0.000	0.004
RMW	45.71%	38.48%	34.48%	29.51%	22.36%	18.05%	0.006	0.012	0.015	0.016	0.014	0.013	0.292
QMJ	46.70%	37.54%	32.16%	26.92%	20.00%	16.01%	0.012	0.009	0.007	0.006	0.005	0.006	0.405
HML_DEVIL	44.46%	36.28%	32.04%	27.27%	20.88%	16.48%	0.018	0.029	0.029	0.026	0.021	0.017	0.356

Posterior probabilities of factors, $\mathbb{E}\left[\gamma_j|\text{data}\right]$, and posterior mean of factor risk premia, $\mathbb{E}\left[\lambda_j|\text{data}\right]$. The last column reports sample average returns for the tradable factors. The data is monthly, 1973:10 to 2016:12. Output is weighted average across the estimates over 25 cross-sections. Numbers denoted with the asterisk in the last column correspond to the weighted average (over cross-sections) return on the factor-mimicking portfolio of the nontradable factor, constructed by a linear projection of its values on the set on test assets, and scaled to have the same volatility as the original nontradable factor. Factors have a prior probability of inclusion of 50%.

Conclusion

Conclusions

I. We provide a unified setting for estimating linear factor models that

1. can handle quadrillion models and directly reflects the underlying uncertainty,
2. is valid for both tradable and non tradable factors,
3. remains valid under misspecification and weak factors,
4. is easy to use and generates confidence bands for <u>everything</u>

II. What did we learn?

1. only a handful of factors are robust sources of risk premia,
2. massive model uncertainty: a substantial concern for risk management,
3. none of the traditional 3-5 factor models beat even top 1000,
4. best models include a large number of factors,
5. averaging across models is the key for reliable inference, both in- and out-of-sample.