The 2nd HKUST IAS-SBM Joint Workshop on Financial Econometrics in the Big Data Era

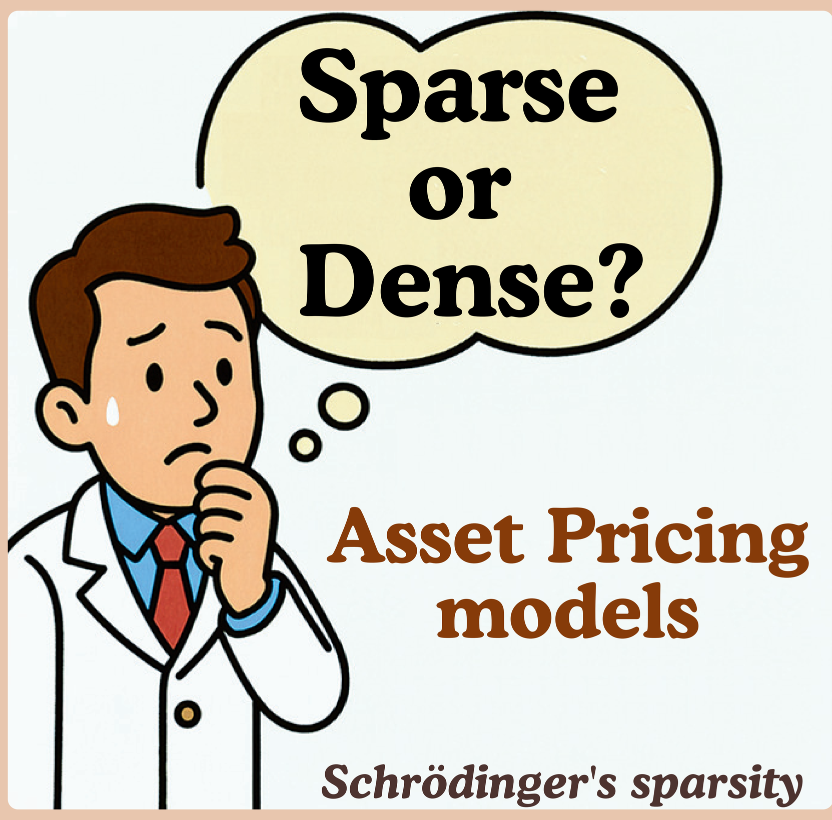
Schrödinger's Sparsity

In the Cross Section of Stock Returns

Doron Avramov | Reichman University
Guanhao Feng
Jingyu He
Shuhua Xiao | City University of HK

SSRN

Personal Web.



Motivation

- Sparse modeling: L_1 penalty, Lasso regression
- Dense modeling: L_2 penalty, Ridge regression

Traditional AP models demand an ex ante decision on sparsity or density. → Empirical findings frequently mirror prior assumptions instead of revealing structure of expected returns.

Can sparsity be treated not as a fixed assumption, but as an inferred property of the data?

The nature of AP models — sparse or dense — are in a state of superposition until empirical data is observed.

Methodology

Conditional latent factor framework of IPCA

$$r_{i,t} = \boldsymbol{\alpha}(\mathbf{Z}_{i,t-1}) + \boldsymbol{\beta}(\mathbf{Z}_{i,t-1})^\top \mathbf{f}_t + \epsilon_{i,t}$$
$$\boldsymbol{\alpha}(\mathbf{Z}_{i,t-1}) = \alpha_0 + \boldsymbol{\alpha}_1^\top \mathbf{Z}_{i,t-1}$$
$$\boldsymbol{\beta}(\mathbf{Z}_{i,t-1}) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1(\mathbb{I}_K \otimes \mathbf{Z}_{i,t-1})$$
$$\epsilon_{i,t} \sim \mathcal{N}(0, \sigma_i^2)$$

$\mathbf{f}_t : K$ latent factors (can be extended to both observable and latent factors).

$\mathbf{Z}_{i,t-1} : L$ lag characteristics

Spike-and-slab prior

$$\beta = \begin{cases} 0 \text{ with prob. } q & \text{Regressor is not chosen} \\ \mathcal{N}(0, \gamma^2) \text{ with prob. } 1-q & \text{Regressor is chosen} \end{cases}$$

Standard spike-and-slab prior: q is a specific value.
Giannone, Lenza, and Primiceri (ECTA 2021):
q has its prior → sample $q \sim \text{Beta}(a, b)$

q: 0 ⟵⟶ 1
lower prob. of sparsity — higher prob. of sparsity

$$r_{i,t} = \alpha_0 + \boldsymbol{\alpha}_1^\top \mathbf{Z}_{i,t-1} + \boldsymbol{\beta}_0^\top \mathbf{f}_t + \boldsymbol{\beta}_1^\top [\mathbf{f}_t \otimes \mathbf{Z}_{i,t-1}] + \epsilon_{i,t}$$

Separate priors

Different prob. of sparsity of alpha and beta.

$$[\boldsymbol{\alpha}_1]_l \sim \begin{cases} \mathcal{N}(0, \gamma_\alpha^2) & \text{if } d_l^\alpha = 1 \\ 0 & \text{if } d_l^\alpha = 0 \end{cases} \qquad [\boldsymbol{\beta}_1]_l \sim \begin{cases} \mathcal{N}(0, \gamma_\beta^2) & \text{if } d_l^\beta = 1 \\ 0 & \text{if } d_l^\beta = 0 \end{cases}$$
$$d_l^\alpha \sim \text{Bernoulli}(1 - q_\alpha) \qquad d_l^\beta \sim \text{Bernoulli}(1 - q_\beta)$$
$$q_\alpha \sim \text{Beta}(a_{q_\alpha}, b_{q_\alpha}) \qquad q_\beta \sim \text{Beta}(a_{q_\beta}, b_{q_\beta})$$
$$\gamma_\alpha^2 \sim \mathcal{IG}(A_{\gamma_\alpha}/2, B_{\gamma_\alpha}/2) \qquad \gamma_\beta^2 \sim \mathcal{IG}(A_{\gamma_\beta}/2, B_{\gamma_\beta}/2)$$

Higher post. mean of q_α or q_β, higher prob. of sparsity.

Separate joint priors

Prior settings of $q \neq$ precise control of sparsity levels!

$$(d_1^\alpha, d_2^\alpha, \cdots, d_L^\alpha) \sim \left[\prod_{l=1}^{L} \text{Bernoulli}(1 - q_\alpha)\right] \times \mathbb{I}\left(\sum_{l=1}^{L} d_l = M_\alpha\right),$$
$$(d_1^\beta, d_2^\beta, \cdots, d_L^\beta) \sim \left[\prod_{l=1}^{L} \text{Bernoulli}(1 - q_\beta)\right] \times \mathbb{I}\left(\sum_{l=1}^{L} d_l = M_\beta\right).$$

Larger M_α or M_β, lower sparsity level.

Extensions

Without mispricing: $r_{i,t} = \boldsymbol{\beta}_0^\top \mathbf{f}_t + \boldsymbol{\beta}_1^\top [\mathbf{f}_t \otimes \mathbf{Z}_{i,t-1}] + \epsilon_{i,t}$

Other factors: $r_{i,t} = \boldsymbol{\alpha}(\mathbf{Z}_{i,t-1}) + \boldsymbol{\beta}(\mathbf{Z}_{i,t-1}) \left[\mathbf{f}_t^O, \mathbf{f}_t^L\right] + \epsilon_{i,t}$

Data

1990-2024

Cross-sectional
- P-Tree (Cong, Feng, He, and He, JFE 2025)
- Portfolios
  - 25 ME/BM portfolios
  - 360 bivariate-sorted portfolios
  - 610 univariate-sorted portfolios
- Individual stocks
  - stocks ranked 1st to 500th by ave ME
  - stocks ranked 501st-1000th by ave ME

Time-series
- Regime1/ Regime2/ Regime3
  - Breakpoints in Smith and Timmermann (RFS 2021): July 1998 and June 2010.
- Normal & Recession period
  - Define recession periods based on the Sahm Rule (88 months)

Empirical Results

Table: Model Performance Under Diff. Priors (K=5)

Panel A: Unrestricted # sel char.		CSR^2	(q_α, q_β)	(M_α, M_β)	Panel B: Fixed #sel char.		CSR^2	Panel C: No sparsity
	0.9 0.9	58.9	0.93, 0.64	1,10		2,2	48.4	(M_α, M_β)
	0.5 0.9	57.0	0.77, 0.64	1,10		10,2	50.0	20,20
	0.1 0.9	56.6	0.63, 0.66	1,10		18,2	37.8	
(q_α, q_β) prior mean	0.9 0.5	**59.9**	0.93, 0.50	**1,10**	(M_α, M_β)	**2,10**	**59.6**	CSR^2
	0.5 0.5	58.8	0.79, 0.50	1,10		10,10	41.1	45.2
	0.1 0.5	58.1	0.64, 0.49	1,10		18,10	39.5	
	0.9 0.1	58.3	0.92, 0.33	1,11		2,18	56.1	
	0.5 0.1	57.9	0.79, 0.34	1,11		10,18	51.0	
	0.1 0.1	53.7	0.62, 0.35	2,10		18,18	42.1	

Probability of sparsity

- Between the extremes of highly sparse (prob→1) and fully dense (prob→0).
- Mispricing: higher sparsity than loading
- Sparsity ∼ number of latent factors K
- Robust across prior settings.

Misspecified Assum. of Sparsity

- Model performance peaks: Fixed inclusion sizes in the constrained model match sparsity levels of probabilistic model.

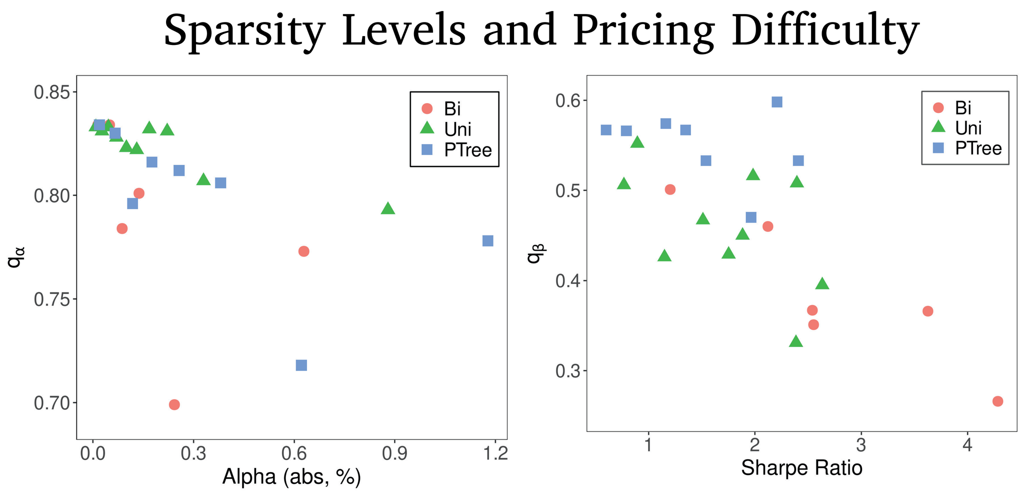
Learn rather than impose sparsity in conditional asset pricing models.

Schrödinger's Sparsity : Test Asset & Marcro Regimes

Table: Sparsity for Diff. Test Assets

	CSR^2	(q_α, q_β)
Panel A: P-Tree		
100	42.4	**0.69,0.43**
200	51.0	**0.60,0.37**
400	45.2	**0.54,0.32**
Panel B: Ind. Stock		
500 big	31.4	0.61, 0.29
500 small	3.9	**0.49,0.38**
Panel C: Others		
ME/BM25	33.6	0.80,0.50
Bi360	7.8	0.50,0.20
Uni610	48.0	0.44,0.20

Sparsity Levels and Pricing Difficulty



Sparsity levels vary across test assets, reflecting pricing difficulty differences.

Table: Sparsity in Diff. Regime

	CSR^2	(q_α, q_β)
Panel A: Seq. seg.		
Regime1	48.5	0.72,0.56
Regime2	24.1	0.71, 0.53
Regime3	59.7	0.77, 0.46
Panel B: Macro-driven. seg.		
Normal	53.8	0.67,0.46
Recession	14.2	**0.76,0.50**

Sparsity Prob. change across both cross-sectional and time-series dimensions.
⇒ i) Test assets / Pricing difficulty
⇒ ii) Time periods / Macro conditions

Assuming AP model to be either sparse or dense ex ante may be wrong.

Model with Observable and / or Latent Factors

	CSR^2	(q_α, q_β)		CSR^2	(q_α, q_β)
Panel A: only obs.			*Panel C: obs+latent.*		
MKT	**14.9**	0.55,0.37	MKT+LF1	**53.9**	0.69,0.35
FF3	27.3	0.65,0.26	MKT+LF5	56.5	0.79,0.48
FF5	50.4	0.74,0.39	FF3 +LF1	41.6	0.67,0.27
Panel B: only latent.			FF3+LF5	57.4	0.80,0.56
LF1	29.5	0.52,**0.47**	FF5+LF1	50.6	0.67,0.35
LF3	45.0	0.68,**0.58**	FF5+LF5	55.8	0.79,0.58
LF5	56.8	0.77,**0.66**			

Table: Augmented Observable Factor Models

Highlights

An important problem: ***How can researchers determine the appropriate model assumption without first examining the data?***

A new approach: ***Flexible Bayesian framework for IPCA***
- Endogenously determine whether the model is sparse or dense
- Exogenously control the sparsity level of the model

Empirical findings: ***How, when, and why firm characteristics matter in the cross section of returns***