Functional Clustering of Financial Time Series in Stata

Penalised B-splines, functional PCA and Ward classification via Stata's Python integration

Carlo Drago[1] Angelo Leogrande[2]

[1]University Niccolò Cusano, Rome, Italy
[2]LUM University "Giuseppe Degennaro", Casamassima, Italy

2026 Italian Stata Conference – Milan, 17–18 September 2026

Values marked [a] come from the authors' archived run and were not recomputed for this version.

Question and contribution

- Many financial series share one calendar: which units behave alike over time?
- Moments (mean, volatility, beta) discard *when* things happen; distances on raw monthly returns are exposed to short-run noise. FDA compares smoothed trajectories instead.

Research question

Can whole return trajectories be grouped inside a reproducible, Stata-orchestrated workflow, and how strong is the resulting grouping?

1. A two-stage pipeline: penalised B-splines (GCV) $\rightarrow$ functional PCA $\rightarrow$ Ward clustering (silhouette).
2. Stata sets parameters, runs Python, imports, labels and archives the results.
3. An application to 35 US-listed technology-related stocks, with explicit evidence on how weak the separation is.

Data and sample design

- 36 requested tickers; 35 downloaded. `SQ` returned no data, consistent with Block, Inc.'s ticker change to `XYZ` announced in January 2025.
- Yahoo Finance adjusted closes via the unofficial `yfinance` package; month-end values; log returns $r_{it} = \log P_{it} - \log P_{i,t-1}$.
- Only dates common to all series are kept (balanced design). Requested start 2013, but SNOW and PLTR were first listed in Sept. 2020, so returns run Oct. 2020 – Jul. 2026.

Series n	35
Monthly returns T	70
Basis functions K	30
Retained FPCs M	5
Candidate k	2–6
Seed	42

Reproducibility caveat

Adjusted prices are revised over time: all results refer to the archived return snapshot, not to a fresh download. The universe is the authors' selection from one provider.

Stage 1: penalised B-spline representation

- Each curve is expanded on $K = 30$ clamped cubic B-splines, $x_i(t) \approx \phi(t)^\top \boldsymbol{c}_i$, with design matrix $\boldsymbol{B} \in \mathbb{R}^{T \times K}$ (full column rank here).
- Roughness is controlled by a second-difference penalty on the coefficients, $\boldsymbol{R} = \boldsymbol{D}_2^\top \boldsymbol{D}_2$ (P-splines, Eilers & Marx, 1996):

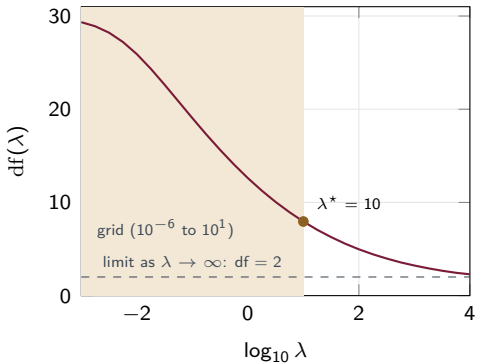
$$\hat{\boldsymbol{c}}_i(\lambda) = \arg\min_{\boldsymbol{c}} \ \|\boldsymbol{y}_i - \boldsymbol{B}\boldsymbol{c}\|^2 + \lambda\, \boldsymbol{c}^\top \boldsymbol{R}\boldsymbol{c} = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{R})^{-1} \boldsymbol{B}^\top \boldsymbol{y}_i, \qquad \hat{\boldsymbol{X}} = \boldsymbol{S}_\lambda \boldsymbol{Y}.$$

- $\boldsymbol{S}_\lambda = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{R})^{-1} \boldsymbol{B}^\top$ is common to all series: one linear solve smooths the whole panel.
- Complexity is measured by the effective degrees of freedom $\mathrm{df}(\lambda) = \mathrm{tr}(\boldsymbol{S}_\lambda)$, from 30 ($\lambda = 0$) down to 2 ($\lambda \to \infty$).
- As $\lambda \to \infty$ the coefficients become linear in k; with clamped knots the fitted curve is then close to, but not exactly, a straight line near the endpoints.

Choosing λ: GCV and a boundary solution

$$\mathrm{GCV}(\lambda) = \frac{\|\boldsymbol{Y} - \boldsymbol{S}_\lambda \boldsymbol{Y}\|_F^2}{T\, n\, [1 - \mathrm{df}(\lambda)/T]^2}$$

- One common λ for the panel; grid $10^{-6}, \ldots, 10^{1}$ (8 values).
- GCV decreases over the whole grid[a]: the minimum is the upper bound, $\lambda^\star = 10$, i.e. $\mathrm{df} = 7.95$.
- So $\lambda^\star$ is *the most smoothing the grid allows*, not a verified optimum. The smoothed curves therefore describe a low-frequency component of monthly returns.



Recomputed for $T = 70$, $K = 30$: df = 12.61 ($\lambda = 1$), 7.95 (10), 4.97 (100), 3.23 (10^3).

Functional PCA and the choice of distance

- On a common equispaced grid, FPCA of the smoothed curves reduces to a standard PCA of $\hat{\boldsymbol{X}}^{\top}$ (series in rows), up to a constant quadrature weight.
- $\hat{x}_i(t) \approx \bar{x}(t) + \sum_{m=1}^{M} \theta_{im}\xi_m(t)$; with $M = 5$ the scores retain 91.4% of the variance of the smoothed curves[a].
- Signs of components are arbitrary; any economic reading of ξ_m requires inspecting the loading curves.

Standardising the scores is a modelling choice

- *Unscaled* scores: Euclidean distance $\approx$ (scaled) L^2 distance between rank-M reconstructions; FPC1 weighs most.
- *Standardised* scores (used here): $d_Z^2(i,j) = \sum_m (\theta_{im} - \theta_{jm})^2 / s_m^2$, a whitened distance that gives low-variance components more weight and may amplify noise.

Comparing standardised and unscaled partitions is part of the robustness agenda (slide 13).

Stage 2: Ward clustering and the choice of k

- Ward merges the pair of groups with the smallest increase in within-group sum of squares, $\Delta(A, B) = \frac{|A||B|}{|A|+|B|} \|\bar{\boldsymbol{z}}_A - \bar{\boldsymbol{z}}_B\|^2$.
- No random start: one tree; every k is a cut of the same hierarchy.
- Average silhouette $\bar{s}(k)$ on the standardised scores (singletons get $s_i = 0$); $k^\star = \arg\max_{k \in \{2,\ldots,6\}} \bar{s}(k)$.

k	$\bar{s}(k)$ [a]
2	0.211
3	0.225
4	0.220
5	**0.246**
6	0.243

Reading

$k^\star = 5$ has the highest value, but the profile is flat: $k = 6$ is almost equivalent (and adds a singleton), and every value is below 0.25. The choice of k is weakly identified.

Implementation: Stata orchestrates, Python computes (schematic)

```
version 18.0
global fda_tickers "MSFT GOOG AMZN ... MU"
global fda_from    "2013-01-01"
global fda_to      "YYYY-MM-DD"  // fixed end date
global fda_nbasis 30
global fda_nharm   5
global fda_k_min   2
global fda_k_max   6
global fda_seed   42
global fda_outdir "`c(pwd)'/fda_stock_output"
python query                    // logs interpreter
python:
  # reads settings via sfi.Macro.getGlobal()
  # download -> GCV -> FPCA -> Ward -> CSV
end
import delimited using "$fda_outdir/stock_clusters.
    csv", clear varnames(1)
label variable hierarchical_cluster "Functional
    FPCA Ward cluster"
save "$fda_outdir/stock_clusters.dta", replace
```

- Python: `scipy` (B-splines, Ward), `numpy` (solves), `scikit-learn` (PCA, silhouette), `yfinance`.
- A dependency check stops the run early with the install command for that interpreter.
- Exported: download status, returns, GCV and silhouette tables, smoothed curves, scores, labels.
- The original do-file sets `fda_to` from the system date; a fixed end date and an archived input file are needed for exact replication.

Results: five descriptive groups

Cluster	n	Members[a]
1	6	PYPL, PANW, CRWD, DOCU, TEAM, NET
2	18	MSFT, GOOG, AMZN, AAPL, TSLA, IBM, ORCL, ADBE, CRM, CSCO, QCOM, TXN, AVGO, INTU, ADSK, NOW, SNOW, DDOG
3	3	INTC, AMD, MU
4	2	NVDA, SMCI
5	6	META, NFLX, SHOP, ZM, PLTR, UBER

- No sector label enters the estimator; groups come from five standardised scores of smoothed return trajectories.
- The universe itself was chosen by sector, so agreement with sector intuition is a plausibility check, not a validation.
- One group holds about half of the sample (18/35); two groups have 2–3 members.

Cluster profiles

Cl.	n	Mean ann. log return[a]	Mean stock vol.[a]	Descriptive label (ex post)
4	2	44.7%	63.5%	"AI infrastructure"
3	3	32.7%	56.3%	"semiconductors"
2	18	9.9%	35.8%	"mega-cap and enterprise core"
5	6	7.8%	49.6%	"digital platforms"
1	6	5.8%	50.0%	"cloud, security and payments"

- Stock-level statistics, $\mu_i = 12\bar{r}_i$ and $\sigma_i = \sqrt{12}\, s_i$, averaged within clusters: not portfolio returns or volatilities.
- Dispersion within clusters is large; e.g. terminal wealth (Sept. 2020 = 1) ranges from 2.16 to 21.62 in cluster 3 and from 0.18 to 12.99 in cluster 5[a].
- Labels are names for groups, not estimated factors. Some memberships are debatable (AVGO in cluster 2, PLTR in cluster 5), consistent with weak separation.

How strong is the grouping?

- Reference scenarios: $n = 35$, $M = 5$ independent coordinates $\rightarrow$ standardise $\rightarrow$ Ward $\rightarrow$ $\max_{k=2..6} \bar{s}(k)$; 5,000 replications each, seed 42. Observed value: 0.2458[a].

Reference distribution	**mean**	**95th pct.**	$\Pr(\max_k \bar{s} \geq 0.2458)$
Gaussian	0.207	0.253	0.076 (MC s.e. 0.004)
Student-t_3	0.299	0.562	0.46 (MC s.e. 0.007)

- The observed silhouette is not beyond the 95th percentile even under the Gaussian scenario, and is typical under heavy tails.
- k-means ($k = 5$, `n_init`$= 10$) gives silhouette 0.204 and adjusted Rand index 0.518 with Ward[a]: partial agreement.
- These scenarios ignore the dependence among stocks; they are not a calibrated financial null.

Conclusion

The partition is a descriptive segmentation. The evidence does not establish distinct latent populations, nor does it rule them out.

What the results do and do not show

Supported

- The pipeline is driven by one Stata do-file and returns a labelled `.dta`.
- Model selection is recorded (GCV and silhouette tables), not only the selected values.
- The groups are interpretable and broadly consistent with sector intuition.
- Separation is weak, and this is quantified.

Not established

- An optimal smoothing level (boundary GCV solution).
- Stability of memberships under resampling, other windows, K, M or scaling.
- That groups reflect factors, betas or causal regimes.
- Any predictive or pricing value of the groups.

Limitations and next steps

Limitations

- λ and k are selected sequentially; their joint uncertainty is not propagated.
- Curves share common shocks: they are not independent draws.
- Smoothed paths are descriptive; volatility dynamics and phase shifts are not modelled.
- Balanced design: 70 months; pre-2020 history is discarded.
- Selected universe, single data provider, survivorship and ticker changes.

Next steps

- Extend the λ grid beyond 10 and report the GCV minimum, if interior.
- Unscaled vs standardised scores; sensitivity to K, M and window.
- Bootstrap co-assignment frequencies for each pair of stocks.
- A dependence-preserving null (e.g. block bootstrap of the panel).
- Model-based alternatives (funHDDC); a Stata command `fdacluster`.

Conclusions

1. Return series can be compared as smoothed curves; this targets timing information that moment-based summaries discard.
2. A transparent pipeline – P-splines with GCV, FPCA, Ward with silhouette – runs inside a Stata-orchestrated workflow with recorded model selection.
3. In 35 technology-related stocks it yields five interpretable groups, but the separation is weak: silhouette 0.246, not exceptional under simple reference scenarios.
4. The main open issues are the boundary choice of λ and the stability of memberships.

Take-away for Stata users

FDA-based clustering fits in a Stata workflow. Archive the input snapshot, fix the sample window, and report diagnostics of cluster strength with the partition.

References

- Bouveyron, C. & Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5(4), 281–300.
- Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31, 377–403.
- Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–121.
- Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.
- Jacques, J. & Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3), 231–255.
- Ramsay, J.O. & Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. Springer.
- Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- StataCorp (2023). *Stata 18 Programming Reference Manual*, [P] `python`. Stata Press.
- Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

Acknowledgements

Carlo Drago gratefully acknowledges the support of the Department of Economic, Psychological, Communication, Educational, and Motor Sciences and of the Università degli Studi Niccolò Cusano, Rome, which made this work possible.

Thank you · carlo.drago@unicusano.it