Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Finanziato nell'ambito del Piano Nazionale di Ripresa e Resilienza PNRR. Missione 4, Componente 2, Investimento 1.3 Creazione di "Partenariati estesi alle università, ai centri di ricerca, alle aziende per il finanziamento di progetti di ricerca di base"

GRINS FOUNDATION

GRINS – Growing Resilient, INclusive and Sustainable

"9. Economic and financial sustainability of systems and territories"

Codice Identificativo: **PE00000018**

Finanziato nell'ambito del Piano Nazionale di Ripresa e Resilienza PNRR Missione 4 – Componente 2

SPOKE 4

D4.1.3 – Robust Latent Markov Modeling for EU SMEs Under Outlier Risk

May 2025

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Catania
GRINS FOUNDATION

Robust Latent Markov Modeling for EU SMEs Under Outlier Risk

Roberto Di Mari, Antonio Punzo

Executive Summary

This policy brief introduces a novel statistical framework designed to capture latent behavioral dynamics in European small and medium-sized enterprises (SMEs), accounting for data contamination and transition uncertainty. The model integrates a bias-adjusted three-step latent Markov estimation procedure with the robust Optimally Tuned Robust Improper Maximum Likelihood Estimator (OTRIMLE). The model is implemented entirely in-house using custom-developed R code, tailored to analyze SME balance sheet data under real-world imperfections. While empirical application is forthcoming, the model has been extensively tested on simulated data, demonstrating strong performance in accurately recovering latent states and transitions in the presence of outliers. Once estimated on the EU SMEs sample, the latent state variable is mapped to ESG (Environmental, Social, and Governance) scores, enabling a forward-looking interpretation of firm sustainability. This methodology opens a pathway toward robust, longitudinal risk segmentation for SMEs, with potential for broader deployment across economic and financial domains.

Context and Importance of the Issue

Small and medium-sized enterprises (SMEs) are vital to the EU economy, driving employment, innovation, and regional integration. However, analyzing their financial trajectories poses two central challenges:

(1) the presence of outliers, due to atypical firm behavior or reporting errors; and

(2) pervasive missing data, arising from irregular reporting or incomplete disclosures.

To address these issues, we develop a robust statistical framework tailored to balance sheet aggregates of EU-listed SMEs. The model combines robust clustering via OTRIMLE and a multiple imputation strategy based on random forests into a time-dynamic hidden (latent) Markov specification.

Multiple imputation yields several completed datasets that preserve complex nonlinear relationships. These are then used to fit the bias-adjusted latent Markov model, enabling inference on firm transitions and covariate effects. Crucially, this process incorporates uncertainty from both imputation and classification, allowing for more accurate estimation of structural parameters and their variability.
By capturing latent firm states and transitions over time—even in the presence of data contamination and missingness—this approach supports the development of statistically sound, temporally consistent firm classifications for use in policy, risk analysis, and sustainability reporting.

Methodological Innovation

This framework advances the state-of-the-art in robust latent modeling by integrating:

- OTRIMLE, a robust estimator for Gaussian mixtures that models the outliers via an improper density component—optimally tuned to balance fit and contamination tolerance.
- A longitudinal latent Markov model, which captures the dynamics of unobserved firm states across time.
- A bias-adjusted three-step estimator (Di Mari, Oberski & Vermunt, 2016), which corrects for classification error when estimating covariate and transition effects.

The implementation was carried out entirely in R using custom code, ensuring full control over simulation, estimation, and diagnostics. The model is designed to handle unbalanced panels, moderate sample sizes, and multiple covariates.
This framework can be generalized to even more complex specifications, including models with cell-wise contamination, where errors affect individual entries rather than full observations.

Key Findings from Simulation Studies

The current development phase has focused on extensive simulation-based evaluation. Simulated datasets were designed to reflect the structure and contamination characteristics observed in real-world SME data. Key results include:

- The model accurately recovers latent states and transition probabilities in the presence of up to 20% contamination.
- Classification accuracy remains high even under low entropy scenarios and moderate sample sizes.

- The bias-adjusted three-step estimator yields consistent and efficient parameter estimates, avoiding distortions typical of naive stepwise approaches.
- The OTRIMLE component offers flexibility in handling noise without requiring explicit trimming or manual exclusion.
- The mapping of latent states to ESG dimensions shows promising alignment in simulations, supporting the interpretability of the state variable as a sustainability proxy.

Policy and Analytical Implications

1. **Toward ESG-Based Firm Typologies.** The ability to link latent behavioral states with ESG outcomes opens opportunities for sustainable finance and regulatory segmentation.
2. **Resilient Risk Classification Under Data Imperfection.** The robust model design ensures that outlier-prone firm data does not distort macro-level inferences or mislead policy actions.
3. **Methodological Infrastructure for Future Applications.** With a modular structure and open-source implementation, the framework can be extended to other domains such as healthcare, public finance, or regional policy.

Recommendations

1. **Integrate Robust Latent Models in SME Analytics.** Policymakers and financial analysts should adopt robust Markov-based segmentation for high-noise corporate datasets.
2. **Support Development and Dissemination of Tools.** Support Development and Dissemination of Tools.
3. **Enable ESG-Behavioral Integration.** Promote research into linking latent financial trajectories with sustainability indicators, ensuring alignment with EU Green Deal objectives.

Implementation Considerations

1. **Data Readiness.** A successful application requires high-frequency balance sheet data across multiple time periods with standardized accounting formats.
2. **Computation.** Although the method is computationally intensive, our R implementation includes optimized routines and scaling strategies.

3. **Capacity Building.** Statistical and econometric training should be strengthened to support adoption in central banks, national agencies, and research institutions.

Conclusion

This policy brief presents a pioneering approach for analyzing longitudinal SME data under real-world imperfections. By integrating robust clustering with temporal modeling and ESG mapping, the method offers a promising tool for policymakers and analysts seeking resilient, interpretable insights into firm behavior. The model has been successfully validated on simulated data, and empirical applications to EU SMEs are forthcoming.

Acknowledgement

This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP E63C22002120006). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

References

[1] Coretto, P., & Hennig, C. (2016). Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. *Journal of the American Statistical Association*, *111*(516), 1648-1659.
[2] Di Mari, R., Oberski, D. L., & Vermunt, J. K. (2016). Bias-Adjusted Three-Step Latent Markov Modeling With Covariates. Structural Equation Modeling, 23(5), 649–660.
[3] Vermunt, J. K. (2010). Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. Political Analysis, 18(4), 450–469.