Fast Super-Paramagnetic Clustering: Mapping Stock Market Interactions to Spin Models

Lionel Yelibi Tim Gebbie

Department of Statistical Science, University of Cape Town, South Africa

Introduction & Objectives

We address the problem of unsupervised statistical learning for feature selection and classification of financial time-series data. By mapping stock market interactions to **Potts Spin Models**, we recover hierarchical structures without pre-defining the number of clusters.

Core comparison:

- **SPC:** Simulated Annealing based Super-Paramagnetic Clustering (Thermodynamic approach).
- **f-SPC:** A modified Maximum Likelihood approach using Parallel Genetic Algorithms (PGA).

Key Finding: f-SPC solutions converge to those found within the Super-Paramagnetic Phase where entropy is maximized, offering qualitatively better results for high-dimensional datasets.

The Potts Model & Phase Transitions

We model data points as interacting spins. The system is governed by the Hamiltonian Energy equation:

$$\mathrm{H_S} = \sum_{\langle \mathrm{i,j} \rangle} \mathrm{J_{ij}}(1 - \delta_{\mathrm{s_i,s_j}}) \tag{1}$$

where S is the spin vector, and $\mathrm{J_{ij}}$ is the interaction strength decreasing with distance $\mathrm{d_{ij}}$ (e.g., $\mathrm{J_{ij}} \propto \mathrm{e}^{-\mathrm{d_{ij}^2}}$).

Phases of Matter as Clustering States

- **Ferromagnetic ($\mathrm{T} < \mathrm{T_c}$):** Spontaneous magnetization. All spins aligned (One giant cluster).
- **Super-Paramagnetic ($\mathrm{T} \approx \mathrm{T_c}$):** The giant cluster breaks into smaller, stable clusters. *This is the region of interest.*
- **Paramagnetic ($\mathrm{T} \gg \mathrm{T_c}$):** Complete disorder ($\langle \mathrm{m} \rangle \approx 0$).

Methodology: SPC vs. f-SPC

1. Super-Paramagnetic Clustering (SPC)

- Uses **Swendsen-Wang MCMC** for sampling configurations.
- Clusters identified via the **Hoshen-Kopelman** algorithm.
- Requires tuning Temperature T. Validity checked via Susceptibility χ.

2. Fast SPC (f-SPC) - Maximum Likelihood

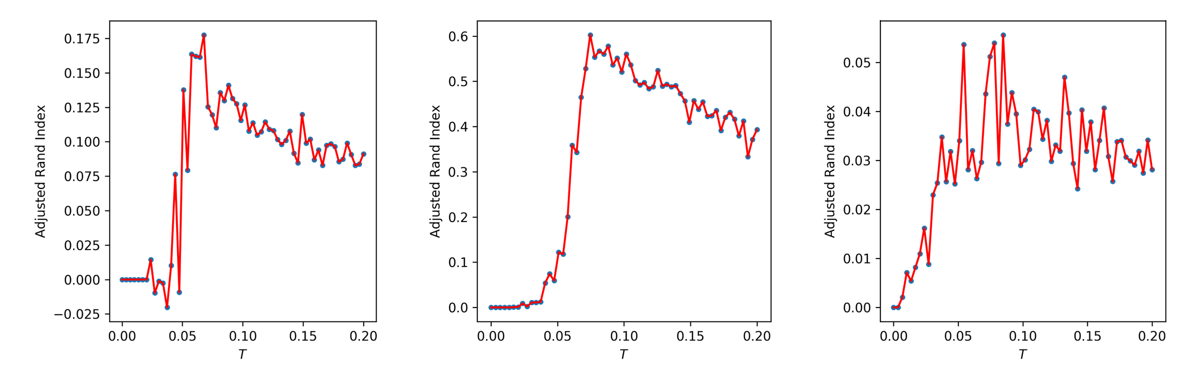
- Parameter-free optimization of the Likelihood $\mathrm{L_c}$:

$$\mathrm{L_c} = \frac{1}{2} \sum_{\mathrm{s:n_s>1}} \ln \frac{\mathrm{n_s}}{\mathrm{c_s}} + (\mathrm{n_s} - 1) \ln \frac{\mathrm{n_s^2} - \mathrm{n_s}}{\mathrm{n_s^2} - \mathrm{c_s}} \tag{2}$$

- Implemented via a **Parallel Genetic Algorithm (PGA)** without crossover to maximize speed.
- Mutations include: Split, Merge, Swap, Scramble, and Flip.

Phase Validation

We validated that maximizing Likelihood ($\mathrm{L_c}$) corresponds to the thermodynamic Super-Paramagnetic phase.



(a) Market Mode (b) IMN (c) RMT

Figure: Adjusted Rand Index (ARI) vs Temperature. Maxima occur in the SP-phase where susceptibility is non-zero

We confirmed that f-SPC maximizes the entropy of the system:

- Free energy $\mathrm{F} = \mathrm{U} - \mathrm{TS}$ reaches a local minimum in the SP-phase.
- f-SPC solutions align with maximum entropy configurations.

Data Pre-Processing & Noise Cleaning

Financial correlation matrices are inherently noisy. We applied two cleaning methods:

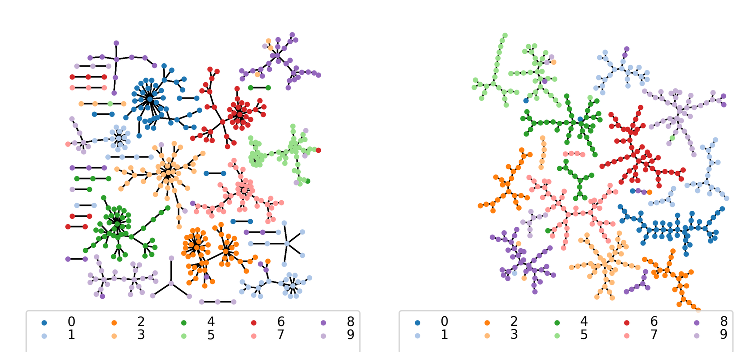
1. **Random Matrix Theory (RMT):** Filters eigenvalues outside the Wishart range ($\lambda_{\max}, \lambda_{\min}$) to remove noise.
2. **Iterative Matrix Normalization (IMN):** Standardizes covariance iteratively to center correlations.

Datasets Tested:

- **Toy:** Concentric circles, Wines, Iris, MNIST digits ($\mathrm{D} = 64$).
- **Real:** NYSE S&P 500 (447 stocks, 1249 days).
- **Real:** BRICS (226 stocks).

Results: High Dimensionality (MNIST)

f-SPC excels in high dimensions. For MNIST ($\mathrm{D} = 64$), f-SPC recovered distinct clusters for digits despite non-linear handwriting styles.

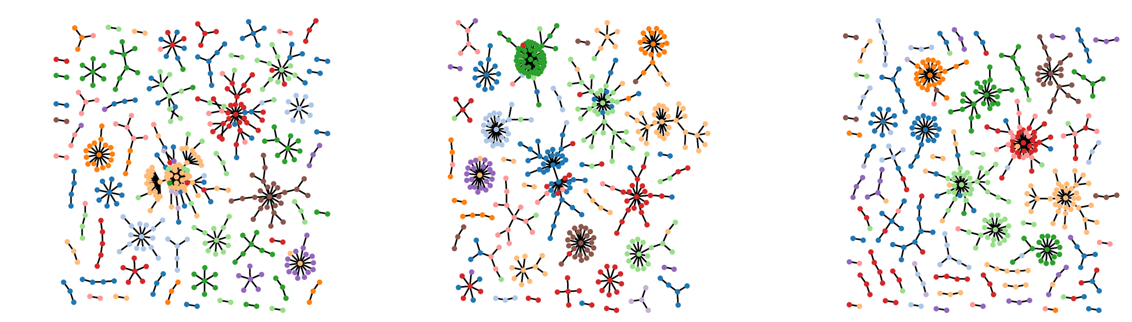


(a) SPC (b) f-SPC

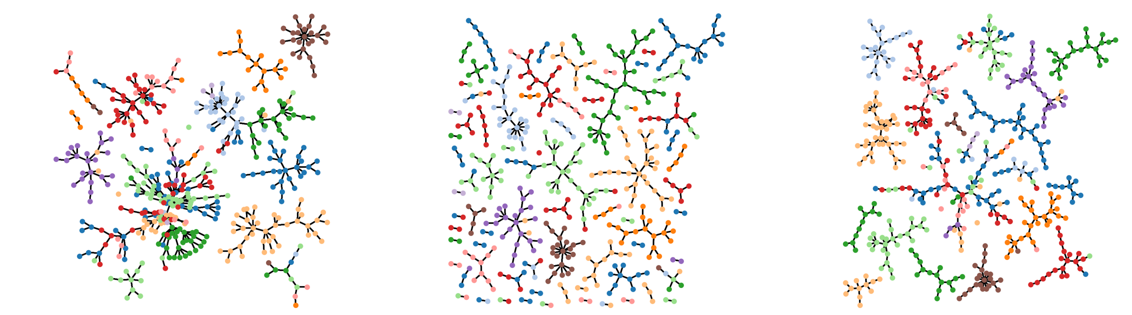
Figure: MST of f-SPC solution on MNIST. ARI = 0.747. It handles varying cluster densities better than DBSCAN (ARI=0)

Application: NYSE S&P 500 Clustering

We compared algorithm outputs against the Global Industry Classification Standard (GICS).



(a) Market Mode - SPC (b) IMN - SPC (c) RMT - SPC



(d) Market Mode - f-SPC (e) IMN - f-SPC (f) RMT - f-SPC

Figure: (Up) SPC and (Down) f-SPC solutions. Colors represent GICS sectors. f-SPC with RMT cleaning produces cleaner separation of industries

Insights:

- **Market Mode:** A large "market" cluster often obscures sector details. RMT cleaning reduces this effect.
- **Adaptive Markets:** Mixed clusters (e.g., Amazon clustering with IT rather than Consumer Discretionary) highlight the adaptive nature of markets, challenging static classifications like GICS.
- **Performance:** f-SPC solutions generally had higher Likelihood values than SPC candidates across the temperature range.

Discussion & Conclusion

- **Convergence:** f-SPC (Maximum Likelihood) naturally converges to the maximum entropy solutions found in the Super-Paramagnetic phase of SPC.
- **High Dimensionality:** f-SPC is more robust for high-dimensional data (e.g., MNIST, clean financial matrices) where thermodynamic approaches may struggle with neighborhood definitions.
- **Unsupervised Utility:** The method effectively recovers economic sectors without prior knowledge, making it a powerful tool for analyzing *Dynamical Cluster Analysis* (DCA) during financial crises.

References

1 L. Yelibi and T. Gebbie, "Fast Super-Paramagnetic Clustering," arXiv:1810.02529v2, 2019.
6 M. Blatt et al., "Superparamagnetic clustering of data," Phys. Rev. Lett. 76, 1996.
19 L. Giada and M. Marsili, "Data clustering and noise undressing," Phys. Rev. E 63, 2001.
24 D. Hendricks et al., "Detecting intraday financial market states," Quant. Finance 16, 2016.
54 F.Y. Wu, "The Potts model," Rev. Mod. Phys. 54, 1982.

Code available at: github.com/tehraio/potts-model-clustering