Causal Regime-Aware Portfolio Allocation under Residual Stress and Tail-Risk Transmission

Andres Schwartz
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
schwartz193@buffalo.edu

Nitin Jha
Department of Computer Science, University of Houston, Houston, TX, USA.
nitin.jha@uh.edu

Yangnan Fan
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
yangnan972@missouri.edu

Abstract

Contemporary portfolio allocation faces fundamental limitations when confronted with non-stationary financial regimes and latent stress dynamics that elude traditional volatility-based risk measures. This paper develops a causal regime-aware allocation framework that integrates residual stress signals and tail-risk transmission channels to improve portfolio robustness under extreme market conditions. We argue that standard mean-variance optimization and its extensions fail to capture the structural persistence of drawdown regimes and the contagion mechanisms through which tail risks propagate across asset classes. By incorporating causal inference techniques to identify regime transitions and residual stress metrics that are leakage-safe against common volatility anomalies, the proposed architecture offers a systems-level reconfiguration of portfolio design. We examine the structural trade-offs between stability, adaptability, and computational tractability, emphasizing the role of governance frameworks and fairness constraints in deployment. The analysis draws on cross-domain comparisons from network epidemiology, critical infrastructure resilience, and macroeconomic stress testing to illustrate how residual stress signals can be operationalized within a causal graph of asset dependencies. Policy implications are discussed in the context of systemic risk regulation, transparency requirements, and the ethical use of predictive signals in automated investment systems. The paper concludes with forward-looking recommendations for integrating causal regime awareness into institutional portfolio governance, highlighting both the promise and the pitfalls of stress-aware allocation in the presence of model uncertainty and data sparsity.

Keywords

causal regime detection, residual stress signal, tail-risk transmission, portfolio allocation, systemic risk, drawdown risk, governance, infrastructure resilience, fairness.

1. Introduction

The financial system has undergone profound structural transformations over the past two decades, with increasing interconnectedness among asset classes, the proliferation of algorithmic trading, and the growing prominence of tail events that challenge conventional

risk management frameworks. Traditional portfolio theory, rooted in the Markowitz mean-variance paradigm, assumes stationary return distributions and symmetric risk preferences that are ill-suited to environments characterized by abrupt regime shifts and cascading losses [1]. While subsequent advances such as the Black-Litterman model and dynamic conditional correlation models have improved estimation efficiency, they remain fundamentally reliant on second-moment statistics that are slow to reflect latent structural stress [2, 3]. The emergence of causal inference methods in financial econometrics offers a new avenue for disentangling the underlying drivers of regime changes, moving beyond mere correlation to identify the mechanisms that transmit stress across assets [4, 5]. This paper proposes a causal regime-aware portfolio allocation framework that integrates a residual stress signal—measuring the deviation from equilibrium expected losses under normal conditions—and explicitly models tail-risk transmission channels as a network of causal dependencies.

The motivation for this work stems from the recurring observation that drawdowns, rather than volatility, constitute the primary source of investor discomfort and institutional failure. Volatility-based risk measures, such as Value-at-Risk and expected shortfall, are pro-cyclical and can exhibit dangerous hysteresis during prolonged stress periods [6, 7]. Recent research has introduced the concept of a leakage-safe residual stress signal that isolates the component of drawdown risk not captured by standard volatility models, thereby providing a more persistent and interpretable indicator of systemic fragility [8]. However, the full potential of such signals can only be realized when embedded within a causal regime detection framework that identifies the structural breaks in asset return dynamics. By treating regimes as latent causal states, the allocation problem becomes one of sequential decision-making under uncertainty about the current regime and the probability of future transitions [9]. This perspective naturally accommodates the transmission of tail risks through feedback loops, such as margin calls, forced deleveraging, and cross-asset contagion, which are absent in standard equilibrium models.

The architecture proposed here is not merely an incremental extension of existing models but a fundamental reconfiguration of the portfolio allocation pipeline. It comprises three interconnected modules: a causal regime detection engine that learns the graph of causal dependencies among assets using historical stress episodes; a residual stress estimator that filters out volatility-induced noise to reveal persistent drawdown pressure; and a tail-risk transmission simulator that propagates stress through the causal graph under hypothetical scenario shocks. The output feeds into a constrained optimization that balances expected returns, residual stress exposure, and network-level connectivity to achieve robustness. This systems-level view aligns with recent advances in critical infrastructure protection, where resilience is measured not only by the ability to absorb shocks but also by the capacity to rewire connections after disruption [10, 11]. By drawing analogies to power grid blackout propagation and epidemic spreading, we can formalize tail-risk transmission as a dynamical process on a weighted directed graph, where nodes represent asset classes or sectors and edges represent causal influence strengths estimated from historical data.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature and positions our contribution. Section 3 develops the causal regime detection methodology and stress modeling approach. Section 4 examines the nature of residual stress and the specific channels through which tail risks are transmitted. Section 5 details the allocation framework and its systems architecture. Section 6 discusses governance, fairness, and policy

implications. Section 7 addresses deployment challenges and sustainability considerations. Section 8 concludes with a synthesis of findings and future research directions.

2. Background and Related Work

Portfolio allocation has historically been dominated by the mean-variance framework proposed by Markowitz, which optimally trades off expected return against variance under the assumption of multivariate normality [1]. Despite its elegance, the framework suffers from well-known instabilities: small changes in input estimates lead to large swings in optimal weights, and it does not account for the fat tails and asymmetries observed in financial data [2]. The Black-Litterman model attempts to mitigate estimation errors by blending subjective views with market equilibrium, but it retains the same second-moment risk measure [3]. Regime-switching models, notably the Markov switching approach of Hamilton, introduced the idea that return distributions evolve according to an unobserved state process, allowing for time-varying parameters that capture bull and bear markets [12]. Subsequent extensions to multivariate settings and stochastic volatility have enriched the empirical toolkit, yet these models typically treat regime transitions as exogenous and do not explicitly model the causal mechanisms that trigger shifts [13, 14].

The concept of tail risk has been central to post-crisis regulatory reforms. The Global Financial Crisis of 2007–2008 exposed the inadequacy of Gaussian-based risk measures, leading to the adoption of expected shortfall and stress testing frameworks [6]. However, these measures are often backward-looking and fail to capture the endogenous buildup of fragility during calm periods. Network approaches to systemic risk, pioneered by Allen and Gale and later developed by Adrian and Brunnermeier, model the propagation of defaults through interbank linkages, but their application to multi-asset portfolios remains limited [15, 16]. The work of Billio et al. used Granger causality and network measures to identify connectedness among financial institutions, yet Granger causality does not imply causal structure in the Pearlian sense [4, 17]. More recently, Peters, Janzing, and Schölkopf have advocated for the use of causal discovery algorithms in finance to uncover the underlying data-generating processes that drive asset returns [5].

A critical gap in the literature is the absence of a metric that jointly captures the persistence of stress and its network-driven amplification. Standard volatility measures are contaminated by short-term noise and are prone to leakage—the inclusion of non-stress-related fluctuations that obscure true drawdown risk. The residual stress signal proposed by Liu addresses this by constructing an anomaly detector that isolates the component of drawdown attributable to structural imbalances rather than normal market noise [8]. This signal, being leakage-safe, offers a more reliable input for regime detection and portfolio rebalancing. However, the integration of such a signal into a causal regime-aware allocation framework has not been systematically explored. Our work fills this gap by embedding the residual stress signal within a causal graph that governs regime transitions, thereby enabling a forward-looking allocation strategy that anticipates tail-risk transmission.

3. Causal Regime Detection and Stress Modeling

The foundation of our framework is a causal model of regime dynamics that distinguishes between exogenous shocks and endogenous amplification. Regimes are defined as stable configurations of the joint distribution of asset returns, parameterized by mean vectors, covariance structures, and crucially, causal connectivity patterns. During normal regimes, asset dependencies are predominantly driven by common risk factors such as interest rates,

inflation, and economic growth. In stress regimes, however, causal links become denser and more directional, as fire sales, flight to quality, and contagion effects dominate [18]. We employ a two-stage detection procedure. First, we use a regime-switching hidden Markov model with time-varying transition probabilities that depend on a set of macroeconomic and financial stress indicators. Second, we apply structural causal discovery algorithms, such as the PC algorithm or GES, on the residuals of the regime-specific models to infer the directed acyclic graph of causal relationships among assets within each regime [5, 19]. This dual approach ensures that the causal graph is not static but adapts to the prevailing regime.

The stress modeling component operationalizes the residual stress signal. Let the expected loss under normal conditions be estimated from a baseline regression that includes lagged returns, volatility factors, and a set of control variables. The residual stress signal is then defined as the difference between the actual drawdown and the baseline-expected drawdown, normalized by the baseline standard deviation of drawdowns [8]. This signal is leakage-safe because it removes the component of drawdown that is predictable from standard volatility information, thereby isolating the abnormal stress that accrues from structural imbalances. We extend this concept to a multi-asset setting by computing a portfolio-level residual stress score as a weighted sum of individual asset signals, where weights are derived from the causal graph to account for transmission effects. This aggregated measure serves as the primary risk input in the allocation optimization, replacing traditional portfolio variance.

An important innovation is the estimation of regime-specific causal strengths. For each edge in the inferred causal graph, we compute a time-varying coefficient using rolling window regression with a regularization penalty that enforces sparsity outside of stress regimes. During normal times, many edges are effectively zero, indicating limited contagion potential. During stress regimes, the number and magnitude of causal edges increase, reflecting the activation of tail-risk transmission channels. This empirical observation is consistent with the notion of "dark matter" in financial networks—latent linkages that become material only during crises [15]. The causal regime detection engine therefore learns not only when regimes change but also how the architecture of risk transmission transforms during those transitions.

4. Residual Stress and Tail-Risk Transmission Channels

Understanding the nature of residual stress requires a careful examination of the mechanisms through which tail risks are transmitted across assets. Traditional models attribute tail risk to common shocks, such as a sudden increase in interest rates or a sovereign default, which simultaneously affect multiple assets. While common shocks are important, they do not explain the cascading patterns observed during events like the 2008 crisis, where the failure of a single institution (Lehman Brothers) triggered widespread contagion across asset classes that had no direct exposure to that institution. The causal regime-aware perspective posits that tail-risk transmission occurs along three primary channels: the leverage channel, the liquidity channel, and the correlation channel [20]. The leverage channel operates when asset price declines force leveraged investors to meet margin calls by selling other assets, thereby transmitting stress from one asset to another via portfolio rebalancing. The liquidity channel emerges when market makers withdraw from trading in stressed assets, causing bid-ask spreads to widen and reducing the ability to hedge, which in turn amplifies price dislocations. The correlation channel reflects the structural increase in cross-asset correlations during crises, driven by the common reaction of investors to heightened uncertainty rather than fundamental linkages.

The residual stress signal is particularly sensitive to the leverage and liquidity channels because these are not captured by standard volatility models. For example, a leveraged hedge fund facing redemptions may be forced to sell liquid assets such as large-cap equities even if the initial shock originates in illiquid credit markets. The resulting price decline in equities is a residual stress event—it cannot be predicted from past volatility patterns because it arises from a non-market-fundamental force. By monitoring residual stress at both the asset and portfolio levels, the framework can detect the early stages of such contagion chains before they amplify into full-blown crises. Moreover, the causal graph learned from data can identify which assets serve as hubs or bridges in the transmission network, enabling the portfolio manager to preemptively reduce exposure to those nodes during stress regimes.

A concrete illustration can be drawn from the COVID-19 market turmoil in March 2020. During that period, the initial shock from lockdowns led to a sharp decline in travel and hospitality stocks, which then transmitted to investment-grade corporate bonds as investors fled to cash. The residual stress signal for corporate bonds would have spiked well before the observed increase in volatility, because the sell-off was driven by liquidity needs rather than a reassessment of credit fundamentals. A causal regime-aware allocation that detected the transition to a stress regime and identified the causal pathway from equity sectors to bond markets could have dynamically reduced exposure to corporate bonds, mitigating losses. This example underscores the practical value of integrating residual stress and transmission modeling into portfolio decisions.

5. Allocation Framework and Systemic Architecture

The allocation framework is built on a hierarchical architecture that separates regime detection, stress estimation, and optimization into distinct but interacting modules. At the top level, the causal regime detection engine continuously monitors a set of macroeconomic variables, market aggregates, and the residual stress signal to classify the current regime as either normal or stressed, with possible subcategories such as early stress, acute crisis, and recovery. Each regime is associated with a precomputed causal graph and a set of regime-specific asset return forecasts. The middle layer consists of the residual stress estimator, which computes the portfolio-level residual stress score using the current causal graph and individual asset signals. This estimator also produces a decomposition of stress contributions by node and by channel, enabling the optimizer to target specific vulnerabilities. The bottom layer is the constrained optimization module, which maximizes expected return subject to a cap on the portfolio residual stress score, a limit on network centrality-based concentration (to avoid excessive exposure to contagion hubs), and a standard set of long-only or leverage constraints.

The optimization is formulated as a convex program when the risk measure is linear in weights, which is the case for residual stress scores that are weighted sums of asset-level signals. This tractability is a deliberate architectural choice: it allows for real-time rebalancing even in high-dimensional spaces, as is common in large institutional portfolios. However, the regime-switching nature introduces non-convexity because the causal graph and stress estimates change discontinuously at regime boundaries. To address this, we implement a smooth transition mechanism that blends the stress estimates of adjacent regimes using a sigmoidal function of the predicted regime probability. This ensures continuity in the optimal portfolio weights and avoids destabilizing jumps during regime transitions. The smooth blending also reflects the realistic gradual nature of regime changes, where the market does

not instantaneously switch from normal to stressed but rather undergoes a period of heightened fragility.

The systemic architecture mirrors that of adaptive control systems in engineering, where a plant (the portfolio) is controlled by a feedback loop that updates its parameters based on observed outputs. In our case, the output is the realized drawdown and residual stress, and the control parameters are the causal graph estimates and regime probabilities. This cybernetic perspective highlights important feedback effects: as the portfolio adjusts its weights to reduce residual stress, it may alter the very causal relationships it seeks to avoid. For instance, reducing exposure to a contagion hub can lower that hub's market impact, potentially shifting the causal graph. This endogeneity implies that the allocation framework must be re-estimated periodically, and the optimization should incorporate a moving horizon that anticipates the portfolio's own effect on the stress landscape. We propose a two-timescale approach: a slow timescale for graph learning and regime estimation (e.g., monthly updates) and a fast timescale for weight adjustments (e.g., daily or intraday). This separation is common in multi-agent systems and avoids the computational burden of fully dynamic causal estimation.

6. Governance, Fairness, and Policy Implications

The deployment of a causal regime-aware portfolio allocation system raises significant governance and fairness challenges that extend beyond traditional risk management. First, the use of causal inference and machine learning models introduces opacity that can undermine accountability. If a portfolio suffers large losses during a regime that the model failed to detect, who is responsible: the model designer, the portfolio manager, or the regulator who approved the methodology? Transparency in the causal graph estimation process is essential. We advocate for a governance structure that requires the publication of the estimated causal graphs for each regime, along with the supporting data and significance tests, so that independent auditors can validate the model's logic. This is analogous to the model risk management guidelines issued by central banks for stress testing models [21].

Fairness concerns arise from the potential for the residual stress signal to disproportionately penalize certain asset classes or sectors that are inherently more vulnerable to liquidity-driven sell-offs. For example, small-cap equities and emerging market bonds typically exhibit higher residual stress during global crises, not because they are fundamentally riskier, but because their market structure makes them more susceptible to contagion. A purely mechanical allocation that minimizes residual stress could systematically underweight these assets, reducing diversification benefits and concentrating holdings in large-cap, highly liquid securities. This could exacerbate inequality in capital allocation, as smaller issuers face higher funding costs. To mitigate this, the allocation framework should incorporate a fairness constraint that ensures minimum weights for each asset class, calibrated to a benchmark that reflects the true economic exposure rather than the liquidity-driven stress component. Moreover, the governance board should include representatives from diverse market segments to ensure that the model's outputs are not inadvertently biased.

Policy implications extend to systemic risk regulation. If multiple large institutional investors adopt similar causal regime-aware allocation strategies, there is a risk of herding behavior: during a stress regime, all systems might simultaneously reduce exposure to contagion hubs, triggering a coordinated sell-off that destabilizes those very hubs. This is a classic macroprudential concern, analogous to the pro-cyclicality of Value-at-Risk models [7]. Regulators should consider requiring that such models include a "systemic stress load," an additional penalty on portfolios that are highly correlated with the aggregate market's residual

stress signal. This would internalize the externality of collective de-leveraging. Furthermore, the causal graph estimation should be required to include explicitly the effects of policy interventions, such as central bank asset purchases, to avoid overstating the persistence of stress regimes. The work of Acemoglu et al. on network resilience provides theoretical foundations for such macroprudential rules [22].

7. Deployment and Sustainability Considerations

Deploying a causal regime-aware allocation system at scale requires careful attention to computational infrastructure, data quality, and organizational change management. The causal graph learning step is computationally intensive, particularly for portfolios with hundreds of assets. We recommend a staged deployment: first, learn the graph on a reduced set of asset classes (e.g., major equity indices, bonds, commodities, and currencies) using daily data over a historical period that spans at least two full business cycles. Then, expand to individual securities within each class using a hierarchical approach, where lower-level graphs are conditional on the higher-level regime. This hierarchical architecture reduces the dimensionality and exploits the natural clustering of assets. The residual stress signal computation is relatively lightweight and can be parallelized across asset silos. To ensure operational resilience, the system should have a fallback mode that reverts to a simple volatility-based allocation if the causal regime detection engine produces unstable or implausible outputs, as determined by a set of pre-defined sanity checks (e.g., the graph must be acyclic and the residual stress metric must be bounded).

Sustainability of the model over time is another critical concern. Financial markets evolve: new asset classes emerge, regulatory changes alter trading dynamics, and the structure of causal relationships shifts due to technological innovation (e.g., the rise of algorithmic trading, the introduction of central bank digital currencies). The model must be re-calibrated periodically, with a retraining cycle that balances responsiveness against overfitting. We recommend a rolling window approach with a minimum ten-year history, updated annually, but with the option for emergency recalibration during a major regime shift (e.g., the onset of a systemic crisis). The re-calibration should include backtesting on out-of-sample stress events to verify that the residual stress signal and causal graph continue to align with observed contagion patterns. This is analogous to the concept of "model drift" in machine learning, but with the added complexity that the ground truth—causal structure—is not directly observable. Therefore, we advocate for the inclusion of a human-in-the-loop in the governance process, where a committee of quantitative analysts and domain experts reviews the updated graphs and approves any changes to the allocation strategy.

Finally, the ethical dimension of using predictive signals in portfolio management must be addressed. The residual stress signal may implicitly encode information about the vulnerability of certain groups of investors (e.g., retail investors in small-cap funds) to forced selling, and acting on that signal could accelerate the very stress it predicts. This is a form of self-fulfilling prophecy. We argue that such concerns are inherent to all risk management systems, but the causal regime-aware framework, by making the transmission channels explicit, at least allows for a transparent discussion of trade-offs. One possible mitigation is to impose a delay on rebalancing decisions during periods of extreme residual stress, akin to circuit breakers in equity markets, to prevent overreaction. Such mechanisms should be codified in the portfolio's investment policy statement and subject to board oversight.

8. Conclusion

This paper has presented a causal regime-aware portfolio allocation framework that integrates a leakage-safe residual stress signal and explicit modeling of tail-risk transmission channels. By moving beyond traditional volatility-based risk measures and incorporating causal inference techniques, the proposed architecture offers a more robust and interpretable approach to portfolio management in non-stationary environments. The system-level design highlights the importance of viewing portfolio allocation as a cybernetic control problem, where the portfolio continuously adapts to changing causal structures and stress levels. Governance and fairness considerations are paramount, given the potential for algorithmic allocation to amplify systemic risks and exacerbate market inequalities. The deployment of such systems must be accompanied by transparent governance, periodic validation, and macroprudential oversight to ensure that they serve the long-term stability of financial markets rather than merely optimizing short-term performance.

Future research should focus on extending the causal regime detection engine to incorporate high-frequency data and alternative data sources, such as news sentiment and satellite imagery, which may provide earlier signals of regime transitions. The integration of reinforcement learning approaches could enable dynamic learning of optimal allocation policies without requiring explicit causal graphs, though at the cost of interpretability. Cross-disciplinary collaborations with network science, epidemiology, and critical infrastructure protection will continue to enrich the theoretical foundations of tail-risk transmission. Ultimately, the goal is to build financial systems that are not only efficient but also resilient, capable of absorbing shocks and recovering quickly—a goal that the causal regime-aware allocation framework moves closer to achieving.

References

1. Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1), 77-91.
2. Black, F., & Litterman, R. (1992). Global portfolio optimization. Financial Analysts Journal, 48(5), 28-43.
3. Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2), 357-384.
4. Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.
5. Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: Foundations and learning algorithms. MIT Press.
6. Adrian, T., & Brunnermeier, M. K. (2016). CoVaR. The American Economic Review, 106(7), 1705-1741.
7. Danielsson, J., James, K. R., Valenzuela, M., & Zer, I. (2018). Model risk of risk models. Journal of Financial Stability, 38, 76-94.
8. Liu, T. (2026). Beyond volatility: A leakage-safe residual-stress signal for drawdown risk monitoring. Available at SSRN 6503179.
9. Ang, A., & Timmermann, A. (2012). Regime changes and financial markets. Annual Review of Financial Economics, 4, 313-337.
10. Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. Quantitative Finance, 1(2), 223-236.

11. Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. The American Economic Review, 105(2), 564-608.

12. Kim, C. J., & Nelson, C. R. (1999). State-space models with regime switching: Classical and Gibbs-sampling approaches with applications. MIT Press.

13. Ang, A., & Piazzesi, M. (2003). A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. Journal of Monetary Economics, 50(4), 745-787.

14. Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics, 20(3), 339-350.

15. Allen, F., & Gale, D. (2000). Financial contagion. Journal of Political Economy, 108(1), 1-33.

16. Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71(2), 579-625.

17. Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics, 104(3), 535-559.

18. Acharya, V. V., Pedersen, L. H., Philippon, T., & Richardson, M. (2017). Measuring systemic risk. The Review of Financial Studies, 30(1), 2-47.

19. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273.

20. Jobson, J. D., & Korkie, B. (1980). Estimation for Markowitz efficient portfolios. Journal of the American Statistical Association, 75(371), 544-554.

21. Schuermann, T. (2014). Stress testing banks. International Journal of Forecasting, 30(3), 717-728.

22. Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. The American Economic Review, 105(2), 564-608.