Research Statement

Justin Sirignano

Large, interacting stochastic systems appear in many facets of today's world and are of broad importance. Examples include the banking system, financial markets, electric power grids, engineering systems, health care, and social networks. Due to their scale, complexity, and large amounts of data, analyzing such systems is inherently challenging and will require new theoretical and computational approaches. My research develops models, computational methods, and statistical tools for such systems. I am particularly interested in new, data-driven approaches for modeling and optimization of these systems.

My research addresses several key challenges arising in this area, with one of my primary focuses being financial systems. The financial sector includes tens of thousands of financial institutions, which are themselves commonly exposed to large pools of default risks such as those associated with corporate loans and bonds, credit cards, auto loans, student loans, mortgages, and derivatives. Pools of loans typical in practice can range from a few thousand to many millions. Although my papers largely focus on financial systems, the approaches are broadly applicable to large systems across many fields. As a consequence of such systems' size and complexity, analysis is computationally expensive and statistical estimation is often intractable. Data available for each component or agent in the system is frequently high-dimensional, further complicating analysis. This data provides critical information, but is challenging to incorporate into models due to the curse of dimensionality. My papers [1-4] develop stochastic models, numerical tools, and statistical inference methods addressing these problems. Loans and securities backed by loans make up a significant fraction of a typical bank's assets. My research provides financial institutions and regulators with efficient data-driven methods to manage and estimate risk in large pools of loans. In ongoing work [6-7], these methods are used to efficiently analyze large quantities of mortgage data with the goal of better understanding risk in the mortgage market, a very important segment of the financial markets. In ongoing work [5], I develop, analyze, and test on real data an approach which allows for tractable, large-scale data-driven optimization of portfolios of loans.

Completed Research Projects

Paper [1] considers a broad class of empirically motivated dynamic point process models describing the correlated default timing in a financial system. The entities in the system are influenced by both idiosyncratic and systematic sources of randomness. (An entity could be a loan, a bank, or some other financial institution.) We also allow for interaction between entities, modeled through a mean-field term. Defaults in the system can adversely affect the remaining entities in the system. Many standard formulations ignore these key factors due to the challenges of analyzing such a complex, interacting system. For the class of models we consider, one must resort to brute-force Monte Carlo simulation for analysis and risk assessment; such simulation can be extremely computationally expensive for large systems. We prove a law of large numbers for this class of models and use this limiting law to approximate the system. The law of large numbers limit satisfies a nonlinear, stochastic partial differential equation. A method-of-moments scheme is devised to numerically solve this stochastic partial differential equation. It is shown that the approximation via the law of large numbers is many orders of magnitude faster than brute-force Monte Carlo simulation of the system.

Paper [2] extends paper [1] by proving a central limit theorem for the class of models. The mathematical analysis involves unique challenges, including serious topological issues. The central limit theorem satisfies a stochastic partial differential equation. The law of large numbers from [1] together with the central limit theorem from [2] yield a second-order accurate approximation for the system. Numerical results demonstrate that the approximation is highly accurate even for very small systems (as little as a few hundred entities) and many orders of magnitude faster than brute-force Monte Carlo simulation. The theory and numerical methods in [1,2] allow for computationally efficient modeling and risk analysis of large financial systems.

The practical implementation of models requires the statistical estimation of model parameters from data on the past behavior of the system. Traditional statistical approaches, such as maximum likelihood inference, are typically computationally intractable for large, interacting stochastic systems. In paper [3], limiting laws, such as the ones proven in [1,2], are exploited in order to develop approximate maximum likelihood estimators for a general class of interacting stochastic systems. Namely, one can approximate the likelihood using the limiting laws and maximize the approximate likelihood to yield approximate estimators. In addition, we prove that the approximate estimators converge to the true parameters and are asymptotically normal as the number of observations and the size of the system become large. A numerical scheme is devised to efficiently calculate the approximate estimators. Numerical studies show that the estimation approach is both computationally tractable and accurate even for moderate-sized systems. The results in [3] will allow for the statistical estimation of many large system models from a diverse set of fields where statistical estimation was previously intractable.

The crisis of 2007-09 highlights the need to better understand the behavior of risk in the mortgage market. The analysis of this market is challenging because the loan-level data available for mortgages and other types of loans is often high-dimensional. Paper [4] proposes a broad class of default and prepayment models for mortgages that takes advantage of this rich loan-level information and then develops an efficient Monte Carlo approximation for the default and prepayment processes in the large pools of loans common in practice. Importantly, the computational cost remains constant no matter how large the loan-level data's dimension becomes. The efficient Monte Carlo approximation is tested on real mortgage data from a data set containing over 25 million subprime and agency mortgages. Numerical studies using this large data set show the improved speed (for the same accuracy) of the approximation in comparison to brute-force simulation of an actual pool. Computational cost is again several orders of magnitude less than brute-force simulation of the actual pool. **Paper [4] won the 2014 SIAM Financial Mathematics and Engineering Conference Paper Prize**.

The research leading to papers [3] and [4] was funded by the NSF's Methodology, Measurement, and Statistics program ($220,000 grant), for which I co-prepared the proposal with my advisor.

Ongoing Research

Project [5], which is nearly complete, builds upon the computational tools of [4] in order to optimally select and design loan portfolios. While there is a vast literature on optimal portfolio selection for equities, little work has been done to address the problem of selecting a portfolio of bonds, loans or other fixed income securities. In light of the recent financial crisis, there is an extremely pressing need to develop ways to optimally design asset-backed security structures in order to reduce risk. A major reason why this very important problem has not been previously addressed is its high computational hurdles; it is a high-dimensional nonlinear integer program. I develop an approach which allows for tractable, large-scale data-driven optimization of portfolios of loans and securities backed by these portfolios. The approach is numerically tested using extensive mortgage and small business loan data sets.

I am also currently working on several empirical projects. I have constructed a unique, comprehensive database of mortgages and mortgage-backed securities (MBS) issued in the US during the past few decades for the purpose of several ongoing projects. Data sources include government-sponsored entities (e.g., Freddie Mac), private financial institutions (e.g., Wells Fargo), government agencies, and data providers (e.g., CoreLogic). The goal is to create a detailed, loan-level data set of the mortgage market that facilitates research into the behavior of risk in the mortgage market. At present, the database contains over 175 million mortgages and as many as one hundred variables per mortgage for every month during the past 20 years. In addition, the database includes detailed real estate data accounting for 97% of all real estate transactions in the United States over the past 20 years. The mortgage data will also be matched with real estate data to identify geographic locations for each mortgage down to street-level.

Project [6] explores the significance of geographical factors influencing default and prepayment risk in mortgages via a geographic network model. Controlling for other factors, it is found that network effects play an important role for the risk profile of mortgages and the securities they back. This has important implications for regulatory agencies and investors. Project [7] investigates risk premia for mortgage-backed securities (MBS) over the past 20 years. First, the default and prepayment model is fit under the historical measure. One can then calculate MBS prices under the historical measure (i.e., the "actuarial price") and find the spread between market prices and the actuarial price (the "risk premium"). The goal is to identify economic factors driving risk premia dynamics by statistically analyzing spreads across thousands of MBS backed by mortgages.

Future Research Directions

I am interested in pursuing several research projects in the near future. Hedging prepayment and default risk for mortgage-backed securities is an incredibly important problem which has not been addressed in the literature. This project will leverage my previous research in the area as well as the unique data set which I have constructed. I am also interested in developing new models and computational methods for the electric power market and grid. My interest was motivated by an internship for British Petroleum's Natural Gas and Power Group where I used machine learning to study the power market. This research will leverage my experience in optimization and energy, developed in research projects [8-10]. In a separate vein, I have some ideas on how to train deep neural networks using trust region optimization methods. Neural networks can be powerful tools for statistical regression and are heavily used across many fields in data science, but are difficult to train due to their non-convexity.

Research Papers

1. "Large Portfolio Asymptotics for Loss from Default" (with K. Giesecke, K. Spiliopoulos, and R.B. Sowers). *Mathematical Finance*, in press, 2013.

2. "Fluctuation Analysis for the Loss from Default" (with K. Spiliopoulos and K. Giesecke). *Stochastic Processes and their Applications*, (124): 2322–2362, 2014.

3. "Likelihood Inference for Large Financial Systems" (with G. Schwenkler and K. Giesecke). To be submitted soon to *Annals of Statistics.* Paper available at http://web.stanford.edu/~jasirign/

4. “Efficient Risk Analysis of Mortgage Pools” (with K. Giesecke). Submitted to *Operations Research*. Paper available at http://web.stanford.edu/~jasirign/

5. “Optimal Selection of Loan Portfolios” (with K. Giesecke and G. Tsoukalas). Work in progress.

6. “Geographic Risk for Mortgages” (with K. Giesecke). Work in progress.

7. “Risk Premia for Mortgage-backed Securities” (with K. Giesecke and M. Ohlrogge). Work in progress.

8. “A Forward-Backward Algorithm for Stochastic Control Problems” (with S. Ludwig, R. Huang, and G. Papanicolaou). *Proceedings of the First International Conference on Operations Research and Enterprise Systems.* Vilamoura, Algarve, Portugal. 4-6 February, 2012.

9. “Optimization of Secondary-Air Addition in a Continuous One-Dimensional Spray Combustor” (with L. Rodriguez, A. Sideris, and W. Sirignano). *Journal of Propulsion and Power*, 26.2: 288-294, 2010.

10. “Machine Learning for the Power Market.” Technical Report prepared for British Petroleum, 2013.