Educational Administration: Theory and Practice
2024, 30(11), 1412-1427
ISSN: 2148-2403
https://kuey.net/

Research Article

New Trends in Discrete Probability and Statistics: A Comprehensive Review

Kapil Dev Pandey[1*], Anupam Narula[2], Richa Singhal[3], Monika Saxena[4]

[1*]Research Scholar, Amity Business School, Amity University, Noida, Mail ID: kdpandey@gmail.com,
ORCID ID: https://orcid.org/0000-0002-8951-8855
[2]Professor, Amity School of Business, Amity University, Noida, Mail ID: anarula@amity.edu
[3]Associate Professor, Department of EAFM, Faculty of Commerce, S.S. Jain Subodh PG College, Jaipur,
Mail ID: richasinghal78@gmail.com
[4]Associate Professor, School of Management, Bennett University, Greater Noida
Mail ID: monika.saxena@bennett.edu.in

Citation: Kapil Dev Pandey, et.al (2024). New Trends in Discrete Probability and Statistics: A Comprehensive Review, *Educational Administration: Theory and Practice*, 30(11) 1412-1427
Doi: 10.53555/kuey.v30i11.9484

ARTICLE INFO

ABSTRACT

In recent years, discrete probability and statistical methods have made impressive developments that keep impacting all kinds of areas which include computer science (in particular, machine learning and finance), epidemiology, cryptography and social network analysis. This paper studies more recent trends of discrete probability via a systematic review on recent trends in Bayesian modeling, non classical distributions, stochastic process, discrete time financial models, as well as probabilistic AI frameworks. Further, we discuss emerging applications in disease modeling, post quantum cryptography, and network science, illustrating how discrete probabilistic methods are converging with deep learning based and hybrid modeling approaches. While highly successful at increasing accuracy, efficiency, and scalability of the predictions, all are still grappling with computational complexity, ethical (as well as interpretable) concerns in probabilistic decision making. This review evaluates the strengths and weaknesses of current model, provides gaps of current research and perspectives on future directions with techniques that are scalable to inference, hybrid probabilistic framework and fairness aware AI model. These studies produce synthesis of important developments and aim to provide researchers, practitioners an overview of modern discrete probability applications and some future research opportunities.

Keywords: Discrete probability, Bayesian inference, stochastic models, Markov chains, probabilistic AI, financial risk modeling, epidemiology, quantum probability

1 Introduction

The probability distributions of interest are called discrete probability and they occur when the outcomes are countable, being either finite or infinite (Grimmett & Stirzaker, 2020). Discrete probability is different than continuous probability in that it deals with smooth distributions, as opposed to Bernoulli trials, Markov chains and Poisson processes (Ross, 2019). Based on this material, each of theses models plays a key role in decision making as well as stochastic modeling and combinatorial analysis. The probability mass function (PMF), expectation, and variance serve as key descriptors of discrete probability distributions (Casella & Berger, 2021). However, discrete statistics is concerned with the collection, analysis, and inference of data that is naturally categorical (discrete) or count based. Discrete data is analyzed either using statistical methods like contingency tables, logistic regression, and chi square tests or using statistical methods (Agresti, 2018). This is an important field in modeling count data in real world, especially in social sciences, healthcare and engineering (McCullagh & Nelder, 2019).

Modern applications of discrete probability and statistics span various domains. Probabilistic graphical models constitute machine learning and artificial intelligence progress in speech recognition, text processing, and robotics (Murphy, 2012). Discrete statistical techniques become fundamental tools of cryptography, information theory and network security (Goldreich, 2019). Discrete probability models are used to track

disease spread or genetic mutations in biostatistics and epidemiology (Anderson & May, 2020). Poisson processes and geometric distribution are used in finance to model rare events, e.g. credit default and market crash (Cont & Tankov, 2021). As discrete probability and statistics is eventually becoming more and more relevant with the growing availability of large scale discrete datasets, the use of discrete probability and statistics is becoming more and more important. Further advances in stochastic computing, high dimensional inference and Bayesian frameworks give rise to their applications to be essential tools in contemporary scientific and technological progress (Jordan, 2018).

Discrete probability and statistics are very important in modern applications in various disciplines. Probabilistic graphical models — Bayesian networks and hidden Markov models for illustration are provided — are used in machine learning and artificial intelligence to advance in speech recognition and natural language processing as well as in decision making algorithms (Murphy, 2012). Discrete Probability plays a fundamental role in the cryptographic algorithms, network security, and the error detection in the digital communications (Goldreich, 2019). Poisson distributions and Markov models are used by biostatistics and epidemiology to model disease spread and survival analysis (Anderson & May, 2020). Discrete stochastic models are used in the financial risk analysis to assess credit default probabilities, model rare financial events, and to optimize investment strategies (Cont & Tankov, 2021). As high dimensional discrete datasets become more widely available, the computational and statistical techniques become new—Bayesian inference, stochastic optimization, and probabilistic machine learning—expand the domain in which discrete probability and statistics are viable. These are indispensable tools for modern scientific and technological innovations (Jordan, 2018).

Recently, the field of discrete probability and statistics has gone through a series of transformations, which are mostly due to the rise in the complexity of modern data, improvement of computing method, and broader diversification of the range of applied disciplines. Consolidation of emerging research, key trends identification and future research directions demand for a comprehensive review of recent advancements. Probabilistic models and inference techniques that use discrete structures to better solve problems such as decision making, optimization, and uncertainty quantification have arisen as a result of the rapid evolution of machine learning, artificial intelligence, and data science (Murphy, 2012). Bayesian networks, probabilistic graphical models and discrete Markov processes now constitute integral methods to AI driven applications, and it is necessary to evaluate their most recent developments and improvements considerations (Jordan, 2018). More recently, growing challenges in big data and computational statistics have also led to the need for new approaches for dealing with high dimensional discrete data, categorical distributions and stochastic modeling (Casella & Berger, 2021). However, many of these datasets are becoming large scale and arise in domains such as genomics, finance, cybersecurity, and epidemiology, and they require statistical techniques to be revised to both bring about efficiency, accuracy, and scalability. For instance, these are now being used to track disease outbreaks, detect fraud, and optimize investment strategies and, therefore, continuous methodological advancements are required (Anderson & May, 2020).

Furthermore, algorithmic and theoretical probability has recently made progress in a few breakthroughs that enable us to analyze random structures, optimize discrete probability distributions, as well as improve computational efficiency. This popularity of the field is highlighted by ways in which these innovations provide a structured review to enable researchers to have an updated synthesis (Goldrich, 2019). This paper attempts to bridge the gap between solid progress in theory and its applications in discrete probability and statistics through a review of recent developments and by providing leads to future research and to innovation in the theory of discrete probability and statistics.

2 Objectives of the Study

I Examine recent theoretical advancements in discrete probability, including developments in Bayesian inference, stochastic processes, non-classical distributions, and probabilistic graphical models.

II Analyze computational techniques that enhance discrete probabilistic modeling, such as Markov Chain Monte Carlo (MCMC), variational inference, hybrid modeling approaches, and scalable deep probabilistic frameworks.

III Explore interdisciplinary applications of discrete probability in machine learning, finance, epidemiology, cybersecurity, and network science, emphasizing its growing impact in real-world problem-solving.

IV Identify challenges and open research problems, including scalability issues, computational complexity, gaps in data availability, and the need for hybrid discrete-continuous probabilistic models.

V By achieving these objectives, this study aims to serve as a valuable resource for researchers, academicians, and practitioners, offering critical insights into the evolving landscape of discrete probability and its potential for driving future innovations.

3 Methodology

This paper is a systematic and systematic approach to the review of the recent progress in the discrete probability and statistics, so that it is comprehensive, unbiased and rigorously academic synthesis of previous research. The methodology includes selecting literature, categorising them in themes, performing comparative

and critical analysis, and this provides for a highly discussed theoretical development, computational techniques, as well as applications in real world.

In this first step, I conducted systematic literature search over the reputable academic database of Google Scholar, IEEE Xplore, Springer link, Science Direct and arXiv. Only publications from the period of the last ten years (2013–2024) were taken into account to maintain the relevance of the review to the current state of affairs in the field. Keywords used were discrete probability, Bayesian inference, Markov chains, probabilistic graphical models, stochastic processes, quantum probability, and statistical decision making, and they were used as a criterion to search. Peer reviewed journal articles, influential conference papers, high impact theoretical contributions were filtered out to prioritise them. Old and redundant studies were filtered out. A criterion of inclusion-exclusion was applied to the dataset to refine it. A list of studies was selected based on introduction of novel theories, computer techniques or real world applications to discrete probability. There have been also included papers with comparative analyses of classical and modern approaches. On the other hand, studies without empirical validation or quantitative analysis, studies without relevance to discrete modeling, and all non English publications without publicly available full texts were excluded.

Based on key themes, they were categorized once relevant studies were identified. A structured analysis was conducted in that they were organized based on the relevant studies identified. Finally, the review focused on theoretical contributions that include progress in discrete probability distributions, stochastic models and Bayesian inference, together with computational progress in Markov chain Monte Carlo, variational inference, and probabilistic deep learning. In addition, considerable effort was spent on applications in real world such as machine learning, finance, cybersecurity, epidemiology, and quantum computing to give a interdisciplinary flavor to the impacts of discrete probability models. Next, a comparison was made between different approaches in order to determine strengths and weaknesses. The areas of trends, computational challenges and open problem were critically investigated by reviewing the selected studies. In cases, where applicable, validations were performed through quantitative comparisons, algorithmic performance metrics and case studies. This ensured that the review was not only theoretical but practical, as discrete probabilistic models applied to real life situations are shown.

After filtering out, the insights were synthesized into a structured discussion at the end point, making it clear, accessible and useful for all the researchers, practitioners and students. This paper provides a coherent account of what emerging trends are, what are their implications and what future research directions should be followed. This in turn provides a critical methodology that should also be described as objective, rigorous and balanced review of discrete probability and statistical advancements, making this review a resource for further research in that field.

4 Emerging Trends in Discrete Probability

4.1 Non-Classical Distributions: New Probability Distributions and Extensions

In the recent years, discrete probability field has been witnessed with the advancement of non classical probability distributions to extend the traditional models to cope with the complexities of contemporary data analysis. This has always been done with classical discrete distributions, like binomial, Poisson, geometric, and negative binomial, in probability theory. But now, with the rising need for flexibility, responsiveness and higher modeling accuracy, researchers have started to come up with new probability distributions and generalizations of those classical models. The development of generalized discrete distributions is one of the key advancements in nonclassical distributions as it includes more parameters to improve the modeling capability. The q-series distributions (e.g., the q-Poisson and q-binomial distributions) provide a deformed form of their classical counterparts by a parameter \(q \) that modifies the extent of departure from the standard form (Charalambides, 2019). The classical models that they have found applications in do not suffice in cases where complex dependencies are present; they have been used in quantum probability, statistical mechanics, and in combinatorial optimization. Major extension in the context of fractional calculus is the discrete Mittag-Leffler distribution, which arises from the same family of distributions (geometric and the negative binomial distribution) as in Pillai and Jayakumar (2021). In particular, this distribution has been used in modeling heavy tailed discrete data in finance, reliability engineering and epidemiology where events show long rang dependence and non exponential waiting times.

Compound and mixture distributions have also become popular because they can model over-dispersed and heterogeneous discrete data. Examples of such distributions that are more flexible in capturing real world count data departing from standard Poisson or negative binomial assumptions are the Poisson-inverse Gaussian (PIG) distribution and the negative binomial-Lindley distribution (Karlis & Xekalaki, 2020). The use of these distributions has been wide spread in actuarial science, biological modeling and risk management. The development of Dirichlet process based discrete distributions resulting from new memoryless (Bayesian) nonparametric approaches to modeling complex categorical data has also occurred (Hjort et al., 2019) since the rise of such approaches. To some extent they have completely changed the game in areas like machine learning, genetics or any other field where a discrete probability distribution was too restricted. These non classical distributions offer powerful extensions as applications of discrete probability research emerges that bridge theory and application. Due to their continued development and refinement they will continue to

contribute highly to addressing new challenges presented in data science, finance, engineering and artificial intelligence, considering the applicability of discrete probability models.

4.2 Advances in Discrete-Time Markov Chains and Hidden Markov Models (HMMs).

DTMCs and HMMs have been standard workhorses in probability modeling since long time and have applications in machine learning, finance, bioinformatics, speech recognition along with other areas of reliability engineering. Over the past few years, these models have been improved in scalability, computational efficiency, and increasing predictive power in order to be applied to more sophisticated and higher dimensional problems. The most important advancements in the DTMCs research have been the introduction of higher order Markov chains and non homogeneous Markov models (Iosifidis & Vlahavas, 2020). The memoryless property of traditional DTMCs is where the probability of transition to the next state solely depends on the current state. However, models of higher order are able to depend on more than one past states and this gives more accurate models for the applications such a finantial time series analysis, genetic sequencing, and reinforcement learning. Moreover, the time varying and non homogeneous Markov models have been introduced to capture dynamic and evolving system, e.g., climate change pattern and market fluctuation (Chen et al., 2022). Recent developments concerning state estimation, model flexibility and interpretability in the realm of Hidden Markov Models (HMMs) are limited by its ability to overcome limitations. In real world applications, the number of hidden states and transition probabilities may need not be fixed, and this may not be indicated while using traditional HMM. In order to address these challenges, Bayesian nonparametric HMMs have been developed by researchers, where the number of states can grow dynamically depending on the complexity of the observed data (Teh & Jordan, 2019). In the area of speech processing, genomics, and topic modeling, such models have proven to be particularly useful, since the true underlying state structure is not known.

Additionally, deep learning enhanced HMMs have proven to be a very powerful alternative to the classical HMM. Researchers have integrated HMMs with neural networks, recurrent neural networks (RNNs), and transformers for great improvement of tasks like speech recognition, handwriting recognition, and financial forecasting (Graves & Jaitly, 2021). By combining the sequential dependencies in HMMs with the representational power of deep learning, these hybrid models surpass the abilities of standard HMMs in difficult time series prediction problems. Reinforcement learning is also applied to HMM training and optimization, which is another important step. EM algorithms for HMM training are traditionally slow, and may experience local optima. In recent years, policy gradient methods and deep Q learning algorithms have been studied in the context of optimization of HMM based decision processes especially in robotics, automated trading and health care diagnostics (Silver et al., 2020). These advancements are extending the applicability of research in discrete time Markov chains and HMMs in more and more domains as research in this area continues to evolve. Markovian modeling, blended with modern computational techniques, has been integrated for the adaptive speech recognition systems and high frequency trading to the next generation of probabilistic modeling and decision making systems.

4.3 Impact of Quantum Probability Models

Quantum probability models have emerged as a transformative extension of classical probability theory, providing new mathematical frameworks for decision-making, machine learning, cryptography, and quantum computing. Unlike classical probability, which relies on Kolmogorov's axioms and assumes that probabilities are real-valued and additive, quantum probability is based on the principles of Hilbert space theory, non-commutative operators, and probability amplitudes, allowing for more flexible representations of uncertainty and complex correlations (Busemeyer & Bruza, 2019). These models have gained attention for their ability to explain paradoxical phenomena in human cognition, optimize quantum computing algorithms, and enhance probabilistic reasoning in artificial intelligence. One of the most significant impacts of quantum probability models has been in cognitive science and decision theory. Classical probability struggles to explain certain inconsistencies in human decision-making, such as the order effects in survey responses, the conjunction fallacy, and violations of the sure-thing principle observed in behavioral economics. Quantum probability provides a superposition-based framework that captures these effects by allowing for context-dependent probability amplitudes, leading to more accurate predictive models in psychology and behavioral economics (Khrennikov, 2020). In machine learning and artificial intelligence, quantum probability has led to the development of quantum-inspired models for pattern recognition, natural language processing (NLP), and probabilistic graphical models. For example, quantum Bayesian networks and quantum Markov models generalize classical probabilistic models by incorporating non-commutative probability spaces, improving inference capabilities in high-dimensional and uncertain environments (Zhang et al., 2021). Additionally, quantum-inspired neural networks leverage quantum probability principles to enhance deep learning architectures, making them more efficient in handling entangled dependencies in sequential and relational data (Haven & Khrennikov, 2019).

Another crucial domain impacted by quantum probability is cryptography and quantum computing. Classical probability models are foundational to conventional cryptographic techniques, such as random number generation, key distribution, and zero-knowledge proofs. However, with the rise of quantum cryptographic protocols, such as quantum key distribution (QKD) and post-quantum cryptography, quantum probability models provide a rigorous mathematical basis for ensuring security in a quantum computing environment (Nielsen & Chuang, 2020). The ability of quantum probability to model superpositions, entanglement, and

uncertainty allows for the design of unconditionally secure cryptographic protocols that are resistant to classical and quantum attacks. Furthermore, in stochastic processes and statistical mechanics, quantum probability is being applied to model non-classical random processes, such as those found in quantum walks, quantum chaos, and open quantum systems. These applications have profound implications for quantum algorithms, material science, and quantum statistical inference, enabling researchers to analyze systems where classical probability fails to capture the underlying stochasticity (Attal et al., 2019). The growing impact of quantum probability models is reshaping multiple fields by providing alternative probabilistic reasoning frameworks that overcome the limitations of classical models. Whether in human cognition, AI, cryptography, or quantum computing, these advancements continue to drive the next wave of innovations in probability theory, information processing, and complex system modeling.

4.4 Entropy-Based Measures: Applications in Data Compression and Signal Processing

Entropy based measures are important in data compression and signal processing, and as a mathematical description of uncertainty, information content and redundancy of discrete data. Entropy is based on Shannon's Information Theory and is the fundamental tool for data encoding and noise reduction which is used in modern computing and communication systems including the optimization of feature extraction (Shannon, 1948). As big data, machine learning, real-time signal analysis, have been growing at a rapid pace, entropy based technique transformed to enable high efficient compression algorithms, better transmission protocols, and better signal reconstruction methods. Entropy is used in data compression to find the minimum redundancy while maintaining the important part of information. The lossless compression algorithms like Huffman coding and Arithmetic coding are based on Shannon's Entropy Coding Principle, meaning that symbols with higher probability are assigned shorter codes, resulting into lesser storage space (Cover & Thomas, 2006). In lossy compression, such as JPEG for images and MP3 for audio, entropy-based methods like Rate-Distortion Theory balance compression efficiency with minimal perceptual quality loss (Sayood, 2017). Yet these ways have made data storage, multimedia streaming, cloud computing to all an efficient bandwidth utilization making methods.

The entropy measures are widely used for feature extraction, pattern recognition and noise filtering in signal processing. Spectral entropy quantifies the signal complexity, which is useful in the EEG brainwave analysis, speech recognition and biomedical imaging (Rosso et al., 2001). The entropy based thresholding improves the denoising techniques in speech and audio processing by discriminating between signal and random noise. Wavelet entropy also facilitates detection of faults, analysis of ECGs, and analysis of geophysical data (Zhao et al., 2011), as well as non-stationary signals. Like that, entropy based methods are also critically important to cryptography security, and in anomaly detection. Shannon entropy and Rényi entropy are applied for detecting irregularities in network traffic, fraud detection and malware analysis, where deviations from expected entropy levels are indicators of potential security threats (Verma & Ranga, 2019). Likewise, entropy prevents the feature representation from becoming robust in biometric authentication systems such as in the Fingerprint, Iris, and Facial recognition systems. With growing volume and complexity in the data, entropy based approaches will further improve compression efficiency, real time signal processing, etc. It is expected that there would be future research towards adaptive entropy based learning models, using deep neural networks with entropy regularization to increase data efficiency, robustness and computational scalability on various technological domains. Following is the table of Trend and Field of application.

Figure 1 - Trend & Field of application

Trend	Description	Impact	Fields of Application
Bayesian Modeling Growth	Expansion of Bayesian inference, including Hamiltonian Monte Carlo (HMC) and Variational Inference (VI), improving computational efficiency.	Enhanced uncertainty quantification, better model interpretability, and improved predictive accuracy.	Machine Learning, Finance, Epidemiology, Genomics
Non-Classical Distributions	Introduction of generalized discrete distributions (e.g., q-series, Mittag-Leffler) to model complex count-based data.	Improved statistical modeling for over-dispersed, heavy-tailed, and non-homogeneous discrete data.	Risk Analysis, Bioinformatics, Stochastic Modeling
Advancements in Hidden Markov Models (HMMs)	Integration of deep learning with HMMs for sequence modeling and dynamic state estimation.	Improved speech recognition, anomaly detection, and time-series forecasting.	NLP, Speech Processing, Cybersecurity, Finance
Quantum Probability Models	Application of quantum-inspired probability in decision-making, machine learning, and cryptography.	Enhanced predictive modeling, improved encryption security, and better representation of cognitive uncertainty.	AI, Cybersecurity, Cognitive Science, Quantum Computing

Discrete-Time Financial Models	Refinements in binomial asset pricing, Markov-based risk models, and Bayesian portfolio optimization.	More accurate financial forecasting, improved risk assessment, and adaptive investment strategies.	Finance, Algorithmic Trading, Risk Management
Graph-Based Discrete Models in Social Networks	Use of stochastic block models (SBMs) and Exponential Random Graph Models (ERGMs) for analyzing network structures.	More accurate detection of communities, influence propagation, and misinformation tracking.	Social Media Analytics, Political Science, Behavioral Economics
Handling Large-Scale Discrete Data	Development of tensor decompositions, probabilistic data structures, and parallel computing frameworks.	Faster processing and analysis of high-dimensional discrete datasets, enabling real-time decision-making.	Big Data Analytics, Genomics, NLP, AI
Discrete Statistical Methods in Disease Modeling	Use of Bayesian epidemiological models, Poisson-based outbreak detection, and agent-based simulations.	Improved real-time pandemic predictions, better resource allocation, and optimized intervention strategies.	Public Health, Epidemiology, Biostatistics
Discrete Probability in Cryptography	Advances in post-quantum cryptographic algorithms (LWE, Ring-LWE) and probabilistic encryption methods.	Increased security against quantum attacks, better key randomness, and more secure communication.	Cybersecurity, Data Privacy, Blockchain

5 Recent Developments in Discrete Statistics

5.1Discrete Probability in Machine Learning and Natural Language Processing (NLP)

Probabilistic predictions as well as the optimization of decision making processes are the fundamental role of discrete probability in machine learning (ML) and natural language processing (NLP), since it provides a mathematical ground for uncertainty modeling. For many real world problems in ML and NLP, these data are inherently discrete, such as categorical, sequential, or count based. Discrete probability is applied for robust solutions of many interesting problems in Bayesian inference and probabilistic graphical models (PGM), hidden Markov models (HMM) and deep generative techniques.

Probabilistic classification, uncertainty estimation and latent variable modeling are crucial parts of machine learning and use discrete probability. Some algorithms such as Naïve Bayes classifiers use discrete probability distributions like multinomial and Bernoulli distributions to solve text, image, and spam detection problems (Murphy, 2012). Finally, discrete probability, Bayesian networks and Markov random fields, are the ones that probabilistic graphical models, including Bayesian networks and Markov random fields, use to model the dependencies among variables in structured data. They are widely applied in medical diagnosis, speech recognition, as well as fraud detection, wherein decisions need to be made under uncertainty (Koller & Friedman, 2009). Discrete probability is critical for language modeling, text generation, sequence prediction, in NLP. For example, classical models such as n-gram language models assume Markov and predict the probability of words in a sequence based on it (Jurafsky & Martin, 2021). The most widely applied HMMs and CRFs are based on discrete probability and have been deployed for speech recognition, part-of-speech tagging, as well as for named entity recognition (NER) (Manning & Schütze, 1999). The first type of models, learn sequential dependence in text and speech and allow structured predictions in NLP applications. Deep learning has brought us to an era where discrete probability remains an important element of probabilistic deep generative models like variational autoencoders (VAEs), restricted Boltzmann machines (RBMs), and discrete latent variable models, and the field has come a long way considering the situation in 2009. These methods combine discrete probability distributions in order to create text, image, and structured data representations that are realistic. Discrete probability is used for token level probability estimation in tasks like text generation, summarization and machine translation (Vaswani et al., 2017) in state-of-the-art NLP models like GPT and BERT, and transformers are used to power those models. Additionally, RL models are very dependent on discrete probability for their policy learning and choice making. Reinforcement learning algorithms are used to optimize conversation in dialogue systems, chatbot's, etc. by using discrete probability distributions over possible actions to choose the optimal conversational response (Sutton and Barto 2018).

Uncertainty quantification also heavily relies on discrete probability in order to use Bayesian deep learning techniques to increase model reliability and meaning. Discrete priors and posterior distributions are used in Bayesian models that use for text classification, sentiment analysis and topic modelling (Blei et al., 2003). Now, as machine learning and NLP evolve, discrete probability is still a subroutine of indispensable importance for designing robust and interpretable probabilistically sound models. It is applicable in sequence modeling, generative AI, uncertainty estimation, and decision making and will remain relevant in the deep learning and AI driven language technologies.

5.2 Methods for Handling Large-Scale Discrete Data

Large scale discrete data are being produced in abundance in a variety of areas such as machine learning, natural language processing (NLP), bioinformatics, and network analysis, and the need is felt to have fast computational and statistical methods for processing and storing as well as analyzing these data. However,

massive discrete dataset often has dimensionality, sparsity, and computational complexity in traditional statistical concept. Recently advanced approaches have advanced a various range of methodologies, like probabilistic modelling, compressed information matters, electronic computing and executable Bayesian proof. The use of probabilistic scheme graphical models, such like Bayesian networks and Markov random fields (MRFs) one of the most important progress in handling high dimensional discrete data, because it permits structured representation dependencies in high dimensional discrete spaces (Koller & Friedman, 2009). However, due to sparsity and conditional independence, these models have low computational complexity and thus are well suited for applications, for example, the social network modeling, fine genomic data analysis and document classification. Approxi mate inference techniques like Markov Chain Monte Carlo (MCMC) and Variational Inference (VI) can be another power method since they can estimate computationally challenging probabilistic model efficiently (Blei et al., 2017). Many discrete probabilistic models have been tackled by the MCMC based methods including Gibbs sampling and the Hamiltonian Monte Carlo. On the other hand, using MCMC requires one to run just that sample many times, which makes it unsuitable for large scale datasets in NLP and AI and VARIATIONAL INFEWERENCE is a scalable alternative which approximates posterior distributions deterministically.

In the case of high dimensional discrete data, dimensionality reduction techniques as random projections, feature hashing, and word embeddings are very important. BERT and word2vec embeddings are used as preprocessors in NLP that transform high dimensional sparse textual data into lower dimensional dense representation in line with the semantic relationships (Vaswani et al., 2017). For example, for such discrete datasets, latent structure can be discovered using the tensor decomposition methods like Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF) for example in the recommendation system and social network analysis (Cichocki et al., 2009). Handling large scale discrete data has also been done by parallel and distributed computing frameworks like Map Reduce, Apache Spark, TensorFlow among others. Due to such frameworks, represented by HDFS, Hive, Pig, SparkSQL, Spark streaming, and Mesos, terabyte scale categorical data in bioinformatics, computational linguistics and large scale probabilistic modeling (Zaharia et al. 2016) can be processed efficiently. A second growing method, which is based on probabilistic data structures like Bloom filters, Count Min Sketch, and HyperLogLog [Cormode & Muthukrishnan, 2012] completely populates the large scale discrete data into a space efficiently representation but allows fast approximate queries. For the studies we performed this is true in network security, streaming data analysis, and real time anomaly detection where exact computation is not possible due to memory and time constraints.

These advanced statistical, mathematical, and computational techniques are becoming indispensable because as data scale and complexity grows, data at such a scale and complexity are always so large. Future research is likely to use hybrid methods of exploitation, such as deep learning, probabilistic models, and distributed computing to improve scalability and efficiency even further in discrete data analysis.

5.3 Growth of Bayesian Modeling Techniques

During the past decade, there have been great improvements in computational methods, efficient inference, as well as interdisciplinary applications of Bayesian modeling. These traditional frequentist approaches often lack the ability for quantifying uncertainty, the ability to infer high dimensional parameters, and the use of small data scenarios, while Bayesian methods offer a proper probabilistic framework, which incorporates prior information and updates their beliefs based on the growing new data (Gelman et al., 2013). Innovations in Markov Chain Monte Carlo (MCMC), Variational Inference (VI), Bayesian deep learning, and probabilistic programming languages have made the growth of Bayesian modeling more scalable and applicable to complex real world problems.

Hamiltonian Monte Carlo (HMC) and Stochastic Variational Inference (SVI) have been major catalysts for the expansion of Bayesian techniques, since they have made HMC computational feasible for inference on the natural parameter of most distributions, while proving effective for draw sampling in variational Bayesian inference (VBGI). Classical MCMC has advanced significantly to the stage where HMC, which improves MCMC by incorporating gradient based sampling, has sped up Bayesian computation in high dimensional parameter spaces to the point of being practical (Neal, 2011). As with SVI, Bayesian inference on large datasets can be done via optimization rather than expensive sampling, and SVI does this for streaming data and large scale probabilistic models as well (Hoffman et al., 2013). Furthermore, Bayesian deep learning has also identified reasons to adopt Bayesian modeling such as in estimation of uncertainty, robustness, and interpretability. Bayesian Neural Networks (BNNs) have been applied successfully in medical diagnostics, autonomous systems and reinforcement learning where the uncertainty quantification is crucial and many other applications (Blundell et al., 2015). Hyperparameter tuning has also been improved through advances in Bayesian optimization especially as it applies to deep learning, very efficient search strategies for deep learning model architectures can be carried out (Snoek et al., 2012). An important advancement is also the creation of probabilistic programming languages (PPLs) like Stan, PyMC3, Edward and TensorFlow Probability which bring the democratization of Bayesian modeling bringing user friendly frameworks for building and inference of complex probabilistic model (Blei et al., 2017). As has happened with the tools, Bayesian adoption has been accelerated through these tools in finance, epidemiology, and the social sciences, and hierarchical Bayesian models are now routinely applied in risk modeling, disease forecasting, and policy evaluation (Vehtari et al., 2021).

In addition to that, there is a trend to use Bayesian nonparametric models that have flexible and data driven structures with no assumptions of fixed numbers of parameters. DPMMs and GPs have been widely used in clustering, regression and spatiotemporal modelling and have been used in many applications, including genomics, natural language processing and image analysis (Rasmussen & Williams, 2006). Looking forward, future research will complement Bayesian modeling with faster ways of implementing it, scalability, interpretability, and making it real-time to bridge the gap between the use of Bayesian method and deep learning while extending the use of Bayesian method in scientific discovery, AI safety, Probabilistic decision making. It is interesting that the integration with modern computational advances and Bayesian principles is paving the way for increasing the influence of probabilistic models in next generation models not only for probabilistic modeling but also for problems much beyond that.

5.4 Analyzing Social Network Structures Using Discrete Models

Since the analysis of social network structures is an important area of research in computational social science, economics, epidemiology and artificial intelligence, we use a set of tools from complex networks and graph theory to analyze the event and actors. Since social networks are composed of discrete entities (nodes) and relations (edges), we have discrete probability and statistical models that form the important foundations for studying topological properties, community structures, and a spread of influence as well as dynamical evolution in social network (Easley & Kleinberg, 2010). As large scale digital networks like social media platforms, citation networks and online communities have grown, so has the need to extract insights from the complex relational data, and this is has been done by means of efficient and scalable discrete probabilistic models. For instance, the fundamental approaches in network analysis are graph based discrete models, namely, the discrete random graph model, often used is the Erdős–Rényi (ER) model and the Barabási–Albert (BA) preferential attachment model (Newman, 2018). In completing this thesis, the ER model was used as a simple yet effective framework to consider degree distributions, clustering coefficients and path lengths in social networks based on the assumption that edges between nodes occur independently with a fixed probability. Real world social networks, however, do not follow this, as their degree distributions are heavy tailed whereby few nodes (influencers) have many more connections than the rest. This is addressed with the BA model that incorporates preferential attachment, where new nodes are drawn to be more likely to connect to already well connected nodes and captures the scale free properties on Twitter and LinkedIn (Barabasi, 2016).

MRFs and ERGMs constitute more flexible models of network dependencies and structural motifs from discrete probabilistic models, while extending the categorical domain. ERGMs generalize classical random graph models as one can define the probability of an edge as a function of network structural properties (such as reciprocity, transitivity, and homophily) (Robins et al., 2007). Such models have been applied frequently in studying friendship networks, online interactions and organizational structures where the relationships are not created freely, but subject to social dynamics. Stochastic block models (SBMs) is another powerful framework to analyze social networks which can define a discrete probabilistic method to cluster the nodes into latent groups based on their connectivity patterns (Abbe, 2017). They have been extensively used in political network analysis, fraud and recommender systems where structures hidden within networks need to be identified. Degree corrected SBMs and hierarchical SBMs extend the modeling accuracy for heterogeneous and multi layered networks (Peixoto 2020). Information diffusion and influence propagation modeling in networks involve also the use of discrete probabilistic techniques. The Independent Cascade (IC) model and the Linear Threshold (LT) model are two models for discrete probability to simulate information, rumors, or innovations spreading over the network (Kempe et al., 2003). Analysis of diffusion dynamics using these models is important and these models are widely used in viral marketing, social contagion studies and epidemiological modeling.

In recent times, as increased amounts of network data become available, there has been a growing effort towards combining Bayesian inference, machine learning, and deep generative models into discrete social network analysis. GNNs have achieved remarkable success in various graph related problems (such as link prediction, anomaly detection, and social recommendation systems) (Zhou et al., 2020). Advanced research of this domain will not stop yet and to solve incomprehensible complex social system problems, there will be a need for integration of classical discrete models with modern computational techniques.

Figure 2- Advancements in Discrete Probability and Bayesian Methods in AI and Data Science

Development	Description	Impact	Fields of Application
Discrete Probability in Machine Learning and NLP	Discrete probability enables probabilistic classification, sequence modeling, and uncertainty estimation in AI and NLP. Methods include Naïve Bayes classifiers, Bayesian networks, and Hidden Markov Models (HMMs).	Improved speech recognition, text generation, and structured decision-making in uncertain environments.	Natural Language Processing (NLP), Speech Recognition, AI-driven Chatbots, Sentiment Analysis
Methods for Handling Large-Scale Discrete Data	Techniques such as probabilistic graphical models, variational	Enhanced computational efficiency in processing	Big Data Analytics, Computational Biology,

	inference, tensor decomposition, and distributed computing frameworks improve scalability.	high-dimensional discrete datasets.	Social Network Analysis, NLP
Growth of Bayesian Modeling Techniques	Advances in Bayesian inference methods, including Hamiltonian Monte Carlo (HMC), Stochastic Variational Inference (SVI), and Bayesian Deep Learning.	More robust probabilistic models for uncertainty quantification, risk analysis, and scientific discovery.	Machine Learning, Healthcare Diagnostics, Financial Risk Management, AI Safety
Analyzing Social Network Structures Using Discrete Models	Use of graph-based probabilistic models, such as Exponential Random Graph Models (ERGMs) and Stochastic Block Models (SBMs), for network analysis.	Better insights into community detection, influence propagation, and fraud detection.	Social Media Analytics, Political Science, Cybersecurity, Behavioral Economics

6 Applications of Discrete Probability and Statistics

6.1 Finance & Risk Analysis: Discrete-time models in quantitative finance.

Quantitative finance and risk analysis are usually full of uncertainty, where financial systems have some certain properties with inherent uncertainty are considered, and the robust probabilistic framework is needed for decision making, hence, discrete probability and statistical models play a key role in these application areas. The discrete time models are used in many financial processes as stock price movements, credit risk assessment and portfolio optimization that provide structure of the risk estimation, derivative pricing and asset management (Shreve, 2004). Discrete time models have evolved a great deal over the years, becoming increasingly able to predict and more efficient in terms of risk management strategies. The most used discrete-time model in finance is the Binomial Asset Pricing Model, as introduced by Cox, Ross and Rubinstein (1979). This is a discrete time stochastic process model in which the price of asset changes with a given probability to go up by a constant amount or down by the same amount in any one step. In the Black-Scholes framework, option pricing proceeds from the binomial model, and is necessary for pricing American options, since early exercise is possible (Hull, 2017). The binomial model is extended to trinomial trees and lattice based methods that provide better approximations for derivative pricing. The second key application is in credit risk modelling, where Markov chains and Hidden Markov Models (HMMs) are applied for the study of the credit rating transitions, default probability of loans, and the corporate bankruptcy risk (Jarrow & Turnbull, 1995). Markov models that take discrete creditworthiness states estimate the likeliness of borrowers residing in different creditworthiness categories and are useful at banks and financial institutions in managing loan portfolios and in the assessment of systemic risk.

Algorithmic trading and high frequency finance are discrete time models where stochastic processes such as Poisson processes and jump diffusion models are used to describe price movements which are not regular (Cont & Tankov, 2004). In particular, discrete event simulations are used in these applications to optimize trade execution strategies as well as minimize the slippage costs incurred by limit order book dynamics and market microstructure modeling. Discrete probability has a central role in portfolio optimization and risk management, where one estimates Value at Risk (VaR), and Expected Shortfall (ES) which are standard risk measures of market risk. Robust tools for stress testing portfolios under extreme market conditions (Glasserman et al., 2002) belong to discrete statistical techniques such as Monte Carlo simulations and bootstrapping methods. Finally, discrete Bayesian models are used for adaptive portfolio management, in which the asset returns are updated continuously based on observed data from the prior beliefs over asset returns (Black & Litterman, 1992). On the one hand, there has been recent progress in machine learning and AI in general, as well as in AI investment in finance in particular, and this further expanded the use of discrete probability models. Nowadays, reinforcement learning algorithms are extensively used in the algorithmic trading and hedging strategies and in the robo-advisors (Fischer, 2018). In addition, discrete probabilistic graphical models, such as Bayesian networks and hidden Markov models have been utilised more and more for fraud detection, anomaly detection in financial transactions, stress testing of banking systems (Bolton & Hand, 2002). With growing financial markets becoming more and more complex as well as data driven, discrete time models will still have a key role to play in risk assessment, trading strategies and financial decisions. By integrating statistical learning techniques, Bayesian inference, and probabilistic deep learning, predictive modeling can be further improved and financial stability in such uncertain environment is foreseen to be more integrated.

6.2 Discrete Models in Genetic Sequencing and Evolutionary Biology

For genetic sequencing and evolutionary biology, the discrete probability and statistical models that have developed have become indispensable (and frankly, indispensable) tools for rigorous mathematical framework

to understand DNA sequences, genetic variation and evolutionary process. As biological sequences (DNA, RNA and proteins) must be discrete in nature, discrete probability distributions and stochastic processes have been used by researchers in order to infer evolutionary relationships, detect mutations and construct phylogenetic trees (Felsenstein, 2004). Feeling of efficiency of discrete models for large scale genetic data comes from rapid advancements in high through put sequencing technologies and computational biology.

The Markov Chain Model is one of the most fundamental discrete models in genetics, and has been applied to many nucleotide transition and transversion problems in DNA sequences. The Jukes-Cantor (JC69) model assumes equal mutation probabilities among all four nucleotides and the Kimura Two Parameter model also includes different rate of transitions (purine-to-purine or pyrimidine-to-pyrimidine) and transversions (purine-to-pyrimidine or vice versa) (Kimura, 1980). To complete phylogenetic inference, more complex models like the General Time Reversible (GTR) model accommodate variable mutation rates for all pairs of the nucleotides (Yang, 1994). Similarly, Hidden Markov Models (HMMs) have also transformed the gene prediction and sequence alignment problem into a problem of hidden biological states, in this case exon-intron boundaries in DNA sequences (Durbin et al., 1998). Probabilistic gene annotation using HMMs is used, for example, to determine protein coding genes in newly sequenced genomes and to detect regulatory motifs in those regions of non-coding DNA. GENSCAN and HMMER are widely used in genomics and proteomics, and these models have been applied successfully in these tools (Eddy, 2011). Discrete coalescent models are used in evolutionary biology to gain understanding of population genetics and inference of ancestors. The Kingman Coalescent Model provides a stochastic model of genealogical trees so that researchers can estimate the population size history, detect evolution of migration patterns and infer genetic bottlenecks (Wakeley, 2009). This framework is extended by the Structured Coalescent Model in order to include spatial and demographic structure, and is hence crucial for understanding pathogen evolution and species divergence (Hein et al., 2005). Bayesian phylogenetics has also become by far the most commonly used method of evolutionary history reconstruction with discrete models. In addition, Bayesian Markov Chain Monte Carlo (MCMC) sampling with software like BEAST or MrBayes allows probabilistic inference of phylogenetic trees based on integrating over uncertainty in model parameter (Drummond & Rambaut, 2007). These Bayesian frameworks are instrumental to analyzing viral evolution (e.g., COVID-19 phyletics) and epidemiological spread, as well as species diversification. Discrete models are also another critical application in which discrete probability distributions are used in genome wide association studies (GWAS) where genetic variants are identified that are associated with diseases. Discrete logistic regression models are used to ascertain the likelihood of a certain genetic marker being connected to a disease phenotype (Visscher et al., 2017). Poisson models and negative binomial distributions are also usually used to model count data and detect differential gene expression between conditions in RNA sequencing (RNA-Seq) data (Love et al., 2014). With advancements in genetic sequencing technologies, discrete probabilistic models will have more and more important roles in personalized medicine, evolutionary genetics, as well as synthetic biology. Deep learning will most likely be used in the future to combine discrete evolutionary models to underpin more precise genetic predictions, evolutionary reconstruction, with large scale genomic datasets easily..

6.3 Applications of Discrete Probability in Encryption and Security

Encrypted, cybersecurity, and cryptographic protocol utilize discrete probability that serves as mathematical foundation for randomness, unpredictability, and secure key generation. Since most cryptographic systems takes discrete structure (finite fields, modular arithmetic, ...), probabilistic techniques are important to provide secure communication channels, detect anomalies, and analyze threats (Goldreich, 2004). The expanded role of discrete probability in modern cybersecurity is due to recent advancements in post-quantum cryptography, probabilistic encryption, and stochastic security models.

Discrete probability is one of the main applications in the context of cryptography for RNG, i.e. for key generation, encryption, and digital signatures, since random numbers are needed in all these tasks. Pseudo random number generators (PRNG) and the true random number generators (TRNG) are used in cryptographic systems depending on unpredictability (Menezes et al., 2018). Discrete probabilistic models like Markov chains and entropy based randomness extraction are used in many PRNGs to produce sequences that look like random but through these PRNGs, the sequences are computationally efficient. Discrete probability distributions are used to model TRNGs (meaning a TRNG generates randomness from physical, i.e. thermal noise or quantum, fluctuations) which guarantees high-security encryption keys. Probabilistic encryption schemes also employ discrete probability in that they add randomness to encrypt a message so that it is not vulnerable to attack. One of the earliest discrete probability models was Goldwasser and Micali's (1982) probabilistic encryption model which supports the notion of discrete probability by ensuring that each plaintext has multiple possible ciphertexts unless using the secret key. Discrete Gaussian distributions are nowadays employed in modern encryption techniques, including homomorphic encryption and lattice based cryptography (Lyubashevsky et al., 2013), while providing the ability of secure computations of encrypted data. In public key cryptography, for instance, the Discrete Logarithm Problem (DLP) or the Integer Factorisation Problem (IFP) which are 'hard' mathematical problems under discrete probability form the basis of the security upon which algorithms such as RSA, Diffie-Hellman key exchange, and Elliptic Curve Cryptography (ECC) (Rivest et al., 1978) are built upon. Solving these problems in the polynomial time is hard as it is dependent on the probabilistic infeasibility of decrypting if we do not have the private key. Recently, discrete noise

distributions have been utilized to protect against quantum attacks by recent advancement on the quantum resistant cryptographic algorithms such as Learning With Errors (LWE) and Ring-LWE (Peikert, 2016).
Indeed, discrete probability would also be crucial in intrusion detection systems (IDS) and anomaly detection in cybersecurity. Only recent machine learning based security systems relies on probabilistic models, such as Hidden Markov Models (HMMs) and Bayesian networks and Poisson distribution are used for detecting irregular login patterns, network anomalies and fraud detection (Denning, 1987), etc. These models allow to make estimates as to how likely it is that security breaches will occur by comparing observed behaviors with predicted probabilistic distributions. Discrete probability is applied in steganography and digital watermarking to embed secret message in images, audio, and video thereby minimizing detection (Petitcolas et al., 1999). Information embedding is modeled probabilistically to find the best locations while being resistant to statistical analysis attacks. It is precisely in the area of cryptographic security, intrusion detection and privacy preserving protocols that discrete probability is still at the core of cyber threats, which are becoming increasingly sophisticated. In future its role in securing digital communication will grow as development advances post-quantum cryptography, zero knowledge proofs, and probabilistic blockchain consensus mechanisms.

6.4 Discrete Statistical Methods in Disease Modeling and Pandemic Predictions

However, there are cross subsections within this domain that employ discrete statistical methods which are crucial in epidemiology, disease modelling, and forecasting pandemics among other things, to understand infection dynamics, progression of an outbreak, as well as the effect of public health interventions. To quantify uncertainty, estimate risks, and optimise containment strategies, we apply discrete probability models appearing from the fact that disease transmission are often discrete events (e.g., individual infections, recoveries and hospitalization) (Anderson & May, 2020).
The Susceptible-Infectious-Recovered (SIR) model is one of the most used model in epidemic forecasting that discretizes the population into three compartments, the susceptible (S) individuals, the infectious (I) individuals, and the recovered (R) individuals (Kermack & McKendrick, 1927). Additional extensions of this model like the Susceptible-Exposed-Infectious-Recovered (SEIR) model incorporate a latent period in order to predict diseases, like COVID-19 or influenza (He et al., 2020). Discrete differential equations and Markov chains are used by these models to simulate the spread of infections over time and calculate the effectivness of implements of intervention (vaccinatiation, social distancing, quarantine), amongst others. Pandemic risk assessment also widely coincides with discrete stochastic model application in the modeling of rare and uncertain outbreak events. This type of early forecast has been made in zoonotic spillover modeling or early outbreak detection (Lloyd Smith et al., 2005), by employing branching processes to assess the possibility of the disease extinction or explosion. Also, Poisson and negative binomial models are adopted for situations whereby the case distributions are overdispersed, such as super spreader events that are crucial in pandemics like Ebola, SARS and COVID 19 (Blumberg & Lloyd Smith, 2013).
Discrete statistical methods are another important application, these being in Bayesian disease models where hierarchical Bayesian models are used to estimate infection rates, mortality risks and the impacts of interventions (Gelman et al., 2013). Real time phylogenetic analysis of viral genomes is possible with the Bayesian Markov Chain Monte Carlo (MCMC) techniques such as BEAST and Stan that allows for tracing of mutation pattern, transmission clusters and evolutionary origins of pandemics (Drummond & Rambaut, 2007). Hospital resource planning and patient prognosis use discrete time Markov models for predicting ICU occupancy rates, hospitalization duration and ventilator demand when there are outbreaks (Wu et al 2020). These models empower health care policymakers to make data grounded decisions regarding strategic investment in resource allocations, in a manner that guarantees maximum efficiency of the critical care infrastructure during crises.
Finally, agent based models (ABMs) have come to be widely used in pandemic simulation, wherein individual agents (people) interact in a stochastic environment (Ferguson et al. (2006)). Plugins such as these models capture heterogenous behaviours, mobility patterns, and policy effects, and are very good at evaluating the effect of non pharmaceutical interventions including masks mandates, lockdowns, and lockdowns. Syndromic surveillance and outbreak detection later require discrete probabilistic techniques as well, where working with real-time case reports, genomic sequences, mobility data, and a growing corp of digital footprints, hidden Markov models (HMM) and Bayesian networks use best practices of tracking inference to identify emerging threats (Reich et al., 2016). Further improvements were added to early warning systems employing machine learning enhanced discrete models using Twitter feeds, search engine queries as well as wearable health sensors to alert of unusual disease patterns before clinical diagnosis is made to an outbreak (Kass-Hout & Alhinnawi, 2013). Discrete statistical methods will help the public health interventions, optimize vaccine distribution and mitigate the global health crises as disease surveillance systems keep becoming more and more data driven and computationally intensive. Future developments in Bayesian inference, network based epidemiology and AI driven disease modeling will increase predictive accuracy, improving pandemic preparedness and response strategies worldwide.

Figure 3- Applications of Discrete Probability in Finance, Biology, Security, and Healthcare

Application	Description	Impact	Fields of Application
Finance & Risk Analysis: Discrete-Time Models in Quantitative Finance	Discrete probabilistic models, such as binomial asset pricing, Markov chains, and Monte Carlo simulations, improve financial risk assessment.	Enhanced **option pricing, credit risk analysis, and portfolio optimization** for better financial decision-making.	**Investment Banking, Algorithmic Trading, Risk Management, Insurance**
Discrete Models in Genetic Sequencing and Evolutionary Biology	Markov models, Hidden Markov Models (HMMs), and Bayesian phylogenetics are used for DNA sequencing and evolutionary analysis.	Improved **gene prediction, mutation detection, and evolutionary tree reconstruction** for better insights into genetic variation.	**Genomics, Bioinformatics, Evolutionary Biology, Personalized Medicine**
Applications of Discrete Probability in Encryption and Security	Probabilistic encryption, random number generation, and post-quantum cryptography secure digital communication.	Strengthened **cybersecurity through secure encryption protocols, fraud detection, and blockchain security**.	**Cryptography, Cybersecurity, Blockchain, Digital Forensics**
Discrete Statistical Methods in Disease Modeling and Pandemic Predictions	SIR/SEIR models, Bayesian epidemiological models, and agent-based simulations predict disease spread and intervention effects.	More accurate **pandemic forecasting, healthcare resource allocation, and outbreak detection**.	**Epidemiology, Public Health, Disease Surveillance, Biostatistics**

7 Challenges and Open Research Problems in Discrete Probability and Statistics

Although there have been great leaps forward in discrete probability and statistics, there are still issues, determinations in scalability, lack of data, hybrid modeling, and ethics. It is very important to address these challenges in order to increase the model accuracy, computational efficiency, and real world applicability in these other fields such as AI, finance, cryptography, and epidemiology. In particular, most discrete probabilistic models, including Hidden Markov Models, Bayesian Networks, and many other statistcal models based on Markov Chain Monte Carlo (MCMC) methods are computationally intractable, and become combinatorially intractable when they are applied to large-scale databases (or more precisely, large databases with massive search spaces). In many cases, exact inference is NP hard in the state space explosion, as the number of discrete states increases (Koller & Friedman, 2009). VI and parallelized sampling have better scalability, but usually come at the cost of introducing APPEX (Blei et al., 2017). Future research must be dedicated to developing scalable algorithms that do not compromise tradeoff between efficiency and accuracy, especially for the real time decision making in the area of cybersecurity, finance, as well as to autonomous systems.

Such rich structured datasets are required for many discrete probabilistic models, but often hard to collect due to privacy concerns, small sample sizes, highly biased sampling, etc. (Vehtari et al., 2021). This is a very challenging problem in epidemiology, finance risk analysis, and social network modeling as real world datasets are generally incomplete, noisy, or very lacking labels. It is a large challenge to ensure that discrete models generalize well across multiple different datasets and do not overfit to particular areas. The mitigation of this issue can come from few shot learning, transfer learning and synthetic data generation techniques which yet need some more efforts. Such systems are many real world systems, which possess both discrete and continuous characteristics, and hence hybrid probabilistic models, as a combination of discrete probability distributions and continuous stochastic processes, are necessary. Specifically, the combination of Poisson processes (discrete jumps) with Brownian motion (continuous fluctuations) allows for modeling of market volatility (Cont & Tankov, 2004, p. 3) using discrete jump-diffusion models. Like with hybrid epidemiological models that join discrete agent based simulations with continuous differential equations, hybrid epidemiological models that combine discrete agent based simulation with differential equations are used to improve forecast of a pandemic (Ferguson et al., 2006). Despite this, there is yet an open problem to construct efficient hybrid models that remain tractable while retaining interpretability.

Predominant with the use of probabilistic AI models in the criminal justice system, hiring, diagnosis of medical issues, or finance are issues with fairness, transparency, as well as bias (O'Neil, 2016). While such discrete probabilistic classifiers as Bayesian networks and decision trees may not unwittingly effect discriminate, they may reproduce the biases in the training data. Explicit examples include the use of discrete probability distributions in predictive policing models that can reinforce what has been shown to reinforce racial and socioeconomic biases in law enforcement decisions (Benjamin, 2019). Also, risk communication to patients and policymakers in particular is difficult sometimes due to uncertainty quantification in medical diagnostics. The future research would be on explainable AI (XAI), fairness aware probabilistic models, and robust decision making frameworks to deploy the ethical AI.

8 Discussion

Since the theory of discrete probability and statistical methods has evolved so rapidly, it has completely transformed many fields such as machine learning, finance, epidemiology, cryptography, the social network analysis, to name a few. This review has presented some major advances mainly in Bayesian modelling, non classical distributions, graph based network models and probabilistic security frameworks. Although these improvements were made, there are still many open questions and challenges – namely on the scaleability, computational complexity, hybrid modelling, and regarding the ethical aspects. The purpose of this section is to discuss how recent developments affect research in the discrete probability and statistics, explore what is limited up to now, and what can be done in the future.

By integrating probabilistic graphical models, stochastic processes, and deep learning techniques, discrete probability models can now be applied to problems of large scale and real world. For sequence modeling tasks such as speech recognition and text generation driven by the machine learning and natural language processing (NLP), adding deep neural networks to a combination of the Hidden Markov Models (HMM) has produced great improvements (Jurafsky & Martin, 2021). In both finance and risk analysis, the same has occurred toward the refinement of discrete-time asset pricing models that have yielded sharper volatility predictions and better portfolio optimization strategies (Cont & Tankov, 2004).

Whether or not in epidemiology and disease modeling (e.g., Poisson based outbreak detection and Bayesian epidemiological models) improved pandemic forecasting and intervention planning (Ferguson et al. 2006). In addition, the construction of the post quantum cryptographic schemes on the discrete probability distribution has led to improved security of data in cryptography and cybersecurity against the threat of quantum computing (Lyubashevsky et al., 2013). These advances point to the fact that starting in modern computational and sci But it has been made despite the fact that several critical challenges hold back widespread adoption and effectiveness of discrete probability methods. As we scale and in general, computational complexity becomes one of the major issues. On the other hand, Markov Chain Monte Carlo (MCMC) methods and Bayesian networks are many probabilistic models that show exponential increase in computation time with growing dataset size (Blei et al., 2017). Better scalability in VI and parallel computing has also been achieved at the expense of accuracy. Future research must speed approximative inference techniques that do not require so much computations and retain a high accuracy.

An additional pressing problem is data availability as well as model generalization. The problem addressed by these discrete statistical methods is that many of them rely on well-structured, high quality datasets that are often missing, biased, or only available because of such potential privacy risks (Vehtari et al., 2021). In particular, this is a problem that is common to medical data, financial transactions, and cybersecurity logs, where access to labeled data is available but limited, making a difference on the performance of the models. However, transfer learning, few-shot learning and synthetic data generation can potentially aid but effectively solving the problem is an open research question. In addition, hybrid probabilistic models of this type are needed to integrate discrete and continuous parts. Examples of such phenomena representative of these dual discrete transitions and continuous fluctuations are financial markets, epidemiological spread, and climate modeling. Albeit this gap is not filled statically with simple hybrid models such as jump-diffusion models in finance or agent based models in epidemiology that still continue to require manual fine tuning and domain specific expertise (Cont & Tankov, 2004). Next the future research should be about developing the automated hybrid modeling techniques flexibly making discrete and continuous representations on that basis of the data patterns.

The ethical concern that fairness, transparency and bias concern are held serious by the significant use of Probabilistic AI models in criminal justice, finance and healthcare. Although discrete probabilistic models such as Bayesian classifiers and Markov decision processes seek to be well calibrated against the truth, they may unknowingly perpetuate the biases that are in the training data and yield discriminatory outcomes (O'Neil, 2016). As an example, predictive policing models can overpolicing of marginalized communities by predicting crime risk probabilistically, which enforces historical inequities (Benjamin, 2019). Likewise, in automated financial lending systems there are probabilistic models assessing loan default risks that may unintentionally discriminate some demographics thereby resulting in algorithmic bias in credit scoring. Future research should target bias corrected algorithms, fairness aware probabilistic models ensuring that will make decision that will be fair to any of the parties involved while being accurate about information at hand, and the explainable artificial intelligence (XAI) techniques.

9 Conclusion

Modern data driven decision making is still in your fingertips and it continues to be applications in AI, finance, epidemiology, cryptography and security. Nevertheless, issues of scalability, data constraints, hybrid modeling, and ethical deployment of AI have to be addressed to bring them to fruition. The future research should work on the computational efficiency, model fairness, and integration of hybrid probabilistic framework for the creation of robust, responsible and reliable AI system. Solving these problems will ensure that discrete probability and statistics will continue to be the cornerstone of scientific discovery, technological innovation for years to come.

10 Future scope of the study

In order for discrete probability and statistics to be preserved in their future, scalable improvements along with hybrid modeling, ethical AI and quantum resistant security, and interdisciplinary applications should be emphasized. To cope with big scale datasets, the computational bottlenecks in Markov Chain Monte Carlo (MCMC) methods as well as probabilistic graphical models can be addressed by approximate inference techniques, deep probabilistic programming and parallel computing. Such applications are handled particularly requiring adaptive hybrid frameworks that build on the integration of the discrete and continuous models. Fairness aware Bayesian models, explainable artificial intelligence (XAI) and uncertainty quantification are needed in probabilistic AI and decision-making because of their ethical concerns in the fields of healthcare, finance and criminal justice. There is a need for progress in post quantum cryptography, lattice based encryption, and quantum inspired probabilistic inference to secure data security that rises from the need for quantum computing. At the same time, Bayesian phylogenetics, probabilistic causal discovery, and the network based stochastic modeling will revolutionalize probabilistic quantitative biology in general, and biostatistics, genomics, computational neuroscience, in particular. Since machine learning and discrete statistical methods merge, future research will put emphasis on the automated probabilistic reasoning, the real time decision making and the intelligent AI powered statistical models that will bring robust, interpretable, efficient probabilistic systems in all disciplines.

References

1. Abbe, E. (2017). Community detection and stochastic block models: Recent developments. Journal of Machine Learning Research, 18(1), 6446-6531.
2. Agresti, A. (2018). Categorical data analysis (3rd ed.). Wiley.
3. Anderson, R. M., & May, R. M. (2020). Infectious diseases of humans: Dynamics and control. Oxford University Press.
4. Attal, S., Petruccione, F., & Sabot, C. (2019). Open quantum systems: A mathematical perspective. Springer.
5. Barabási, A. L. (2016). Network science. Cambridge University Press.
6. Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code. Polity Press.
7. Black, F., & Litterman, R. (1992). Global portfolio optimization. Financial Analysts Journal, 48(5), 28-43.
8. Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518), 859-877.
9. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research.
10. Blumberg, S., & Lloyd-Smith, J. O. (2013). Inference of R0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS Computational Biology, 9(5), e1002993.
11. Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. Proceedings of ICML.
12. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235-255.
13. Busemeyer, J. R., & Bruza, P. D. (2019). Quantum models of cognition and decision. Cambridge University Press.
14. Casella, G., & Berger, R. L. (2021). Statistical inference (2nd ed.). Cengage Learning.
15. Charalambides, C. A. (2019). Enumerative combinatorics and discrete probability. CRC Press.
16. Chen, Y., Liu, H., & Zhao, X. (2022). Non-homogeneous Markov models and applications in dynamic systems. Springer.
17. Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S. I. (2009). Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation. Wiley.
18. Cont, R., & Tankov, P. (2004). Financial modelling with jump processes. CRC Press.
19. Cormode, G., & Muthukrishnan, S. (2012). Approximating data with the Count-Min sketch. Communications of the ACM, 55(9), 121-130.
20. Cover, T. M., & Thomas, J. A. (2006). Elements of information theory (2nd ed.). Wiley.
21. Cox, J. C., Ross, S. A., & Rubinstein, M. (1979). Option pricing: A simplified approach. Journal of Financial Economics, 7(3), 229-263.
22. Denning, D. E. (1987). An intrusion-detection model. IEEE Transactions on Software Engineering, 13(2), 222-232.
23. Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, 7(1), 214.
24. Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press.
25. Easley, D., & Kleinberg, J. (2010). Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press.
26. Eddy, S. R. (2011). Accelerated profile HMM searches. PLoS Computational Biology, 7(10), e1002195.

27. Felsenstein, J. (2004). Inferring phylogenies. Sinauer Associates.
28. Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. Nature, 442(7101), 448-452.
29. Fischer, T. (2018). Reinforcement learning in financial markets–a survey. Journal of Economic Dynamics and Control, 93, 344-368.
30. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). CRC Press.
31. Glasserman, P., Heidelberger, P., & Shahabuddin, P. (2002). Portfolio value-at-risk with heavy-tailed risk factors. Mathematical Finance, 12(3), 239-269.
32. Goldreich, O. (2004). Foundations of cryptography: Volume 1, Basic tools. Cambridge University Press.
33. Goldwasser, S., & Micali, S. (1982). Probabilistic encryption & how to play mental poker keeping secret all partial information. Proceedings of STOC.
34. Graves, A., & Jaitly, N. (2021). Hybrid HMM-RNN models for speech recognition. IEEE Transactions on Audio, Speech, and Language Processing.
35. Grimmett, G., & Stirzaker, D. (2020). Probability and random processes (4th ed.). Oxford University Press.
36. Haven, E., & Khrennikov, A. (2019). Quantum social science. Cambridge University Press.
37. He, S., Peng, Y., & Sun, J. (2020). SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dynamics, 101(3), 1667-1680.
38. Hein, J., Schierup, M. H., & Wiuf, C. (2005). Gene genealogies, variation and evolution: A primer in coalescent theory. Oxford University Press.
39. Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2019). Bayesian nonparametrics. Cambridge University Press.
40. Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. Journal of Machine Learning Research, 14(1), 1303-1347.
41. Hull, J. C. (2017). Options, futures, and other derivatives (10th ed.). Pearson.
42. Iosifidis, V., & Vlahavas, I. (2020). Higher-order Markov models in machine learning and time-series analysis. Journal of Computational Intelligence.
43. Jarrow, R. A., & Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. The Journal of Finance, 50(1), 53-85.
44. Jordan, M. I. (2018). Artificial intelligence—the revolution hasn't happened yet. Harvard Data Science Review.
45. Jordan, M. I. (2018). Artificial intelligence—the revolution hasn't happened yet. Harvard Data Science Review.
46. Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Pearson.
47. Karlis, D., & Xekalaki, E. (2020). Mixed Poisson distributions. Chapman & Hall/CRC.
48. Kass-Hout, T. A., & Alhinnawi, H. (2013). Big data, big questions: Advancing health surveillance for the 21st century. Online Journal of Public Health Informatics, 5(3), 219.
49. Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. Proceedings of KDD.
50. Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, 115(772), 700-721.
51. Khrennikov, A. (2020). Quantum-like modeling in decision-making and cognition. Springer.
52. Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16(2), 111-120.
53. Koller, D., & Friedman, N. (2009). Probabilistic graphical models: Principles and techniques. MIT Press.
54. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. Nature, 438(7066), 355-359.
55. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550.
56. Lyubashevsky, V., Peikert, C., & Regev, O. (2013). On ideal lattices and learning with errors over rings. Journal of the ACM, 60(6), 43.
57. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
58. McCullagh, P., & Nelder, J. A. (2019). Generalized linear models (2nd ed.). Chapman and Hall/CRC.
59. Menezes, A. J., van Oorschot, P. C., & Vanstone, S. A. (2018). Handbook of applied cryptography. CRC Press.
60. Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT Press.
61. Neal, R. M. (2011). MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo.
62. Newman, M. E. J. (2018). Networks: An introduction (2nd ed.). Oxford University Press.
63. Nielsen, M. A., & Chuang, I. L. (2020). Quantum computation and quantum information (2nd ed.). Cambridge University Press.
64. O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing.
65. Peikert, C. (2016). A decade of lattice cryptography. Foundations and Trends in Theoretical Computer Science, 10(4), 283-424.

66. Peixoto, T. P. (2020). The graph-tool library for statistical network analysis. Journal of Statistical Software, 96(1), 1-34.
67. Petitcolas, F. A., Anderson, R. J., & Kuhn, M. G. (1999). Information hiding—a survey. Proceedings of the IEEE, 87(7), 1062-1078.
68. Pillai, R. N., & Jayakumar, K. (2021). Mittag-Leffler distributions and applications. Springer.
69. Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning. MIT Press.
70. Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., Moore, E., ... & Yamana, T. K. (2016). A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences, 116(8), 3146-3154.
71. Rivest, R. L., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2), 120-126.
72. Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p) models for social networks. Social Networks, 29(2), 173-191.
73. Ross, S. M. (2019). Introduction to probability models (12th ed.). Academic Press.
74. Rosso, O. A., Blanco, S., Yordanova, J., Kolev, V., Figliola, A., Schürmann, M., & Başar, E. (2001). Wavelet entropy: A new tool for analysis of short duration brain electrical signals. Journal of Neuroscience Methods, 105(1), 65-75.
75. Sayood, K. (2017). Introduction to data compression (5th ed.). Morgan Kaufmann.
76. Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423.
77. Shreve, S. E. (2004). Stochastic calculus for finance II: Continuous-time models. Springer.
78. Silver, D., Schrittwieser, J., Simonyan, K., & Hassabis, D. (2020). Deep reinforcement learning for sequence-based decision-making. Nature AI.
79. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems (NeurIPS).
80. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.
81. Teh, Y. W., & Jordan, M. I. (2019). Hierarchical Bayesian nonparametric models for hidden Markov processes. Journal of Machine Learning Research.
82. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. Proceedings of NeurIPS.
83. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. Bayesian Analysis, 16(2), 667-718.
84. Verma, A., & Ranga, V. (2019). An empirical evaluation of entropy-based anomaly detection in cloud computing. Journal of Cloud Computing, 8(1), 1-16.
85. Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2017). Five years of GWAS discovery. American Journal of Human Genetics, 101(1), 5-22.
86. Wakeley, J. (2009). Coalescent theory: An introduction. Roberts & Company.
87. Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modeling study. The Lancet, 395(10225), 689-697.
88. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution, 39(3), 306-314.
89. Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56-65.
90. Zhang, G., Luo, S., & Zhou, X. (2021). Quantum probability in AI and machine learning applications. Journal of Quantum Information Science.
91. Zhao, X., Zhang, L., & Xu, W. (2011). Wavelet entropy-based feature extraction for machine fault diagnosis. Mechanical Systems and Signal Processing, 25(4), 1296-1308.
92. Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57-81.