Thursday, 6th June 2024 - Morning Session (MS.01)

Sarah Filippi (Imperial College London) Variational Bayes for high-dimensional proportional hazard models (10:30-11:00)
Abstract. Few Bayesian methods for analysing high-dimensional sparse data provide scalable variable selection, effect estimation and uncertainty quantification. Most methods either sacrifice uncertainty quantification by computing maximum a posteriori estimates, or quantify the uncertainty at high (unscalable) computational expense. In this talk, we will focus on a specific problem setting: variable selection for high dimensional sparse survival data. For this setting, we develop an interpretable and scalable model for prediction and variable selection. Our method, based on a variational approximation, overcomes the high computational cost of MCMC whilst retaining the useful features, providing excellent point estimates and offering a natural mechanism for variable selection via posterior inclusion probabilities. We compare our methods against other state-of-the-art Bayesian variable selection methods on simulated data and demonstrate their application for variable selection on real biomedical data.

Hamid Rahkooy (University of Oxford) - Algebraic identifiability of partial differential equation models (11:00-11:30)
Abstract. Differential equation models in biology often include parameters and finding the values of the parameters is crucial for understanding the behaviour of the model. The parameter identifiability problem asks if the values of the parameters can be determined from the input and output functions. Identifiability is an important step for parameter estimation. For ordinary differential equation models, the identifiability problem has been widely studied using various approaches. However, a systematic algebraic approach for the identifiability problem for partial differential equations has not been yet developed. We present an algebraic approach to the identifiability problem for PDEs and demonstrate our method on several standard biological models

Short Talk session, 12:00-13:00

Vahid Shahrezaei (Imperial College London) - Bayesian model discovery for revers-engineering biochemical networks from data
Abstract. The reverse engineering of gene regulatory networks based on gene expression data is a challenging inference task. A related problem in computational systems biology lies in identifying signalling networks that perform particular functions, such as adaptation. Indeed, for many research questions, there is an ongoing search for efficient inference algorithms that can identify the simplest model among a larger set of related models. To this end, in this talk, I introduce SLInG, a Bayesian sparse likelihood-free inference method using Gibbs sampling. I demonstrate that SLInG can reverse engineer stochastic gene regulatory networks from single-cell data with high accuracy, outperforming state-of-the-art correlation-based methods. Furthermore, I show that SLInG can successfully identify signalling networks that

execute adaptation. Finally, I discuss application to inference of a cell fate gene regulatory network from real data. Sparse hierarchical Bayesian inference thus provides a versatile tool for model discovery in systems biology and beyond.

Andrea Maria Vergani (Human Technopole) - Prediction of incident cardiovascular events using cardiac MRI-derived latent factors
Abstract. Risk scores for cardiovascular disease (CVD) traditionally rely on a limited set of vascular risk factors (e.g., age, sex, smoking status). Cardiac magnetic resonance imaging (MRI) is known to provide additional information potentially predictive for CVD; however, recent studies only evaluated the impact of conventional cardiac measures and radiomics features from MRI in the forecast of incident cardiovascular events, without considering more comprehensive representations of cardiac magnetic resonance acquisitions. To bridge the gap, we extracted deep latent representations of long axis heart MRI with autoencoder-based techniques, and integrated them into risk prediction models for CVD (survival analysis study involving approximately 30k UK Biobank subjects healthy at imaging visit, with incidence of about 5% in the follow-up). In this way, we were able to evaluate the predictivity of cardiac MRI-derived latent factors for CVD endpoints and interpret them in terms of CVD risk. To the best of our knowledge, this is the first work about the integration of a comprehensive latent representation of cardiac MRI into prediction models for incident CVD, with the goal of developing more accurate risk scores.

Nicolas Rubido (University of Aberdeen) - Small-worldness favours network inference in synthetic neural networks
Abstract. A main goal in the analysis of a complex system is to infer its underlying network structure from time-series observations of its behaviour. The inference process is often done by using bi-variate similarity measures, such as the cross-correlation (CC) or mutual information (MI), however, the main factors favouring or hindering its success are still puzzling. In this work, we use synthetic neuron models to reveal the main topological properties that frustrate or facilitate inferring the underlying network from CC measurements. Specifically, we use pulse-coupled Izhikevich neurons connected as in the Caenorhabditis elegans neural networks as well as in networks with similar randomness and small-worldness. We analyse the effectiveness and robustness of the inference process under different observations and collective dynamics, contrasting the results obtained from using membrane potentials and inter-spike interval time-series. We find that, overall, small-worldness favours network inference and degree heterogeneity hinders it. Success rates in C. elegans networks – that combine small-world properties with degree heterogeneity – are closer to success rates in Erdös-Rényi network models than those from Watts-Strogatz network models. These results are relevant to understand better the relationship between topological properties and function in different neural networks.

Petar Jovanovski (Chalmers University of Technology and University of Gothenburg) - Towards Data-Conditional Simulation for ABC Inference in Stochastic Differential Equations

Abstract. We develop a Bayesian inference method for discretely-observed stochastic differential equations (SDEs). Inference is challenging for most SDEs, due to the analytical intractability of the likelihood function. Nevertheless, forward simulation via numerical methods is straightforward, motivating the use of approximate Bayesian computation (ABC). We propose a conditional simulation scheme for SDEs that is based on lookahead strategies for sequential Monte Carlo (SMC) and particle smoothing using backward simulation. This leads to the simulation of trajectories that are consistent with the observed trajectory, thereby increasing the ABC acceptance rate. We additionally employ an invariant neural network, previously developed for Markov processes, to learn the summary statistics function required in ABC. The neural network is incrementally retrained by exploiting an ABC-SMC sampler, which provides new training data at each round. Since the SDE simulation scheme differs from standard forward simulation, we propose a suitable importance sampling correction, which has the added advantage of guiding the parameters towards regions of high posterior density, especially in the first ABC-SMC round. Our approach achieves accurate inference and is about three times faster than standard (forward-only) ABC-SMC. We illustrate our method in four simulation studies, including three examples from the Chan-Karaolyi-Longstaff-Sanders SDE family.

Thursday, 6th June 2024 - Afternoon Session (MS.01)

Cathals Mills (University of Oxford) - A multi-disciplinary approach for wavelet analysis, climate- based modelling, and probabilistic ensemble forecasting of dengue epidemic dynamics (14:00-14:30)

Abstract. Understanding the past, current, and future dynamics of dengue, the world's most common arthropod-borne disease, is of increasing importance to public health authorities. Wavelet methods and Bayesian climate-based modelling have successfully characterised dengue epidemic dynamics, whilst modern statistical, machine learning, and probabilistic ensemble methods have demonstrated strong predictive capabilities for infectious disease forecasting. Here, using data from fourteen provinces in northern Peru across 2010 to 2021 as a case study, we provide a new, unifying multi-disciplinary approach for analysing and forecasting dengue incidence.

Our wavelet analysis allowed us to study spatiotemporal patterns in epidemic dynamics, and we identified that large outbreak years were associated with climatic forcing and enhanced spatial similarity in both epidemic synchrony and coherence. Spatially, in more northerly coastal provinces, climatic influences, shorter pairwise distances, and greater connectivity induced greater similarity in epidemic dynamics. We quantified the timing, structure, and intensity of climatic relationships with dengue incidence rates (DIRs) via our well-fitting Bayesian climate-based model. Elevated DIR risk was uncovered for greater temperatures, sustained drought, and heavy precipitation, thus supporting aforementioned peaks in climatic forcing during large outbreak years.

In our probabilistic ensemble forecasting analysis, demonstrating trained and untrained model weighting approaches, we found that multi-disciplinary ensemble frameworks, which borrow strengths from diverse fields, yielded stable prediction accuracy, superior proper scoring rule values, robust uncertainty quantification, and reliable outbreak detection, irrespective of forecast representation either by samples or (less computationally intensive) quantiles. Indeed, an equally-weighted ensemble outperformed individual forecasters in terms of probabilistic calibration, sharpness, predictive accuracy, and ability to anticipate months with DIR ≥ 50 per 100,000 (True Positive = 0.85, False Positive = 0.06) one month ahead of time.

Looking forward, our multi-disciplinary approach could be used by authorities to probabilistically estimate time-and-space-dependent environmental risks, current and past epidemic dynamics, and likely future trajectories

Hannah Bensoussane (University of Warwick) - Bayesian individual-level infectious disease modelling: heterogeneous transmission and dealing with costly likelihood evaluation when estimating missing data (14:30-15:00)

Abstract. Fitting mathematical models to epidemic data is challenging because the transmission process is largely unobserved. Add into the mix that the characteristics of an individual can affect their ability to catch and transmit infectious diseases and resulting models can be both complex and computationally costly. We introduce two models that allow for heterogeneous disease transmission in a discrete time setting and develop an inferential approach that allows accurate inference about model parameters with minimal computational burden.

We use an adaptive MCMC algorithm to estimate model parameters and identify characteristics that affect transmission. In both models, likelihood evaluation is expensive and so by developing a grouped approach to updating missing infection and infectious times, we reduce the number of likelihood evaluations and overall computational cost. In addition, we develop an algorithm that automatically tunes this group size based on the user's desired acceptance rate. The grouped nature of the updates allows us to explore novel proposal mechanisms and for each algorithmic variant we find the desired acceptance rate for groups of infection and infectious times that results in the highest mean square jumping distance (MSJD) per second. Model performance is assessed through a variety of simulation work and the results are promising – both models have the ability to successfully identify covariate effects on both susceptibility and infectiousness.

In the case of emerging diseases, the models introduced will prove invaluable in their ability to quickly determine the key characteristics that drive epidemics on an individual level. A key strength of both models is their ability to account for the effect of virtually any characteristic of interest. Additionally, the novel algorithm developed for the efficient estimation of infection and infectious times is easily implementable and requires minimal input from the user.

Catalina Vallejos Mills (University of Edinburgh) - Using routine healthcare data to predict future health (15:30-16:00)

Abstract. Can we identify who will experience an adverse health event (e.g. disease onset) weeks, months or even years before it happens? Questions like this are at the core of health data science research and have been empowered by the increasing ability to securely access routinely collected electronic health records (EHR). EHR is a rich source of data which contains detailed information about entire patient populations. However, extracting robust insights from EHR is not a trivial task for multiple reasons including recording errors, unstructured data collection and observational biases, amongst others. In this talk, I will provide an overview of our research in this area, highlighting how EHR can be used to stratify individuals according to their risk profiles and disease trajectories. I will also describe the development of SPARRAv4 (Scottish Patients at Risk of Readmission and Admission version 4) a risk score developed in collaboration with Public Health Scotland and that will be shortly deployed at national level in primary care settings. Finally, I will describe some of the practical and methodological challenges that we have encountered throughout these projects.

Huizi Zhang (University of Edinburgh) - Bayesian modelling of RNA velocity from single-cell RNA sequencing data (16:30-17:00)

Abstract. The notion of RNA velocity has provided a new way to understand dynamic information from the snapshot of single-cell RNA sequencing (scRNA-seq) data, garnering significant attention and leading to numerous extensions. RNA velocity is defined as the vec-

tor of the rates of change of spliced RNA for different genes. However, existing methods often lack uncertainty quantification and many adopt unrealistic assumptions or employ complex black-box models that are difficult to interpret. Here we present a Bayesian hierarchical model to estimate RNA velocity, with a time-dependent transcription rate and non-trivial initial conditions, which provides appropriate calibration of the uncertainty. There are various challenges in implementing an MCMC algorithm such as multimodality of the posterior distribution and weak identifiability of some parameters. The method has been validated in a simulation study and applied to sc-RNAseq data from mouse embryonic stem cells. The model provides a reasonable fit to the considered real data which was done using mixed posterior predictive checks. As well as estimating uncertainty of RNA velocity and model parameters, for the considered data, our model provides a better fit to the data than a model with a constant transcription (scVelo).

Poster session (Thursday, 6th June, 17:00-19:00)

Tarek Alrefae (University of Oxford) - Heterogeneity in Models of Infectious Disease.

Abstract. When modelling infectious disease spread, one must make several assumptions to characterise the interactions between infected and susceptible individuals in a population1. The mathematical underpinnings of these interactions have great implications on several key epidemiological metrics, including the final attack size and herd immunity threshold2. Traditional compartmental models which assume homogeneous mixing and exponentially distributed transition times between susceptibility, latency, and infection are analytically tractable but oversimplified. Computationally intensive individual-based models, while accounting for potentially limitless heterogeneity, are often too complex for practical use and are not amenable to statistical inferencing. The exact role that population-level heterogeneities play in shaping epidemic outbreaks remains poorly-defined. Here we show that the inclusion of two types of heterogeneities (susceptibility and connectivity) in an age-dependent network model yields more realistic dynamics of disease spread compared to traditional compartmental models. Building on previously published work3,4, we amplify the versatility and applicability of these methods for modelling more realistic disease spread by using moment closure and generalising from previously-assumed Gamma distributions of heterogeneous susceptibility and connectivity. This more general framework allows for arbitrary distributions of heterogeneity across two major population-level characteristics to be plugged-in by the user, providing modellers and policy makers with the tools to tackle specific disease dynamics under minimal assumptions. We present this work as a proof-of-concept that the interplay between analytical tractability, computational efficiency, and realism can be more appropriately balanced in modelling infectious diseases. The adaptability of this general framework and its potential to be fit to real-time epidemiological data could significantly enhance the forecasting ability of infectious disease models, and thereby increase the efficacy of potential non-pharmaceutical interventions.

Jake Carson (University of Warwick) - Inference of Infection Disease Transmission through a Relaxed Bottleneck Using Multiple Genomes per host.

Abstract. In recent times, pathogen genome sequencing has become increasingly used to investigate infectious disease outbreaks. When genomic data is sampled densely enough amongst infected individuals, it can help resolve who infected whom. However, transmission analysis cannot rely solely on a phylogeny of the genomes but must account for the within-host evolution of the pathogen, which blurs the relationship between phylogenetic and transmission trees. When only a single genome is sampled for each host, the uncertainty about who infected whom can be quite high. Consequently, transmission analysis based on multiple genomes of the same pathogen per host has a clear potential for delivering more precise results, even though it is more laborious to achieve. Here, we present a new methodology that can use any number of genomes sampled from a set of individuals to reconstruct their transmission network. Furthermore, we remove the need for the assumption of a complete transmission bottleneck. We use simulated data to show that our method becomes more accurate as more genomes per host are provided, and that it can infer key infectious disease parameters such as the size of the transmission bottleneck, within-host growth rate, basic reproduction number, and sampling fraction. We demonstrate the usefulness of our method in applications to real datasets from an outbreak of Pseudomonas aeruginosa amongst cystic fibrosis patients and a nosocomial outbreak of Klebsiella pneumoniae.

Yu Chen (Imperial College London) - Bayesian rate consistency model to uncovering the hidden structure of contemporary sexual networks in Africa
Abstract. Sexual network analysis is crucial for comprehending and forecasting HIV spread, guiding effective public health interventions. We developed a non-parametric Bayesian model to characterise age, gender and occupation specific pairwise sexual contact patterns similar to spatiotemporal models with Matern covariance structure. Critically though, contact flows must add up and we are able to leverage the resulting consistency constraints to correct for under-reporting and age-heaping, and reveal the true extent and heterogeneity in contact patterns. The model can also recover (not predict) contact intensities for unseen participant age groups based on contacts reported from younger participants to these age groups. Our approach remains scalability based on Kronecker product kernels and Hilbert Space Gaussian process approximations. We showcase results on large-scale sexual contact network data from the Rakai Community Cohort Study in East Africa. This study advances understanding of latent under-reporting factors in sexual networks, crucial for informing public health strategies against HIV transmission.

Luca Del Core (University of Nottingham) - Accounting for stochastic gating whilst estimating ion channel kinetics from whole-cell patch-clamp recordings.
Abstract. The heartbeat is coordinated by ion-channels in cell membranes that change their conformation, a process known as gating, allowing ions to pass through them. Mathematical models of cardiac ion channel gating provide insight into this process and are the basis of multi-scale models of the heart. These models are usually defined as biochemical reactions describing the transitions between the different ion channel configurations, modelled as Markov latent states, as a result of the electrical activity of the cell membrane. Whole-cell voltage-clamp data allows us to calibrate the parameters of such mathematical models.

However, standard approaches do not distinguish between stochastic noise and measurement errors, and the resulting parameter estimates can be biased. To overcome these limitations, we propose a state-space model including a set of Itô-type stochastic differential equations describing ion channel gating, coupled with an Ohmic equation linking the noisy measurements of the current to the underlying Markov states. Parameter inference is based on an expectation-maximization algorithm.

Synthetic studies show that our proposed method can infer the unknown parameters with a low degree of uncertainty and high predictive power. Our proposed state-space formulation is also able to distinguish between stochastic noise and measurement errors, thus allowing to estimate the number of ion channels in the cell membrane that are contributing to stochastic gating. These results will improve models of ion channel dynamics by accounting for stochastic gating and measurement errors during fitting.

Richard Everitt (University of Warwick) - Ensemble Kalman inversion approximate Bayesian computation.

Abstract. Approximate Bayesian computation (ABC) is the most popular approach to inferring parameters in the case where the data model is specified in the form of a simulator. It is not possible to directly implement standard Monte Carlo methods for inference in such a model, due to the likelihood not being available to evaluate pointwise. The main idea of ABC is to perform inference on an alternative model with an approximate likelihood, sometimes known as the ABC likelihood. The ABC likelihood is chosen such that an unbiased estimator of it is easy to construct from simulations from the data model, allowing the use of pseudo-marginal Monte Carlo algorithms for inference under the approximate model. The central challenge of ABC is then to trade-off bias (introduced by approximating the model) with the variance introduced by estimating the ABC likelihood. Stabilising the variance of the ABC likelihood requires a computational cost that is exponential in the dimension of the data, thus the most common approach to reducing variance is to perform inference conditional on summary statistics.

In this talk we introduce a new approach to estimating the ABC likelihood: using ensemble Kalman inversion (EnKI). Ensemble Kalman algorithms are Monte Carlo approximations of Bayesian inference for linear/Gaussian models. These methods are often applied outside of the linear/Gaussian setting being used, for example, as an alternative to particle filtering for inference in non-linear state space models. Loosely speaking, EnKI can be used as an alternative to an SMC sampler on a sequence of annealed likelihoods. We see that EnKI has some appealing properties when used to estimate the ABC likelihood. It circumvents the exponential scaling with dimension of standard ABC, and does not require the reparameterisation imposed by the rare-event SMC approach of Prangle et al. (2018). It is able to achieve this with no additional simulations from the data model, thus it is likely to bring the most benefit in cases where this model is very expensive to simulate.

Kit Gallagher (University of Oxford) - Using a Kalman filter to infer biological parameters from imperfect time series data.

Abstract. Stochastic models used to fit biological time series traditionally assume that the data is perfect, but this is rarely the case in practice. For example, when tracking the emergence of a pandemic, data on the number of new infections are typically imperfectly recorded, but used to inform policy-relevant parameters, such as the time-varying reproduction number, Rt. We introduce a Kalman-filter-based method which can estimate Rt when case data have arbitrary measurement imperfections. Our method is fast, straightforward to adapt to a given model of case miscounting and can handle diseases with or without superspreaders. Using real and simulated case data for a range of infectious diseases, we demonstrate how our method can be used to mechanistically model reporting delays and simultaneously estimate Rt and the true infection count for a variety of measurement imperfections. By fitting a mechanistic reporting delay model to COVID-19 case data with marked weekly variation in case counts, we find drastically different reporting delays across England, Israel and Spain and estimates of Rt with substantially more uncertainty than if the case data are smoothed. Our results showcase the benefits of thinking mechanistically about the processes that generated disease case data and that failing to account for measurement imperfections can lead to seriously overconfident estimates of Rt.

Alicia Gill (University of Warwick) - Bayesian Inference of Reproduction Number from Epidemic and Genomic Data using MCMC Methods.
Abstract. Typically, reproduction number is inferred using only epidemic data, such as prevalence per day. However, prevalence data is often noisy and partially observed, and it can be difficult to identify whether you have observed many cases of a small epidemic or few cases of a large epidemic. Genomic data is therefore increasingly being used to understand infectious disease epidemiology, and inference methods incorporating both genomic and epidemiological information are an active area of research.

We use Markov chain Monte Carlo methods to infer parameters of the epidemic using both a dated phylogeny and partial prevalence data to improve inference compared with using only one source of information. To do this, we have implemented a sequential Monte Carlo algorithm to infer the latent unobserved epidemic, which is then used to infer the reproduction number as it varies through time. We then analyse the performance of this approach using simulated data. Finally we present case studies applying the method to real datasets.

Andrii Krutsylo (Institute of Computer Science Polish Academy of Science) - The Forward-Forward Algorithm: Biologically Inspired Optimization for Continual Learning.
Abstract. Continual Learning (CL) remains a significant challenge in neural networks, largely due to the risk of catastrophic forgetting during sequential task learning. Traditional gradient-based optimization methods struggle to retain knowledge from previous tasks while adapting to new ones. This paper introduces an innovative approach that replaces gradient-based optimization with the Forward-Forward (FF) algorithm, combined with complementary CL techniques, to address this challenge. The FF algorithm diverges from conventional backpropagation by employing two forward passes for positive and negative data, eliminating the need for separate backward passes and promoting efficient learning. Task-specific objective functions for positive and negative data facilitate the retention of task-specific information while accommodating new tasks. Despite the Forward-Forward algorithm's limited advantages over standard SGD and its lower performance on established benchmarks, its biologically inspired nature makes it an intriguing choice for Continual Learning. While vulnerable to catastrophic forgetting when used in isolation, FF shows promise in enhancing existing CL methods and potentially catalyzing the development of novel approaches tailored to its unique characteristics.

Alessia Mapelli (Politecnico di Milano) - Graphs for representation of interacting biological systems and prediction of complex diseases.
Abstract. As diseases progress, the human body transforms, affecting genetic and observable characteristics. With the advent of molecular biology, omics data, including genomics, proteomics, and metabolomics, provide insights into disease progression and have become pivotal

in predicting disease states, enabling early diagnosis, and understanding disease mechanisms. Therefore, they continue to attract growing research interest. Recent studies suggest that co-functional gene modules or activated pathways better reflect disease status and biological processes. Methods exploring the topological information within omics networks have shown promising results in identifying biomarkers and uncovering disease mechanisms, providing a new direction for omics data analysis. Despite the significant advancements in leveraging omics data for disease diagnosis, existing methods often rely either on general biological graph databases or solely on data-driven approaches. However, protein-protein interaction databases favor well-studied proteins and genes, affecting the analysis's thoroughness and accuracy. At the same time, confounding factors may influence co-expression networks, leading to biased associations. This work introduces a novel data representation method to integrate biological knowledge from databases into an estimation-based approach, minimizing noise in co-expression network construction and addressing the impact of confounding factors. The method has been employed to create a protein-protein network starting from a subset of the UK Biobank cohort and to study the biological impact of diabetes via differential network analysis.

Thomas Morrish (University of Warwick) - An approximate likelihood framework motivated by the irregular movement of animals.
Abstract. Motivated by challenges in animal movement modelling and drawing inspiration from rough path theory, we present an approximate likelihood framework designed for discretely observed sets of controlled differential equations. The framework is developed with flexibility in mind so that inference is facilitated in a diverse range of scenarios. These can include nontrivial driving paths such as a fractional Brownian motion (where $h > 1/4$) or other Gaussian Processes, and further, in multivariate SDE settings where a Lamperti transform is not admitted i.e. where the set of equations cannot be reduced to one with unit diffusion coefficient. The methodology may have a variety of applications, including but not limited to finance, computational biology and population genetics.

Michael Plank (University of Canterbury) - A compartment-based model of Covid-19 in New Zealand: exploiting model structure to improve inference methods.
Abstract. In this talk, I will present a compartment-based epidemiological model for Covid-19 that was designed and used for policy advice in Aotearoa New Zealand during the pandemic. The model includes various processes and variables, including age-dependent susceptibility and contact rates, vaccine- and infection-derived immunity to infection and severe disease, and waning immunity. The model was fitted to epidemiological data on reported Covid-19 cases, hospital admissions and deaths. This was done using an approximate Bayesian computation method for a number of uncertain model parameters, including a time-dependent transmission coefficient. However, this method was inefficient for a relatively high-dimensional parameter space. In the second part of the talk, I will discuss work in progress that aims to exploit known aspects of model structure to develop a more efficient inference algorithm. This approach could be applicable to a range of model types where

there is a subset of the target parameters that have a known and predictable effect on the model output.

Ian Roberts (University of Warwick) - Bayesian Inference for the Structured Coalescent.

Abstract. The structured coalescent models the common ancestry of genetic material sampled from a spatially structured population. A realisation of this process can be represented as a structured tree consisting of a phylogenetic tree giving ancestral genetic relationships, and a migration history identifying the geographic location of each lineage at all times. The space of structured trees is notoriously large, with (2n-3)!! possible phylogenetic tree topologies, n-1 coalescent times and any finite number of migration events placed across the tree. This results in a highly multimodal, non-Euclidean space which is difficult to sample from.

Current inference methods either rely on approximations to the structured coalescent, or attempt to simultaneously infer the structured tree at great computational expense. I will present a Markov Chain Monte Carlo (MCMC) scheme which strikes a balance between these extremes by sampling migration histories on a fixed phylogenetic tree using a localised approximation of the structured coalescent to generate MCMC proposals.

Kristian Romano (University of Warwick) Hidden Markov Models for Real Time Telemetric Monitoring of the Circadian Rhythm.

Abstract. Motivated by the MultiDom clinical trial (NCT04263948) we developed methodology based on the Hidden Markov Model (HMM) for detecting alterations in the rhythmicity of time series data after an intervention. Disruptions of the Circadian rhythm in cancer patients are associated with poorer treatment outcomes, and short progression-free and overall-survival. MultiDom telemonitors patients with advanced pancreatic cancer undergoing standard chemotherapy aiming to reduce the rate of patients undergoing toxicity-related emergency admissions. The use of telecommunicating wearable devices in the trial provides a continuous flow of physical activity and body temperature data that allows for near-real time analyses and estimation of circadian parameters dynamics to inform proactive decision making by Medical Experts. For that aim we developed a HMM with a time changing transition matrix estimated via Adaptive Metropolis. As we assumed Zero Inflated Gamma emissions, we will employ a Metropolis-Hastings for the shape parameter with an educated proposal distribution. The work presented provides a first step into the reproducible and fully automatic Bayesian computation in real time of Circadian parameters of interest for the purpose of telemonitoring.

Elena Sabbioni (Politecnico di Torino) Regularized MANOVA test for zero-inflated semicontinuous high-dimensional data

Abstract. Dealing with extremely high-dimensional data is common in many biological research frameworks, for example when working with omics data. Moreover, data often exhibits a substantial number of missing observations or values equal to zero. To deal with this kind of data, we propose a MANOVA test tailored for semicontinuous measurements with a certain fraction of zero observations, that remains applicable even when the dimensionality exceeds the sample size, enabling the comparison of both means and the probability of missing data across different groups. The test statistic is derived as a regularized likelihood ratio test, with both the numerator and denominator computed at the maxima of penalized likelihood functions under each hypothesis. The use of closed-form solutions for the regularized estimators reduces computational costs. We establish the null distribution through a permutation scheme and assess the power and significance level of the resulting test through a simulation study. To illustrate this novel methodology, we apply the proposed test on an original dataset, where we examine the expression of microRNAs in human blastocyst cultures, comparing data from the Inner Cell Mass (ICM) with the remaining cells forming the outer TrophoEctoderm (TE). The ICM develops into a human being, while the TE evolves into a placenta and other annexes and these data exhibit numerous zeros due to the very early stage of development.

Joseph Shuttleworth (University of Nottingham) - Using many different protocols to characterise discrepancy in mathematical ion channel models.
Abstract. Model discrepancy, the underlying differences between mathematical models and the behaviour of some real-world process, is a particular challenge when training accurate predictive models. We use multiple experiments to thoroughly validate our models and produce predictions that are more robust to model discrepancy.

We explore model discrepancy in models of IKr—the total current through a cell's hERG channels. This current is of particular interest because it plays a crucial role in the propagation of electrical signals through the heart and its inhibition (whether by mutations or drugs) is associated with an increased risk of arrhythmia. By accounting for model discrepancy, we may improve our models and better classify pro-arrhythmic risk.

We show that, due to model discrepancy, the choice of experiment used to train the model has a significant impact on resulting parameter estimates and the accuracy of subsequent predictions. We propose an approach to account for these effects by using multiple experiments to independently train the same model. This results in an ensemble of parameter sets which we use to quantify uncertainty in our predictions of IKr.

This work provides a way to select better models, and may aid design of future experimental designs.

Nicholas Steyn (University of Oxford) - Sequential Monte Carlo methods for reproduction number estimation.
Abstract. Estimation of the instantaneous reproduction number is a well-researched problem in infectious disease epidemiology. Popular methods, like EpiEstim and EpiFilter, use a renewal model to estimate this quantity from reported case data. However, within these models, it can be important to account for various, sometimes disease-specific, confounding factors such as reporting delays, missing data, or sporadically reported data. Modifications to current methods can account for one or two of these factors at a time, but the complexity of these methods increases as additional factors are accounted for. We present a sequential Monte Carlo framework that allows for the adjustment of as many factors as required, consolidating best practices from existing literature into a simple unified framework, while maintaining a constant level of complexity. Our approach can be implemented in fewer than

100 lines of code, negating the need for external software packages and allowing researchers to understand the nuances of their model.

Nenad Šuvak (University of Osijek) - Time-changed SIRV model for epidemic of SARS-CoV-2 virus.
Abstract. The stochastic version of the SIRV (susceptible-infected-recovered-vaccinated) model for the epidemic of the SARS-CoV-2 virus in the population of non-constant size and finite period of immunity is considered. Among many parameters influencing the dynamics of this model, the most important is the contact rate, i.e. the average number of adequate contacts of an infective person, where an adequate contact is one which is sufficient for the transmission of an infection if it is between a susceptible and an infected individual. It is expected that this parameter exhibits time-space clusters which reflect interchanging of periods of low and steady transmission and periods of high and volatile transmission of the disease. The stochastics in the SIRV model considered here comes from the noise represented as the sum of the conditional Brownian motion and Poisson random field, closely related to the corresponding time-changed Brownian motion and the time-changed Poisson random measure. The existence and uniqueness of positive global solution of the stochastic SIRV process is proven by classical techniques. Furthermore, persistence and extinction of infection in population in long-run scenario are analyzed. In particular, conditions depending on parameters of the model and the underlying measure, under which the persistence and the extinction of the disease appear, are derived. The theoretical results are illustrated via simulated examples. In particular, transmission coefficient is simulated as the mean-reverting CIR diffusion with jumps with different propositions for the absolutely continuous time-change process. The recovery problem of the transmission coefficient and the corresponding time-change process from numbers of susceptible, vaccinated, infected, and recovered individuals is briefly discussed.

Jia Le Tan (University of Warwick) - Pareto Smoothed Sequential Monte Carlo.
Abstract. Importance Sampling (IS) is recognised for its theoretically unbiased nature. However, challenges arise when the proposal distribution markedly deviates from the target distribution, primarily due to the high variance in the weights it generates. In scenarios where a sampled point from a low-likelihood region in the proposal distribution aligns with a high-likelihood in the target distribution, there is a disproportionate 'weight stealing' effect, leading to a reduced effective sample size. This issue persists in higher dimensions, even when the proposal and target distributions are relatively similar. To address this, Vehtari et al. (2022) introduced Pareto-Smoothing (PS) within IS, culminating in the development of Pareto-Smoothed Importance Sampling (PSIS). This technique has been effective in substantially reducing weight variance with only a minimal increase in bias, especially when compared to methods like Truncated Importance Sampling (TIS).

Building on the foundation laid by Vehtari et al., our research integrates the PS methodology within the Sequential Monte Carlo (SMC) framework. Our goal is to replicate the enhancements observed in IS for SMC. In particular, we selectively apply Pareto-Smoothing to overcome challenges to SMC identical to those of IS aforementioned, such as sample

impoverishment and weight degeneracy.

A pivotal aspect of our study is the investigation of diverse methods for embedding Pareto-Smoothing into the SMC framework. We showcase various integration strategies and evaluate the efficacy of these novel methods using environmental models. Through this comparative analysis, we aim to pinpoint the most effective Pareto-Smoothing techniques for optimising SMC, specifically tailored to varied environmental modelling contexts.

Joseph Lok Hei Tsui (University of Oxford) - Optimal disease surveillance with graph-based active learning.
Abstract. Timely detection of emerging pathogens and precise tracking of their spread is critical to the design of effective public health responses. In resource-constrained settings, policymakers face the challenge of allocating limited testing resources across space, with the goal of providing accurate estimates of the underlying disease spread in a cost-effective manner. This study considers the task of designing optimal policies for disease surveillance in undirected and unweighted graphs, with nodes representing countries and edges representing human mobility between countries. We design our policies sequentially via information gain-based criteria: at each time step, a pre-specified number of countries are selected for testing, feedback from which is then used to update estimates of the probability of the pathogen being present at each location and to guide the selection of countries for testing at the next time step. We evaluate and compare the performance of standard active learning policies, e.g. node-entropy and BALD. Using data from simulated epidemics on empirical human mobility networks, we demonstrate the advantages and disadvantages of various policies under different outbreak scenarios and network structures. With these insights, we propose a novel policy that takes explicit consideration of the impact of different testing outcomes and selects countries for testing with the objective of maximising expected impact.

João Pedro Valeriano Miranda (Instituto de Física Teórica, Universidade Estadual Paulista Júlio de Mesquita Filho) - Recovering the dynamics of unobserved quantities in stochastic processes.
Abstract. Studying the dynamics of biological systems most frequently requires dealing with incomplete data: we can only keep track of a limited number of degrees of freedom necessary to fully characterize a system, while possibly many others go unobserved. The description of phenomena in such scenarios remains an open question, given we cannot measure every quantity we may wish for. However, we can expect that if some unobserved variables have their dynamics tightly correlated to that of observed variables, it may be possible to start from an incomplete dataset and recover a mathematical model describing the dynamics of both observed and latent variables. The recovered model would allow the reconstruction of unobserved trajectories using the observed ones, opening new paths for investigating complex biological systems. Here, we discuss the main challenges associated with considering hidden variables, and we present some preliminary results trying to learn stochastic differential equation models from incomplete dynamical data on simple biological systems. We rely on the filtering and smoothing formalism, which permits a precise definition of dynamical and measurement models, separating observed and unobserved variables. The same formalism

provides likelihoods for these incomplete datasets that can be approximated analytically or through high-dimensional numerical integration.

Sarah Vollert (Queensland University of Technology) - Constructing constraint-informed prior distributions for inference in data-limited scenarios: a case study in ecosystem population models.
Abstract. Ecosystem population models are a useful tool for analysing the potential effects of conservation actions. However, they can be challenging to parameterise in the face of limited data. Time series data is often sparse, noisy, or unavailable, leading ecosystem modellers to rely on priors informed by theoretical assumptions about how ecosystems function. Still, building these informed priors can be computationally intensive due to the cost of simulating the model and the low probability of finding ecosystems which meet the constraints. We develop novel calibration methods inspired by sequential Monte Carlo approximate Bayesian algorithms to parameterise ecosystem models without the computational cost, by steadily and adaptively incorporating the constraints. Our new algorithms unlock the capability to model the large and complex ecosystems that occur in nature – a task previously not possible – and to explore the theoretical assumptions that underpin ecosystem modelling. These algorithms can be broadly applied in mathematical biology to incorporate qualitative knowledge into parameterisation processes.

Kate Woolley-Allen (University of Warwick) - Multiplicative transposase cutting bias in ATAC-seq data.
Abstract. ATAC-seq is a key data type where an enzyme called Tn5 transposase is used to reveal open chromatin that normally corresponds to regions of regulatory DNA, informing the underlying mechanisms that drive transcriptional changes. However, inherent enzyme cutting bias associated with these technologies distorts raw data. This bias can interfere with downstream analyses such as footprint detection methods that elucidate the location of transcription factor binding sites. We consider the structure of the Tn5 homodimer and the interaction of the pair of single-strand cut sites that are both required for DNA fragmentation to occur. Whilst bias correction methods exist for ATAC-seq data, none consider the interplay of the bias across the paired sites, investigating the sequence between the pair of cut sites instead. We find the combined effect of the bias at the two cut sites of a single Tn5 homodimer is multiplicative, not additive as may be anticipated. Having established this characteristic of the Tn5 transposase, we have been able to mathematically model the enzyme cutting bias of the homodimer and propose associated methods for the deep correction of ATAC-seq data. Since the recognition of footprints within ATAC-seq data is particularly sensitive to the enzyme cutting bias, we believe our bias correction methods will improve the accuracy of footprinting methods.

Mengxin Xi (King's College London) - Extrapolation methods for Bayesian Inverse Problems
Abstract. Accurate cardiac models, based on systems of coupled and parametrized differential equations, are vital in guiding personalized treatments, necessitating robust inference of model parameters. The Bayesian framework allows to naturally convey uncertainty

in the parameters and account for system identifiability. However, as we integrate more phenomenological details, model complexity increases. This emphasizes the importance of optimizing the computation of posterior distributions of the parameters of cardiac models, where the balance between computational efficiency and accuracy is essential. First, We will focus on sample posteriors produced by interacting particle sampling methods, such as Stein Variational Gradient Descent (SVGD), for effective parameter estimation. We identify 'meta-parameters' related to the inference procedure. First, likelihood evaluation requires computing numerical solutions the differential equations describing cardiac models, which converge towards the exact solution as the discretization parameter approaches zero. Achieving such convergence is limited by computational capacity, given the high cost associated with small values of this parameter. Moreover, inference could be targeting a tempered posterior, where the tempering parameter handles a trade off between accuracy and computational complexity. Finally, the performance of the chosen sampling method (SVGD) depends on the number of particles and on running time, with larger values working towards recovering more precisely the target posterior, at the expense of higher cost. In this work, we consider the case where a given quantity of interest (QoI), such as the posterior mean or variance, is specified by the user. Instead of running a single expensive inference procedure (EIP), we suggest running multiple cheap inference procedures (CIP), based on sub-optimal tuning of the above meta-parameters. We speculate that the QoI of the EIP can be extrapolated from the corresponding QoI of the CIPs. The main advantage is offered by running the CIPs on parallel computational setups, which enables the compression of working time, aligned with clinical application needs.

Dominic Zhou (University of Warwick) - Adaptive MCMC inference in the Kingman coalescent model – propriety, ergodicity, efficiency.
Abstract. The Kingman coalescent model underpins population genetics study, but performing inferences from its complex posterior distribution is computationally expensive. We apply adaptive MCMC algorithms to coalescent-based targets defined on a combinatorial state space consisting of both discrete tree topologies and continuous branch lengths. A geometric embedding, is used to construct an applicable state space for sampling and adaptation on the structure among discrete variables. In order to tackle the low efficiency of adapting large covariance matrix, algorithms are equipped with a class of Adapted Increasingly Rarely Markov Chain Monte Carlo. In order for maintaining propriety of posterior, we provide some conditions on prior distributions of the mutation rate and corresponding bounds of tempering schedules in parallel tempering algorithms. Another challenge we identify and prove rigorously is that the random walk Metropolis type algorithms on coalescent models will not be geometrically ergodic under the standard choice of prior distributions. A practical remedy is based on preconditioned Crank–Nicolson algorithms.

Friday, the 7th June 2024

Alex Browning (University of Oxford) - “Little data”in mathematical oncology (09:30-10:00)

Abstract. Complicating the effective use of mathematical models in cancer research is the quantity and quality of available data. In the clinic, data collection is inherently invasive and often a low priority. Meanwhile, in the lab, high-throughout experimental designs come at the cost of granularity. These limitations are juxtaposed with the complexity and heterogeneity of most cancers, and the corresponding level of intricacy required by models to answer important biological questions. Through two pieces of work, I present and discuss novel approaches and future directions to tackle the problem of ”little data” in mathematical oncology in the context of both prediction and drawing biological insight. First, by developing a minimal mathematical model and simple statistical learning algorithm to draw clinical predictions of tumour volume in head-and-neck cancer patients. Secondly, by exploiting a hierarchy of models to draw insights from otherwise non-identifiable mechanistic models using simple, albeit overwhelming common, experimental data of tumour growth.

Helena Coggan (University College London) - An agent-based modelling framework to study cell plasticity in non-small cell lung cancer (10:00-10:30)

Abstract. Recent breakthroughs in phylogenetic analysis of bulk tissue have allowed researchers to reconstruct the evolutionary histories of tumours. The hope is that knowledge of the order and timing of genetic mutations will allow us to characterise the properties of early-stage cancer cells, to better identify and target them. This is complicated by the phenomenon of tumour plasticity, the ability of cells to acquire the hallmarks of cancer via mechanisms other than heritable genetic mutation. In a tumour with only genetic heritability, changes in cell phenotype would occur only as a result of genetic mutations. By contrast, non-genetic heritability could lead to a disconnect between genotype and phenotype. Whilst both mechanisms are likely to impact on cancer development, we ask whether one has a significantly stronger influence than the other by simulating tumour growth under each regime and comparing the distribution of mutations obtained in either case to patient data.

Here, we address this question in a real patient cohort by adapting a mathematical model of lung tumour evolution, which we use to test the validity of two scenarios. In a ‘low-plasticity’ scenario, changes to reproductive fitness occur when cells acquire mutations in the exome. Mutations may be ‘drivers’ (beneficial) or ‘passengers’ (deleterious or neutral). In a ‘high-plasticity’ scenario, mutations have no effect on cell fitness. Cells experience ‘driver-like’ or ‘passenger-like’ cell fitness changes on division, without leaving a genetic mark. The model describes three-dimensional growth of a tumour from a single cell and incorporates biologically-informed death patterns and local competition for space and resources. When the simulation has reached a realistic size, cells on the surface are sampled and sequenced to predict the relatedness of mutations present at detectable frequencies in each of several regions. The outputs are designed to allow comparison with those of the TRACERx cohort of 421 non-small- cell lung cancer patients (NSCLC), comprising multi-region whole-exome (WXS) and bulk RNA sequencing.

In this ongoing work, we show results from a large cohort of simulations under both scenarios and predict corresponding patterns of genetic similarity. This allows us to use approximate Bayesian computation (ABC) to predict the mechanisms at play in the TRACERx cohort of 421 non-small-cell lung cancer patients (NSCLC). We present a novel 'meta-inference' approach, where evolutionary parameters are fit to each patient's data using well-chosen summary statistics. Given the size of the TRACERx cohort, this enables the evaluation of each scenario by examining the plausibility and similarity of output parameters obtained across patients. We hope that this work will shed light on the role of heritability in lung cancer development and guide future research into therapeutic approaches.

Qiquan Wang (Imperial College London) - A Topological Gaussian Mixture Model for Bone Marrow Morphology in Leukaemia (10:30-11:00)
Abstract. Acute Myeloid Leukaemia (AML), a cancer of the blood and bone marrow, is characterized by the proliferation of abnormal clonal haematopoietic cells in the bone marrow leading to bone marrow failure and death, if untreated. Angiogenic factors released by leukemic cells drastically alter the bone marrow vascular niches in support the prolonged proliferation and suppression of normal haematopoiesis. In addition to increased angiogenesis, increased vascular permeability and various structural abnormalities are observed as part of the alteration as the disease progresses.

Snapshots of the bone marrow vasculature obtained through confocal microscopy at different stages of AML progression capture the combinations of morphological features specific to the particular stages of development. In this work, we use Persistent Homology (PH), an integral tool in topological data analysis (TDA), to quantify the images and infer on the disease through the imaging morphological features. Analysis using PH uncovers succinct dissimilarities between the early and late stages of AML development. We then construct a class of stage-dependent Gaussian Mixture Models (GMMs) applied to PH summaries to both infer patterns in morphological change between different stages of progression as well as provide a basis for prediction.

Guglielmo Gattiglio (University of Warwick) - Nearest Neighbor GParareal: Improving Scalability of Gaussian Processes for Parallel-in-Time Solvers (12:00-12:30)
Abstract. With the advent of supercomputers, multi-processor environments and parallel-in-time (PiT) algorithms provide ways to integrate ordinary differential equations (ODEs) over long time intervals, a task often unfeasible with sequential time-stepping solvers within realistic timeframes. This is particularly evident in the simulation of molecular dynamics, demanding computation over extensive trajectories. A recent approach, GParareal, combines machine learning (Gaussian Processes) with traditional PiT methodology (Parareal) to achieve faster parallel speed-ups. Unfortunately, the applicability of the model is limited to a small number of computer cores and ODE dimensions. We present Nearest Neighbor GParareal (NN-GParareal), a data-enriched parallel-in-time integration algorithm that builds upon its predecessor, GParareal. NN-GParareal improves the numerical stability and scalability properties of GParareal for high-dimensional systems and moderate sample

sizes. Moreover, the cost complexity is reduced from cubic in the sample size to loglinear, yielding a fast, automated procedure to integrate initial value problems over long intervals. The practical utility of NN-GParareal is demonstrated theoretically and empirically through its evaluations on nine different systems. Our analysis offers tangible evidence of NN-GParareal's behavior, advantages, and validity.

Andonis Gerardos (AMU) - MiSFI, a robust algorithm to select a minimal model for dynamical data with large sampling intervals (12:30-13:00)
Abstract. Biological systems, from individual proteins to groups of animals, have noisy dynamics. Their dynamics are frequently acquired with measurement noise and large time intervals. Thus, finding and learning a model, that reveals the key mechanisms, is a hard task. Often, adding parameters to a model is tempting, but it could rapidly lead to overfitting. How can one avoid that? To tackle this challenge, we develop an algorithm for systems that can be modeled by a Stochastic Differential Equation (SDE). Our algorithm selects, among a class of SDE models, the one that best captures the dynamics with the minimum number of parameters. We successfully benchmarked our algorithm against others on synthetic data (Lorenz, Lotka-Volterra model) with dynamical noise, measurement noise and large time intervals. Looking ahead, as an example, we envision one potential application to ecology systems modeled by the Lotka-Volterra equation to learn the interaction between species.

Heba Sailem (King's College London) - Deep learning approaches for identifying predictive biomarkers from the tumour microenvironment (14:00-14:30)
Abstract. Tumour microenvironment plays a critical role in cancer progression and resistance. Histopathology and cellular imaging approaches allow for capturing the spatial organisation of different cell types in the tumour microenvironment and their potential interactions. While convolutional neural networks demonstrated great performance in classifying histopathological datasets, they are difficult to interpret. We evaluated several deep learning architectures for histopathology data including weakly supervised and vision transformer approaches. We demonstrate how graph neural networks allow creating explainable deep learning models that provide insights into the diversity and complexity of the tumour microenvironment.

Hong Ge (University of Cambridge) - TBC (14:30-15:00)
Abstract.