List of Projects 2019/20

MSc in Complex Systems Modelling

PROJECT 1: The role of Kemeny constant in Markov State Models

SUPERVISOR: A. Annibale

PROJECT DESCRIPTION:
Markov State Models are widely used models for investigating kinetic networks and interpreting data of molecular simulations in different applications. However, their dimensionality is typically very large and makes them prohibitively expensive to work with, hence coarse graining methods have been introduced to reduce their dimensionality, capturing in particular the slowest kinetic processes [1].

In this project we aim to derive exact relations for a clusting protocol of the Smoluchowski process (i.e. diffusion in a one-dimensional potential) that entails minimization of the Kemeny constant. Kemeny constant has attracted a large interest, over the years, since its introduction in 1960 by Kemeny and Snell, and it represents the sum of all relaxation timescales in a kinetic network or Markov chain. Remarkably, the Kemeny constant is also equivalent with the weighted sum of all mean first passage times from a selected state i to all other states j, where the weights are the equilibrium populations of states j. Surprisingly, the Kemeny constant is independent on the starting state i. This intriguing constancy has been the subject of several studies [2,3]. The Kemeny constant has also attracted considerable interest in the field of graph theory and networks science.

It was recently shown in Ref. [4] that the relaxation time in a coarse-grained system is always smaller than or equal the relaxation time in the original system and that there exists a relation between the relaxation time of a coarse-grained Smoluchowski process and the mean first passage times to the boundaries of its clusters, when starting from inside the clusters. Hence, the optimal position of the boundaries was determined by maximizing the relaxation time. The analysis was limited to two clusters and three clusters with symmetric potentials. This project can take multiple directions:

(a) One can first start by re-producing the results from Ref. [4]. Then, one can derive analogous relations for the Kemeny constant and find the optimal position, of the boundaries, by minimizing the Kemeny constant, as opposed to the relaxation time. In particular, it will be interesting to look at the three-state clustering of a double well potential and compare with results in [4]. Generalizations to non-symmetric potentials for the three-state clustering or to four-state clustering with symmetric potential may also be considered.

(b) A second direction for the project would be to consider network processes (e.g. diffusion on networks) instead of Smoluchowski processes. A non-exhaustive list of tasks that could be undertaken here includes:

- Studying the impact of network structure on Kemeny constant.
- Use the framework derived recently in [5] to infer the rates along the links of the given networks, from mean first passage times. Alternatively, infer the network structure from mean first passage times, assuming a diffusion process. Propose measures for node centrality based on MFPTs and compare with other measures that have been proposed in the literature.
- Define a suitable 1-dimensional projection of the network, cluster according to the protocols defined for the Smoluchowski process, check properties of the nodes at the boundaries, like betweenness centrality or closeness centrality. For a review on network clustering see [6].

(c) A third, completely different, direction would be to apply the results recently derived in [5] in milestoning techniques [7], aimed at sampling rare events efficiently. In particular, [5] provides a recipe to infer the equilibrium distribution of a system from MFPTs. One may envisage to run suitably biased simulations to efficiently compute MFPTs and then unbias the results.

PREREQUISITES:
Familiarity with 7CCMCS02 and 7CCMCS04 and programming skills.

REFERENCES:

1 G. Hummer, A. Szabo, Optimal Dimensionality Reduction of Multistate Kinetic and Markov-State Models, J. Phys. Chem. B 2015, 119, 29, 9029-9037

2 Peter G. Doyle, The Kemeny constant of a Markov chain, preprint arXiv:0909.2636

3 D. Bini, J. Hunter, G. Latouche, B. Meini, P. Taylor (2018). Why is Kemeny's constant a constant? Journal of Applied Probability, 55(4), 1025-1036. doi:10.1017/jpr.2018.68

4 A. Kells, S. Mihálka, A. Annibale, E. Rosta, Mean first passage times in variational coarse graining using Markov state models, J. Chem. Phys. 150, 134107 (2019)

5 A Kells, E Rosta, A Annibale, Correlation functions, mean first passage times and the Kemeny constant (2019) preprint arXiv:1911.01729

6 S. Fortunato, Community Detection in Graphs, Physics Reports Volume 486, Issues 3–5, February 2010, Pages 75-174.

7 A. M. Berezhkovskii and A. Szabo, Committors, first-passage times, fluxes, Markov states, milestones, and all that, J. Chem. Phys. 150, 054106 (2019); https://doi.org/10.1063/1.5079742

PROJECT 2: Neural networks and bipartite graphs models for gene expression

SUPERVISOR: A. Annibale

PROJECT DESCRIPTION:
Cellular differentiation has for long been thought to be an irreversible process, whereby an embryonic stem cell evolves into a variety of somatic cell types. However, recent experiments have shown that nearly all human cell types can be reprogrammed to a stem cell state, by using four transcription factors (TFs), known as the Yamanaka factors. Cell types have for long been regarded as stable attractors of a high-dimensional gene expression dynamics, similarly to 'memories' stored in a neural network. The analogy between neural networks (NN) and gene-regulatory networks (GRN) has been known for decades, but it has only started being exploited recently [1,2]. However, there are also important differences between NN and GNR models. While NN models have usually dense and symmetric (Hopfield-like) interactions [3,4,5,6], GRNs have sparse and directed interactions. Recently, sparsity in neural networks patterns has been shown to lead to parallel retrieval [7,8]. However, parallel retrieval would be an undesired feature in gene-regulatory networks. In addition, NN models are mostly formulated with two-body interactions, while gene-regulatory networks have higher-order interactions. Recently, a bipartite graph model for gene regulation, comprising genes and TFs, has been fomulated, which takes into account both the sparsity and the high-order nature of gene interactions [9]. However, no multiplicity of attractors is embedded in this model.

This project combines analytical calculations and numerical work to better understand how GRNs process information. The project can go in several directions. One consists in further developing the model introduced in [1], to incorporate the mechanism that induces transitions between pluripotent and fully differentiated cells in developmental biology, and, possibly, sparsity in the interactions. Another direction is to suitably modify the choice of interactions in model [9] to embed a multiplicity of attractors, corresponding to stable cell-types. A further direction would be to use gene knock-out experiments data and other datasets to infer the interations and calibrate model [9].

PREREQUISITES:
7CCMCS03 and ideally 7CCMCS04. Programming and numerical skills are an advantage.

REFERENCES:

1 R. Hannam, A. Annibale, R. Kühn (2017). Cell reprogramming modelled as transitions in a hierarchy of cell cycles. *J. Phys. A: Math. Theor.* 50 425601.

2 S. Huang, G. Eichler, Y. Bar-Yam, D.E. Ingber (2005). Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* 87(87).

3 Hopfield, Proc. Natl. Acud. Sci. USA, 79 (1982) 2554; 81 (1984) 3088.

4 D.J. Amit, H. Gutfreund and H. Sompolinsky (1985) *Phys. Rev. Lett.* 55 1530; H. Sompolinsky, I. Kanter (1986), *Phys. Rev. Lett.* 57, 2861.

5 N. Parga, M.A. Virasoro (1986). The ultrametric organization of memories in a neural network. *Journal de Physique*, 47 (11), pp.1857-1864.

6 S. Bos, R. Kuhn and J. L. van Hemmen (1988), *Z. Phys. B*, 71 261.

7 E Agliari, A Annibale, A Barra, ACC Coolen, D Tantari (2013) Immune networks: multi-tasking capabilities at medium load. J. Phys. A: Math. Theor: 46 (33).

8 P. Sollich, D Tantari, A Annibale, A Barra (2014). Extensive Parallel Processing on Scale-Free Networks. *Phys. Rev. Lett.* 113 (23), 238106.

9 R Hannam, R, Kuhn, A Annibale Sustaining like: Percolation in Gene Regulatory Networks (2019) J. Phys. A: Math. Theor. 52 334002

PROJECT 3: Comparing Landscapes: Deep Neural Networks versus Glassy Systems

SUPERVISOR: C. Cammarota

PROJECT DESCRIPTION:
The training process of a deep neural network (DNN) shares strong similarities with the physical dynamics of disordered systems: the loss function plays the role of the energy, the weights are the degrees of freedom, and the dataset corresponds to the parameters defining the energy function. This project is aimed at studing this similarity by using toy models of DNN along the lines traced in [1]. The plan is to use a toy dynamics as close as possible to the physical dynamics of disordered systems, for the training of DNN to reveal the similarities between the landscape of loss functions of DNNs and the one of the energy of paradigmatic disordered systems.

PREREQUISITES:
Numerical skills in any programming language, Path integrals, Langevin dynamics

REFERENCES:

1 M.Baity-Jesi, L.Sagun, M.Geiger, S.Spigler, G.Ben Arous, C.Cammarota, Y.LeCun, M.Wyart, G.Biroli, 'Comparing Dynamics: Deep Neural Networks versus Glassy Systems' Proceedings of the 35th International Conference on Machine Learning, PMLR 80, 314 (2018) http://proceedings.mlr.press/v80/baity-jesi18a.html

2 L.F.Cugliandolo, J.Kurchan, 'Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model' PRL 71, 173 (1993) https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.71.173

3 C.D.Freeman and J.Bruna, 'Topology and geometry of deep rectified network optimization landscapes' preprint on arXiv:1611.01540 (2016) https://arxiv.org/abs/1611.01540

4 F. Draxler, K. Veschgini, M. Salmhofer, F. Hamprecht 'Essentially No Barriers in Neural Network Energy Landscape' Proceedings of the 35th International Conference on Machine Learning, PMLR 80:1309-1318, 2018

5 A. Barrat, M. Mézard. Phase Space Diffusion and Low Temperature Aging. Journal de Physique I, EDP Sciences, 1995, 5 (8), pp.941-947. 10.1051/jp1:1995174 . jpa-00247118

PROJECT 4: A spin glass model for ecosystems with a large number of interacting species.

SUPERVISOR: C. Cammarota

PROJECT DESCRIPTION:
The debate about whether ecosystems (described as ensembles of large number of species living in the same environment and interacting constructively/destructively with each others) are dominated by stochastic behaviour (neutral description) or by competition/selection rules (niche description) is a long standing one. Nice recent results suggest that these two descriptions could represent the two regimes (phases) of the same ecosystem alternatively emerging when a few important ecological parameters are changed. This project is aimed at developing a detailed analysis of this behaviour by the study of a disordered model, called presence/absence model, that provides a schematic description of the ecological systems. The analogy of the presence/absence model of ecological inspiration with more traditional and better known spin glass models is striking and will be exploited to obtain new results on the ecological neutral/niche transition through the analysis of the outcomes from old studies of pure spin glass paradigms and the extension to the ecological problem of analytical tools developed for spin glasses (involving the use of replica analysis).

PREREQUISITES:
Familiarity with Gaussian integrals and saddle point integration. It would be useful to have seen at least once the replica trick and its use before the project starts.

REFERENCES:

1 Charles K. Fisher and Pankaj Mehta (2014) 'The transition between the niche and neutral regimes in ecology' Proceedings of the National Academy of Sciences 111 (36) 13111 with Supporting Information http://www.pnas.org/content/111/36/13111.full.pdf?with-ds=yes

2 R. F. Soares, F. D. Nobre, and J.R.L.de Almeida (1994) 'Effects of a Gaussian random field in the Sherrington-Kirkpatrick spin glass' Physical Review B 50 (9) 6151

3 M.Mezard, G.Parisi, M.Virasoro 'Spin glass theory and beyond' World Scientific Lecture Notes in Physics Vol.9 (1987)

4 Rosindell J, Hubbell SP, He F, Harmon LJ, Etienne RS (2012) The case for ecological neutral theory. Trends Ecol Evol 27(4):203-208

5 Macarthur R, Levins R (1967) The limiting similarity, convergence, and divergence of coexisting species. Am Nat 101:377-385

PROJECT 5 Spherical spin glass in the magnetic field and the simplest random optimization problems.

SUPERVISOR: Yan Fyodorov

PROJECT DESCRIPTION:
Finding the global minimum:

$$E_{min} = \min_{|\mathbf{x}|=R} \{E_h(\mathbf{x})\}\,,\ E_h(\mathbf{x}) = -\frac{1}{2}\,(\mathbf{x}, H\mathbf{x}) - (\mathbf{h}, \mathbf{x})\,, \quad \mathbf{h}, \mathbf{x} \in \mathbb{R}_N,$$

of a function given by the sum of a quadratic and a linear form in N real variables over $N-1$-dimensional sphere $\mathbb{S}^{N-1}$ is one of the simplest, yet paradigmatic problems in Optimization Theory known as the "trust region subproblem" [1]. When both H and h in the cost function are random Gaussian this amounts to studying the lowest energy of the simplest spherical spin glass [2] in a random magnetic field [3,4,5]. This can be studied both statically and dynamically.

Another closely related problem is to understand as much as possible the long-time dynamics of the following model:

$$\frac{dx_j}{dt} = -\left(\mu + \alpha\frac{\mathbf{x}^2}{N}\right) x_j + h_k + \sum_{k=1}^{N} J_{jk}\, x_k + \xi_i(t),, \quad j = 1, \ldots, N\,, \tag{1}$$

where the parameters $\mu > 0$ and $\alpha > 0$ ensure global stability of the system due to presence of the term $\mathbf{x}^2 = \sum_{i=1}^{N} x_i^2$, growing at infinity, whereas $\xi_i(t)$ is the Gaussian white noise: $\langle \xi_i(t_1)\xi_j(t_2)\rangle = 2T\delta(t_1 - t_2)$. The coupling matrix is symmetric $J_{jk} = J_{kj}$, and otherwise chosen to be random Gaussian variables with mean zero and variance $< J_{jk}^2 >= \frac{J^2}{N}(1 + \delta_{jk})$. The field components h_k can be chosen Gaussian random as well. The model resembles in many respects the spherical $p = 2$ spin-glass model studied in [4,5,6,7], but the energy landscape controlling the dynamics is somewhat different, and needs to be investigated as a function of μ, α and J, together with analysis of statics/thermodynamics which is expected to describe the equilibrium properties of the model. The goal of the project will be studying various aspects of those problems both analytically and numerically, using results from random matrices, replica trick, etc. and comparing the results.

PREREQUISITES:
Good analytical and numerical skills, understanding of the probability concepts (probability density, Gaussian distribution in many variables, etc., saddle-point method, mean-field models).

REFERENCES:

1 Hager, W.W. Minimizing a quadratic over a sphere. *SIAM J. Optim.* **12**(1), 188—208 (2001)

2 Fyodorov Y V and Le Doussal P. Topology trivialization and large deviations for the minimum in the simplest random optimization. *J. Stat. Phys.* **154** 466—90 (2014)

3 Kosterlitz J M, Thouless D J and Jones R C. Spherical model of a spin-glass. *Phys. Rev. Lett.* **36** 1217 (1976)

4 Cugliandolo L F and Dean D S. On the dynamics of a spherical spin-glass in a magnetic field. *J. Phys. A: Math. Gen.* **28** (1995) L453–L459.

5 Cugliandolo L. F., Dean D. S., and Yoshino H. Nonlinear Susceptibilities of spherical models. *J. Phys. A: Math. Gen.* **40**, 4285–4303 (2007)

6 Cugliandolo L. F. and Dean D. S. Full dynamical solution for a spherical spin-glass model. *J. Phys. A: Math. Gen.* **28** 4213—34 (1995)

7 De Dominicis C. and Giardina I. chapter 4 in the book: "Random Fields and Spin Glasses".

PROJECT 6: Signal Reconstruction, Random Matrices and Replica Method.

SUPERVISOR: Yan Fyodorov

PROJECT DESCRIPTION:
A paradigmatic problem in the communication theory is the reconstruction of a source signal from its encrypted form corrupted by an additive noise when passed from a sender to a recipient. Signals are represented by $N-$dimensional source (column) vectors $\mathbf{s} = \begin{pmatrix} s_1 \\ \dots \\ s_N \end{pmatrix} \in \mathbb{R}^N$. The encryption is a transformation $\mathbf{s} \mapsto \mathbf{y} = (y_1, \dots, y_M)^T \in \mathbb{R}^M$ known both to the sender and a recipient. In the simplest case of *random linear* encryption such transformation is linear: $\mathbf{y} = A\mathbf{s}$ where A is a $N \times M$ matrix. Due to imperfect communication channels the recipients however get access to the encrypted signals only in a corrupted form $\mathbf{z}$. In the simplest corruption mechanism the encrypted images $\mathbf{y}$ are corrupted by an additive random noise, i.e. $\mathbf{z} = \mathbf{y} + \mathbf{b}$. The noise vectors $\mathbf{b}$ are further assumed to be normally distributed: $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{1}_M)$, i.e. components b_k, $k = 1, \dots, M$ are i.i.d. mean zero real Gaussian variables with the covariance $\langle b_k b_l \rangle = \delta_{kl}\sigma^2$, where the notation $\langle \dots \rangle$ stands for the expected value $\mathbb{E}[\dots]$ with respect to all types of random variables. The recipient's aim is to reconstruct the source signal $\mathbf{s}$ from the knowledge of $\mathbf{z}$. In the presence of noise such reconstruction can be only approximate, and reconstructed signals are known in the signal processing literature as 'estimates' of the source signals. Their properties depend on the reconstruction scheme (or the 'estimator') used. The strength of the noise, i.e. the parameter σ^2, may or may not be known to the recipient. We consider the input signal through the reconstruction procedure as a *fixed* vector, and then employ the Least-Square reconstruction scheme, which for a given set of observations $z_k = A_k(\mathbf{s}) + b_k$ returns an estimate of the input signal as

$$\mathbf{x} := Argmin_{\mathbf{w}} \frac{1}{2}||\mathbf{z} - A\mathbf{w}|^2 + \lambda||\mathbf{w}||^2 \tag{2}$$

where λ is a tuning parameter and $||\mathbf{x}|| = \sqrt{(\mathbf{x}, \mathbf{x})}$ is the Euclidean L_2 norm of the vector $\mathbf{x}$. Quality of the signal reconstruction under our scheme is then characterized by the value of a distortion parameter measuring the difference between the fixed source signal $\mathbf{s}$ and the estimate $\mathbf{x}$. For this one may use the Euclidean distance normalized to the signal strength: $d(\mathbf{x}; \mathbf{s}) = \frac{(\mathbf{x}-\mathbf{s}, \mathbf{x}-\mathbf{s})}{(\mathbf{s}, \mathbf{s})}$. Our goal will be to characterize $d(\mathbf{x}; \mathbf{s})$ statistically assuming that A is a random Gaussian matrix, most importantly when the sizes $N.M$ are big: $N \to \infty, M \to \infty$ and $\mu = M/N > 1$ is fixed. For this we will use both the techniques of Random Matrix theory and methods of statistical mechanics, the so-called replica trick.

Similar problems can be formulated assuming that the signal norm $||\mathbf{s}||$ is known, so the signal recovery is confined to a sphere. Closely related problems appear in statistics and has been addressed very recently in [2].

PREREQUISITES:
Good analytical and numerical skills, understanding of the probability concepts (probability density, Gaussian distribution in many variables, etc.)

REFERENCES:
1 Y.V. Fyodorov. A spin glass model for reconstructing nonlinearly encrypted signals corrupted by noise. *J Stat. Phys.*, **175**(5), 789–818(2019)

2 Fyodorov YV, Tublin R. Counting stationary points of the loss function in the simplest constrained least-square optimization. *arXiv:1911.12452*

3 Bereihi A., Müller R., Schulz-Baldes H. Replica symmetry breaking in compressive sensing. *arXiv:1704.08013*.

PROJECT 7: Jam formation in traffic processes through networks

SUPERVISOR: I. Neri

PROJECT DESCRIPTION:
Traffic jams are a nuissance: in 2017 the economic cost of traffic jams in London was more than £9.5 billion [1]. Traffic jams occur because vehicles cannot overtake. They are a generic feature of active transport processes with excluded volume interactions and appear in different contexts. For example, traffic jams of motor proteins along molecular high ways disrupt the fast intracellular transport of cargoes, which is necessary for cellular signalling processes, and are associated with neurodegenerative diseases.

How network architecture influences the formation of traffic jams is poorly understood. If we reconnect roads in a network, can we reduce the total amount of traffic jams and their negative impact on transport? In this project you will address this question through the study of a stochastic transport process through a network. To successfully complete this project, you will need to follow four steps: (1) construct a model for a transport network (2) choose a transport model that exhibits traffic jams (3) develop a measure that quantifies the influence of traffic jams on transport through a network (4) calculate how this measure depends on the parameters of the network and the transport process. See background references [2-7] and the review papers [8,9] for more information.

PREREQUISITES:
Completion of the project requires good knowledge of CS02. Basic programming skills and familiarity with the theory of Markov processes are an advantage.

REFERENCES:

1 http://inrix.com

2 K Nagel, M Schreckenberg, *A cellular automaton model for freeway traffic*, Journal de Physique I **2**, (1992).

3 B S Kerner, P Konhauser, *Cluster effect in initially homogeneous traffic flow*, Physical Review E **48**, R2335 (1993).

4 A Schadschneider, *The Nagel-Schreckenberg model revisited*, The European Physical Journal B-Condensed Matter and Complex Systems **10**, 573-582 (1999).

5 I Neri, N Kern, A Parmeggiani, *Totally asymmetric simple exclusion process on networks*, Physical Review Letters **107**, 068702 (2011).

6 D V Denisov, D M Miedema, B Nienhuis, P Schall, *Totally asymmetric simple exclusion process simulations of molecular motor transport on random networks with asymmetric exit rates*, Physical Review E **92**, 052714 (2015).

7 L Zhang, G Zeng, D Li, H J Huang, H E Stanley, S Havling, *Scale-free resilience of real traffic jams*, Proceedings of the National Academy of Sciences **116**, 8673-8678 (2019).

8 D Helbing, *Traffic and related self-driven many-particle systems*, Review of Modern Physics **73**, 1067 (2001).

9 M Barthelemy, *Spatial networks*, Physics Reports **499**, 1-101 (2011).

PROJECT 8: Spectra of large networks

SUPERVISOR: I. Neri

PROJECT DESCRIPTION:
Spectra of large networks reveal important information about the processes governed through them. For example, the stability of the dynamics of a system to an initial perturbation is determined by the eigenvalue with the largest real part of the associated adjacency matrix, and the nature of the destabilising mode is determined by the associated eigenvector. The purpose of this project is (i) to determine the spectra of adjacency matrices of networks that characterise a real-world system of choice and (ii) to use the mathematical methods from sparse random-matrix theory to understand the features observed in the empirically obtained spectra. For example, you could gather data about an ecological foodweb and subsequently compare its empirical spectrum with the one predicted with mean-field methods. Finally, you are expected to critically evaluate the discrepancies observed between theory and experiment and draw some conclusions on how the spectral features characterise the system dynamics. Relevant background references are [1-7].

PREREQUISITES:
Completion of the project requires good knowledge of CS02. Basic programming skills are an advantage.

REFERENCES:

1 G J Rodgers, A J Bray, *Density of states of a sparse random matrix*, Phys. Rev. B **37**, 3558 (1988).

2 R Kühn, *Spectra of sparse random matrices*, Journal of Physics A: Mathematical and Theoretical **41**, 295002 (2008).

3 T Rogers, I P Castillo, R Kühn, K Takeda, *Cavity approach to the spectral density of sparse symmetric random matrices*, Physical Review E **78**, 031116 (2008).

4 F L Metz, I Neri, D Bollé, *Spectra of sparse regular graphs with loops*, Physical Review E **84**, 055101 (2011).

5 S Allesina, J Grilli, G Barabás, S Tang, J Aljadeff, A Maritan, *Predicting the stability of large structured food webs*, Nature communications **6**, 7842 (2015).

6 J Grilli, T Rogers, S Allesina, *Modularity and stability in ecological communities*, Nature communications **7**, 12031 (2016).

7 G T Cantwell M E J Newman, *Message passing on networks with loops*, Proceedings of the National Academy of Sciences **116**, 23398 (2019).

PROJECT 9: Stochastic dependence between manifold-valued data

SUPERVISOR: D. Pigoli

PROJECT DESCRIPTION:
The investigation of the relationship between two (or more) variables is at the very core of statistics as a discipline. In classical univariate or multivariate statistics, a well-established measure for this relationship is the covariance, whose definition relies on the inner product between the two random variables. However, this concept does not extend in a straightforward way to model the relationship between variables that are more complex than numbers or vectors, such as images and curves or particularly manifold-valued data like points on a sphere, shapes, positive (or negative) definite matrices, etc. The general aim of the project consists in evaluating how other existing measures of dependence can be used in the estimation of the dependence between two samples of manifold-valued data and/or between a sample of manifold-valued data and a sample of univariate data. In recent years, metrics have been proposed either based on Riemannian distance ([1],[2]) or on local tangent approximations ([3],[4]). The former can be defined globally on the manifold but it is not always clear what are they telling us about the stochastic dependence. The latter are locally well-defined and easily interpretable but the local approximation may not be accurate if the dispersion on the manifold is large. A third approach is to embed the manifolds in a larger Euclidean space ([5]). The project will explore advantages and disadvantages of these approaches in theory and/or with simulated examples, considering a few examples of manifolds relevant for applications, such as positive-definite matrices, shapes/hyperspheres, probability distributions or graphs.

PREREQUISITES:
Basic knowledge of manifolds and probability theory. Programming skills to run simulation studies.

REFERENCES:

1 Lyons, R. (2013) Distance covariance in metric spaces. The Annals of Probability, 41:3284-3305. Arxiv url: https://arxiv.org/abs/1106.5758

2 Dubey, P., & Mueller, H. G. (2019). Functional models for time-varying random objects. Arxiv url: https://arxiv.org/abs/1907.10829

3 Zhang, M., & Fletcher, T. (2013). Probabilistic principal geodesic analysis. In Advances in Neural Information Processing Systems (pp. 1178-1186).

4 Pigoli, D., Menafoglio, A., & Secchi, P. (2016). Kriging prediction for manifold-valued random fields. Journal of Multivariate Analysis, 145, 117-131., open access url: https://www.repository.cam.ac.uk/handle/1810/253013

5 Lin, L., St. Thomas, B., Zhu, H., & Dunson, D. B. (2017). Extrinsic local regression on manifold-valued data. Journal of the American Statistical Association, 112(519), 1261-1273. Arxiv url: https://arxiv.org/abs/1508.02201

PROJECT 10: Bagging and aggregation for prediction of spatial data

SUPERVISOR: D. Pigoli

PROJECT DESCRIPTION:
The statistical analysis of non-stationary data (i.e. data whose distribution properties are not homogeneous) that are observed over spatial domains with complex topology (such as aquifer basins, islands, etc...) poses relevant challenges in many geophysical applications. A useful approach to prediction in this setting is the one based on random domain decomposition, where data are randomly partitioned in small neighbourhoods (or tiles) where local (possibly linear) prediction is used. To overcome the arbitrariness of the partition and the discontinuities that are introduced at the boundaries between neighbourhoods, multiple partitions are randomly sampled and the predictions are then aggregated via (possibly weighted) averaging. The aim of the project is to analyse the statistical properties of these procedures either in theory or via simulation studies and explore the connections with the existing bagging and aggregation procedures for independent data on the one hand and with the mixture of experts approach in the machine learning literature on the other hand.

PREREQUISITES:
Some programming skills in any language, notions of probability theory. Knowledge of prediction methods (linear and nonlinear regression) is advantageous but not necessary.

REFERENCES:

1 Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. The Annals of Applied Statistics, 2(3), 916-954.

2 Robert B. Gramacy and Herbert K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. Journal of the American Statistical Association, 103(483):1119– 1130, 2008.

3 Alessandra Menafoglio, Giorgia Gaetani, and Piercesare Secchi. Random domain decompositions for object-oriented kriging over complex domains. Stochastic Environmental Research and Risk Assessment, Aug 2018.

4 Carl E. Rasmussen and Zoubin Ghahramani. Infinite mixtures of Gaussian process experts. In Advances in Neural Information Processing Systems 14, pages 881–888. MIT Press, 2002.

PROJECT 11: Metropolis proposals for large data

SUPERVISOR: Michael Pitt

PROJECT DESCRIPTION:
The aim of this project is to understand two papers using Langevin proposals for statistical inference, i.e. [1] and [2]. These papers examine what happens as the dimension of the parameters, d, becomes large. However, researchers have not investigated either empirically or analytically what happens as the size of the data increases at a faster rate than d. The project would explore these open questions.

PREREQUISITES:
This project requires programmming skills in C or matlab.

REFERENCES:

1 G.O. Roberts and J.S. Rosenthal (1998), Optimal scaling of discrete approximations to Langevin diffusions, *J. Royal Soc. Series B* **60**(1): 255–268.

2 M Girolami and B Calderhead (2011), Riemann manifold Langevin and Hamiltonian Monte Carlo methods, *J. Royal Soc. Series B* **73**(2): 123–214.

PROJECT 12: The estimation and filtering for the Stochastic Volatility model

SUPERVISOR: Michael Pitt

PROJECT DESCRIPTION:
The stochastic volatility model (see [1]) is commonly used to model the changing variance of stock returns. Whilst the statistical properties of the model are easy to establish, the estimation and filtering is more involved. The exploration of Markov chain monte Carlo is important, see [2]. Additionally, the exploration of different methods for filtering in discrete time (see [3]) and in continuous time (see [4]) can be explored.

PREREQUISITES:
This project requires programmming skills in C or matlab.

REFERENCES:

1 A. C. Harvey, Time Series Models, Harvester Wheatsheaf, 1993.

2 N Shephard and M. K. Pitt (1997), Likelihood analysis of non-Gaussian measurement time series, *Biometrika* **84**, 653-667.

3 M.K. Pitt and N Shephard (1999), Filtering via simulation: auxiliary particle filters *JASA*, **94**(446): 590-599

4 S Malik and M.K. Pitt (2011), Particle filters for continuous likelihood evaluation and maximisation, *J. of Econometrics* **165**(2):190-209.

PROJECT 13: Design of experiments for estimating parameters in differential equations

SUPERVISOR: S. Gilmour

PROJECT DESCRIPTION:
Different methods are used for estimating the parameters of differential equations from data collected from processes governed by these equations. One approach, based on generalized smoothing, has become popular in some areas of application, since being introduced in a discussion paper [1] by Ramsay et al. In the discussion of this paper, SG Gilmour pointed out that the success of the parameter estimation depends crucially on how the experiments are designed and he briefly described how the design of such experiments could be approached. This suggestion has not been followed up. The aim of this project is to implement the design of experiments for parameter estimation in differential equation models, based either on the generalized smoothing approach, or any other approach for which we can approximate the covariance matrix of the parameter estimators. Implementation will be for a small experiment, such as that described in [1], or any other that interests the student.

PREREQUISITES:
This project requires some programming skills and a command of some 7CCMCS06T material (linear regression models).

REFERENCES:

1 JO Ramsay, G Hooker, D Campbell and J Cao (2007), Journal of the Royal Statistical Society, Series B **69**(5), 741-796 *Parameter estimation for differential equations: a generalized smoothing approach (with discussion).*