Group projects for SM2 and SC2

Below we propose a few topic for SM2 and SC2 projects. Please discuss your plans with Mathieu Gerber and Matteo Fasiolo to make sure that project plan covers the requirements for SM2 and SC2.

1) Modeling household electricity demand

- *Data*: the Irish household demand data set provided here: https://github.com/mfasiolo/electBook/blob/master/data/Irish.RData. It contains demand and survey data from over 2000 households. The demand is observed half-hourly, so there are 48 observations per day per customer.
- *Research questions*: some interesting things you might do with the data are (note that that these are alternatives, the project could not cover more than one point):
  1. Predict the full (48 dimensional) household demand one day ahead.
  2. Predict, one day ahead, the time and/or size of the daily demand peak (the daily max), for each customer.
  3. Aggregate the customers according to some criteria (e.g. daily demand profile similarity) and predict the demand of these larger groups.
  4. Study the correlation between customers. For example, is the demand of the customers more correlated during weekends or during the working days, how does it change with the time of day, etc?
- *Methods*: to answer questions such as a., b. or c. you can use methods such penalised (ridge) regression, Gaussian processes or similar. You are encouraged to use Bayesian methods to fit such models to data. For a.-b. you will need to deal with several thousand time series, hence it would be advantageous to do model fitting in parallel (for example, using BlueCrystal). For c. you will deal with fewer time series, hence you could consider implementing a fully Bayesian approach based on, e.g., MCMC or HMC sampling. Implementing such samplers in (e.g.) Rcpp or STAN would clearly be advantageous. To study the correlations in d., you might use (e.g.) local or weighted likelihood methods. You will need to be able to handle large matrices or many small matrices, hence using HPC and/or compiled code would speed up computation.
- *Contact*: Matteo Fasiolo.

2) Simulation-based inference for intractable models

- *Description*: there are many practically useful models with intractable likelihoods. The likelihood might be intractable because it requires solving a high-dimensional integral (as is often the case in state space models) or for other reasons. In this project you will consider an intractable model and you will fit it to data by implementing Sequential Monte Carlo (SMC), Approximate Bayesian Computation (ABC) or Synthetic Likelihood (SL) methods. All three methodologies are simulation-based, hence a pure R or Python implementation would be very slow and using compiled code would be highly advantageous. Also, simulations from the model can often be performed independently, hence parallelization via OpenMP or HPC is generally straightforward.
- *Model and data*: some interesting models and data sets that you could consider are:
  - the *pomp* R package contains a wealth of model and data set. For instance, the *bsflu* data frame contains data on an outbreak of influenza in an all-boys boarding school, *LondonYorke* is a data frame containing the monthly number of reported cases of chickenpox, measles, and mumps from two American cities, *ewmeas* and *ewcitmeas* are data frames containing weekly reported cases of measles in England and Wales. See the package documentation for more details. Such data could be modelled via, e.g., a Susceptible Infective Recovered (SIR) model. If you don't want to work on epidemiological data, have a look at the *parus* data frame, for example. Note that you are expected to develop your own code to fit such models, not to use the code already provided by *pomp*.

  - you could consider Zombie epidemics. Fortunately, real data is not available in this case, and you would have to simulate it from your model and then to fit the model to the simulated data. For ideas on potential models see, for example:
    * https://loe.org/images/content/091023/Zombie%20Publication.pdf
    * https://people.maths.ox.ac.uk/maini/PKM%20publications/384.pdf
    * https://arxiv.org/pdf/1503.01104.pdf

- *Methods*: you should aim at fitting the chosen model(s) to data using either SMC methods or approximate methods such as ABC or SL. For an overview in an ecological and epidemiological setting see:
  - Fasiolo, Matteo, Natalya Pya, and Simon N. Wood. "A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology." Statistical Science (2016): 96-118.

- *Contact*: Matteo Fasiolo.

3) State-space model representation of univariate Gaussian process regression

- *Description*: in this project you will use a Gaussian process regression model for online prediction of a stationary time series $(Y_t^0)_{t\geq 1}$.

  More precisely, you will consider a model with $p = 1$ predictor, and use as prior a $\text{GP}(0, k_\gamma)$ where $k_\gamma(x,x') = \exp(-|x-x'|/\gamma)$ is the exponential kernel. You will start by showing that if $f \sim \text{GP}(0, k_\gamma)$ then $(f(t))_{t\geq 1}$ is a homogenous Markov chain (computing its initial distribution and transition). From this result, you will write down the state-space representation of the GP model and implement a Kalman filter for computing exactly the gradient (w.r.t. λ and γ) of the marginal likelihood of the data and a Kalman filter for predicting $(Y_t)_{t\geq 1}$ in on online fashion. In a second step you will apply these two algorithms on a dataset of your choice, using a training set for choosing (λ, γ) with the empirical Bayes approach and a test set to assess the prediction errors.

  Some references:

  - Chapter 11 of Särkkä, Simo, and Arno Solin. "Applied stochastic differential equations. Vol. 10." Cambridge University Press, 2019.
  - Chapter 7 of Chopin, Nicolas, and Omiros Papaspiliopoulos. "An introduction to sequential Monte Carlo." Vol. 4. New York: Springer, 2020.
  - Koopman, Siem Jan, and Neil Shephard. "Exact score for time series models in state space form." Biometrika (1992): 823-826.

- *Contact*: Mathieu Gerber.

4) Sequential Bayesian parameter inference

- *Description*: In this project you will use Sequential Monte Carlo (SMC) to recursively approximate a sequence of posterior distributions. In addition to allow to "see" the evolution of the posterior distribution as the sample size increases, an important advantage of SMC over Markov Chain Monte Carlo methods is that an estimate of the model evidence (on which Bayesian model choice relies) is directly obtained as a by-product of the algorithm.

- *Model*: Sequential Monte Carlo can be applied to approximate the posterior distribution of a parameter in a wide range of statistical models, including state-space models, models with latent variables and Gaussian process regression models. The choice of the model (and of the data) will depend on your own interests, but the model should lead to a non trivial implementation of SMC.

- *Main objectives:* understand and implement SMC on a non trivial model.

- *Contact*: Mathieu Gerber.