**Title: Folkhalsomyndigheten's daily data download**

Folkhalsomyndigheten (FHM) has once again changed the way it provides COVID19 data concerning Sweden. The aim of the project is to implement a (preferably R) method that combines data from FHM's three data structures into a single database (preferably saved as an R file, or collection of csv files). The current way of providing data by FHM seems redundant and one part is to identify the non-redundant information to save.

As FHM backcorrects data, it is important to treat data provided each day/week as a separate entity.

**Supervisor: Krzysztof Bartoszek**

**Title: Arctic microbiome analysis**

Microbiome communities dominate the oceans in terms of diversity and abundance, and play a key role in the global biogeochemical cycle. To date, many studies have focused on characterizing microbiome composition in marine sediments around the world. Unfortunately, data on pelagic microbiome composition in the Arctic Ocean, which is particularly important due to its sensitivity to climate change, remain puzzling. Here, we obtained next-generation reads of microbiome community from the Fram Strait and we clustered all reads into operational taxonomic units (OTUs). Overall, high sequencing coverage was obtained, which allowed to produce a reliable description of the number of OTUs present in the samples and quantitative estimates of microbial community composition. We obtained a large number of singletons and identified many unique OTUs. It has previously been suggested that there is an underestimation of microbial taxonomic diversity in polar seawaters and that 70% of OTUs are unique to the Arctic Ocean. However, despite the development of different statistical algorithms (e.g. to perform phylogenetic analysis), there is no single universal statistical method. It is therefore important to develop suitable framework/scripts to explore different statistical algorithms for differential abundance analysis of microbiome data. All OTUs of microbiome community to develop statistical algorithms will be shared by the Department of Marine Plankton Research, University of Gdansk.

**Supervisor: Krzysztof Bartoszek**

**Title: R package for modelling birds' migratory behaviors from circular orientation cages**

Modelling birds' migratory behaviours from circular orientation cages To study birds' migratory behaviour, so called Busse cages are used. Birds are captured and placed in these cages. Their desired migration directions are deduced from marking left on these cages. The idea of the project is to develop an R software package for proposed models and statistical descriptions (especially in terms of polar coordinates) for this type of data. Measurements will be provided by the Ornithology Unit, Department of Vertebrate Ecology and Zoology, University of Gdansk. A research project from HT22 investigated models for this data and contains some starting R code.

**Supervisor: Krzysztof Bartoszek**

**Title: Cohort simulation study**

Some risks of psychiatric disorders in children and adolescents with type 1 diabetes can be seen in Sweden but not in Denmark. Is there a real difference or is difference between the two countries due to the smaller population of Denmark? This can be investigated through a cohort simulation and power calculation.

References:

Agnieszka Butwicka, Louise Frisen, Catarina Almqvist, Bjorn Zethelius and Paul Lichtenstein. “Risks of Psychiatric Disorders and Suicide Attempts in Children and Adolescents With Type 1 Diabetes: A Population-Based Cohort Study”. *Diabetes Care* (2015) 38:453–459

<http://care.diabetesjournals.org/lookup/doi/10.2337/dc14-0262>

Daniel Dybdal, Janne S. Tolstrup, Stine M. Sildorf, Kirsten A. Boisen, Jannet Svensson, Anne Mette Skovgaard and Grete K. Teilmann. “Increasing risk of psychiatric morbidity after childhood onset type 1 diabetes: a population-based cohort study”. *Diabetologia* (2018) 61:831–838

<https://link.springer.com/article/10.1007/s00125-017-4517-7>

**Supervisor: Annika Tillander**

**TITLE: Visualization of high dimensional data in sports (two projects, one for each sport)**

BRIEF: Within team sports there are an abundance of players available for a team to choose to draft or trade for. Each player will contribute something unique to the team, but there might be similarities in player attributes or performances that can help a team succeed equally well. A great example of this is the book/movie Moneyball about baseball. Finding these underlying and hidden relations between players is difficult as the data is high-dimensional (>20 variables) and might change with time. The dataset consists of information about a player's skills (their attributes) and performance metrics (impact in a game) from a simulated hockey (18 seasons) / football (9 seasons) league. The hockey data has information about 3-400 players every season while the football data has information about 150-180 players every season.

Specific research questions:

* Find a method that can identify similarities between players and project them in two (or in worst case three) dimensions to create simple visualizations.
* Compare methods that reduce the dimensionality of data.

**Supervisor: Isak Hietala**

**TITLE: Identifying attributes that affect player performance (two projects, one for each sport)**

BRIEF: In various team sports such as hockey or football, there are usually an abundance of attributes that can describe a player’s ability and their performance. Finding the relation between the attributes can provide vital information about how a player might progress during their career and which attributes are more important than others for different outcomes. To represent a player's skill in video games, the player attributes are usually grouped together to different areas (offensive, defensive, physical, mental, etc.) but it is not clear which ones are relevant to be a good player at a set position.

The dataset consists of repeated measures data for a set of players (ca 500 hockey players, 250 football players) from a simulated league. For each sport, approximately 20-30 attributes describe the level of skill for the player at doing certain aspects of the sport.

Specific research questions:

* Find the importance of different attributes, within their attribute group, with the target variable being the performance of a player at their set position.
* Find a model that can accurately predict the performance of a player.

I could maybe do four sub-projects (8 total) instead of two, if the students focus on one of the research questions each.

**Supervisor: Isak Hietala**

**Title 1**

Recently, a deep learning model called scDEAL [1] has been proposed to predict drug sensivies in single cell data from bulk repositories. This model is an autoencoder that uses transfer learning to merge two data domains. However, this model does not take into account uncertainty in the botleneck layer. This uncertainty can actually be modelled by switching to Variaonal Autoencoder architecture (through the reparametrizaon trick) and changing MMD discrepancy towards a new funcon [2]

Your task is

* Modify scDEAL into VAE by reparametrizaon trick
* Replace MMD measure by measure proposed in [2]
* Study how performance of the updated model compares to the original method for some of the datasets used in the original publicaon.

1. Chen, J., Wang, X., Ma, A., Wang, Q. E., Liu, B., Li, L., ... & Ma, Q. (2022). Deep transfer learning

of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nature Communications*, *13*(1), 6494.

1. Wu, F., & Zhuang, X. (2021). Unsupervised domain adaptation with variational approximation for cardiac segmentation. *IEEE Transactions on Medical Imaging*, *40*(12), 3555-3567.

**Supervisor: Oleg Sysoev**

**Title 2**

A deep learning model called DeepDRK [1] has been proposed to predict drug effect based on sample characteristics and drug characteristics. Your task is to study

1. How much the predicon accuracy (AUC) is affected when
   1. Each of sample characteriscs is removed as an input, one characterisc at a me
   2. Each of drug characteriscs is removed as an input, one characterisc at a me
2. How much the predicon accuracy (AUC) is affected when
   1. Only *p* Highly Variable Genes are chosen (p to be varied)

Wang, Y., Yang, Y., Chen, S., & Wang, J. (2021). DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. *Briefings in Bioinformatics*, *22*(5), bbab048.

[1] Wang, Y., Yang, Y., Chen, S., & Wang, J. (2021). DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. *Briefings in Bioinformatics*, *22*(5), bbab048.

**Supervisor: Oleg Sysoev**

**Title: Comparing fuzzy and non-fuzzy clustering methods based on quantile and mean on image segmentation**

Image segmentation is the clustering of an image into different groups, and it is an important part of image analysis. In medical applications, the detection of tumors is the most important purpose of image segmentation. In this research, the purpose is to investigate the performance of several clustering methods in the case of image segmentation. In fact, it is expected to compare fuzzy and non-fuzzy clustering methods based on quantile and mean in the case of image segmentation. To do so, it is required to answer the following questions:

1. Evaluate and compare the performance of clustering algorithms based on mean (k-means and c-means) and algorithms based on quantile (k-quantile and c-quantile) on image segmentation.
2. Provide an R or Python package for quantile clustering.

References:

1. Seidpisheh, M., Bamdadi, R. Fuzzy and non-fuzzy k-quantile clustering for high-variance data. *Pattern Anal Applic* **26**, 517–528 (2023). <https://doi.org/10.1007/s10044-022-01127-7>.
2. A. Kumar and S. S. Sodhi, "Comparative Analysis of Fuzzy C- Means and K-Means Clustering in the Case of Image Segmentation," 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2021, pp. 194-200.

**Supervisor: Mohammad Seidpisheh**

**Title: A hybrid model including independent component analysis and neural network for forecasting temperature in Sweden**

Independent component analysis (ICA) is used to separate blind sources of data. In fact, independent component analysis converts complicated and dependent variables to independent variables whose prediction would be easier. Then the idea of this research is to use a hybrid model including ICA and neural network to forecast data. In complicated and dependent data, ICA can first be used as a technique to recognize the independent source of data and then use different models like neural networks to forecast independent data. The main purpose of this research is to investigate the following research questions:

1. Evaluate and compare the performance of different hybrid model to forecast temperature in Sweden.
2. Apply this model to forecast other Applications (financial dataset) if applicable.
3. Adopting algorithms to spark framework to make them more efficient.

J. Henríquez, W. Kristjanpoller, A combined Independent Component Analysis–Neural Network model for forecasting exchange rate variation, Applied Soft Computing, 83,2019,105654, https://doi.org/10.1016/j.asoc.2019.105654.(https://www.sciencedirect.com/science/article/pii/S156849461930434X)

**Supervisor: Mohammad Seidpisheh**

**Title: Learning input-dependent label noise using self-supervision**

**Description:** Large annotated data sets typically contain faulty labels, or label noise. As supervised learning relies on labelled data, such label noise inevitably risks hurting model performance. While methods for handling label noise exist, many relies on non-realistic assumptions such as the noise rate being known or the noise being input-independent, see e.g. [1]. In practice, we seldomly know the true error rate. Moreover, in many applications, the probability of obtaining a faulty label will depend on the observation that is being labelled. For example, in the case of image annotation, an annotator will likely find it harder to identify an object in an image if the image were taken in poor lighting and the object is far away, compared to if the image was taken in broad daylight and the object is close by.

The aim of this project is to investigate the possibility of learning input-dependent label noise in classification. The idea is to make use of unsupervised or self-supervised learning (possibly starting from e.g. a pre-trained DINO model [2]) to obtain a latent representation of the data and then to train an additional model (e.g. a small neural network) to predict the noise distribution given the latent representation. The final aim would be to combine the two models to produce a noise-robust classifier. Although the ultimate goal is to learn to predict, and be robust to, input-dependent noise, the project could also include an investigation of how well we can learn simpler noise models (e.g. input-independent noise) by combining an unsupervised/self-supervised model with a noise predictor.

**References:**

[1] Song, H., Kim, M., Park, D., Shin, Y., & Lee, J. G. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

[2] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9650-9660).

**Supervisor: Amanda Olmin**

**Title: Regularization of the covariance matrix to structured covariance matrix**

It is well known that covariance matrix has to be symmetric positive semi-definite. Often it is

beneficial to assume that covariance matrix has some particular pattern (for example that

covariance matrix has CS structure). Implement method that having an arbitrary covariance

matrix calculate closest matrices with particular (i.e. AR, CS, ...) structures. Which structured

covariance matrix is closest to the original matrix? In which sense are you measuring it? Is the

obtained structured matrix positive semi-definite? If not, find closest matrix that is positive

semi-definite. Write a short description of the methods used. Observe that theoretical results

differ in difficulty depending on considered structure. Students’ capacity to analyze and

implement not trivial methods will lead to the higher grade.

Some references:

Lijing Lin, Nicholas J. Higham, Jianxin Pan (2014) Covariance structure regularization via

entropy loss function

Xiangzhao Cui, Chun Li, Jine Zhao, Li Zeng, Defei Zhang, Jianxin Pan (2016) Covariance

structure regularization via Frobenius-norm discrepancy

**Supervisor: Jolanta Pielaszkiewicz**

**Title: Evaluation of machine learning methods for predicting the initial ability in Computerized Adaptive Tests (CAT)**

**Description:** In a CAT, the questions (also called items), are adapted to the examinee’s ability level. The next item in the test is given based of the estimated latent ability of the examinee, which is calculated from the response of the previous item. This means that a high performing examinee will be given more difficult items and a low performing examinee will be given less difficult items, resulting in more precise ability estimates. In comparison to a fixed test where all examinees are given the same items, a CAT will adjust to the level of the examinee’s abilities. Examinees with different abilities will, therefore, not be given the same test.

Before the test starts, there is no available estimates of the examinee’s abilities. This problem can be tackled by using background variables of the examinees to estimate an initial ability that is used for selecting the first item of the test. The aim of this project will be to choose and compare a couple of machine learning methods for predicting the initial ability using background variables such as age, sex, and educational level. The methods can be e.g., linear regression, random forest, decision tree, or support vector machines. The methods can be evaluated with cross validation. You will be given a data set of estimated abilities (from a fixed test of 40 items) and background variables for the examinees. The data set has missing values which need to be handled in some way.

**References:** Pliakos, K., Joo, S.H., Park, J.Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education.*

**Supervisor: Jonas Bjermo**

**Title: The effect of using ability estimates when predicting the initial ability in Computerized Adaptive Tests (CAT)**

**Description:** See the above project for a description of a CAT and the initial ability prediction problem. When predicting the initial ability in the project described above, the estimated abilities come with some uncertainty which can be considered as a dependent variable measurement error model. They will differ from the true unobserved latent abilities of the examinees. The aim of this project is to quantify the loss of prediction accuracy when the ability estimates are considered without measurement error compared to when the error is considered.

The task can be done by using a simulation study where the true abilities are assumed to be known. R-code will be given that generates a response matrix using Item Response Theory (IRT). For simplicity, linear regression (or some other method of your choice) can be used to predict the examinee’s initial ability when the ability measurement error is considered vs. not considered. It can be useful to use the fact that the variance of the ability estimates is asymptotically equal to the inverse of the test information when the test length is large enough. The test information is dependent on the item parameters of the items (questions) of the test and is given by a specific formula. Even though R-code is given for generating a response matrix it is not necessary to use R for the study.

**References:** Baker, F. B., Kim, S.H. (2017).The Basics of Item Response Theory Using R. Springer.

**Supervisor: Jonas Bjermo**

**Title: Evaluation and comparison of variable selection methods for linear regression**

There exists many different methods about how to select the important covariates in a linear regression problem. The aim of this project is to evaluate and compare different frequentist and Bayesian methods for variable selection in linear regression problems. The idea is to evaluate how good the different methods discriminate on different settings.

**The different settings for the simulated datasets should at least be to**

1. Vary the number of covariates p in the model
2. Vary the number of parameters that are 0 and not equal to 0
3. Vary the variance of the error term
4. Vary the correlations between the independent variables

**The different variable selection methods for the simulated datasets should at least be**

1. *Frequentist methods*
   1. Stepwise backward selection
   2. Stepwise forward selection
   3. Stepwise forward- and backward selection
   4. Best subset selection
2. *Bayesian methods*
   1. Lasso regression (Bayesian perspective)
   2. Ridge regression (Bayesian perspective)
   3. Variable selection with horseshoe priors

Stochastic Search Variable Selection (SSVS)

**Supervisor: Bertil Wegmann**