Thesis topic proposals, spring 2022, for *One Year Master Thesis in Statistics*

On the following pages you will find several (19) topic proposals listed together with the names of the persons suggesting them. Please feel free to contact them if you want further information. Note that some persons have suggested several topics but may only have time to supervisor one (or a few) of them.

Summary of thesis proposals

Master's Thesis Topics

For more information please contact Priyantha: Email: priyantha.wijayatunga@umu.se

Topic 1: Uncertainty quantification in deep neural networks

Often a deep neural network can have tens of thousands of parameters, therefore quantifying its prediction uncertainty through its parameters can be hard. See the paper Abdar et al. (2021) for a comprehensive review on the topic. However, to our surprise researchers often follow this line for doing it! But mathematically, we can define the feature space of the neural network classifier through the output signal space of its last layer. Note that these signals are the input to the classification node of the network, that often uses so-called soft-max activation function to generate the classifying probabilities. Since input to a deep neural network is high-dimensional (such as images, etc.) we can consider the whole network as a dimension reduction tool in this way. The reduced feature space is the output signal space of the last layer of the network.

Once we have derived "the best" feature space we can define a reduced parameter set for the whole network, whose values ranges over the feature space. Note that these parameters are virtual, but their state spaces are not, i.e., they are realistic. For a given training dataset, we should be able to count different configurations of these virtual parameters and that of the output variable. Thus, we can define "precision" of the soft-max probabilities that can be transformed into a Dirichlet distribution. By this way, we have reduced a deep neural network classification task to a multinomial-Dirichlet probability model. The uncertainty quantification can be done with simple probabilistic calculations as in the case of probabilistic networks.

This project deals with implementing deep neural networks with built-in software functions for testing the proposed method compared to existing methods.

Reference

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavanmzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. 2021. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion* 76: 243–297.

Topic 2: Uncertainty quantification in probabilistic Bayesian network inferences

Probabilistic Bayesian networks (BN) are better tools for uncertainty reasoning in artificial intelligence systems. But there is a strong need to quantify prediction uncertainty in those systems, e.g., in automated vehicles, in medical classifiers, etc. In this project we test a simple way to express the uncertainty in their predictions which are otherwise being done in complicated ways (See Allen et al. 2008). Generally, uncertainty of parameters of BN is specified the "precision" of the respective parameter's posterior distribution. For discrete variables it is done using the Dirichlet distribution. For a prediction query, the answer is generally not a posterior expectation a single parameter, but a function of many of them. Here we argue that we can express it as an expectation of a virtual parameter, whose virtual distribution can be taken as a Dirichlet distribution with an unknown hyperparameter. Then we need to specify precision of the virtual hyper-parameter for a given training data set,

which can be used as the uncertainty of the answer to the given probabilistic query. The simplest and logical way to find it is to apply the smallest value among the precisions of all posterior distributions of the parameters in the expression for the answer to the query. And there may be other ways too.

In this project we build BNs for given set of training datasets and evaluate various probabilistic queries for testing and comparing their uncertainties. We use built-in software functions and packages for the purpose.

References

Allen, T. V., Singh, A., Greiner, R., and Hooper, P. 2008. Quantifying the Uncertainty of a Belief Net Response: Bayesian Error-bars for Belief Net Inference. *Artificial Intelligence* 172: 483–513.

Topic 3: On P-value Problems in Statistical Inference

P-values are the mostly used, yet highly abused statistical measures (also called statistics) in the frequentist framework of statistical inference. As stated in Gao (2020); "Without any exaggeration, humankind's wellbeing is profoundly affected by p-values: Health depends on prevention and intervention, ascertaining their efficacies relies on research, and research findings hinge on p-values." Currently many science journals have prohibited using them for statistical inferences. This is mainly because *p*-values are viewed by many as the root cause of the so-called replication crisis, which is characterized by the prevalence of positive scientific findings that are contradicted in subsequent studies (it is difficult to get the same positive results again and again for the inferences that are based on p-values, see Higgins et al 2020). So, it is high time to study about p-values. In fact, there are hundreds of papers written on p-value problems. This project is about analysis of variations in p-values in given contexts. First of all we try to adjust the calculated p-value for a given context in accordance with any uncertainty present in the context. Next, we do simulation studies to see the variations. Ideally, we look for any improvement that we can do empirically (not necessarily theoretically!). Can we devise better tests that avoids current problems in p-values, in specific situations such as, the population mean test (t-test), the regression coefficient test, the chi-squared test of independence, etc. In this project, students get deep knowledge about frequentist hypothesis/significance testing procedures and how to do the tasks of computational experiments on statistical inferences. It is ideal for getting deeper knowledge on statistical inference theory.

References

Gao, J. (2020). P-values – a chronic Conundrum. **BMC Medical Research Methodology**, 20:167 https://doi.org/10.1186/s12874-020-01051-6

Higgins, J. J., Higgins, M. J. & Lin, J. (2020) From One Environment to Many: The Problem of Replicability of Statistical Inferences. **The American Statistician**. https://doi.org/10.1080/00031305.2020.1829047

Topic 4: Measures of statistical dependence for feature selection – a computational study

Feature selection is an important problem in statistics and machine learning for interpretable predictive modeling and scientific discoveries. Straightforward method of feature selection is done by measuring the dependence between feature variables and response variable. So-called mutual information between two random variables is the most commonly used dependency measure. But it is not a perfect measure of dependence and also it is difficult to estimate. In fact, measuring dependence between two random variables is probably the most addressed and applicable aspect of statistics and data science. Due to the advent of Big Data, currently it is more intriguing (see Reshef, et al., 2011 that claims a universal dependence measure!). It is well–known that Pearson's correlation coefficient is measuring only linear dependencies accurately whereas mutual information, Kendall's tau, Spearman's correlation, etc. are used for measuring non-linear dependencies. But they are accurate for monotonic dependencies. Recently in Wijayatunga (2016) and in Wijayatunga (2017), a generalization of the Pearson's correlation coefficient to any non-linear dependency is defined. Especially in feature selection for machine learning and statistical model learning, various dependence measures are proposed, for e.g., Sobolev Independence Criterion (Mroueh, et al. 2019), Hilbert-Schmidt Dependence Criterion (Song, et al. 2012), etc.

In this project, we investigate the cases of feature selection that can be done computationally by using approximations. And of course, we need to compare our results with those obtained by applying distance correlation that is regarded as the best measure of any type of dependency so far. Comparisons with other measures should also done. So, it deals with using R packages, coding in R and using both bench-mark datasets and simulations.

References

1. Mroueh, Y., et al. (2019). Sobolev Independence Criterion. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. IBM Research https://www.ibm.com/blogs/research/2019/12/sobolev-independence-criterion/
2. Reshef, D. N., et al. (2011) Detecting novel associations in large data sets. **Science** 334(6062), pp.1518–1524.
3. Song, L. et al. (2012). Feature Selection via Dependence Maximization. Journal of Machine Learning Research 13, pp. 1393-1434
4. Wijayatunga, P. (2016). A geometric view on Pearson's correlation coefficient and a generalization of it to non-linear dependencies. **Ratio Mathematica**, 30, pp. 3–21. doi: 10.23755/rm.v30i1.5 http://eiris.it/ojs/index.php/ratiomathematica/article/view/5
5. Wijayatunga, P. (2017). Discussion on the Paper "Sparse graphs using exchangeable random measures" by Caron and Fox. ***Journal of the Royal Statistical Society: Series B (Statistical Methodology)***, Vol. 79, Iss. 5, pp. 1359–1359.

Topic 5: Associative confounder bias in causal models – a simulation study

Conditioning on or controlling for 'sufficient' set of confounding variables that affect both treatment and outcome variables causally is necessary for eliminating confounding bias in

tasks of estimating causal effects of the treatment on the outcome. However, things get rather confused when the modeler finds a variable that is non-causally associated with both the treatment and the outcome. Such variables can sometimes induce so–called M–bias or collider-bias (Pearl, 2009 and Sjölander 2009) when they are controlled for. We call this associative confounder bias.

Some researchers who use potential outcome causal model have argued that those associative factors (colliders) should also be included in propensity score calculations for removing bias, i.e., they should also be controlled for, whereas others who use causal graphical models have argued that they cause no bias when they are ignored, and therefore conditioning on them introduces spurious dependence between the treatment and the outcome, thus resulting in some bias in the causal estimates. But in Wijayatunga (2015) has shown that there can be errors in both the arguments, especially in different contexts, if we assume the common cause principle to replace non-causal associations with unobserved hidden common causes. We argue that when such a non-causal confounder is observed, in some cases neither of the actions is appropriate. That is, causal effect estimates are biased either way! And in some other cases, one of the arguments is more correct than the other, and it can be found by observing strengths of associations among variables in the context.
Therefore, here we try to characterize these cases using strengths of associations between associative confounder and, the treatment and the outcome. In literature, some simulation studies are done but they ignore some of the important features of the problem (See an extensive simulation study done in Luque-Fernandez, M. A. et al. (2018) where R code of it is available). Also see "Common Structure Bias" using R (https://cran.r-project.org/web/packages/ggdag/vignettes/bias-structures.html). New simulation studies are needed to verify the characterization of different contexts, especially when exact proofs are too complicated. So, this is a simulation study where we try to find some systematic ways to advice on when to condition on associative confounders in order to obtain unbiased causal effect estimates. Note that an earlier study done by Thoemmes (2015) also offers no advice!

References

1. Ding, P. and Miratrix, L. W. (2014). To Adjust or Not to Adjust? Sensitivity Analysis of *M*-Bias and Butterfly-Bias**. Journal of Causal Inference**. 2193-368
2. Luque-Fernandez, M. A. et al. (2018). Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application. **International Journal of Epidemiology**, pp. 1-14. doi: 10.1093/ije/dyy275
3. Pearl J. (2009). Letter to the editor. **Statistics in Medicine**, 28:1415–16.
4. Sjölander A. (2009). Propensity scores and M-structures**. Statistics in Medicine**, 28:1416–20.
5. Thoemmes, F. (2015). M-bias, Butterfly Bias, and Butterfly Bias with Correlated Causes – A Comment on Ding and Miratrix (2015). **Journal of Causal Inference**, 3(2): 253–258.
6. Wijayatunga, P. (2015). On Associative Confounder Bias. In *Nowaczyk, S. (Ed.)***,** Proceedings of The Thirteenth Scandinavian Conference on Artificial Intelligence, Halmstad Sweden. pp. 157–166. doi:10.3233/978-1-61499-589-0-157.

Thesis supervisor: **Priyantha Wijayatunga**

Topic 6: Dynamic Clustering of Accelerometer Data

Accelerometers are used to detect and recognize human physical activities and motions. They can be especially used for helping old people to change their activities and motions during the day, for e.g., if the person is walking too long, then he/she can be instructed to sit down after detecting the activity through accelerometers (see References). Generally, data created by them are three dimensional in space. And often the activity patterns are not tagged, as walking, running, etc. So, classification of activities in such data streams are about unsupervised learning, such a clustering, etc. For example, when a person is wearing an accelerometer, in order to detect his/her activity pattern, we need to cluster activities dynamically over time. For that we need good methods that uses feature selection methods for clustering of accelerometer data to detect physical activity. Since there can be many features, principal component analysis feature selection and correlation feature selection have been used to remove redundant features. Then hierarchical clustering, k-mean cluster, etc. methods can create refined clusters. This project is about using raw accelerometer data, obtained from smartphones and smartwatches, to extract both time and frequency domain features for efficient dynamic clustering. The project deals with computational aspects and feature selection methods. So, programming tasks are highly involved.

References

Dobbins C, Rawassizadeh R. (2018). Towards Clustering of Mobile and Smartwatch Accelerometer Data for Physical Activity Recognition. Informatics, 5(2):29. https://doi.org/10.3390/informatics5020029

Thesis supervisor: **Priyantha Wijayatunga**

Thesis proposals, spring 2022, Anders Lundquist

I have a few proposals with the joint theme “analysis of longitudinal data”, using longitudinal data(!) from the OLIN-study, a population-based study on lung disease, for see https://www.norrbotten.se/Utveckling-och-tillvaxt/Utveckling-inom-halso--och-sjukvard/Forskning/Forska-i-landstinget/OLIN-studierna/

Here, we have repeated measurements of the subjects over a 15-year time-period where subjects start out diseased and remain so for the study period, start out healthy and remain so, or start out healthy but develop disease. There are several possible ways of looking at and analyzing this data. One way is to look at actual measurement of various lung capacity measurements using linear mixed models, where one may further pursue “specializations”, such as

1. Generalized additive mixed models.
2. Joint longitudinal/time-to-event-modeling.

Another possibility is to use the clinical definition of disease as a binary outcome, where one then uses generalized linear mixed models, and again a deeper dive can be made, into:

3. Generalized (additive) mixed models.
4. Joint longitudinal/time-to-event-modeling.

Besides OLIN-related topics I have one other topics:

5. Predicting autism from resting-state fMRI, using time series classification methods.

Out of the suggestions 1-4, suggestion number 3 is mostly relevant for the clinicians.

Master thesis proposal

Xijia Liu

December 10, 2021

Topic I: Study based on a new coefficient of correlation

The classic correlation coefficients, such as Pearson's correlation coefficient, Spearman's ρ, and Kendall's τ, are often criticized for being not effective for detecting associations that are not monotonic. For this reason, this classic topic has been repeatedly studied all the time, for example, [13], [15], [8], [6], and so on. For a comprehensive review, check [1]. However, these tests suffer from different critics, e.g. simple asymptotic theory is not available, the computational cost is very higher, inconsistent in non-monotonic cases, and so on. Recently, Chatterjee [3] proposed a new coefficient of correlation. The author claim that the new coefficient of correlation is simple, computationally efficient, and has neat asymptotic theory. Most importantly, it is a consistent estimator of some measure of dependence even for monotonic case. Based on this paper, we may ask some research questions, for example,

1. Investigate the performance of the statistical test of independence in the case of a limited sample size. In the paper, the author only shows the performance of the test based on asymptotic distribution when the number of observations is 100. If the sample size is reduced, how much will the effectiveness of the test be lost? Can resampling methods be applied to hypothesis testing?

2. In the simulation study, author only applied simple measurable function to generate dependent samples to check the performance of the test. If we generate the dependent samples using more complicated non-linear function, e.g. neural networks, can the coefficient efficiently detect the dependence?

3. Can we apply the new coefficient of correlation for feature selection problem? We may compare it with different classical variable selection tools, e.g. LASSO, Random Forest, and so on.

Topic II: Study on Hotelling's T-sqaured test

Hotelling's T-squared is famous in multivariate data analysis and process control chart. T squared test is very recommended because it considers the correlation between variables and therefore well controls the first type of error rate, see [4]. Regarding to Hotelling's T-squared test, we have the following two research questions.

1. How about the performance of Hotelling's T-squared test when the dataset contains missing values? In the case of a missing value, the maximum likelihood estimation of mean vectors and covariance matrix can be obtained by the EM algorithm and therefore make Hotelling's T-squared test feasible. Meanwhile, [17] proposed an alternative, scaled Chi-squared test, and claim that their test has an advantage over Hotelling's T-squared test from the missing value point of view. However, in their simulation studies, Hotelling's t-square test was only applied to incomplete data cases. In this study, we may investigate factors that influence the performance of Hotelling's T-squared test on the dataset with missing values, and compare it with the scaled Chi-squared test.

2. Hotelling's T-sqaured test is a global test on the overall hypothesis. Therefore, it is necessary to further test the significance of each variable after rejecting the overall hypothesis and then directly face the multiple test problem. Although there is a confidence interval based on T-squared that can well control the first type of error rate, at the same time, the too wide intervals also make the test very conservative, and people have to go back to one-at-a-time interval to find a significant result. People seek guarantees for it, e.g. the authors claim that "Moreover, if the one-at-a-time intervals are calculated only when the T^2-test rejects the null hypothesis, some researchers think they may more accurately represent the information about the means than the T^2-intervals do." in [10]. However, there is no research that quantifies this gain using Hotelling's T-squared test. In this study, we may consider designing a good simulation study aiming for results that can provid some evidence to this strong statement.

Topic III: Further developing RUSBoost algorithm

Balanced class is a crucial factor for training a machine learning algorithm. However, imbalanced class is prevalent in the real world, especially in medical research. Many methods have been proposed to solve this issue and a set of popular solutions are based on resampling methods, for example, random under sampling (RUS). Simply speaking, one can embed the resample procedure into any existing algorithm.

RUSBoost algorithm [14] is a modified version of Adaboost algorithm [5] with RUS embedding. It has been successfully applied in many empirical works in machine learning, for example, [16] [2]. However, there are some obvious issues inside the algorithm and that inspire several research ideas. In this project, we will further develop RUSBoost algorithm accordingly and implement it in R environment. Simulation studies and real cases studies will be performed to compare the performance of our methods with the original RUSBoost algorithm.

Topic IV: Functional data analysis on fMRI data

Functional data analysis (FDA), as an active branch of statistics, analyzes some continuous observations that vary over a continuum, e.g. curves, surfaces, and so on. In other words, we treat the entire observation varying over the continuum as one point and capture the statistical feature of the subject in a general vector (Hilbert) space. FDA has been intensively studied during the last two decades. For theoretical knowledge, you are recommended to read [12], Regarding practical usage, [11] provides a good tutorial based on R statistical package. In addition, a solid mathematical background have been summarized in [9]. On another hand, the main research object of this study is Functional magnetic resonance imaging (fMRI) data. fMRI can be applied to measure the brain activities by detecting changes associated with blood flow. As an important complementary technique to other medical images, fMRI has been widely used in many research, for example, neuroscience.

For this study, we apply FDA approaches to fMRI data with a neuroscience background. The fMRI data that will be analysised is a published data [1] provided in [7]. The research questions can be very open and this study can be regarded as an exploratory analysis, that is, mining the statistical characteristics of the data through the modelling and analysis of foundational data and seeking a reasonable scientific explanation from the perspective of neuroscience. Meantime, we are also interested in a specific research question that applies FDA clustering analysis to segment the brain image. In neuroscience, most studies based on fMRI data rely on a strong a priori hypothesis that the biological brain can be divided into different functional areas. Most neuroscience researches are based on pre-determined functional areas segmentation for data analysis, and this just provides us with a good motivation. Therefore, one can naturally ask whether data-driven methods can be developed to replace this strong a priori hypothesis? FDA approaches provide us with a direction to the potential solutions to this question.

[1] In order to analysis the data, we need to preprocess the data using ready-made Matlab code that provided in a github repository Github link.

References

[1] A. G. Asuero, A. Sayago, and A. Gonzalez. The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1):41–59, 2006.

[2] F. L. Bayisa, X. Liu, A. Garpebring, and J. Yu. Statistical learning in computed tomography image estimation. *Medical physics*, 45(12):5450–5460, 2018.

[3] S. Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 0(0):1–21, 2020.

[4] J.-Q. Fang. *Handbook of Medical Statistics*. # N/A, 2017.

[5] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[6] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, A. J. Smola, et al. A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer, 2007.

[7] D. Gutierrez-Barragan, M. A. Basson, S. Panzeri, and A. Gozzi. Infraslow state fluctuations govern spontaneous fmri network dynamics. *Current Biology*, 29(14):2295–2306, 2019.

[8] R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

[9] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

[10] R. A. Johnson, D. W. Wichern, et al. *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:, 2014.

[11] J. Ramsay and B. W. Silverman. *Functional data analysis (Springer series in statistics)*. 1997.

[12] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer, 2002.

[13] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[14] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2009.

[15] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

[16] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. An empirical evaluation of repetitive undersampling techniques. *International Journal of Software Engineering and Knowledge Engineering*, 20(02):173–195, 2010.

[17] Y. Wu, M. G. Genton, and L. A. Stefanski. A multivariate two-sample mean test for small sample size and missing data. *Biometrics*, 62(3):877–885, 2006.

Innovative statistical methodologies for missing data imputation, noise reduction and bias correction in simulation study

Jianfeng Wang

Missing, noisy and biased measurements are three common occurrence and can have a significant effect on statistical conclusions that can be drawn from the data. In literature, there are a number of statistical methods for missing data imputation and noise reduction. However, the performance could be heavily affected by researcher's subjective choice. As for bias, there is very little about bias correction. In this project, new statistical methodologies will be developed to improve the state of the art methods in missing data imputation, noise reduction and bias correction. Comparisons will be made between the new developed methodologies and traditional statistical methods.

The frame work of the project could be done under the compressive sensing through convex-optimization problem, for example,

$$\min_{M} \frac{1}{2}||O - M||_F + \lambda_1||\sigma(M)||_1 + \lambda_2||\sigma(M)||_2, \tag{1}$$

where O is the observed Matrix with missing, noise and bias, M is the matrix we want to estimate, and $\sigma(M)$ denotes the singular value of M.

Supervisor: Hamed Sabahno

Suggested Title:
''The combined effects of measurement errors and autocorrelation on the performance of a simultaneous 'mean' and 'variability' monitoring scheme for multivariate normal processes''

Abstract
''Statistical Control charts are very easy and effective ways of monitoring any process to see if it is in-control or not. If we have only one process characteristic, we use univariate control charts, on the other hand, if we have more than one process characteristic, we use multivariate control charts. There are many control charts for different situations and conditions.

Sabahno et al. (2020a) developed a new multivariate control chart for adaptively and simultaneously monitoring of the process parameters (mean and variance-covariance). Later, they extended their scheme in case of measurement errors (Sabahno et al., 2020b) and autocorrelation between the observations (Sabahno et al., 2020c). Although in most real case applications, measurement errors and autocorrelation both exist in the process at the same time, for simplicity, they only considered one of them in each paper.

The goal is to combine Sabahno et al. (2020a and b)' works and design a new scheme which contains measurement errors and autocorrelation together. We want to use Markov chains in order to compute performance measures. We will also present an illustrative example to show how the scheme can be applied in practice.''

References:
1-Sabahno H, Amiri A and Castagliola P(2020a). A New Adaptive Control Chart for the Simultaneous Monitoring of the Mean and Variability of Multivariate Normal Processes. Computers & Industrial Engineering, https://doi.org/10.1016/j.cie.2020.106524.
2-Sabahno H, Castagliola P and Amiri A (2020b). A Variable Parameters Multivariate Control Chart for Simultaneous Monitoring of the Process Mean and Variability with Measurement Errors. Quality and Reliability Engineering International, Vol.36, No.4, pp.1161–1196.
3. Sabahno H, Castagliola P and Amiri A (2020c). An Adaptive Variable-Parameters Scheme for the Simultaneous Monitoring of the Mean and Variability of an Autocorrelated Multivariate Normal Process. Journal of Statistical Computation and Simulation, Vol.90, No.8, pp. 1430–1465.

Sleep apnea prediction in a Swedish cohort

Background: In sleep apnea, there are repeated pauses in breathing followed by poor oxygenation (hypoxia). The disease is defined as more than 5 breathing pauses per hour sleep (= apnea-hypopnea index> 5), with mild sleep apnea AHI 5-15, moderate sleep apnea 15-30, and severe sleep apnea AHI> 30. People with sleep apnea are snoring, day tired, and have an increased risk of cardiovascular disease including high blood pressure, stroke, and atrial fibrillation. Obesity and a small lower jaw are risk factors for sleep apnea. Up to 50% of adult men and women have sleep apnea, but only about 10% of these have been investigated and diagnosed. CPAP is the most effective treatment.

It is important to identify patients with sleep apnea before surgery as they have an increased risk of complications after surgery. STOP-Bang is a questionnaire for predicting the risk of sleep apnea based on risk factors such as snoring, apnea, male gender, obesity, hypertension, and large throat circumference. We have found in ongoing studies that sleep apnea is very common among patients who are operated on for cancer of the colon and rectum and that STOP-Bang can not identify these patients.

Project:

1) Create one or more prediction models that, based on data at the personal level, predict the degree of sleep apnea defined as mild or moderate to severe sleep apnea. The predictive power of the prediction models must be quantified.
2) Simplify the proposed model into a useful form similar to "Stop Bang". Quantify how the prediction properties of simplification change.
3) If possible, with access to data: Compare the predictability of the models above with the predictability of the "Stop Bang" form on the Swedish cohort http://www.stopbang.ca/translation/pdf/seswe.pdf

Special consideration must be given to gender during model building as factors are expected to affect the incidence of sleep apnea differently depending on whether the person is male or female.

Data: Available data comes from SCAPIS study of sleep apnea in Umeå, Gothenburg and Uppsala and contains data on Swedish citizens aged 50-65 years. Relevant variables in the data include age, gender, weight, height, BMI, neck circumference, incidence of hypertension, etc.

Supervisor: Johan Svensson Senior Lecturer at the Department of Statistics.

References and relevant links:

Franklin KA, Axelsson S, Rehnqvist N. SBU. Obstructive sleep apnoea syndrome. Report of a joint Nordic project. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2007. ISBN 978-91-85413-16-4.

apnea ASoATFoPMopwos. Practice guidelines for the perioperative management of patients with obstructive sleep apnea: an updated report by the American Society of Anesthesiologists Task Force on Perioperative Management of patients with obstructive sleep apnea. *Anesthesiology* 2014;**120**(2): 268-286.

Chung F, Yegneswaran B, Liao P, Chung SA, Vairavanathan S, Islam S, Khajehdehi A, Shapiro CM. STOP questionnaire: a tool to screen patients for obstructive sleep apnea. *Anesthesiology* 2008;**108**(5): 812-821.

Singh M, Liao P, Kobah S, Wijeysundera DN, Shapiro C, Chung F. Proportion of surgical patients with undiagnosed obstructive sleep apnoea. *Br J Anaesth* 2013;**110**(4): 629-636.

Heinzer R, Vat S, Marques-Vidal P, Marti-Soler H, Andries D, Tobback N, Mooser V, Preisig M, Malhotra A, Waeber G, Vollenweider P, Tafti M, Haba-Rubio J. Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study. *Lancet Respir Med* 2015;**3**(4): 310-318.

Franklin KA, Sahlin C, Stenlund H, Lindberg E. Sleep apnoea is a common occurrence in females. Eur Respir J. 2013; 3: 610-15.

Franklin KA, Lindberg E. Obstructive sleep apnea is a common disorder in the population. -A review on the epidemiology of sleep apnea. J Thoracic Disease 2015; 10: 1311-22

Methods for outlier detection for functional data

In many applications it is of interest to detect outliers in data before performing analysis. It is not trivial how to do this when the data of interest are functions/curves. The suggestion for this thesis topic is to review existing literature and describe available methods for outlier detection, including studies of depth- and centrality measures for functional/curve data. The identified methods could be applied to human movement data. In biomechanics, it is of interest to identify errors and/or anomalies (real values) arising from the participant or the experiment to get a more correct estimate of the movement variability. Also, anomalies may need to be analyzed separately to understand the causes of the abnormal pattern.

Supervisor: Johan Strandberg (Lina Schelin)