Chair for Quantitative Business Administration

Topics for a Master Thesis

8th July 2019

Contents

1 Topics in Structural Estimation
   (Supervisor: Gregor Reich) 2
   1.1 Statistical Efficiency in the Estimation of Dynamic Discrete Choice Models . . . 2
   1.2 Estimation of Agent-Based Models ....................................................... 3
   1.3 Parallel Global Optimization for Estimation Problems .............................. 4
   1.4 Estimation of Poorly- or Non-Identified Parameters ............................... 5

2 Topics in Portfolio Optimization
   (Supervisor: Robert Erbe) 6
   2.1 Applied Topics in Asset Management .................................................... 6

3 Topics in Blockchain Mechanism Design
   (Supervisor: José Parra Moyano) 6
   3.1 Blockchain Mechanism Design ................................................................. 6

4 Topics in Optimization Models for Portfolio Selection
   (Supervisor: János Mayer) 7
   4.1 Portfolio Selection with Equal Risk Contribution according to
       Conditional Value–at–Risk ....................................................................... 7

5 Topics in Machine Learning
   (Supervisor: Vanessa Kummer) 8
   5.1 Manipulating Voice by Sliding Attributes ................................................. 8
   5.2 Uncovering and Mitigating Latent Algorithmic Bias in Natural Language Pro-
       cessing ................................................................................................... 9
   5.3 Testing the Robustness of the Multiple, Iterative Imputation Approach proposed
       by Egger, Kummer, Meusel, and Schmedders (2017) ............................. 9
1 Topics in Structural Estimation  
(Supervisor: Gregor Reich)

1.1 Statistical Efficiency in the Estimation of Dynamic Discrete Choice Models

Background

When applying dynamic economic models, usually the models first have to be solved before they can be analyzed, which is often a challenging task of its own from a computational perspective. However, when the goal is to estimate such a model, given data on the state and the corresponding decisions of the agents (for example, the state and age of a car, and the corresponding decision on whether to carry out an expensive repair, or to scrap it instead), the solution of the model can be circumvented under some conditions, by directly inferring the solution from the observed actions in the data. This concept is called two-step estimation, and goes back to Hotz and Miller (1994). As a downside, this method is known to have a lower statistical efficiency, as the parameter estimates usually face a substantial small sample bias.

A more recent development in the literature on estimating dynamic discrete choice models (DDCM) is the incorporation of state variables that are serially correlated, and at the same time unobserved by the econometrician who estimates the model. An important example for such a state variable is learning: While it takes place in the head of the decision maker and is therefore often not part of any measurement nor data, it is (hopefully) serially correlated as there is usually progress in learning over time. To compute the likelihood function for such a model, a large “average” has to be computed, taking into account all possible values of the unobserved state variables (e.g. learning states), which corresponds to—mathematically speaking—computing a high-dimensional integral. For both estimation approaches, recent methods exist to incorporate such variables, namely Arcidiacono and Miller (2011) for the two-step estimators, and Reich (2016) for the maximum likelihood estimation.

Research Question

To date, nothing is known about the statistical efficiency of two-step estimators for DDCMs with serially correlated unobserved states in practice. Therefore, in this project, you pick a (simple) DDCM, featuring serially correlated unobserved states, and estimate it both using (i) the maximum likelihood estimation approach with recursive likelihood function integration by Reich (2016), and the two-step expectation-maximization approach by Arcidiacono and Miller (2011). Using simulated datasets and thus having control over the data generating process, you can extensively analyze and compare the two methods for statistical as well as computational efficiency.
Literature


1.2 Estimation of Agent-Based Models

Background

Agent-based computational economics is a simulation-based approach to explain aggregate dynamics from actions on the micro-level, e.g. explaining stylized facts of financial markets by modeling the behavior of individual investors and their environment. For example, the model of Kirman (1991) generates herding behavior through simple agent interaction that replicates some standard stylized facts such as volatility clustering and fat tails of the return distribution. However, agent-based models usually have many parameters, some of which have no real-world equivalent and are thus hard to set. Given data on the particular market that is modeled, one can choose the parameters such that the real data is “well predicted” by the model in some sense; several approaches exist using minimization of a moment function (see, for example, Gilli and Winker, 2003, or Franke, 2009), to match the statistical moments of real data (e.g. stock returns) to the ones from the simulated data.

However, the fundamental problem of most existing methods face is that the evaluation of the moment function involves simulating the model, and thus is subject to noise: if the moment function is evaluated several times for identical parameter values, it will always return slightly different values, since the model simulation always produces different paths. As a consequence, the objective function has many local extrema—both, minima and maxima—that are purely artificial, and make it difficult for the optimization solver to identify the true, “natural” solution to the moment fitting problem.

Research Question

In this project, you will first review the literature on more recent heuristic approaches using simulation techniques to find parameter values that produce a good fit to the data; this includes studying a master’s thesis on this topic previously written at the chair (Foster, 2014). Second, a different approach to the estimation of simulation models is implemented as a prototype, and then applied to the models of Kirman (1991) and Franke (2009); finally, the estimation results
and their robustness are compared to the findings of the original papers.

**Literature**


### 1.3 Parallel Global Optimization for Estimation Problems

**Background**

The process of estimating a economic model, i.e. fitting its parameter such that the data predicted from the model fits the real data as good as possible, is tightly linked to (mathematical) optimization: Either, a function called the likelihood function, which maps the model parameters to the probability of observing the real data given the model, is maximized, or, moment function, which maps the model parameters to the statistical moments of the data predicted by the model and compared them to the moments of the real data, is minimized. However, in “real” problems, these functions often exhibit difficult numerical properties, most prominently multiple solutions, and non-differentiability, which are hard to overcome by traditional optimization solvers.

More recently, a class of methods called heuristics has been established, solving optimization problems based on less rigorous concepts, therefore often not relying on particular mathematical properties such as differentiability or convexity. One class of heuristics to find the globally optimal solution are the population based heuristics, with its most prominent member being the genetic algorithms: Here, the solution is not obtain from following a trajectory of candidate solutions obtained from purely local improvements—such as prototypically done in the steepest decent methods; rather, a “population” of candidate solutions is maintained, and each individual is evolved based on the “knowledge” from the whole population. By means of a fitness comparison, good candidates are favored over bad ones, and the population as a whole is evolutionarily evolved. This procedure can overcome the global solution problem, by making sure that solutions that are locally, but not globally optimal will eventually be replaced by global solution candidates; the application of heuristics in econometrics to overcome the globality problem has triggered some substantial interest in the past, see e.g. Dorsey and Mayer (2012).
Another recent development in computational sciences is the fact that single compute units are not becoming faster anymore; rather, the number of compute units, often the cores of a processor, is increasing instead. Therefore, computing-intensive applications—statistical estimation being one example!—are more and more required to make use of parallelism.

**Research Question**

In this project, you will first review the literature on parallel population based heuristics, in particular regarding the limited degree of parallelism arising from small population sizes, and the load-imbalancing problems if the evaluation time of the objective function varies. Second, you will conceptually derive an asynchronous population based method, implement a prototype, and compare the results to the standard algorithms from the literature.

**Literature**


1.4 Estimation of Poorly- or Non-Identified Parameters

**Background**

When estimating the parameters of economic models, such as dynamic discrete choice models (DDCM), care has to be taken whether the parameters are identified at all, i.e. whether the model is set up such that given the data, it is possible to recover the parameters using an estimation technique such as maximum likelihood estimation. It is well known that for many models, important parameters such as the time discount factor cannot be recovered together with the other parameters of the model. For example, Rust (1987)—pioneering the DDCM estimation literature—notes that “I was not able to precisely estimate the discount factor $\beta$. Changing $\beta$ to .98 or .999999 produced negligible changes in the likelihood function and parameter estimates of [the other parameters]. The reason for this insensitivity is that $\beta$ is highly collinear with [one of the other parameters]” which holds true for many other models as well.

**Research Question**

In this project, you will first compare different approaches to estimate models for all values of a particular critical parameters, and argue under which circumstances the “naïv” approaches fail. Second, you will derive a novel approach, implement and apply a prototype, and compare your results both qualitatively and quantitatively.
Literature


2 Topics in Portfolio Optimization
(Supervisor: Robert Erbe)

2.1 Applied Topics in Asset Management

We have established a fruitful collaboration with a Swiss-based asset manager that allows master students to work on real-world portfolio optimization problems. The general research topics are typically quite involved and require solid mathematical knowledge along with good programming skills. The necessary datasets as well as supplementary R code will usually be provided to the student—thus having already worked with R often makes things a bit easier but it is no prerequisite. We invite ambitious students to proactively approach us by sending us their CV as well as their latest university transcript, highlighting their areas of interest.

The following topics have been assigned to students in the past (non-exhaustive list):

- Minimum Variance Portfolio with Market Impact Costs
- Estimation of High-Dimensional Covariance Matrices
- Portfolios from Sorts – Bayesian Methods for Portfolio Optimization

3 Topics in Blockchain Mechanism Design
(Supervisor: José Parra Moyano)

3.1 Blockchain Mechanism Design

Background

Smart contracts are computer protocols that facilitate, verify, and enforce contracts. Blockchain Technology empowers the implementation of smart contracts that can substantially improve existing inter- and intra-corporate structures, making them more efficient and avoiding the conduction of parallel tasks. These new rules can open a new field of study in the Information Systems literature.

Research Question

The research question is to analyze how the business logic of an existing corporate structures can be improved by means of implementing smart contracts in a Blockchain. Specifically, the objective is to find out how “exporting” an existing business model to a Blockchain structure can improve the efficiency of the business logic, avoid failures and inefficiencies. Answering
these questions can serve as a basis for the development of the Applied Information System Literature.

Literature


4 Topics in Optimization Models for Portfolio Selection
(Supervisor: János Mayer)

4.1 Portfolio Selection with Equal Risk Contribution according to Conditional Value–at–Risk

Background and Research Question

The financial portfolio selection literature reflects an increasing popularity of the Equal Risk Contribution (ERC) approach. Generally ERC means a portfolio selection technique, which is based on the decomposition of the total portfolio risk into an equal risk contribution of each of the component positions, see Maillard, Roncalli and Teiletche (2010). ERC has been usually implemented with the standard deviation chosen as risk measure. In the recent literature this approach is extended by employing Conditional Value–at–Risk (CVaR) as the underlying risk measure, see Boudt, Carl and Peterson (2013) and Cesarone and Colucci (2015). For CVaR see, e.g., Rockafellar and Uryasev (2002).

The main constituents of this project are as follows: Carrying out a literature review and discussing both CVaR and ERC from the theoretical point of view, as well as summarizing the practical aspects of ERC. Subsequently comparative numerical studies are being carried out with monthly return data sets according to two aspects. On the one hand, the portfolios resulting from the CVaR–ERC approach, with– and without the assumption of normally distributed returns, are to be compared. On the other hand, an out–of–sample comparison has to be carried out, by comparing the portfolios from the CVaR–ERC approach with the equally weighted portfolio and with the portfolios obtained from the global minimum variance model and from the global minimum CVaR model. The portfolio performance measures used for this purpose should include besides the Sharpe-Ratio also some additional portfolio performance measures, not yet utilized in Cesarone and Colucci (2015). In all of the portfolio selection models short sales are to be excluded.
Literature


5  Topics in Machine Learning
   (Supervisor: Vanessa Kummer)

5.1 Manipulating Voice by Sliding Attributes

Background

Voice recognition is becoming increasingly ubiquitous and powerful. In 2017, Google has reported that 20% of their searches are already made by voice queries. In addition, they reported that their speech recognition system had a 95% accuracy rate. While that’s an impressive number, it begs the question: 95% accurate for whom? Recent analyses of various machine learning (ML) systems (e.g. of classification models that determine if a person is qualified to receive a loan or not, but also of the widely used task of facial recognition) have shown that there exists an extreme amount of algorithmic bias among selected demographics. The reason for those biases is biased training data. Bias often occurs unintentional and sometimes it is even latent. However, bias can also arise from sample size disparity.

Research Question

Since it is often not possible to just collect more data in order to balance the training set, and under-sampling leads to a lot of information getting lost, the goal of this project is to artificially create data of minority groups (can for example be women with Irish accent). Lample et al. (2017) have introduced such an architecture for images, allowing to swap genders or to add/remove glasses. By balancing the training set, the resulting system should become fairer.

Literature

5.2 Uncovering and Mitigating Latent Algorithmic Bias in Natural Language Processing

Background

In today’s world, many decisions—such as hiring, advertising, criminal sentencing, or lending—are entirely made, or at least influenced, by machines. Using training data, these systems learn a model which can be applied to other people and make predictions about what the correct outputs should be for them. This becomes problematic though if the training data is biased. Online retailer Amazon recently discontinued the use of a recruiting algorithm after discovering gender bias. The data that engineers used to create the algorithm were derived from the resumes submitted to Amazon over a 10-year period, which were predominantly from white males. The algorithm was taught to recognize word patterns in the resumes and these data were benchmarked against the company’s predominantly male engineering department to determine an applicant’s fit. As a result, the AI software penalized resumes of women. The problem is that, even if gender is not included as a feature itself, certain words are connected to certain groups and so the model picks up on existing racial and gender biases shown by humans.

Research Question

Amini et al. (2019) have developed an algorithm that uncovers and mitigates hidden, and potentially unknown, bias in facial detection systems in order to address the issue of algorithmic racial and gender bias. The goal of this project is to apply their (or a similar) algorithm to text data in order to reduce the algorithmic bias of natural language processing models that are trained with that text data.

Literature


5.3 Testing the Robustness of the Multiple, Iterative Imputation Approach proposed by Egger, Kummer, Meusel, and Schmedders (2017)

Background

Missing data arise in almost all statistical analyses. A common approach is to delete the missing values. The problem that comes along with listwise deletion is the loss of power and sample size, as well as potential bias in the data that comes along with listwise deletion. A better approach is therefore to fill in (impute) the missing values. An often used imputation method is mean imputation. The advantage of mean imputation is that it is easy to implement and that it does not change the mean. The drawback is that the relationships with other variables are ignored. Fortunately, there are exist other methods.
Research Question

In order to impute a data set having missing values in all variables, one can make use of unsupervised learning. In our working paper (Egger, Kummer, Meusel, and Schmedders (2017)), we describe an approach that combines multiple state-of-the-art imputation methods. So far, this method has only been applied on real-world data. It would be interesting to see though, how robust the method performs on simulated data.

Literature