Thesis for the Degree Doctor of Philosophy of Engineering in Mathematical Statistics

Probabilistic modeling in Sports, Finance and Weather

Jan Lennartsson

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden, 2014

Probabilistic modeling in Sports, Finance and Weather

Jan Lennartsson

© Jan Lennartsson, 2014

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE–41296 Göteborg, Sweden
Phone: +4631–772 1000

ISBN 987-91-628-9179-4 (printed)
ISBN 987-91-628-9176-3 (e-publication)

Printed in Sweden, 2014.

Abstract

In this thesis, we build mathematical and statistical models for a wide variety of real world applications. The mathematical models include applications in team sport tactics and optimal portfolio selection, while the statistical modeling concerns weather and specifically precipitation.

For the sport application, we define an underlying value function for evaluating team sport situations in a game theoretic set-up. A consequence of the adopted setting is that the concept of game intelligence is concretized and we are able to give optimal strategies in various decision situations. Finally, we analyze specific examples within ice hockey and team handball and show that these optimal strategies are not always applied in practice, indicating sub-optimal player behaviour even by professionals.

Regarding the application for finance, we analyze optimal portfolio selection when performance is measured in excess of an externally given benchmark. This approach to measure performance dominates in the financial industry. We assume that the assets follow the Barndorff-Nielsen and Shephard model, and are able to give the optimal value function explicitly in Feynman-Kac form, as well as the optimal portfolio weights.

For the weather application, we analyze the precipitation process over the spatial domain of Sweden. We model the precipitation process with the aim of creating a *weather generator*; a stochastic number generator of which synthesized data is similar to the observed process in a weakly sense. In Paper [C], the precipitation process is modeled as a point-wise product of a zero-one Markov process, indicating occurrence or the lack of rainfall, and a transformed Gaussian process, giving the intensities. In Paper [D], the process is modeled as a transformed censored latent Gaussian field. Both models accurately capture significant properties of the modeled quantity. In addition, the second model also possesses the substantial feature of accurately replicating the spatial dependence structure.

Keywords: Mathematical modelling; Game theory; Team sport tactics; Modern portfolio theory; Gaussian fields.

Acknowledgements

First, I would like to express my gratitude to my supervisor, Patrik Albin, for the support, guidance and belief in me. Thank you for always having time with my questions and ideas and for sharing your experience and for pushing me forward with great enthusiasm. You also deserve a rose for just being the lovely, unique man that you are.

Secondly thanks go to my co-advisors Carl Lindberg and Anastassia Baxevani, both of whom (independently) helped me by proposing to me research problems to work on while I still was a professional athlete, helping me to get back into the exciting world of mathematics. Thanks to Carl for being a great inspirer; aiding and pushing me to develop my scientific skills to levels beyond those which I thought were possible when I reentered the mathematical department. You are also a great visionary, venturing out into realms of great ideas, some of which actually come true! Thanks to Anastassia for aiding me, and also placing demands for stringency with my many ideas. It is due to your precision and thoroughness that our joint projects have come to scholary, scientific fruition.

Thanks also go to the co-writer, Nicklas Lidström, for great insights in the field of sports and for showing me why one should also use the inside of the head in competitive interactions.

Thanks to my students, who always keep me on my toes.

Thanks to former colleagues in the mathematical sciences.

And thanks also to by far the most important person in my life, the one that stole my heart many years ago, Elin. You are my *ezer kenegdo*, supporting me through thick and thin. And lastly thanks go to my three wonderful children, Wilda, Maxian, and Winston; you are my sanctuary.

*

This work was financially supported by Göteborgs MiljöVetenskapliga centrum (GMV).

Contents

Abstract i

Acknowledgements iii

Introduction 1

- What is a mathematical model? 1
- Data 2

Motivation for Paper [A] 5

- Description 6
- My contribution 6

Motivation for Paper [B] 9

- Description 11
- My contribution 11

Motivation for Papers [C] and [D] 13

- Description 14
  - Paper [C] 14
  - Paper [D] 15
- My contribution 16

Bibliography 17

Appended papers 19

Introduction

This thesis is based on the four papers:

A Lennartsson, J., Lidström, N., and Lindberg, C., Game intelligence in team sports (2014).
(Submitted).

B Lennartsson, J. and Lindberg, C., Merton's problem for an investor with a benchmark in a Barndorff-Nielsen and Shephard market (2014).
(Submitted).

C Lennartsson, J., Baxevani, A. and Chen, D., Modelling precipitation in Sweden using multiple step Markov chains and a composite model, *Journal of Hydrology* (2008), Volume 363, Issue 1-4, Pages 42-59.

D Baxevani, A. and Lennartsson, J., A Spatio-temporal precipitation generator based on a censored latent Gaussian field (2014).
(Work in progress).

The papers are appended after this introduction. Papers [A] and [B] represent distinct research topics in mathematical modeling, while Papers [C] and [D] both concern statistical modeling of precipitation (rain, snow, hail etc.). Here I present an introductory motivation and a non-technical description of the three topics of the thesis. The aim is to present a brief survey where I convey the ideas behind the tools and procedures that underlie the research. Further I which to provide the reader with an understanding of the key aspects of the utilized mathematical and statistical modeling techniques in form of principles, concepts, and methods. Appended after each introduction is also a statement of my personal contributions to each of the Papers as compared with that of my co-authors.

What is a mathematical model?

Scientific observations are never completely conclusive for real world systems, whether it is the field of team sports, the stock market, precipitation or any other object of interest. In order to grasp the complex nature of a subject, models are introduced to describe the observed relationships. A model can in this sense be seen as a thought experiment, a simplified version of reality, in which exact calculations and conclusions can be made.

Models are constructed based on empirical observations and general considerations such as simplicity, generality, interpretability and explanatory power. Then, since exact computations within the defined framework are possible, consequences of the given model are evaluated in a statistical sense, or a purely logical fashion, and conclusions of the underlying system are drawn. Note that any conclusions extracted from a model are based on the premises of the models applicability and one should be careful in extrapolating results in directions where the premises may be placed under question.

A naive perspective on modeling is to ask if the model, in an absolute sense, represents the real world. But it is rarely the case that a model is true in all aspects – which does not mean that it is useless. A more sophisticated perspective acknowledges that no model, be it mathematical, statistical, or a model in any other setting, can take into account all the complex dependence structures of the underlying reality. Further, here we accept that we build models for a reason: To understand and decide on specific actions and how to deal with the system. While we recognize that the models represent a necessary simplification of the stunningly complex world, if we choose them wisely then they will enable us to draw considerable conclusions about the specific system they describe. Further, we construct models that are good enough for their designed purposes. In particular, regarding the Papers [C] and [D], we do not intend to deductively claim that the daily amount of precipitation is governed in precisely the specific ways that the Papers conjecture. Note also that the models proposed in these Papers are non-overlapping; i.e., both can not be correct, in an absolute sense, at the same time. However, both models suit their purposes – as stochastic weather generators, and there are no apparent differences between synthetically generated data and actual observations.

Models enable us to give a simplified description of the relationships between several (possibly a large number) of variables, based on the fact that many of the variables are caused by intermediate relationships. In Papers [B] and [D], we conjecture the existence of an underlying unobserved variable, which is related to each of the stochastic quantities modeled, and from which a dependence structure is induced. For example in Paper [B], the very idea is to introduce the latent (unobserved) variables of news processes, in order to model shifts in the volatility matrix. However, two aspects need to be clearly stated here; first, the actual value of the latent variable will never be observed. It actually is an unobservable quantity in terms that no-one explicitly will be able to calculate other than in situation-based estimates of it. Secondly, empirical observations of asset price processes show a strong tendency to react to the same background information. Further, as contradicting as it may sound, eventhough the news processes are unobservable and only exist in books, is it a intelligible quantity for most investors.

Data

For the statistical part of this thesis, Papers [C] and [D], the data is of significant importance. The data we use is measurements of the amount of daily precipitation at specific locations. The data utilized in Paper [C] is recordings at 20 weather stations distributed all over Sweden where each station features 44 years of data. In Paper [D] we had access to the daily recordings for 51 years and 14 specific weather stations distributed in a more dense area of Sweden, i.e., the northern portion of the region of Götaland.

In general, as the complexity of a model increases, so does its goodness of fit to the available sample, – however the goodness of fit to other samples drawn from the same distribution may deteriorate. If a model is too elaborate, it may represent the observed sample very well indeed – but will fail to explain the underlying process. A model that is too simple on the other hand risks missing out on potential key aspects of the system. The aim of statistical modeling is to generalize from the sample, i.e., to accurately model the underlying process, in contrast to just modeling the sample. In order to deal with these issues, the principle of cross validation is utilized. Cross validation is a computationally intensive approach, where the available sample is divided into two sets: a design set, and a validation set. The model is then trained on the design set, where point estimation is performed for the model parameters. In principle, a model with parameters estimated from the design set, from which synthesized data retains a high degree of similarity with observations of validation set, is the most accurate and desired model. Models that meet this criteria feature an appropriate level of complexity.

When utilizing data to build models one has to consider how appropriate and accurate the samples are, whether the samples represent the quantity modeled, and whether they are clean or soiled with noise. Here the entity measured (the daily amount of precipitation) is fairly accessible and all statistical analysis is based on the assumption that the data is not soiled with noise. However, data for individual weather stations is partially missing for long periods of time, which obviously reduces the representativeness of the sample. In Papers [C] and [D], we assume that the data is missing completely at random. That means that events that lead any particular station and/or specific time to feature incomplete data are independent of the precipitation process. Note in particular that the extreme value analysis may severely suffer from missing data since the data set of extremes is typically small even for samples without missing observations. In addition, the possibility of extreme weather affecting the devices of measurement in such a way that a malfunction is caused, which then, in turn, induces missing data, is not extremely small, but in these Papers is completely disregarded.

Motivation for Paper [A]

The concepts and ideas applied in Paper [A] originates from game theory. Game theory is a branch of mathematics devoted to the logic of decision making and has been widely applied in research areas such as psychology, economy, evolution and biology. We extend this list to also include the field of decision making in team sports.

We are particularly interested in zero-sum two-player games which are defined by a structure consisting of three objects:

- two decision makers, here called team A and team B,
- multiple ways of acting, called strategies, for each team such that the outcome of the game depends upon the strategy choices,
- well defined preferences among the possible outcomes quantified by a utility function, u, such that the utility, or payoff, for team B is the negative of the payoff for team A.

A significant property of game theory models is that each participant only has partial control over the outcome. The principal objective is to determine what strategies the teams ought to choose in order to pursue their own interests rationally. Equivalently, it can be seen purely as an instrument for avoiding inconsistency in decision making. Given that an outcome for team A features higher expected payoff conditional on a specific strategy, then the rational decision for them is to stick with this strategy. Any other choice would be inconsistent with the predefined preferences defined by the utility function.

In order to find solutions that are stable in some sense, alternative approaches have been published. The standard such solution is the Nash equilibrium, which is the pair of strategies, such that any one-sided deviation will not increase the expected payoff for the deviator. It can also be shown that the Nash equilibrium in a zero-sum two-player game is the strategy that replicates the mini-max of expected utility for team A and correspondingly the negative of maxi-min of expected utility for team B. In other words, the optimal strategies are the ones that minimize the expected payoff of the best alternative for the opposing team.

In Paper [A], we analyze general situational tactics in team sports. More precisely the subset of team sports where two opposing teams each have a goal to defend, and the team that scores the most points win. In games like these, there is an almost magical aura surrounding game intelligence. Some players, that repeatedly pick the winning strategies, are considered blessed by it as a natural ability, while less prosperous players struggle to learn the ability to pick winning strategies over the course of their entire careers. Usually, the only unifying factor regarding game intelligence is accredited to experience, where there is no clear definition of what kind of experience is alluded to. In general, we mean that a player's overall ability can be categorized into two conceptually different parts. First, the ability to decide on a strategy, which is in some sense, optimal in each encountered game situation. Secondly, the ability to carry out the chosen strategy. The second category is decided upon by the player's skill set - technique, strength, agility, endurance, etc., while the first category is what defines the game intelligence of the player. In Paper [A], we focus on analyzing the first category.

In order to analyze specific game situations, we define a natural utility for the outcomes – the potential. The potential is the probability that team A will score next, minus the probability that the next goal will be scored by team B. Further, we analyze how players should act based on formal reasoning alone in order to be consistent regarding potential. This reasoning gives a strategic advantage in competitive interactions and hence it is closely related to game intelligence. In this setting, we need to recognize the fact that one can never know which choice would have resulted in the best outcome in each particular instance. However, we show that we can find strategies which are optimal in the mean. As a result, we claim that the optimal strategies correspond to game-intelligent strategies. Hence game intelligence is an acquired skill, and in particular consistent behavior according to the potential is equivalent to high game intelligence.

Description

We set up a game theoretical framework to analyze a wide range of situations from team sports. By using the potential as utility function, we model specific game situations, and solve these using standard game theoretic methods. The first part of the Paper gives the theoretical foundation underlying our analysis, including the concept of potential fields, and derives some results with applications to game intelligence. The following sections focus on game situations where the players make decisions based on a given set of strategies. Here, we apply principles from game theory to determine which decisions are optimal. A main consequence of our problem set-up is that the optimal defensive strategy is to make the best offensive choice, in terms of potential, as low as possible. Further, the optimal strategy for the offense is to distribute shots between the players so that all players take all shot opportunities that have a potential larger than a certain threshold. This threshold is the same for all players. It is important to note that the optimal strategy does not guarantee a successful outcome on each occasion. Rather, the optimal strategy for a specific situation gives the best outcome in terms of potential. We develop categorical as well as continuous models, and obtain optimal strategies for both offense and defense.

My contribution

My co-advisor – Carl Lindberg – was very convincing when proposing the idea of scientifically investigating decision making within sports. In particular, he meant that the former ice hockey player Nicklas Lidström played according to a different algorithm compared to other players. By reversed engineering, when analyzing his strategies we found that he solved the problem of maximizing utility in terms of the potential, a fact that he confirmed when questioned personally about his strategies.

Together we worked on unfolding the nature of winning strategies in a mathematical setting and ended up with Paper [A].

Due my background as a professional athlete, I have a profound understanding of key aspects in team sports. Even so, I was surprised to gain several new insights concerning fundamental game intelligence, altering my previous comprehension of the game. Old ”truths”, that had gone unchallenged during my career, suddenly appeared completely irrational.

Motivation for Paper [B]

A portfolio is a grouping of financial assets such as stocks, bonds and cash equivalents. By allocating the investments within a portfolio an investor may maximize the momentary return, minimize the risk of loss of wealth over a day or month, or optimize according to any other criteria. In general, the fundamental concepts of high return and low risks are conflicting, and there exists no perceptionless optimal portfolio.

Classical portfolio optimization was founded by Markowitz (1952) where the conflicting fundamental concepts were quantified. Markowitz operated in a one-period discretized setting and his optimal portfolio is also known as the mean-variance portfolio since it is based on the means and covariances of the various assets. Since then, the field has been generalized and sophisticated stock market models have been proposed where stock prices are modeled as random processes, i.e., time dependent random variables. A particularly successful stock market model is the Black Scholes model, where stock prices are given by a geometric Brownian motion, i.e., the continuous time version of an exponential random walk. If W_t is a Brownian motion and S_t is the price of the stock at time t, then the dynamics of a stock in the Black Scholes market is defined by the stochastic differential equation

$$dS_t = S_t(\mu dt + \sigma dW_t), \tag{1}$$

where in particular the volatility, σ is constant. In a continuous time setting, also continuous re-allocation of the portfolio may be performed and we need the concept of self-financing strategies; any investment strategy that only invests the specific value of the portfolio at each time point is called a self-financing strategy. In this setting, Merton (1969) introduced the concept of utility to be able to solve the problem of optimal re-allocation strategies. The utility function assigns the experienced profit by an extra token (monetary unit). Most investors assign more profit to the first won token and assign less profit to each extra token than to the one that went before. This behavior is called "risk averse". In particular, the optimal continuously rebalanced self-financing portfolio given by applying the negative exponential utility function in a Black Scholes market is equivalent to Markowitz mean-variance portfolio re-allocated in continuum. This means that investing in the mean-variance portfolio over short time horizons is actually equivalent to the long term investment strategy defined by the optimal allocation according to the utility function of negative exponential.

There are many upsides for the widely applied Black Scholes model but there are also some limitations. For example, by the market dynamics (1), the log-returns

$$\log \frac{S_t - S_{t-1}}{S_{t-1}}$$

are independent and identically normally distributed, a property that generally fits poorly with observations of true markets. Typically, log-returns of the market feature heavy tails and large observations, either positive or negative, tend to group. In order to form a general model, that incorporates the properties mentioned, Barndorf-Nielsen and Shephard (B-NS) introduced a stochastic volatility market, see Barndorff-Nielsen and Shephard (2001), where specifically the volatility, σ, is driven by a non-Gaussian Ornstein-Uhlenbeck process. In this model the log-returns are time dependent, such that large observations, positive or negative, have higher probability to occur in groups. Further, the tails of the marginal distribution of the log-return in B-NS model are heavier than in the corresponding Black Scholes model. In Lindberg (2006) the optimal allocation problem for a portfolio was solved for a multi-stock B-NS market.

A benchmark is any item used to mark a point as an elevation reference. In the financial industry, a benchmark is the performance of a predetermined set of securities, used for comparison purposes. The dominating approach in the industry is to measure performance in excess of an externally given benchmark index, the so called alpha, at a deterministic future time. Classical portfolio optimization literature, see e.g., the aforementioned Merton (1969) and Lindberg (2006), focus on the problem of maximizing expected utility of wealth, called beta. While the classical concept of measuring results in beta may be an intuitive problem setting it is not put into practice in the real world. Measuring performance in alpha better captures the skill of the individual investor. A single investor hardly dictates the performance of the entire market and absolute wealth depends directly on overall market fluctuations. Instead relative wealth, i.e., the difference of wealth between the portfolio and the benchmark index, is the interesting unit of measure for industry practitioners. It gives an isolated measure of the investor's performance which does not depend on the underlying market. In order for relative wealth to be independent of the benchmark, it is necessary that the benchmark is continuously rebalanced. The level of wealth of the benchmark is set to be equal to that of the personal investor. This could, in principle, be done as often as every day, but for practical reasons it is often done when the wealth between the personal investor and the wealth of the benchmark has crossed some predetermined upper or lower level. The adaptation of the benchmark wealth is made to avoid the situation where the investor's performance will be affected by the so called beta effect. The term beta effect is used to describe the situation when the performance of the investor, in terms of alpha, starts to be affected notably by the absolute performance of the benchmark. This happens if the investor's capital is considerably larger, or smaller, than the capital held in the benchmark. Further, this implies that we are evaluating performance in excess of a non-self-financing portfolio. It sums over the difference between the investor's daily profit minus the daily profit of the current benchmark. Hence, the concept of relative wealth is merely an abstraction, but nevertheless an industry ubiquity.

In practice, the financial industry standard procedure is to rebalance one-period mean variance portfolios over short consecutive time horizons. It has been unknown to what extent this "local" optimization approach actually yields good results even in the long run. Recently Korn and Lindberg (2013) solved the problem in a Black Scholes market. In Paper [B] we consider the corresponding problem in the B-NS market. We show that the optimal portfolio in terms of exponential utility of relative wealth in a B-NS market replicates the optimal portfolio of the corresponding Markowitz mean-variance problem in continuum. That is, by continuously rebalancing one-period benchmark relative mean-variance solutions one replicates the optimal portfolio for an investor maximizing expected utility of terminal wealth. This is actually completely analogous to Merton (1969), and the differences between his Paper and Paper [B] are small (aside from the fact that we use a stochastic volatility model). Mainly, the difference is that Merton aims at finding an optimal beta portfolio while in Paper [B] we consider an optimal alpha portfolio. Merton considers strategies as being fractions of wealth and his optimal strategy - the local mean-variance strategy the investor should apply continuously - has the constraint that the sum of all portfolio weights should be equal to one. We, on the other hand, view the strategies in terms of capital, and use the constraint that the sum of all portfolio weights should equal zero, i.e., that the net exposure relative to the benchmark should be zero. With Merton's problem, the optimal strategy amounts to solving the stochastic control problem of maximizing expected utility of terminal wealth. Analogously, for our problem the optimal strategy amounts to maximizing the expected utility of terminal wealth in excess of the benchmark.

By solely considering relative wealth, it is natural to ask whether we have almost sure non-negativity of the investor's total wealth. The answer to this question is affirmative, since one is able to choose the bounds on the portfolio weights in such a way that the investor's total portfolio holdings remain positive. In practice, this constraint is often active for stocks that have small index weights.

Description

In order to deduct the results of Paper [B], we mimic standard literature and set up the market specific stochastic control problem by the Hamilton-Jacobi-Bellman (HJB) equation. Further, the technical part of the Paper consists of proving that the optimal value function is well defined. Then we proceed by setting up a tailor-made verification theorem that shows the uniqueness of the solution, given that it is well defined. Furthermore, we suggest an explicit optimal value function, a solution to the optimization problem, and devote considerable effort showing that it is well defined and actually solves the HJB equation. The presented optimal value function is then represented as a conditional expected value in Feynman-Kac form.

My contribution

My co-advisor, Carl Lindberg, proposed the problem of finding an optimal alpha portfolio in a Barndorf-Nielsen and Shephard market. I formatted the notation, and created the proofs, with the close supervision of Carl which resulted in Paper [B].

Motivation for Papers [C] and [D]

The distribution of precipitation is a key research area on both a national and an international level. For example, it is clear that the effects of drought or anomalously wet weather conditions may have devastating consequences on agriculture. However, the entire distribution of rainfall is also of great interest and therefore, realistic sequences of meteorological variables such as precipitation are key inputs in many hydrologic, ecologic and agricultural models.

The physical principles that govern climate, and in particular precipitation, are well studied. These principles may be represented in mathematical language as differential equations see e.g., Das et al. (2014). In meteorology, where only a short time span is considered, these differential equations are used to predict precipitation. However, the equations feature high complexity which has the consequence that solutions deteriorate over time. This, in turn, results in the situation where variability and uncertainty for predicting or modeling precipitation during longer time periods, e.g., over a month or a year, grows to an unmanageable extent. In order to model precipitation for longer time spans or when historical records are of insufficient duration or inadequate spatial and/or temporal coverage, the precipitation process is usually considered a stochastic process. In these cases, synthetic sequences may be used to fill in gaps in historical records, to extend the historical record, or to generate realizations of weather that are stochastically similar to the historical record. Whereby, here the actual physical processes that are influencing the climate are of less importance than what they accomplish, that is, the daily amount of precipitation.

A weather generator is a stochastic numerical model that generates a daily weather series with the same statistical properties as observed in a real world weather series. In this setting, an appropriate stochastic generator is one in which the synthesized data replicates the observations in a weak sense, by the distributions of different precipitation measures e.g., probability of precipitation on a specific day of the year or correlation in the amount of precipitation between locations, etc. In contrast to meteorology, which focuses on short term weather systems, interest lies in unfolding the structure of the underlying process. When watching the news, one is usually more interested in knowledge about the opposite – the prediction of precipitation in the very near future.

However, even if the measured information of interest (i.e., the amount of precipitation) is a fairly accessible entity, the actual building of a weather generator is made difficult by the long time periods and the complex processes that govern it. From a statistical point of view, modeling of precipitation is complicated by the semi-continuous nature of the process. Distribution of daily amounts of rainfall have a mass at zero, indicating a dry day with no precipitation, and a continuous distribution for wet days, i.e., days with a positive amount of rainfall. These properties make the, otherwise standard, normal assumption, inappropriate. In addition, rare events of extreme precipitation have a tremendous impact on the community. Further, any model attempting to properly represent the precipitation process needs to replicate the dry/wet day behavior and the extreme amounts of daily precipitation. Throughout both these Papers emphasis is put on these fundamental features of the process.

Description

The aim is to create a weather generator that can produce realistic sequences of daily precipitation for sites in Sweden. Here, ”realistic” is set into the statistical framework in the sense that the artifical data should replicate different measures of precipitation such as mean intensity, monthly maximum and length of number of consecutive dry or wet days, etc. In order to quantitatively measure weather, the Expert Team of World Meteorological Organization/Climate Variability and Predictability, see Peterson et al. (2001) and Karl et al. (1999), have defined a standardized set of specific weather indices. These indices make a comparison of extreme weather conditions possible across regions. The models proposed in these Papers are also validated by their performance in regard to these weather indices.

A common feature for both Papers is that the dependence structure is separated from the marginal distributions. For the marginal distributions, and in particular for the extremeal part, there exists a well-developed theory. In both Papers, we adopt a peak over threshold approach to model the extreme amounts of rainfall. By the extremal theorem, see e.g., Coles (2001) (Theorem 4.1), there is only one possible parametric distribution modeling the conditional probability for the excesses over a high enough threshold. This is a very useful property, but note that it is only applicable above some inexplicit threshold. This implies that we may have an elementary decision to accurately pick a suitable model, but meanwhile the data to design the parameters may be sparse which hampers the certitude of the point estimation.

The dependence structure of the precipitation process and field (Paper [D]) is potentially a very complex mathematical object. In both Papers we model this by a Gaussian structure. By adopting this structure we derive a model that is both simple and general, producing very good results in terms of weak measures.

Paper [C]

In Paper [C] the temporal variability of the precipitation process is the pivotal point. The applied statistical model is a chain-dependent model which consists of two conceptually separated parts. The first part is a model for the dichotomous sequence of wet or dry days, for which we employ a multiple order Markov chain. The second part of the model gives the intensity of precipitation given there is a wet day. The intensities are modeled by a composite model for the marginal distribution that incorporates the empirical distribution together with an extreme value distribution for high amounts of rainfall. Further, for the temporal dependence of the intensities we introduce a Gaussian copula, implicitly modeling the process as a one-step auto-regressive process.

Here, the mathematical part revolves, in large part, around Markov processes. Briefly put, a Markov process is a process that features the property of amnesia: The evolution of the process is independent of the past conditioned on the present state.

A multiple m-step Markov chain is a process with the memory property strengthened. Conditional on the m last steps the evolution of the process is independent of the history, excluded the m last steps, of observations. The Markov property is very useful in terms of explicitly finding the risk or chance of interesting events such as drought, consecutive dry days, and anomalously wet weather.

The temporal dependence structure of intensity of precipitation is modeled by a Gaussian copula. Where, instead of estimating the parameter for the copula directly from the data (a computationally very hard assignment) we sidestep potential numerical issues by means of using the underlying empirical ranks. Under the assumption that the Gaussian copula governs the temporal dependence, by transforming the empirical ranks to their corresponding standard normal quantile the transformed bivariate random variables of consecutive amount precipitation in wet days are bivariate normally distributed. Estimating the parameter for the Gaussian copula is then reduced to the elementary task of estimating the correlation coefficient of a sample of bivariate (normally distributed) random variables.

The validation part is performed by a Monte Carlo approach. Artificial precipitation is simulated, which is the element-wise product of the dichotomous Markov chain and the transformed Gaussian auto-regressive temporal process, and distributions of the weather indices are computed. The stochastic generator replicates the observed indices at a satisfactory level.

The upside of the proposed model, besides that it replicates weather indices at a satisfactory level, is that we both accurately and easily can model the risks of long dry- and wet spells. The drawbacks of the model is that additional features such as the intermittence effect (smooth transitions of intensity of precipitation between wet and dry days and that given there is anomalously wet weather then the risk of an excess of precipitation is higher than during longer periods of dry days) are not represented at all. The model proposed in the Paper also implicitly constrains the model to be of an auto-regressive form, implying that the dependence of intensity decays quickly with the time lag. In addition, the substantially most considerable limitation of the model is that it features no venture to extend into a spatial dependence structure.

Paper [D]

In Paper [D] the entire spatio-temporal variability of the precipitation process is investigated. The aim is, as in Paper [C], to build a stochastic number generator that replicates synthesized data with corresponding dependence structure and appropriate marginal distributions. Note here that by introducing spatial dependency, the level of complexity is many times doubled. For example, in contrast to time domain, there is no natural spatial ordering and two neighboring sites may influence a third neighbor differently at various times.

The model is constructed of a transformed Gaussian field, which models both the dry/wet behavior and the intensities of the precipitation. The observation of a negative value of the field at a location in a specific day corresponds to a dry day at the specific point in space. For positive values of the field, the intensity of precipitation is given by the location specific transformation of the field value at the location. The key intention of the model was to extend the successful copula approach from Paper [C] to cover the entire process, not just the intensity part. By this approach we also take the intermittence effect into account (the property of higher probability of large amount of precipitation given that close points in space and/or time feature wet weather). Further, by a model like this we can replicate the property of the precipitation field's smooth transitions between shifts of weather regimes in terms of start and ending of time periods or spatial domains with drought or anomalous wet weather, which is a frequently occurring feature of weather data.

We model the location specific marginal distributions as a composite model of a gamma distribution, with temporal dependent parameters, and a generalized Pareto distribution, with stepwise constant parameters, for the excesses of a high threshold. Mean intensities, qq-plots and weather indices concerning extreme amounts illustrates a very good fit of the model.

We use empirical transforms of the marginals to find observations of the censored latent field. Then we estimate the covariance structure pointwise, a task that is computationally demanding since the field is censored. In order to perform this estimation we introduce new methodology to estimate the underlying covariance of censored Gaussian variables. Then we model the dependence structure by a non-separable parametric model. A non-separable covariance structure implies that the dependence of temporal lag and spatial lag cannot be separated – a property the data gave clear indications of.

Model validation is based on cross-validation, where simulations show that the model replicates historical data well, by means of statistical test with respect to fundamental measures and the established weather indices. Hence, the dependence structure of the precipitation process is modeled with a satisfactory level of similarity with observed data.

My contribution

The problem of creating a weather generator, together with suggestions for suitable data sets, was proposed to me by environmentalist Deliang Chen. Regarding Paper [C], my head advisor, Patrik Albin, got me started by suggesting dividing the precipitation process into two parts and specifically modeling the dry/wet behavior part as a multiple step Markov chain and the intensity as a combination of empirical distribution and an appropriate extreme value distribution. By the fruitful tutoring of my second co-advisor, Anastassia Baxevani, we moved the problem forward and ended up with Paper [C] at hand. Except for the above mentioned crucial advisory contributions, everything in this thesis has been done by myself, the undersigned.

Regarding Paper [D], Anastassia proposed the problem and has consistently steered me in the direction of our final destination. Her contribution is highly appreciated in terms of forcing me to find suitable framework and scientific foundations to the proposed ideas. However, the idea of modeling the precipitation as driven by a censored Gaussian variable as well as the programming and building of the mathematical model were done by myself.

Bibliography

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-gaussian ornstein-uhlenbeck-based models and some of their uses in financial economics. *J. R. Statist. Soc B. 63, Part 2*, pages 167–241.

Coles, S. (2001). *An introduction to statistical modeling of extreme values.* Springer-Verlag, ISBN 1-85233-459-2.

Das, S., Devi, M., and Talukdar, P. H. (2014). Model computation reliability of precipitation over ne india using parallel computing technique: A comparative assessment between varsha, imd and trmm derive precipitation. *International Journal of Intelligent Computing in Science Technology*, 1(1).

Karl, T., Nicholls, N., and Ghazi, A. (1999). Clivar/GCOS/WMO workshop on indices and indicators for climate extremes: Workshop summary. *Climatic Change*, 32:3–7.

Korn, R. and Lindberg, C. (2013). Portfolio optimization for an investor with a benchmark. *Decisions Econ Finan DOI 10.1007/s10203-013-0148-8.*

Lindberg, C. (2006). Portfolio optimization and a factor model in stochastic volatility market. *Stochastics: An International Journal of Probability and Stochastics Processes*, 78(5):259–279.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.

Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: the continuous time case. *Rev. Econ. Stat.*, 51:247–257.

Peterson, T., Folland, C., Gruza, G., Hogg, W., Mokssit, A., and Plummer, N. (2001). Report on the activities of the working group on climate change detection and related rapporteurs 1998-2001. *World Meteorological Organisation, WCDMP-47, WMO-TD 1071.*

Appended papers

A Lennartsson, J., Lidström, N., and Lindberg, C., Game intelligence in team sports (2014).
(Submitted).

B Lennartsson, J. and Lindberg, C., Merton′s problem for an investor with a benchmark in a Barndorff-Nielsen and Shephard market (2014).
(Submitted).

C Lennartsson, J., Baxevani, A. and Chen, D., Modelling precipitation in Sweden using multiple step Markov chains and a composite model, *Journal of Hydrology* (2008), Volume 363, Issue 1-4, Pages 42-59.

D Baxevani, A. and Lennartsson, J., A Spatio-temporal precipitation generator based on a censored latent Gaussian field (2014).
(Work in progress).