my Technical Toolbox

An Overview

Tarak Kharrat[1, 2]

[1]University of Liverpool, London Campus, UK.
[2]Kickdex Limited, Lodon, UK.

April 23, 2018

Abstract

The main motivation of this document is to give the reader an overview of my technical toolbox and the projects I have been involved in. It is meant to complement my (short) resume.

1 Research Activity

I am currently a (part-time) research fellow at the University of Liverpool where I undertake research projects that fall under the umbrella of *'Computational Statistics'*. In this context, together with colleagues, we developed methods and software to solve problems in:

- **Counting Processes:** We developed a new family of counting processes that generalise the standard Poisson and negative binomial models. These models have the nice property to allow fitting over as well as under-dispersed data. The theoretical properties are discussed in Baker and Kharrat (3). An `R` package R Core Team (17) has been published on CRAN and described in a dedicated paper (14).
- **Time Series:** We developed new methodologies to fit non-Gaussian non-linear state space models. The methods are implemented in the `R` package `GKF` (for internal use only).
- **Statistical Distributions:** We created new estimation techniques for the 4 parameters of Stable law distributions. These methods converge faster and still enjoy the normal asymptotic properties. This work is discussed in (13).
- **Survival Analysis:** I have been involved in some research with time to event data. In particular, I have a large experience with competing risks models and their multi-state generalisation.

2 Sports Analytics

As co-head of R&D at Kickdex Limited, a start-up specialised in predictive modelling in football, I developed many tools and metrics to analyse, in a purely quantitative way, the game of football. Also the applied side of my research is dedicated to football. In fact, some chapters in my PhD thesis (12) discuss some models suggested mainly to forecast the match score grid. In the rest of this section, I detail some of the projects I worked on in this area:

- I created a unique database (from scratch) by web-scrapping the internet. These database contains event-by-event data (similar to opta F24), video game players information (EA

Sports FIFA and Konami PES), betting prices (both pre-match and in play for different markets) and historical injury record. The database is updated weekly and is stored on the mongo-Atlas cloud.

- My primary focus at Kickdex is on predictive models. We developed, together with a colleague, an industry leading forecasting model which is used by a multi-millions betting syndicate. Although technical details cannot be disclosed, I can say that we built families of basic models using standard machine learning classifiers (Random Forest, boosted trees, support vector machine, neural networks, k-nearest neighbours, naive Bayes, ...) and other models using more classic statistical techniques based on counting processes, copula and survival analysis (for example, you can see this paper (4)). These basic models have been combined in an *ensemble* which is know to perform better than any single model taken individually. We also leveraged techniques such as multi-task and transfer learning. Besides, some effort has been spent on feature engineering, feature selection, model performance testing ...
- I also did some work on players evaluation. In this context, I created several metrics under the *REAL Analytics* (RA) label which is meant to help football clubs use sounds mathematics to answer relevant questions for the game of football:
  - *Plus-Minus Rating*: How important is the player for his team? The theory has been published in a paper (15). In particular, we measure the player's importance in terms of goal differential (`PM`), expected goal differential (`xGPM`) and expected points differential (`xPPM`).
  - *valuing actions*: We developed an algorithm to compute the contribution of every action in football to the probability of scoring/conceding goals. This algorithm allow us to measure the importance of players' action and to derive two objective ratings: *performance* rating to answer the question how did the player perform in a specific match? and *overall* rating to answer the question how good is the player right now (in terms of football skills)?
  - *potential*: It is important to know how good a player will be in the future leveraging some modern time series forecasting techniques. We developed an algorithm to project the *overall* rating in the future. Depending on the player's age and his recent performance, we can estimate how good he is likely to be.
  - *reliability*: The RA reliability index estimates how often a player is available to play. It takes into account information such as historical injury and red card records and gives an estimate of the probability of a player being available for selection of a given random match.
  - *players' likely impact*: Together with a colleague at Salford business school, we developed a set of algorithms labelled *Sports Analytics Machine* (SAM). The BBC is a prime user of SAM and among other things, SAM is able to compute the impact of new signing in a team. In fact, including the new signing in the squad and simulating the league path allow us to objectively quantify the change of probability (with and without the player) of winning the league, finishing in champions league positions ... An example of application is given here.

All these ratings and tools have been exposed in a web-app I maintain (data updated on regular basis). The different functionality are explained in this introductory-video.

- **Tracking-data**: we have been mandated by a company (name cannot be disclosed) to compare the quality of different tracking data (Prozone, inStat, TRACAB, STATS). This project allowed us to familiarise ourselves with these new generation of data and to prototype a new model to extend our *valuing actions* model which used only event-by-event ball data. However, the confidentiality agreement and the small sample size didn't allow

us to publish our findings.

3 Algorithmic Trading

- solid experience designing, back-testing and deploying trading strategies in the US equity and football betting markets.
- good experience working with high frequency (tick-by-tick, seconds, minutes ...) price and volume data.
- ability to discuss, explain and present trading results to (non technical) stake holders audience.

4 Consultancy

Over the past 10 years, I have been involved in several consultancies as a team leader or a technical expert in a specific subject:

- *TOTAL (Oil and Gas Trading & Shipping)*: tested and improved the STAGE simulator, a software to simulate ships trips between loading and unloading terminals taking into account real life constraints (weather, traffic, dry-dock, ...).
- *major UK bookmaker*: auditing and improving the in-house forecasting models for weak leagues and helping optimising the cash-out algorithm.
- *Atomic Weapons Establishment*: provided an algorithm to solve a dynamic Poisson regression problem.
- *Thales*: created a software to model a gap in the coating of a submarine (9).
- *UEFA*: developed Aalgorithms to detect fixed games (joint work with sporting index).
- *Nottingham Forest F.C*: specific statistical reports on a list of target players.
- *BBC sports*: different applications of SAM (most likely score forecast, end of season league table simulations, likely impact of a signing, players importance ...).
- *major tracking data provider*: compare the accuracy of different tracking data sources objectively (on going).
- *major UK law firm*: estimate the likelihood of the client (a footballer who got injured in 2012) to make it to the top level in England (on going).

5 Programming

- **Compiled languages**: strong knowledge of C++:
  - contributed to the oomph lib library (8): author of the Helmholtz module.
  - good working knowledge of the Armadillo (18) and Eigen (10) libraries for linear algebra, NLopt (Johnson et al.) for nonlinear optimisation and openMP (7) for parallel computing.
- **Interpreted languages**:
  - expert knowledge in `R`: author of several packages on CRAN.
  - strong experience interfacing C++ code from R (to improve code performance).
  - ability to build comprehensive web apps using shiny and shiny dashboard.
  - basic understanding of python (learning it at the moment).
  - intermediate knowledge of Julia.
- **Machine learning**:

  - solid understanding of machine learning algorithms for classification, regression and clustering.
  - solid working experience with ML libraries such as h2o.ai (5) and keras (6)(R interface with CPU and GPU backend).
  - extensive experience with the `caret` R package (16) for model testing.
  - ability to deploy models in production environment leveraging new technologies such as Ducker, version control (git) and conda manager.
- **Database**:
  - ability to build, maintain and store large amount of data using MongoDB (no sql) either locally or on the cloud (using mongo-Atlas).
- **Parallel computing**
  - good experience with multithreading programming (using the foreach package (1) in R or openMP (7) in C++) as well as GPU computing (using the RCUDA package (2)).

References

[1] Analytics, R. and Weston, S. (2014). doparallel: Foreach parallel adaptor for the parallel package. *R package version*, 1(8).

[2] Baines, P. (2014). Rcuda: General programming facilities for gpus in r.

[3] Baker, R. and Kharrat, T. (2017). Event count distributions from renewal processes: fast computation of probabilities. *IMA Journal of Management Mathematics*.

[4] Boshnakov, G., Kharrat, T., and McHale, I. G. (2017). A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466.

[5] Candel, A., Parmar, V., LeDell, E., and Arora, A. (2016). Deep learning with h2o. *H2O. ai Inc.*

[6] Chollet, F. et al. (2015). Keras: Deep learning library for theano and tensorflow.(2015).

[7] Dagum, L. and Menon, R. (1998). Openmp: an industry standard api for shared-memory programming. *IEEE computational science and engineering*, 5(1):46–55.

[8] Heil, M. and Hazel, A. L. (2006). oomph-lib–an object-oriented multi-physics finite-element library. In *Fluid-structure interaction*, pages 19–49. Springer.

[9] Heil, M., Kharrat, T., Cotterill, P. A., and Abrahams, I. D. (2012). Quasi-resonances in sound-insulating coatings. *Journal of Sound and Vibration*, 331(21):4774–4784.

[10] Jacob, B. and Guennebaud, G. (2012). Eigen is a c++ template library for linear algebra: Matrices, vectors, numerical solvers, and related algorithms.

[Johnson et al.] Johnson, S., Joannopoulos, J., and Soljačić, M. Nlopt library.

[12] Kharrat, T. (2016). *A Journey Across Football Modelling with Application to Algorithmic Trading (PhD Thesis)*. University of Manchester.

[13] Kharrat, T. and Boshnakov, G. (2015). StableEstim: An r package for estimating the stable laws parameters and running monte carlo simulations. *Journal of Statistical Software*.

[14] Kharrat, T. and Boshnakov, G. (2018). Flexible regression for count data based on renewal processes: The Countr package. *Journal of Statistical Software*.

[15] Kharrat, T., McHale, I. G., and Lopez Pena, J. (2017). Plus-minus player ratings for soccer. *Journal of Machine Learing*.

[16] Kuhn, M. et al. (2008). Caret package. *Journal of statistical software*, 28(5):1–26.

[17] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[18] Sanderson, C. and Curtin, R. (2016). Armadillo: a template-based c++ library for linear algebra. *Journal of Open Source Software*.