Information in Local Curvature: Three Papers on Adaptive Methods in Computational Statistics

by

Berent Ånund Strømnes Lunde

Thesis submitted in fulfilment of
the requirements for the degree of
PHILOSOPHIAE DOCTOR
(PhD)

University
of Stavanger

Faculty of Science and Technology
Department of Mathematics and Physics
2020

University of Stavanger
NO-4036 Stavanger
NORWAY
www.uis.no

ISBN:
ISSN:
PhD: Thesis UiS No.

Preface

This thesis is submitted in partial fulfilment of the requirements for the degree of Philosophiae Doctor (PhD) at the University of Stavanger, Faculty of Science and Technology, Norway. The research has been carried out at the University of Stavanger from September 2017 to August 2020.

The present work is divided in two. The first part gives a brief introduction and background information to the most important topics and ideas of the work. The second part consists of the following papers:

Paper I

Lunde, Berent Ånund Strømnes, Tore Selland Kleppe, and Hans Julius Skaug (2020). Saddlepoint adjusted inversion of characteristic functions. *Journal of Journals* 57, 80-93.

Paper II

Lunde, Berent Ånund Strømnes, Tore Selland Kleppe, and Hans Julius Skaug (2020). An information criterion for automatic gradient tree boosting. *To be submitted for publication in Journal of Journals.*

Paper III

Lunde, Berent Ånund Strømnes, Tore Selland Kleppe, and Hans Julius Skaug (2020). aGTBoost: Adaptive and Automatic Gradient Tree Boosting Computations. *To be submitted for publication in Journal of Journals.*

Acknowledgements

I would like to thank my supervisor, Professor Tore Selland Kleppe, for his constant support and invaluable guidance. You have allowed me to pursue ideas, to fail, and so many times steered me in the right direction with detailed feedback and questions that would uncover flaws, but which would eventually lead me closer towards truth. Thank you for your encouraging words, enthusiasm and genuine thoughtfulness.

Thanks are also due to my co-supervisors, the professors Hans Julius Skaug and Jan Terje Kvaløy. Professor Skaug has co-authored two of the papers in this thesis, allowed a stay at the University of Bergen, and always shared of his time and wide experience, for this I am grateful. Professor Kvaløy has been presence of constant cheerfulness and inspiration. Thank you for always taking a genuine interest in people and their ideas. I extend my thanks to my fellow PhD-students. Utmost appreciation goes out to Kjartan Kloster Osmundsen and Birthe Aarekol, for the many discussions and several trips all around the world to academic conferences and meetings.

Finally, I want to give many thanks to my friends, in particular Kjetil for his unlimited accommodating spirit. And also my family, in particular my mother, Katrin, for invaluable advice, and my wife, Saeron Min, which has been a force of continued support. You have let me dive into hours of silent thoughts, calculations and coding when it was needed, but have also pulled me away and forced upon me a more balanced life when I would encounter a runtime error of the mind, but of the kind which I would not be able to see or solve by myself.

Berent Ånund Strømnes Lunde
Stavanger, August 2020

Abstract

Advanced statistical computations have become increasingly important, as with the increased flexibility of models capturing complex relationships in new data and use-cases, comes increased difficulties of fitting procedures for the models. For example if the model is complex, involving multiple sources of randomness, then the probability density function used in maximum likelihood estimation typically does not have a closed form. On the other hand, in regression type problems the closed form of the conditional distribution of the response is often known. However, the relationship between features and response can be complex, high dimensional and is generally unknown, motivating non-parametric procedures with new sets of fitting problems.

This thesis explores techniques utilizing the local curvature of objective functions, and using the information inherent in this local curvature, to create more stable and automatic fitting procedures. In the first paper, a saddlepoint adjusted inverse Fourier transform is proposed. The method performs arbitrarily accurate numerical inversion, even in the tails of the distribution. This allow practitioners to specify their models in terms of the characteristic function which often exists in closed form. The second paper proposes an information criterion for the node-splits, after greedy binary splitting, in gradient boosted trees. This removes the need for computationally expensive cross validation and expert opinions in tuning hyperparameters associate gradient tree boosting. The third paper focuses on the implementation of the theory presented in the second paper into the R-package aGTBoost, and also builds on the information criterion to suggest an adjustment of ordinary greedy-binary-splitting, adapted to gradient tree boosting.

Table of Contents

Appendix

1 Introduction

Advanced statistical methods and procedures have seen increased and widespread usage in later years. This is backed by access to more data and new use-cases, cheaper computational power, and adaption into mainstream languages such as Python and R. Underlying this trend is also the increased usability of said algorithms, in regards to training on data and putting them into production. The goal of this thesis is to further the usability of computational methods in statistics with regards to stability, speed, and automatic functionality.

The main approach of the present work is to, in some loose and wide sense, approximate some objective function with a local quadratic approximation to either solve stability issues, create dynamic step-lengths, or measure the uncertainty of estimators. The hope is then that the iterative methods that the local quadratic approximation is applied to, will see an increased adaptivity to individual data and problems, and corresponding decrease in manual tuning performed before applications to the problem at hand.

The first part of this thesis will give a brief and informal introduction to the concepts and techniques that are used in papers I-III. The basis is the objective of maximum likelihood and supervised learning, which are presented in the first section. The second section introduces the local quadratic approximation, and showcase it in the relevant use-cases for papers I-III, i.e. maximum-likelihood numerical optimization, the saddlepoint approximation, gradient tree boosting, asymptotic theory and model selection. The final section summarises the papers of the thesis.

2 Maximum likelihood and supervised learning

Maximum likelihood estimation and supervised learning are briefly introduced in an informal manner. This is done to provide intuition to the fundamental objectives of the algorithms that are presented, and as motivation to the research on the algorithms's problems presented in this thesis.

2.1 Maximum likelihood estimation

Let $\mathbf{x}$ denote an n-dimensional vector of observations from a parametric distribution, with density denoted $p(\mathbf{x};\theta_0)$, where $\theta_0 \in \Theta$, $\Theta \subseteq \mathrm{R}^p$ is a p-dimensional vector. Is often the case that a reasonable parametric family of functions, $p(\mathbf{x};\theta)$, $\theta \in \Theta$, can be inferred from the problem and from inspection of the data. However, θ_0 will be unknown, and it is reasonable to estimate it using the observed data $\mathbf{x}$. To this end, maximum likelihood estimation is a popular approach. The maximum likelihood estimate (MLE) is the value of θ in Θ which maximizes the probability of the data, i.e. the likelihood,

$$\hat{\theta} = \arg\min_{\theta}\{-\log p(\mathbf{x};\theta)\}. \tag{2.1}$$

The maximum likelihood estimate, $\hat{\theta}$ is, under suitable regularity conditions, the asymptotically unbiased minimum variance estimate, and asymptotically normal. See -Van der vaart- for a treatment of their asymptotic properties.

2.2 Supervised learning

The supervised learning objective is perhaps easiest stated as "regression", but also bears resemblance to maximum likelihood estimation. Assume now that $\mathbf{x} \in R^{n\times m}$ is a matrix of p covariates or features for n observations. Let $y \in R^n$ be an n-vector of response observations.

In general, individual response observations, y_i, $i = 1 \dots n$, could also be multidimensional, but throughout this thesis they are assumed one-dimensional. Let $\hat{y}_i$ be a prediction for y_i and let the loss function $l(y_i, \hat{y}_i)$ be a function measuring the difference between a response and its prediction. The supervised learning objective is to find the best possible predictive function, $f(x) = \hat{y}$, which takes a feature vector (row-vector of $\mathbf{x}$) as its argument, and outputs a prediction $\hat{y}$. "Best possible" is here in reference to the loss l over observations not part of the training data $(\mathbf{x}, y)$. More formally, we seek f so that

$$\hat{f} = \arg\min_{f} \left\{ E\left[l(y^0, f(x^0)) \right] \right\}, \tag{2.2}$$

where the superscript (y^0, x^0) indicates an observation unseen in the training data, and E denotes the expectation. Notice that, if the search is constrained over a parametric family of functions indexed by $\theta \in \Theta$, and the loss function is taken to be the negative log-likelihood, $l = -\log p$, then the supervised learning objective is closely related to the objective of maximum likelihood estimation (2.1) in a regression setting. In fact, the objective in (2.1) is the sample estimator of the expected value in (2.2), but biased downwards in expectation, as evaluation is done over observations in the training set.

3 Quadratic approximations in statistics

The maximum likelihood objective (2.1) and supervised learning objective (2.2) are, except for the most trivial of cases, not straightforward, and must be solved numerically. This then typically involve some iterative algorithm, which may require substantial manual tuning and trial and error before successful application. However, a local quadratic approximations to some otherwise intractable function can often be of help in making these algorithms more automatic and adaptive to the data and problem at hand.

When referring to a local quadratic approximation, as is frequently done in this thesis, it is meant to refer to a 2'nd order Taylor approximation of a function $f(x)$, about some point x_0. For example, the quadratic approximation of the negative log-likelihood loss function $l = -\log p$ about some value of θ, say θ_k, gives

$$\begin{aligned} l(y_i, f(x_i;\theta)) &\approx l(y_i, f(x_i;\theta_k)) + \nabla_\theta l(y_i, f(x_i;\theta_k))(\theta - \theta_k) \\ &+ \frac{1}{2}(\theta-\theta_k)^T \nabla_\theta^2 l(y_i, f(x_i;\theta_k))(\theta-\theta_k). \end{aligned} \tag{3.1}$$

Example 3.0.1 (Newton-Raphson) *The MLE $\hat{\theta}$ in* (2.1) *typically has to be found numerically, as the score equations, $0 = \nabla_\theta l(y_i, f(x_i;\theta_k))$, are not possible to solve analytically. Assuming that l is differentiable and convex in θ, the Newton-Raphson algorithm will converge to the MLE $\hat{\theta}$. The iterative Newton-Raphson algorithm is constructed by employing the r.h.s. of* (3.1) *iteratively to the current value of θ, say θ_k, the next value in the iterative algorithm is then given by*

$$\theta_{k+1} = \theta_k - \left[\nabla_\theta^2 l(y_i, f(x_i;\theta_k))\right]^{-1} \nabla_\theta l(y_i, f(x_i;\theta_k)),$$

the MLE if l indeed was equal to the quadratic approximation on the r.h.s. in (3.1).

There are many problems in computational statistics that may be helped by (3.1), however only a few, the ones relevant to papers I-III, are

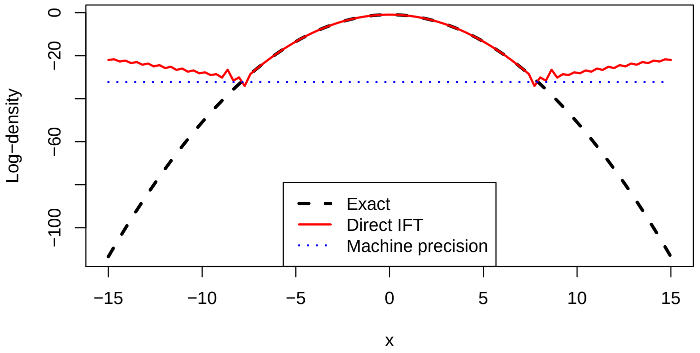

Figure 3.1: Figure included from Paper 1. Illustrating the dominance of inaccuracies of the IFT (3.2), calculated with quadrature, at machine precision at $\log(1.0 \times 10^{-14})$ indicated by the dotted horizontal line.

discussed here. The following sections discuss applications of local quadratic approximations as helpful tools in dealing with some of the problem associate/in dealing numerical optimization of (2.1) and (2.2).

3.1 The saddlepoint approximation

It is often the case that the density $p_X(x;\theta)$ of a random variable X, is not available in closed form when there are multiple sources of randomness present in X. Direct optimization of (2.1) is therefore difficult. However, the characteristic function or the Fourier transform of the density, $\varphi_X(s) = E[\exp(isX)]$ often still exhibits closed form, even in situations with more than one source of randomness. The density might then be retrieved by numerically evaluating the inverse Fourier transform

$$p_X(x;\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_X(s;\theta)e^{-isx}ds = \frac{1}{2\pi}\int_{\infty}^{\infty} e^{K_X(is;\theta)-isx}ds, \quad (3.2)$$

where $K_X(s;\theta) = \log \varphi_X(-is;\theta)$ is the cumulative generating function (CGF).

However, consider the case of numerical MLE optimization, for example by using the Newton-Raphson algorithm in Example 3.0.1. Here, at the first iteration, say θ_1, the initial estimate is likely to start at values far from the population MLE, θ_0. Necessarily, observations x will take place in low-density regions of $p(x;\theta_1)$, and this will continue to be the case at subsequent iterations, until θ_k is close to θ_0. This constitutes a problem to direct numerical inversion of (3.2) using quadrature schemes (weighted sum of integrand evaluations), as numerical inaccuracies related to the (binary) representation of floating-point numbers will dominate. More specifically, considering double precision at order 1.0×10^{-16}, if x is in a region with log-density $\log p(x;\theta_k)$ smaller than this value, the inaccuracy of the representation is sure to dominate. Even more is that such behaviour/pathologies in practice happens a few orders of magnitude higher than the theoretical limit given above. In Figure 3.1, the error dominates already at 1.0×10^{-14}.

An inversion technique that does not suffer from erroneous computations in low-density regions, and in fact is renown for its tail-accuracy, is the saddlepoint approximation (SPA) -cite daniels, Butler-. It is developed in paper 1 through an argument of exponential tilting, which takes place on the "time-domain" side of the Fourier transform. Complimentary, an argument on the "frequency-domain" side is given here, that closely follows the derivation in -citet Butler, chapter-. First, notice that the value of the integral is unchanged if we integrate through a line parallel to the imaginary axis, say τ,

$$p_X(x;\theta) = \frac{1}{2\pi} \int_{\infty}^{\infty} e^{K_X(\tau+is;\theta)-(\tau+is)x} ds \tag{3.3}$$

Now, apply the quadratic approximation (3.1) to the log-integrand, $K_X(\tau + is;\theta) - (\tau + is)x$, locally about the value of τ solving the saddlepoint equations

$$\hat{\tau} = \arg\min_{\tau} \{K_X(\tau;\theta) - \tau x\}, \tag{3.4}$$

henceforth called the saddlepoint. This then gives the approximation of the log-integrand

$$K_X(\hat{\tau}+is;\theta)-(\hat{\tau}+is)x \approx K_X(\hat{\tau})-\hat{\tau}x-\frac{1}{2}\frac{d^2}{d\tau^2}K_X(\hat{\tau})s^2. \tag{3.5}$$

Inserting this into the integral, and performing the transformation $u = \sqrt{\frac{d^2}{d\tau^2}K_X(\hat{\tau})}s$, then gives the SPA as

$$\begin{aligned} p_X(x;\theta) &\approx \frac{\exp(K_X(\hat{\tau})-\hat{\tau}x)}{2\pi\sqrt{\frac{d^2}{d\tau^2}K_X(\hat{\tau})}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}du \\ &= \frac{\exp(K_X(\hat{\tau})-\hat{\tau}x)}{\sqrt{2\pi\frac{d^2}{d\tau^2}K_X(\hat{\tau})}} = spa_X(x;\theta). \end{aligned} \tag{3.6}$$

The SPA (3.6) is often accurate, and is asymptotically exact in n if there is some asymptotic normality underlying X, for example for $X = n^{-1}\sum_i X_i$. In particular, its relatively fast computation if implemented using automatic-differentiation software to solve the inner problem (3.4), and its previously mentioned tail-accuracy, are properties that are highly tractable. On the down-side, the SPA does not integrate to one, except for in a few special cases. Also, the approximation is often unimodal even if the target density is multimodal, which could very well be the case when X consists of multiple sources of randomness. A common technique is to multiply the SPA with a constant value c, where $c^{-1} = \int spa_X(x;\theta)dx$, that ensure it is a density. This is immediately more computationally costly, require bespoke implementation, and does not solve the problems of unimodality. These problems are the subject of Paper I.

3.2 Gradient tree boosting

The idea behind gradient boosting emerged in Friedman (2001) and in particular in Mason et al. (1999) as a way to approximate functional gradient descent for the optimization problem in (2.2), similarly to

Algorithm 1 Second-order generic gradient boosting Hastie et al. (2001)

```
Input:
    - A training set $\{(x_i, y_i)\}_{i=1}^n$
    - A differentiable loss function $l(\cdot, \cdot)$
    - A learning rate $\delta \in (0, 1]$
    - Number of boosting iterations $K$
    - Type of statistical model $\mathcal{F}$
1. Initialize model with a constant value:
    $f^{(0)}(x) \equiv \arg\min_{\eta} \sum_{i=1}^n l(y_i, \eta)$
2. for $k = 1$ to $K$:
    $i$) Compute derivatives (3.9)
    $ii$) Fit a statistical model $\tilde{f} \in \mathcal{F}$ to derivatives using (3.10)
    $iii$) Scale the model with the learning rate
        $f_k(x) = \delta \tilde{f}(x)$
    $iv$) Update the model:
        $f^{(k)}(x) = f^{(k-1)}(x) + f_k(x)$
end for
3. Output the model: Return $f^{(K)}(x)$
```

If needed explaining text

how the Newton-Raphson algorithm from Example 3.0.1 solves the optimization problem in (2.1): Given an initial function $f^0(x) = f_0(x)$, one ideally seeks a function $f_1(x)$ minimizing

$$\hat{f}_1(x) = \arg\min_{f_1} E\left[l\left(y, f^0(x) + f_1(x) \right) \right]. \tag{3.7}$$

If this is difficult, a reasonable substitute to $\hat{f}_1$ is to find the functional derivative of this objective and add the negative direction to the model, say $f^1 = f_0 + f_1$, and then repeat the procedure until convergence at iteration K, which would yield the final model $\hat{f} = f^0 + \cdots + f^K$.

Difficulties arise to this procedure, as the joint distribution of (y, x) is generally unknown. Therefore, the expectation cannot be computed

Algorithm 2 Greedy recursive binary splitting, from Paper II

```
Input:
    - A training set with derivatives and features $\{(x_i, g_{i,k}, h_{i,k})\}_{i=1}^n$
Do:
1. Initialize the tree with a constant value $\hat{w}$ in a root node:
    $\hat{w} = -\frac{\sum_{i=1}^n g_{i,k}}{\sum_{i=1}^n h_{i,k}}$
2. Choose a leaf node $t$ and let $I_{tk}$ be the index set of observations
    falling into node $t$
    For each feature $j$, compute the reduction in training loss
    $\mathcal{R}_t(j, s_j) = \frac{1}{2n} \left[ \frac{\left(\sum_{i \in I_L(j,s:j)} g_{ik}\right)^2}{\sum_{i \in I_L(j,s_j)} h_{ik}} + \frac{\left(\sum_{i \in I_R(j,s_j)} g_{ik}\right)^2}{\sum_{i \in I_R(j,s_j)} h_{ik}} - \frac{\left(\sum_{i \in I_{tk}} g_{ik}\right)^2}{\sum_{i \in I_{tk}} h_{ik}} \right]$
    for different split-points $s_j$, and where
    $I_L(j, s_j) = \{i \in I_{tk} : x_{ij} \leq s_j\}$ and $I_R(j, s_j) = \{i \in I_{tk} : x_{ij} > s_j\}$
    The values of $j$ and $s_j$ maximizing $\mathcal{R}_t(j, s_j)$ are chosen as
    the next split, creating two new leaves from the old leaf $t$.
3. Continue step 2 iteratively, until some threshold on
    tree-complexity is reached.

If needed explaining text
```

explicitly and neither can the functional derivative. The immediate solution, if having access to a dataset $\mathcal{D}_n = \{y_i, x_i\}_{i=1}^n$ of independent observations, is to average the loss over these observations, and instead, at iteration k, seek

$$\hat{f}_k(x) = \arg\min_{f_k} \frac{1}{n} \sum_{i=1}^n l\left(y_i, f^{k-1}(x_i) + f_k(x_i)\right). \tag{3.8}$$

To learn a function that is as close as possible (given the information in the sample) to the functional derivative, use the predictions resulting from the current model, f^{k-1}, to compute predictions $\hat{y}_i^{(k-1)} = f_0(x_i) + \cdots + f^{(k-1)}(x_i)$, and then derivatives for observations in the sample as

$$g_{i,k} = \frac{\partial}{\partial \hat{y}_i} l(y_i, \hat{y}_i^{(k-1)}), \;\; h_{i,k} = \frac{\partial^2}{\partial \hat{y}_i{}^2} l(y_i, \hat{y}_i^{(k-1)}). \tag{3.9}$$

These are used in a 2'nd order approximation to the original loss

$$\begin{aligned} \hat{f}_k(x) &= \arg\min_{f_k} \frac{1}{n} \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(k-1)}) + g_{i,k} f_k(x_i) + \frac{1}{2} h_{i,k} f_k(x_i)^2 \\ &= \arg\min_{f_k} \frac{1}{n} \sum_{i=1}^{n} g_{i,k} f_k(x_i) + \frac{1}{2} h_{i,k} f_k(x_i)^2 \end{aligned} \tag{3.10}$$

which is quadratic-type loss amenable to fast optimization. Now, a function that completely minimizes the above sample loss (both the original and/or the approximate), is likely to adapt to the inherent randomness in the sample. Furthermore, a search over all possible functions is obviously infeasible. For these reasons, the search is constrained to a family of functions that admits fast fitting routines/optimization, and that are somehow constrained and thus less likely to overfit. Typical families include linear functions, local regression using kernels, or most popularly, trees.

Gradient boosting emerges as the collection of the above-mentioned ideas: Iteratively, compute derivative information or pseudo-residuals through (3.9), and fit a statistical model $\tilde{f}_k$ to these observations using (3.10). The final ingredient of gradient boosting is to shrink the model by some constant $\delta \in (0, 1]$, $\hat{f}_k = \delta \tilde{f}_k$ to make space for new models, and add it to the ensemble model $f^{(k)} = f^{(k-1)} + \hat{f}_k$. The gradient boosting pseudo-algorithm is given in Algorithm 1. Note that this is the modern-type gradient boosting algorithm, slightly different from the original algorithm of Friedman (2001) which is a first-order type algorithm, that would fit the model using mean-squared-error loss, then scale the model with an optimized constant value and finally shrink it.

The choice of statistical model to fit to the pseudo-residuals is not arbitrary. The popular choice is to use classification and regression trees

(CART) (Breiman et al., 1984), which gives gradient tree boosting, the boosting type that has dominated in many machine-learning competitions since the introduction of xgboost (Chen and Guestrin, 2016). Using CART as week learners can be motivated by first considering linear functions, $\hat{y} = \beta^T x$, where only a portion of the full estimate of the $\hat{\beta}_j$ decreasing (3.10) the most is used. This then resembles a type of shrinked foreward stagewise procedure, which is closely related to computing LASSO solution paths (Hastie et al., 2001). Boosting thus holds the possibility of efficiently building sparse models in the face of high-dimensional problems, by excluding features x_j that does not contribute significant decreases in (3.10). If a week learner is used that fits and use all features simultaneously, this pathology of boosting is likely to disappear.

If CART is fit using greedy binary splitting (Algorithm 2), then features are used sequentially in the learning procedure. Furthermore, in contrast to linear functions, CART can learn non-linear functions and interaction effects automatically. In essence, gradient tree boosting may learn sparse, non-linear models with complex interaction effects efficiently, while the shrinkage δ applied to each tree will smooth out the piecewise constant functions. Model complexity may range from the constant model, to high-dimensional non-linear functions, and in between is often something that may decrease the objective in (2.2) more than other types of (fixed complexity) models.

Gradient boosting originally has two hyperparameters, namely the number of boosting iterations K, and shrinkage or the learning rate δ, which is usually set to some "small" value. Using CART as week learners introduces additional tuning to control the complexity of individual trees. Friedman et al. (2000) suggests global hyperparameters fixed equally for all trees, as the greedy binary splitting algorithm is optimized as if the current boosting iteration is the last iteration. Important such hyperparameters are a parameter for the maximum depth of trees, a maximum number of leaves or terminal nodes in a tree, and a threshold for minimum reduction in training loss (3.10) if a split is to be performed.

Hyperparamters are typically learned using k-fold cross-validation (CV) (Stone, 1974), which increase computation times significantly.

3.3 Distribution of estimated parameters

A key to solving problems of the type in (2.1), which there are multiple of in the boosting solution to (2.2), is to use the distribution of estimated parameters to evaluate the significance of the model. This can for-example be used to control complexity of the model, or to reject alternative hypothesis.

There are multiple ways of approximating the distribution of estimated quantities. The perhaps most straight-forward method, if observations are independent, is the bootstrap (Efron, 1992). The idea is that if the distribution, $P_X(x)$, behind the true data-generating process is known, then a large number, say B, of size-n datasets, i.i.d. of the training data $\{y_i, x_i\}_{i=1}^n$, could be sampled. Then, the fitting procedure could be performed for each sampled-dataset, and finally statistical methods could be used to investigate the sampled quantities. However, the true distribution P_X is of course generally unknown. The idea of the bootstrap is to exchange P_X with the empirical distribution, $P_X^*(x) = n^{-1}\sum_{i=1}^n 1(x_i \leq x)$, and then perform the above mentioned procedure.

Sampling procedures are in general, however, quite expensive, and this is no different for the bootstrap. At early experimental stages of this work, the SPA was used together with the idea of the empirical distribution P_X^* to retrieve necessary density approximations while avoiding costly sampling. The idea is to use the empirical CGF, $K_X^*(s) = \log\left(\sum_{i=1}^n e^{sx_i}\right) - \log n$, in the SPA (3.6), from which desired results can be drawn. This is explained in Butler (2007, Chapter 14), and in particular Chapter 12.2 for the ratio estimators appearing in Algorithm 2. The goal was a computationally efficient version of the Efron Information Criterion (EIC) (Ishiguro et al., 1997), but was abandoned due to concerns regarding stability and speed of computations, in comparison to analytical asymptotic results.

Often the most computationally efficient procedures are analytical results, which may be obtained for estimated quantities through asymptotics. The central limit theorem may be applied to the score equations, $0 = \nabla_\theta l(y, f(\mathbf{x}; \hat{\theta})) = n^{-1} \sum_{i=1}^n l(y_i, f(x_i; \hat{\theta}))$, to obtain asymptotic normality, and from this, asymptotic normality of estimated parameters (under certain regularity conditions, see van der Vaart (1998)) can be obtained through the delta method as

$$\begin{aligned} n^{-1}(\hat{\theta} - \theta_0) &\sim N\left(0, J(\theta_0)^{-1} I(\theta_0)\right), \\ J(\theta) &= E[\nabla_\theta^2 l(y, x; \theta))], \\ I(\theta) &= E\left[\nabla_\theta l(y, f(x); \theta)) \nabla_\theta l(y, f(x); \theta))^T\right]. \end{aligned} \tag{3.11}$$

Estimates of J and I can be obtained through averaging and by using $\hat{\theta}$ in place of θ_0, computation is usually highly efficient, and stability is a non-issue. Furthermore, Gaussian results are highly tractable, as a Gaussian empirical process often converge asymptotically to known and well-studied continuous-time stochastic processes. As such, using asymptotic normality emerged as the preferred solution in Paper II and III. Much more can be said about different asymptotic results in statistics, conditions under which normality emerge, and its applications. For an overview see van der Vaart (1998).

3.4 Model selection

Denote the true distribution of X as $G_X(x)$ with density $g_X(x)$, which is attempted modelled by density $p_X(x)$, then the Kullback-Leibler divergence (KLD) -cite KL 1951-, denoted D is given by

$$D(g, p) := \int g_X(x) \log g_X(x) dx - \int g_X(x) \log p_X(x; \hat{\theta}) dx. \tag{3.12}$$

Since the first integral in (3.12) is constant w.r.t. different choices of models $p_X(x; \theta)$, only the negative remaining integral is relevant, and is commonly referred to as relative KLD. The negative log-likelihood objective of the optimization problem in (2.1) is a sample version of relative KLD, and appears as the natural objective for optimization over different values of θ, when minimizing KLD is the the overarching goal.

Selection between models is more difficult when candidate models are of different functional form and complexity. This becomes clear if we rewrite the fitted sample negative log-likelihood as the expectation integral with respect to the empirical distribution,

$$-n^{-1}\log p_X(\mathbf{x};\hat{\theta}) = -\int \log p_X(x;\hat{\theta})dG_X^*(x). \tag{3.13}$$

The empirical distribution G_X^* corresponds more closely towards fitted models $p_X(x;\hat{\theta})$ with higher complexity, than does true G_X. Therefore, naively using the negative log-likelihood as the basis for model selection will result in unfairly consistent choices of models with high complexity over parsimonious models. This is termed the "optimism" of the training loss (Hastie et al., 2001). This is taken into consideration in the supervised-learning optimization objective (2.2). If the loss function l appearing in (2.2) is a negative log-likelihood $-\log p$, then the objective of (2.2) is exactly the negative last integral of the KLD (3.12), as evaluation is over data (y^0, x^0) unseen in the fitting of $\hat{\theta}$.

In the coming discussion, consider the loss-based regression setting $l(y, f(x;\theta))$. The idea of generalization-based information criteria is to adjust for the bias induced by integrating the model w.r.t. the empirical distribution instead of the true distribution G_X in (3.13). Denote this bias or the optimism by $C(\hat{\theta})$, making the dependence upon fitted parameters $\hat{\theta}$ explicit. Then $C(\hat{\theta})$ is given by

$$C(\hat{\theta}) = E\left[l(y^0, f(x^0;\hat{\theta}))\right] - E\left[l(y, f(x;\hat{\theta}))\right], \tag{3.14}$$

where in the first expectation, (y^0, x^0) is independent of data using in fitting of $\hat{\theta}$, while in the second expectation (y, x) is part of the training set. Information criteria like the celebrated Akaike Information Criterion (commonly known as AIC) (Akaike, 1974), Takeuchi Information

Criterion (TIC) (Takeuchi, 1976) and Network Information Criterion (NIC) (Murata et al., 1994) targets this bias $C(\hat{\theta})$. A detailed development of AIC and TIC is found in Burnham and Anderson (2003), and the case for NIC is almost completely analogous.

Common for all three information criteria mentioned above, is that they rely on the asymptotic approximation

$$C(\hat{\theta}) \approx \mathtt{tr}\left(E\left[\nabla_{\theta}^{2} l(y, f(x;\theta_0))\right] Cov(\hat{\theta})\right), \tag{3.15}$$

which is developed from two quadratic approximations 3.1 of the loss l about θ_0 and $\hat{\theta}$. The approximation is applicable when the loss is appropriately differentiable in θ, and $\hat{\theta}$ is a consistent estimator. In the case of AIC, the true model g is assumed as an interior point in the space of θ. Under this assumption, the covariance in (3.15) is the inverse of the expected hessian. Thus the right hand side of (3.15) reduces to the number of dimensions of θ, say d. If g is not assumed to be an interior point, the Sandwich-estimator due to Huber (Huber et al., 1967) can be used for the covariance. The trace in (3.15) then gives $C(\hat{\theta}) \approx \mathtt{tr}(J(\theta_0)^{-1} I(\theta_0))$, and estimation of these as discussed in 3.3 results in TIC and NIC. Notice that (3.15) is not directly applicable to the tree-models at different stages of Algorithm 2. This is because l is generally not differentiable in the different split points being profiled over to maximize reduction in loss.

Finally, note the existence of many other information criteria, that may seek to improve on the above mentioned criteria, or that targets other objectives than KLD and expected generalization loss. Notorious is the Corrected Akaike Information Criterion (AICc), the Bayesian information criterion (BIC), which may also be developed using a quadratic approximation together with a close-cousin of the SPA called the Laplace approximation, and the Focused Information Criterion (FIC). See -citet Hjort model selection and model averaging- for an overview.

4 Summary of the papers

The first paper of the thesis, "Saddlepoint adjusted inversion of characteristic functions", written in collaboration with the professors Tore Selland Kleppe and Hans Julius Skaug, the SPA is developed through exponential tilting in the time-domain side of the Fourier transform. This development admits a deconstruction of the density of some random variable X as the SPA and the density of a standardized random variable Z evaluated at zero. The variable Z is specified through its CGF $K_Z(s)$, a function of the CGF of X. The representation of p_X is exact, and necessarily takes care of issues regarding renormalization and unimodality of the SPA. Furthermore, as evaluation of the density of Z is only necessary at the, by-design, high-density point zero, inversion using quadrature is accurate and does not suffer from the numerical issues of direct IFT using quadrature, even at low-density regions of X where the SPA will dominate. The methodology is illustrated using the Negative Inverse Gaussian distribution, and applied to financial data through the Merton Jump Diffusion model.

In the second paper, "An information criterion for automatic gradient tree boosting", written in collaboration with the professors Tore Selland Kleppe and Hans Julius Skaug, an information criterion is constructed that takes into consideration the optimism induced into the training loss by the greedy recursive binary splitting in Algorithm 2. The bias (3.14) is found by relating the asymptotic dynamics of the loss under split-profiling, to that of a Cox-Ingersoll-Ross process (CIR). The asymptotic normality (3.11) of estimators is key to establish a familiar continuous-time stochastic process, and eventually the CIR, that makes it possible to build upon the approximate equation (3.15) for generalization-loss based information criteria. Finally, to take into consideration the optimism induced by selecting the maximum reduction in loss over features j in Algorithm 2, extreme-value theory is employed. The criterion is built into an algorithm for gradient tree boosting which is automatic, removing hyperparameters such as the number of boosting iterations

K and constraints regularizing trees. The underlying assumptions are tested on simulated data, and the algorithm is validated on a collection of real data by measuring both speed and accuracy versus comparative methodologies.

The third paper, "aGTBoost: Adaptive and Automatic Gradient Tree Boosting Computations", written in collaboration with professor Tore Selland Kleppe, focuses on the implementation of the theory in Paper II in an R-package named aGTBoost. In addition, the paper propose a modification of the splitting procedure in Algorithm 2, to be more adept to gradient boosting by taking into consideration the possible root-split produced in the coming boosting iteration. Usage of aGTBoost is illustrated, and its functionality on the large Higgs dataset is compared versus the xgboost R-package. Results suggests superior performance of aGTBoost versus the default settings of xgboost. Furthermore, the aGTBoost package lowers the threshold for users to employ gradient tree boosting, as detailed knowledge on hyperparameter tuning and setting up $k-$fold CV in code no longer is a necessity.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and Regression Trees*. CRC Press.

Burnham, K. P. and D. R. Anderson (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media.

Butler, R. W. (2007). *Saddlepoint approximations with applications*. Cambridge University Press.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 569–593. Springer.

Friedman, J., T. Hastie, R. Tibshirani, et al. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics 28*(2), 337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics New York, NY, USA:.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 221–233. University of California Press.

Ishiguro, M., Y. Sakamoto, and G. Kitagawa (1997). Bootstrapping log likelihood and eic, an extension of aic. *Annals of the Institute of Statistical Mathematics 49*(3), 411–434.

Mason, L., J. Baxter, P. Bartlett, and M. Frean (1999). Boosting algorithms as gradient descent in function space (technical report). *RSISE, Australian National University*.

Murata, N., S. Yoshizawa, and S.-i. Amari (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks 5*(6), 865–872.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 111–147.

Takeuchi, K. (1976). Distribution of information statistics and validity criteria of models. *Mathematical Science 153*, 12–18.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

Appendix

Paper I

Saddlepoint adjusted inversion of characteristic functions

The First Paper

Ole Olesen[1], Geir Geirsen[1], and Jens Jensen[2]

[1]Department of Mathematics and Physics, University of Stavanger, Norway
[2]Department of Safety, Economics and Planning, University of Stavanger, Norway

January 6, 2020

Abstract

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam eget interdum mauris. Suspendisse ut odio sed urna molestie sollicitudin at sit amet nisl. Vivamus vel libero vel metus eleifend pulvinar ac eget eros. Maecenas lobortis sem nisl, sed ultrices metus elementum eget. Cras leo urna, rhoncus a volutpat at, dapibus in eros. Sed sit amet nulla placerat, vulputate diam non, tempus nisl. Sed quis nibh ut lectus accumsan mollis ut et mi. Ut et egestas mauris. Etiam vel pharetra enim. Donec a laoreet justo, eu ultrices enim.

Nunc maximus eu quam in iaculis. Nulla eget metus et sem varius aliquam. Morbi a ipsum aliquet, vestibulum quam eget, commodo tortor. Praesent purus orci, fermentum at tellus eu, aliquam consectetur mauris. Nullam quis diam ligula. Etiam at ipsum vel diam hendrerit ornare quis ut lectus. Donec elementum at ipsum vel ullamcorper. Etiam in magna non mauris lobortis semper ac a leo. Nam ultricies purus a pretium ultrices. Donec a mi vitae leo egestas imperdiet. Vestibulum faucibus urna suscipit tincidunt suscipit. Duis id velit nec magna eleifend facilisis pellentesque dapibus nunc. Integer interdum nunc orci, a dictum magna pulvinar vitae. Nam ac maximus felis.

Paper II

An information criterion for automatic gradient tree boosting

The Second Paper

Ole Olesen[1], Geir Geirsen[1], and Jens Jensen[2]

[1]Department of Mathematics and Physics, University of Stavanger, Norway
[2]Department of Safety, Economics and Planning, University of Stavanger, Norway

January 6, 2020

Abstract

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam eget interdum mauris. Suspendisse ut odio sed urna molestie sollicitudin at sit amet nisl. Vivamus vel libero vel metus eleifend pulvinar ac eget eros. Maecenas lobortis sem nisl, sed ultrices metus elementum eget. Cras leo urna, rhoncus a volutpat at, dapibus in eros. Sed sit amet nulla placerat, vulputate diam non, tempus nisl. Sed quis nibh ut lectus accumsan mollis ut et mi. Ut et egestas mauris. Etiam vel pharetra enim. Donec a laoreet justo, eu ultrices enim.

Nunc maximus eu quam in iaculis. Nulla eget metus et sem varius aliquam. Morbi a ipsum aliquet, vestibulum quam eget, commodo tortor. Praesent purus orci, fermentum at tellus eu, aliquam consectetur mauris. Nullam quis diam ligula. Etiam at ipsum vel diam hendrerit ornare quis ut lectus. Donec elementum at ipsum vel ullamcorper. Etiam in magna non mauris lobortis semper ac a leo. Nam ultricies purus a pretium ultrices. Donec a mi vitae leo egestas imperdiet. Vestibulum faucibus urna suscipit tincidunt suscipit. Duis id velit nec magna eleifend facilisis pellentesque dapibus nunc. Integer interdum nunc orci, a dictum magna pulvinar vitae. Nam ac maximus felis.

Paper III

aGTBoost: Adaptive and Automatic Gradient Tree Boosting Computations

The Third Paper

Ole Olesen[1], Geir Geirsen[1], and Jens Jensen[2]

[1]Department of Mathematics and Physics, University of Stavanger, Norway
[2]Department of Safety, Economics and Planning, University of Stavanger, Norway

January 6, 2020

Abstract

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam eget interdum mauris. Suspendisse ut odio sed urna molestie sollicitudin at sit amet nisl. Vivamus vel libero vel metus eleifend pulvinar ac eget eros. Maecenas lobortis sem nisl, sed ultrices metus elementum eget. Cras leo urna, rhoncus a volutpat at, dapibus in eros. Sed sit amet nulla placerat, vulputate diam non, tempus nisl. Sed quis nibh ut lectus accumsan mollis ut et mi. Ut et egestas mauris. Etiam vel pharetra enim. Donec a laoreet justo, eu ultrices enim.

Nunc maximus eu quam in iaculis. Nulla eget metus et sem varius aliquam. Morbi a ipsum aliquet, vestibulum quam eget, commodo tortor. Praesent purus orci, fermentum at tellus eu, aliquam consectetur mauris. Nullam quis diam ligula. Etiam at ipsum vel diam hendrerit ornare quis ut lectus. Donec elementum at ipsum vel ullamcorper. Etiam in magna non mauris lobortis semper ac a leo. Nam ultricies purus a pretium ultrices. Donec a mi vitae leo egestas imperdiet. Vestibulum faucibus urna suscipit tincidunt suscipit. Duis id velit nec magna eleifend facilisis pellentesque dapibus nunc. Integer interdum nunc orci, a dictum magna pulvinar vitae. Nam ac maximus felis.