HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Information and Natural Sciences

Emil Eirola

Variable Selection with the Delta Test in Theory and Practice

Master's thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology in the Degree Programme of Engineering Physics and Mathematics.

Espoo, November 3 2009

Supervisor: Professor Olli Simula

Instructor: Docent Amaury Lendasse

HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Information and Natural Sciences
Degree Programme of Engineering Physics and Mathematics

ABSTRACT OF MASTER'S THESIS

Author:	Emil Eirola	
Title of thesis:	Variable Selection with the Delta Test in Theory and Practice	
Date:	November 3 2009	**Pages:** 40 + 2
Professorship:	Computer and Information Science	**Code:** T–115
Supervisor:	Professor Olli Simula	
Instructor:	Docent Amaury Lendasse	

The importance of variable selection procedures in non-linear regression analysis is becoming increasingly important as the size of data sets which can be gathered and handled continues to grow. In addition to reducing the size of the problem, variable selection can improve the performance of regression models by discarding noisy data. Furthermore, variable selection provides valuable interpretability of the data by specifying which variables are more relevant than others. This thesis assesses some of the currently available state-of-the-art methods and presents the use of the "Delta test" noise variance estimator for input variable selection.

The use of the Delta test for variable selection is studied in a theoretical framework, and a theorem is derived which shows that, under reasonable assumptions, the expectation of the Delta test is minimised by the optimal selection of variables. The method is also analysed from a practical standpoint, including some simulated experiments to investigate its behaviour under specific conditions.

The Delta test is compared to two alternative methods for variable selection: *mutual information* and *least angle regression*. The performance of each method when used with a *Least Squares Support Vector Machines* non-linear regression model is evaluated on a total of five real world data sets, and it is found that the Delta test excels on average. The conceptually simple and computationally efficient method outputs a good, model-independent selection of variables, and can consequently be considered a viable competitor among the currently commonly used methods.

Keywords:	Delta test, variable selection, noise variance estimation, nearest neighbour, non-linear regression
Language:	English

TEKNISKA HÖGSKOLAN SAMMANFATTNING
Fakulteten för informations- och naturvetenskaper AV DIPLOMARBETET
Utbildningsprogrammet för teknisk fysik och matematik

Utfört av:	Emil Eirola	
Arbetets namn:	Val av variabler med Delta-testet i teori och praktik	
Datum:	3 november 2009	**Sidantal:** 40 + 2
Professur:	Informationsteknik	**Kod:** T–115
Övervakare:	Professor Olli Simula	
Handledare:	Docent Amaury Lendasse	

Betydelsen av att välja rätta variabler inom icke-linjär regressionsanalys har blivit allt väsentligare då storleken på datamängder som kan samlas in och hanteras fortsätter att öka. Förutom att minska problemets storlek, kan valet av variabler förbättra resultaten för regressionsmodeller genom att avlägsna meningslös data (brus). Dessutom tillför variabelvalet en värdefull tolkning av datamängden genom att ange vilka variabler som kan anses vara mer relevanta än andra. I detta diplomarbete analyseras några moderna metoder och användingen av brusvariansestimatorn "Delta-testet" presenteras som ett alternativ för val av variabler.

Användningen av Delta-testet för val av variabler undersöks från en teoretisk synvinkel, och det härleds en sats som visar att under vissa rimliga antaganden minimerar det optimala valet av variablerna Delta-testets väntevärde. Metoden analyseras också ur ett praktiskt perspektiv, med hjälp av några konstgjorda experiment som åskådliggör dess beteende under säskilda förhållanden

Delta-testet jämförs med två andra metoder för val av variabler: gemensam information (*mutual information*) och minsta-vinkelsregression (*least angle regression*). Prestationen av varje metod i samband med en minsta-kvadrats-stödvektormaskiners (*Least Squares Support Vector Machines*) icke-linjär regressionsmodell utvärderas på sammanlagt fem datamängder som baserar sig på praktiska tillämpningar. Resultaten visar att Delta-testet utmärker sig i genomsnitt. Den lättbegripliga och beräkningsmässigt effektiva metoden ger ut ett lämpligt och modelloberoende val av variabler, och kan därmed anses vara en kraftig konkurrent bland de oftast använda metoderna.

Nyckelord:	Delta-test, val av variabler, brusvariansestimation, närmaste-grannemetod, icke-linjär regression
Språk:	Engelska

TEKNILLINEN KORKEAKOULU
Informaatio- ja luonnontieteiden tiedekunta
Teknillisen fysiikan ja matematiikan tutkinto-ohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Emil Eirola	
Työn nimi:	Muuttujien valinta Delta-testillä teoriassa ja käytännössä	
Päiväys:	3. marraskuuta 2009	**Sivumäärä:** 40 + 2
Professuuri:	Informaatiotekniikka	**Koodi:** T–115
Työn valvoja:	Professori Olli Simula	
Työn ohjaaja:	Dosentti Amaury Lendasse	

Muuttujien valinnan tärkeys epälineaarisessa regressioanalyysissä on korostunut kerättävissä ja käsiteltävissä olevan mittaustiedon koon kasvaessa. Mallintamistehtävän pelkistämisen lisäksi muuttujien valinta voi parantaa tehokkuutta erottamalla datasta kohinaa sisältäviä komponentteja. Lisäksi muuttujien valinta auttaa tulkitsemaan tietomäärää erittelemällä mitkä syötemuuttujat vaikuttavat tärkeimmiltä. Tässä diplomityössä katsastetaan alan kehityksen nykytasoa vastaavia menetelmiä, sekä esitellään kohinan varianssin estimointiin perustuvan "Delta-testi" -menetelmän soveltuvuutta muuttujien valintaan.

Delta-testin käyttöä muuttujien valinnassa tutkitaan teoreettisella tasolla, ja johdetaan lause, joka kohtuullisten olettamusten alla osoittaa, että Delta-testin odotusarvon minimi saavutetaan optimaalisella valikoimalla muuttujia. Menetelmää tarkastetaan myös käytännön näkökulmasta, ja työssä esitellään simuloituja kokeita jotka havainnollistavat sen käyttäytymistä tietynlaisissa tilanteissa.

Delta-testiä verrataan kahteen vaihtoehtoiseen menetelmään: keskinäinen informaatio (*mutual information*) sekä pienimmän kulman regressio (*least angle regression*). Menetelmien toimintaa vertaillaan viidessä eri mittauksiin perustuvassa mallinnusongelmassa käyttämällä epälineaarista pienimmän neliösumman tukivektorikoneiden (*Least Squares Support Vector Machines*) mallia. Tulosten perusteella Delta-testi suoriutuu keskimäärin parhaiten. Käsittellisesti yksinkertaista sekä laskennallisesti kevyttä menetelmää voidaan siten pitää varteenotettavana kilpailijana nykyisille yleisessä käytössä oleville menetelmille.

Avainsanat:	Delta-testi, muuttujien valinta, kohinan varianssin estimointi, lähimmän naapurin menetelmä, epälineaarinen regressio
Kieli:	Englanti

Acknowledgements

The work for this Master's thesis has been been conducted at the Department of Information and Computer Science at the Helsinki University of Technology over the past couple of years, and I want to thank every one of my co-workers there for the enjoyable environment.

Most of all, I wish to thank my supervisor, Professor Olli Simula, for the experience, expertise, and support he has provided, and Docent Amaury "Momo" Lendasse for his invaluable guidance and trust. This work would not have been possible without them. My gratitude also goes to all of the other current and former members of the Time Series and Chemoinformatics research group who have helped me during the years, Francesco Corona, Elia Liitiäinen, Federico Montesino Pouzols, Antti Sorjamaa, Yoan Miché, Dušan Sovilj, Mark van Heeswijk, Tuomas Kärnä, and Qi Yu. Working with you has been a blast. I also wish to mention Professor Michel Verleysen, whose tough questions kept me on track and made sure I did not take any unjustified shortcuts.

I also want to thank my parents, Timo and Stina, for believing in me, and finally, I would like to thank all of my friends, particularly those in *Cause of Death* and *Metal Club Mökä*, for enabling an outlet to balance the frustrations occasionally induced by these scientific endeavours.

Otaniemi, November 3 2009

Emil Eirola

Contents

Abbreviations

DT	Delta test
LARS	Least angle regression
LOO	Leave-one-out (cross-validation)
LS-SVM	Least squares support vector machines
MI	Mutual information
MSE	Mean squared error
NN	Nearest neighbour
RBF	Radial basis function

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Scope

With evolving technology and the continuing development of more efficient data mechanisms, the size and complexity of interesting regression modelling tasks has grown considerably. The number of input variables which can be measured might be large, and it may be difficult to recognise which variables are important for the task at hand. Occasionally variables are irrelevant for the output, and at other times they contain redundant information already available in other variables. Hence the concept of *variable selection* has become increasingly important. Being able to discard the unnecessary ones is beneficial for performance and stability of the model. Identifying the most essential variables also provides a better understanding of the problem and interpetability of the data. In a sense, variable selection improves the effective signal-to-noise ratio by getting rid of some of the noisy components of the data.

For linear problems, the issue can be solved by simple covariance or correlation-based methods. In the case of non-linear problems, the situation is less straightforward, and there are several specialised methods to pick from, many of which require parameters which are non-trivial to tune. Other methods only rank variables, but cannot tell you how many to choose. Such ranking schemes may also easily fail to identify situations where some variables are useful only in combination with others. There appears to be a unfilled need for a simple and entirely non-parametric, model-independent alternative. A method based on a noise variance estimator known as the *Delta test* is one such method, however, it has not seen extensive use since its properties have not been carefully analysed until now.

This thesis investigates the use of Delta test for variable selection both from a

theoretical as well as an experimental perspective. Its advantages and disadvantages are studied, and the current state-of-the-art of variable selection for regression is explored by comparing the Delta test to other methods.

The Delta test itself is based on a nearest-neighbour approach which is generally applicable and in its purest form has no parameters to tune, making it robust and easy to use. Essentially, a noise variance estimation procedure is applied to each nonempty subset of variables, and that subset which minimise the estimate is chosen. The estimate given by the method in a sense represents the lowest attainable mean squared error (MSE) by quantifying the extent of the "random" component of the data. Although the method has been successfully used in situations as [1, 2]—and in an adapted form in [3]—little theoretical basis for its use has been established. In other words, while the method itself is not entirely new, this thesis presents a more formal and explicit treatment of the Delta test for variable selection than what has been seen before.

Before reaching the main matter of the thesis work, Chapter 2 reviews the concept of regression analysis in order to establish some notation and conventions, and introduces the problem of variable selection. Some popular methods are presented, including variable selection by mutual information, and least angle regression.

Chapter 3 discusses noise variance estimation and describes the Delta test algorithm in its original context, including some important convergence properties of the estimator.

The main contribution of this thesis is in Chapter 4, which contains some intuitive as well as theoretical—mathematically sound—justification for the use of the Delta test for variable selection. Another issue investigated is why in particular the Delta test is an appropriate estimator to use for this task, when it is well known that there are more sophisticated methods for actual noise estimation—such as those presented in [4, 5, 6, 7].

In Chapter 5, the use of the method from a practical point of view is considered. As the computational load increases exponentially with the number of variables, the method can become impractical with huge datasets, and hence some efficient search schemes have been developed for finding near-optimal solutions. Experiments which illustrate the behaviour of the Delta test in different situations are included, and the performances of the different methods are compared on several publicly available real world data sets.

1.2 Publications

This thesis is based on the research originally published in the following articles:

1. [8], where the Delta test was first explicitly introduced as a method for variable selection at the European Symposium on Artificial Neural Networks 2008 in Bruges, Belgium.

2. [9], the journal article which expands on the theoretical and experimental arguments—currently under review for the IEEE Transactions on Neural Networks.

Chapter 2

Problem Definition

2.1 Regression Analysis

Regression analysis refers to the process of modelling the dependencies between a *response variable* (or *output*) y and one or more *explanatory variables* (or *inputs*) $x^{(k)}$ from a limited set of data (measurements). The goal is to build a model $f(x^{(1)}, \ldots, x^{(d)})$, based on the inputs, which follows y as accurately as possible. For statistical processing, the available output data y_i and corresponding input data $\boldsymbol{x}_i$ are generally arranged in a column vector $\boldsymbol{y}$ and matrix $\boldsymbol{X}$:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_M^T \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(d)} \\ x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ x_M^{(1)} & x_M^{(2)} & \ldots & x_M^{(d)} \end{bmatrix}.$$

Here M is the number of data points (*samples*) available, and d is the number of input variables.

After a model has been built on the available measurements, the goal is to use it to predict the output for *new* data by applying the model f to values of $\boldsymbol{x}$ not present in the initial set.

2.1.1 Linear Regression

The simplest form of regression is *linear regression*, where the relationship between the output and inputs is assumer to follow the form

$$y_i = \alpha_0 + \sum_{k=1}^{d} \alpha_k x_i^{(k)} = \begin{bmatrix} 1 & \boldsymbol{x}_i^T \end{bmatrix} \boldsymbol{\alpha} \,.$$

This can be described as a linear system

$$\boldsymbol{y} = \boldsymbol{\chi}\boldsymbol{\alpha} \,,$$

where $\boldsymbol{\chi} = [\mathbf{1} \quad \boldsymbol{X}]$. The optimal values of α_k in the least-squares sense are found by solving the linear system using the Moore-Penrose pseudoinverse:

$$\boldsymbol{\alpha} = \left(\boldsymbol{\chi}^T\boldsymbol{\chi}\right)^{-1}\boldsymbol{\chi}^T\boldsymbol{y} \,,$$

provided there is no collinearity between the inputs, and the matrix $\boldsymbol{\chi}$ is of full rank.

Most interesting regression problems include more intricate dependencies between the variables, and hence the scope of situations where a linear model is applicable is limited. Consequently, a variety of more sophisticated models have been devised.

2.1.2 Least Squares Support Vector Machines

One widely used non-linear model is *Least Squares Support Vector Machines* (LS-SVM) [10]. It is a variation of the original support vector machines [11], designed to be computationally lighter without sacrificing accuracy. The technique is closely related to that of *Gaussian processes* [12]. The LS-SVM is of particular interest in the context of this thesis, since my experience has shown that it benefits greatly from appropriate selection of input variables, and hence this is the model which is used in the experimental part of the thesis to evaluate the performance of the different variable selection methods.

This section presents a brief summary of the method, see [10] for a detailed exposition. The model can be represented in its primal space as

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}) + b \,,$$

where $\boldsymbol{\varphi} : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is a mapping to a higher dimensional *feature space* (possibly even infinite dimensional), $\boldsymbol{w}$ is a corresponding weight vector, and b a bias term.

Training of the model is performed by the minimisation problem

$$\min_{\boldsymbol{w},b,\boldsymbol{e}} J(\boldsymbol{w},\boldsymbol{e}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma \left\|\boldsymbol{e}\right\|^2$$
$$\text{s.t.} \qquad y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i \qquad i \in \{1,\ldots,M\}$$

The function J is the sum of a regularisation term and the fitting error. The relative weights of the two terms, and the extent of the regularisation, is determined by the positive, real parameter γ. The problem is impractical in the primal space, since $\boldsymbol{\varphi}(\boldsymbol{x})$ and $\boldsymbol{w}$ are potentially infinite dimensional, and for this reason it is studied in the dual space, where $\boldsymbol{\varphi}(\boldsymbol{x})$ does not have to be explicitly constructed. Instead, it suffices to define a kernel K such that

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{\varphi}(\boldsymbol{x}_j) \qquad \forall i,j \in \{1,\ldots,M\}$$

The most common choice for K is the radial basis function (Gaussian) kernel:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\left\{-\frac{\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2}{\sigma^2}\right\},$$

where the parameter σ determines the kernel width. The model can eventually be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i K(\boldsymbol{x},\boldsymbol{x}_i) + b$$

and the parameters b and $\boldsymbol{\alpha}$ can be solved from a linear system. Using LS-SVM with the RBF kernel then requires the user to choose two real valued parameters: γ and σ. The selection of these is non-trivial, and the parameters can not be optimised separately from each other. One suggested method to perform this tuning is by a grid search to minimise the random k-fold cross-validation error (see Section 2.1.3) of the resulting model.

2.1.3 Validation

The crucial step of evaluating the performance of a model is not an obvious issue. For many models, it is sensible to examine the output error $y_k - f(\boldsymbol{x}_k)$ for each sample, and take the average of the square of these to obtain the *mean squared error* (MSE). When the same data is used both for building the model and evaluation, this is known as *training error*.

The truly interesting measure would still be the error that the model produces

for such hypothetical data that is not in the training data, but distributed similarly (in some sense). As the number of such independent samples tends to infinity, the limiting MSE is called the *generalisation error*. In many cases, the training error significantly underestimates the generalisation error, as the model is optimised on the same data as it is evaluated on. This is an example of *overfitting*.

To recognise and avoid overfitting, the samples which are used for training and evaluation must be separated. The available data can be split into two complementary sets, the *training* and *test* (or *validation*) sets. If the model is trained on the training set and evaluated on the test set, the resulting MSE is likely to be a better indicator of the generalisation error. The process can be performed repeatedly to increase the confidence on the estimate.

One way to structure the repetition is *k-fold cross-validation*, where the data is (usually randomly) partitioned into k equally sized sets. Each of the k sets is sequentially chosen to be the test set, and the model is trained on the union of the remaining $k - 1$ sets. Averaging these test MSEs then provides a nice estimate for the generalisation error, as every sample has been used for testing exactly once.

A special case of k-fold cross-validation is called *leave-one-out* (LOO) cross-validation, when $k = M$. As the name implies, here each single sample is sequentially left for the test set while the model is trained on the remaining samples, and the squared errors are averaged. As this generally requires the training of M models, it is often too inefficient to be practical, but for certain methods (such as the LS-SVM) it is possible to obtain the LOO error exactly without explicitly performing the repeated training of the model [13].

2.2 Variable Selection

In modern modelling problems it is not uncommon to have an overwhelming number of input variables. Many of them may turn out to be irrelevant for the task at hand, but without external information it is often difficult to identify these variables. *Variable selection* (also known as *feature extraction*, *subset selection*, or *attribute selection* [3]) is the process of automating this task of choosing the most representative subset of variables for some modelling task.

Variable selection is a special case of dimensionality reduction. It can be used to simplify models by refining the data through discarding insignificant variables. As many regression models and other popular data analysis algorithms suffer from the so-called *curse of dimensionality* [14, 15, 16] to some degree it is necessary to perform some kind of dimensionality reduction to facilitate their effective use.

In contrast to general dimensional reduction techniques, variable selection provides additional value by distinctly specifying which variables are important and which are not [17]. This leads to a better intuitive insight into the relationship between the inputs and outputs, and assigns interpretability to the input variables. In cases where the user has control over some inputs, variable selection emphasises which variables to focus on and which are likely to be less relevant. Furthermore, discarding the less important inputs may result in cost savings in cases where measuring some properties would be expensive (such as chemical properties of a substance).

Variable selection techniques are generally based on either *variable ranking* or *subset selection*. While subset selection methods attempt to return a single optimal subset of variables, the ranking methods only provide an ordering of the variables' estimated relevance for predicting the output. For regressions tasks, it is then left up to the user to select how many of the top ranked variables to choose. Due to their nature, ranking methods often fail to recognise situations where certain variables are useful only when combined with specific other variables.

2.2.1 Correlation

The simplest effective variable ranking method is to calculate and rank each input $x^{(k)}$ by the *Pearson product-moment correlation coefficient* between it and the output y:

$$\rho_k = \frac{\operatorname{cov}(x^{(k)}, y)}{\sigma_{x^{(k)}}\sigma_y} = \frac{\sum_{i=1}^{M}\left(x_i^{(k)} - \overline{x^{(k)}}\right)\left(y_i - \overline{y}\right)}{\sigma_{x^{(k)}}\sigma_y}\,.$$

Here $\overline{x^{(k)}}$ and $\sigma_{x^{(k)}}$ are the *sample mean* and *sample standard deviation*, respectively. The measure can only account for linear dependence of the output on the inputs, and is unable recognise more intricate connections between the variables.

Correlation can also be used to evaluate subset selections [18], based on the idea that

> *A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.*

leading to the expression

$$\frac{k\overline{r_{yi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}\,,$$

where k is the number of selected variables, $\overline{r_{yi}}$ is the mean of the correlation coefficients between the selected variables and the output, and $\overline{r_{ii}}$ is the mean of the correlation coefficients among the selected variables. Although the simplicity of the

method is appealing, it has not been widely used and [18] finds several notable concerns and lacking performance.

2.2.2 Least Angle Regression

A recent improvement to the method of variable ranking by correlation is *least angle regression* (LARS) [19]. While the first variable is still selected based on the correlation coefficient alone, this method instead proceeds to consider the residual of the output and a linear model based on the selected variable, and studies the correlation of the variables with this residual. More precisely, the method moves in the direction of the highest correlated variable (say, $x^{(i)}$) until some other variable ($x^{(j)}$) has as much correlation with the current residual. Interpreting $\boldsymbol{y}$ and the $\boldsymbol{x}^{(i)}$'s as vectors in $\mathbb{R}^M$, the method then continues equiangularly between $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ until some third variable has the same correlation. At that stage, it proceeds equiangularly (along the "least angle" direction) between the three vectors, and so on.

The order in which the variables are selected provides a ranking of their usefulness for predicting the output. Compared to the simple ranking by correlation, LARS is better as it specifically chooses the variables based on how much of the *residual* they can explain, i.e., how much *new* information they bring. This avoids the selection of an undesired variable in situations were a variable is highly correlated with the output only because it is highly correlated with some of the other highly correlated inputs.

As the method only ranks the input variables, it does not explicitly specify the number of top-ranked variables to select for optimal results, and this must somehow be chosen by the user.

2.2.3 Mutual Information

One popular method specifically for variable subset selection is evaluation by *mutual information* (MI) [20, 21]. Assuming that the input and output points originate from random variables X and Y, the mutual information is a quantity that measures the mutual dependence of the two random variables. If they have a joint probability distribution function $\mu(x, y)$, the MI between X and Y is defined by

$$I(X,Y) = \int\int \mu(x,y) \log \frac{\mu(x,y)}{\mu_x(x)\mu_y(y)} \, dx \, dy \,,$$

where $\mu_x(x) = \int \mu(x,y) \, dy$ and $\mu_y(y) = \int \mu(x,y) \, dx$ are the respective marginal probability densities of X and Y. The subset selection then amounts to choosing that

subset of input variables which maximises the mutual information with the output. As $\mu(x, y)$ is generally unknown, the value of $I(X, Y)$ can not be explicitly calculated. Instead, methods have been developed to estimate the mutual information from a limited data set. One of these is Kraskov's method [22], which will be used in the experimental part of thesis. Kraskov's method is based on the concept of nearest neighbours, and it is not entirely automatic, since it involves a parameter p which must be selected by the user to specify the number of neighbours to take into account.

Chapter 3

The Delta Test

3.1 Noise Variance Estimation

The Delta test is traditionally considered a method for residual noise variance estimation. In the kind of regression tasks considered here, the data generally consist of M input points $(\boldsymbol{x}_i)_{i=1}^M$ and associated scalar outputs $(y_i)_{i=1}^M$ [4]. The assumption is that there is a functional dependence between them with an additive noise term:

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i \,.$$

The function f is assumed to be smooth, and the residual variance—or *noise*—terms ε_i are independent and identically distributed with zero mean. Noise variance estimation is the study of how to give an *a priori* estimate for $\mathrm{Var}(\varepsilon) = \sigma^2$ given some data *without* considering any specifics of the shape of f. Having a reliable estimate of the amount of noise is useful for choosing and building an appropriate regression model as well as determining when a model may be overfitting.

The original formulation [23] of the Delta test was based on the concept of variable-sized neighbourhoods, but an alternative formulation [24, 5] with a first-nearest-neighbour (NN) approach has later surfaced. In this treatment, specifically this 1-NN formulation will be used. Its advantages are that it is entirely non-parametric, conceptually simple, and computationally efficient.

The nearest neighbour of a point is defined as the unique point which minimises a distance metric to that point *in the input space*:

$$N(i) := \arg\min_{j \neq i} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \,.$$

Theoretically—or due to limited machine precision—it may occur that the nearest neighbour is not unique, but in most practical situations this is rare, and if it does happen it is sufficient to merely pick one from the set of nearest neighbours, for instance, randomly or by choosing whichever is indexed first.

In this context, the Euclidean distance is used, but other metrics can also be applied. As the Euclidean metric considers each variable with equal importance it is often recommended to normalise the input variables for best results when using the Delta test. In some cases it may be justified to use other metrics to get better results if some input variables are known to have specific characteristics that the Euclidean metric fails to account for appropriately.

The Delta test itself is usually written as

$$\delta = \frac{1}{2M} \sum_{i=1}^{M} \left(y_i - y_{N(i)} \right)^2 \approx \mathrm{Var}(\varepsilon) \,,$$

i.e., the differences in the outputs associated with neighboring (in the input space) points are considered. This is a well-known estimator and it has been shown—e.g., in [25]—that the estimate converges to the true value of the noise variance in the limit $M \to \infty$. Although it is not considered to be the most accurate noise variance estimator, its advantages include robustness, simplicity, and computational efficiency [5].

3.2 Properties of the Delta test

The properties of the estimator have been extensively studied. In [25], it is shown that under reasonable assumptions, the Delta test is asymptotically unbiased in the sense that

$$\lim_{M \to \infty} \left| \mathrm{E}\left[\delta\right] - \sigma^2 \right| = 0 \,.$$

For a finite M, however, there is a positive bias which depends mostly on the gradient of the underlying function and the distribution of samples in the input space. In [26], the bias is calculated to be of order $\mathcal{O}(M^{-2/d})$. The implications of the strictly positive bias are explored further in Chapter 4.

The variance is shown to be of order $\mathcal{O}(M^{-1/2})$ in [25], and as such it converges to 0, implying that the estimator converges to the variance of the noise with probability 1.

3.3 The Delta Test for Variable Selection

The Delta test was originally intended to be used for estimating the residual variance. Following [1, 2, 27] this thesis examines a different use: to use it for variable selection by choosing that selection of variables which minimises the Delta test. That is, the estimate would be calculated on each subset of variables, and the selection resulting in the lowest estimate is chosen. More specifically:

1. Assuming d input variables, consider the $2^d - 1$ non-empty subsets of variables.
2. For each such subset, calculate the Delta test in the subspace spanned by the selected variables (i.e., so that the nearest neighbours are determined by considering only those variables).
3. Select that subset which provides the lowest Delta test.

The remainder of this thesis intends to investigate whether this constitutes an effective variable selection procedure for regression modelling. Note that it is not always necessary to search all $2^d - 1$ candidates, as is discussed in Chapter 5.

Chapter 4

Theory

In this section, a theoretical treatment is provided to support the claim that the Delta test is able to identify the best subset of input variables for modelling under certain conditions. As the purpose of the Delta test is to deal with noisy data, it is impossible to formulate a mathematically solid statement showing that the Delta test could *always* choose the appropriate variables. Hence the assertions presented here consider the *expectation* of the Delta test, and show that the expectation is minimised for the best selection of variables when the number M of data points is finite and sufficiently large.

Some assumptions concerning the distribution of the data are required in order for the results to hold true. These continuity assumptions detailed below are designed to be similar to and compatible with the assumptions many popular non-linear modelling techniques make about the data. This enhances the usability of the Delta test as a preprocessing step for practically any non-linear regression task.

Assume a set $\{X_i\}_{i=1}^M$ of random variables which are independent and identically distributed according to some probability density function $p(x)$ for $1 \leq i \leq M$. Here $p(x)$ is a continuous probability density on some open, bounded $C \subset \mathbb{R}^d$ and $p(x) > 0$ for $x \in C$.

Let $f : C \to \mathbb{R}$ be differentiable and the random variables $Y_i = f(X_i) + \varepsilon_i$, where ε_i are independently distributed according to some distribution with mean 0 and $\mathrm{Var}[\varepsilon_i] = \sigma^2$. Denote by x_i a realisation of X_i and by y_i a realisation of Y_i. A component k of x is denoted by $x^{(k)}$.

Let $I = \{1, \ldots, d\}$ denote the full set of input variables, and consider subsets $\tilde{I}$ of I corresponding to possible selections of input variables. Define the Delta test

$\delta : \mathcal{P}(I) \to \mathbb{R}$ as

$$\delta(\tilde{I}) := \frac{1}{2M} \sum_{i=1}^{M} \left(y_i - y_{N_{\tilde{I}}(i)} \right)^2$$

where

$$N_{\tilde{I}}(i) := \arg\min_{j \neq i} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\tilde{I}}^2 ,$$

and the seminorm

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\tilde{I}}^2 := \sum_{k \in \tilde{I}} \left(x_i^{(k)} - x_j^{(k)} \right)^2 .$$

4.1 Linear Study

As the Delta test is based on a nearest neighbour search, which is a local phenomenon, the method can be analysed locally to give an intuitive overview of its behaviour. On a sufficiently small scale, any differentiable function is linear, and any continuous probability distribution is flat.

First assume that the data $\boldsymbol{x}_i \in (0,1)^d$ for $i \in \{1, \ldots, M\}$ are i.i.d. uniformly distributed on the unit hypercube (sans boundary). Consequently the components $x_i^{(k)}$ of each $\boldsymbol{x}_i$ are i.i.d. on the open interval $(0,1)$. Let $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ for $i \in \{1, \ldots, M\}$, where $f(\boldsymbol{x}_i) = a_0 + \sum_{k=1}^{d} a_k x_i^{(k)}$. The points $(\boldsymbol{x}_i)_{i=1}^{M}$ and $(y_i)_{i=1}^{M}$ comprise the imitation data set.

In general, there will be some inputs for f which are not significant, so denote by $D \in \mathcal{P}(\{1, ..., d\})$ the set of variables which truly affect the output:

$$D = \{k \mid \partial_k f \text{ is non-zero somewhere}\} = \{k \mid a_k \neq 0\}$$

Lemma 1. *The correct selection of variables uniquely minimises the expected value of the Delta test.*

$$S \neq D \implies \mathrm{E}[\delta(S)] > \mathrm{E}[\delta(D)]$$

Proof.

$$\begin{aligned}
\mathrm{E}[\delta(S)] =& \mathrm{E}\left[\frac{1}{2M} \sum_{i=1}^{M} \left(y_i - y_{N_S(i)} \right)^2 \right] = \frac{1}{2} \mathrm{E}\left[\left(y_i - y_{N_S(i)} \right)^2 \right] \\
=& \frac{1}{2} \mathrm{E}\left[\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{N_S(i)}) + \varepsilon_i - \varepsilon_{N_S(i)} \right)^2 \right] \\
=& \frac{1}{2} \mathrm{E}\left[\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{N_S(i)}) \right)^2 \right] + \sigma^2,
\end{aligned}$$

since the ε terms are independent from the $\boldsymbol{x}_i$'s and each other. It then suffices to

show that

$$S \neq D \implies \mathrm{E}\left[\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{N_S(i)})\right)^2\right] > \mathrm{E}\left[\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{N_D(i)})\right)^2\right].$$

With the linear f,

$$\mathrm{E}\left[\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{N_S(i)})\right)^2\right] = \mathrm{E}\left[\left(\sum_{k\in D} a_k\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)\right)^2\right]$$

and since the components are uncorrelated:

$$\begin{aligned}
&= \mathrm{E}\left[\sum_{k\in D} a_k^2\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)^2\right] = \sum_{k\in D} a_k^2\,\mathrm{E}\left[\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)^2\right] \\
&= \sum_{k\in D\cap S} a_k^2 \underbrace{\mathrm{E}\left[\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)^2\right]}_{=g(\#S)} + \sum_{k\in D\setminus S} a_k^2 \underbrace{\mathrm{E}\left[\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)^2\right]}_{=1/6}
\end{aligned}$$

Here the second term is $1/6$ because $x_i^{(k)}$ and $x_{N_S(i)}^{(k)}$ are independent and uniformly distributed on $(0,1)$ when $k \notin S$. The function $g(\#S)$—which measures the expected distance (squared) along one component in S from a point to its nearest neighbour in the subspace of S—however, should clearly be far less than $1/6$, as long as M is large enough so that nearest neighbours can be expected to be considerably closer than randomly chosen points.

Still, $g(\#S)$ is an increasing function of the number of variables in S, since the distance to the nearest neighbour naturally increases with dimensionality [28]. This means that the expression is minimised by the *smallest* selection which includes D, so it is minimised by $S = D$. □

The spirit of the above treatment can be extended to differentiable functions and continuous distributions. Apply the mean-value theorem to give a point $\hat{\boldsymbol{x}}_i$ on the line segment between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{N_S(i)}$ for which

$$f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{N_S(i)}) = \nabla f\left(\hat{\boldsymbol{x}}_i\right)\left(\boldsymbol{x}_i - \boldsymbol{x}_{N_S(i)}\right).$$

Since the components are uncorrelated, it is possible to proceed as in the linear case.

$$\begin{aligned}
&\mathrm{E}\left[\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{N_S(i)})\right)^2\right] = \mathrm{E}\left[\left(\nabla f\left(\hat{\boldsymbol{x}}_i\right)\left(\boldsymbol{x}_i - \boldsymbol{x}_{N_S(i)}\right)\right)^2\right] \\
&\quad = \mathrm{E}\left[\left(\sum_{k\in D} \partial_k f\left(\hat{\boldsymbol{x}}_i\right)\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)\right)^2\right] = \sum_{k\in D} \mathrm{E}\left[\left(\partial_k f\left(\hat{\boldsymbol{x}}_i\right)\right)^2\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)^2\right] \\
&\quad = \sum_{k\in D\cap S} \mathrm{E}\left[\left(\partial_k f\left(\hat{\boldsymbol{x}}_i\right)\right)^2\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)^2\right] + \sum_{k\in D\setminus S} \mathrm{E}\left[\left(\partial_k f\left(\hat{\boldsymbol{x}}_i\right)\right)^2\left(x_i^{(k)} - x_{N_S(i)}^{(k)}\right)^2\right]
\end{aligned}$$

As above, the second term here will be considerably large if $D \setminus S \neq \emptyset$, since those particular variables are not considered in the minimisation but do affect the output. Hence we need $S \supset D$ for S to minimise the expression. As for the first term, the differences $x_i^{(k)} - x_{N_S(i)}^{(k)}$ will on average grow slightly with the size of S as there are more variables to take into account in the nearest neighbour search. So again, the minimising selection is the smallest set which contains D.

4.2 Analysis of the Delta Test

As the previous section only presented a simplistic argument justifying the use of the Delta test, this section provides the more solid treatment for a more general class of functions and distributions.

If $\tilde{I} \subset I$ is a candidate selection of variables, then $\tilde{\boldsymbol{x}}_i = (x_i^{(\tilde{I}_1)}, x_i^{(\tilde{I}_2)}, \dots)$ is the projection of each point to the subspace corresponding to the selected variables. Similarly, $\tilde{\boldsymbol{x}}'_i$ includes the components *not* in $\tilde{I}$ and is the projection onto the subspace corresponding to $I \setminus \tilde{I}$. Define

$$\tilde{f}(\tilde{\boldsymbol{x}}) = \int_C f(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}')\tilde{p}(\tilde{\boldsymbol{x}}')\, d\tilde{\boldsymbol{x}}'$$

where $\tilde{p}(\tilde{\boldsymbol{x}}')$ is the marginal density

$$\tilde{p}(\tilde{\boldsymbol{x}}') = \int_C p(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}')\, d\tilde{\boldsymbol{x}}.$$

Now $\tilde{f}(\tilde{\boldsymbol{x}}_i)$ can be thought of as the best possible approximation of y_i when using only the variables in $\tilde{I}$. In particular, if $\tilde{I}$ holds all the information for determining the noiseless part of y, it holds that $\tilde{f}(\tilde{\boldsymbol{x}}_i) = f(\boldsymbol{x}_i)$ for all i. In other words, $\tilde{f}(\tilde{\boldsymbol{x}}_i)$ is the conditional expectation with partial information:

$$\tilde{f}(\tilde{\boldsymbol{x}}_i) = \mathrm{E}\left[Y_i \mid \tilde{X}_i = \tilde{\boldsymbol{x}}\right].$$

The argument is split into two lemmas which together imply the main result.

Lemma 2. *If $\tilde{I}$ is such that $\exists \boldsymbol{x}_0 \in C$ for which $\tilde{f}(\tilde{\boldsymbol{x}}_0) \neq f(\boldsymbol{x}_0)$ (i.e., the variables in $\tilde{I}$ are not sufficient to explain f) then for any sufficiently large M*

$$\mathrm{E}\left[\delta\left(\tilde{I}\right)\right] > \mathrm{E}\left[\delta\left(I\right)\right] .$$

Proof. According to [25], the estimate when using an incomplete selection of variables converges to the residual noise

$$\mathrm{E}\left[\delta\left(\tilde{I}\right)\right] \to \mathrm{E}\left[\left(Y_i - \tilde{f}\left(\tilde{X}_i\right)\right)^2\right]$$

and, correspondingly,

$$\mathrm{E}\left[\delta\left(I\right)\right] \to \mathrm{E}\left[\left(Y_i - f\left(X_i\right)\right)^2\right] .$$

Furthermore,

$$\begin{aligned}\mathrm{E}\left[\left(Y_i - \tilde{f}\left(\tilde{X}_i\right)\right)^2\right] &= \mathrm{E}\left[\left(Y_i - f(X_i) + f(X_i) - \tilde{f}(\tilde{X}_i)\right)^2\right] \\ &= \mathrm{E}\left[\left(Y_i - f(X_i)\right)^2\right] + \mathrm{E}\left[\left(f(X_i) - \tilde{f}(\tilde{X}_i)\right)^2\right]\end{aligned}$$

since the cross terms cancel due to independence of the noise:

$$\mathrm{E}\left[\left(Y_i - f(X_i)\right)\left(f(X_i) - \tilde{f}(\tilde{X}_i)\right)\right] = \mathrm{E}\left[\varepsilon_i\left(f(X_i) - \tilde{f}(\tilde{X}_i)\right)\right] = 0 .$$

Now

$$\mathrm{E}\left[\left(f(X_i) - \tilde{f}(\tilde{X}_i)\right)^2\right] = \int_C \left(f(\boldsymbol{x}) - \tilde{f}(\tilde{\boldsymbol{x}})\right)^2 p(\boldsymbol{x})\, d\boldsymbol{x} > 0$$

where the integral is positive because the continuity of f, $\tilde{f}$ and p means there is an open subset of C around $\boldsymbol{x}_0$ where $\tilde{f}(\tilde{\boldsymbol{x}}) \neq f(\boldsymbol{x})$ and $p(\boldsymbol{x}) > 0$. Since the term is independent of M, the difference

$$\mathrm{E}\left[\delta\left(\tilde{I}\right)\right] - \mathrm{E}\left[\delta\left(I\right)\right] \to \int_C \left(f(\boldsymbol{x}) - \tilde{f}(\tilde{\boldsymbol{x}})\right)^2 p(\boldsymbol{x})\, d\boldsymbol{x} > 0$$

is strictly positive even in the limit $M \to \infty$, implying there exists an M_0 such that the expression is positive for any $M \geq M_0$. Hence, for sufficiently large M, the first term is larger, proving the lemma. □

Lemma 3. *If $\tilde{I}$ and $\hat{I}$ are such that $\forall i : \tilde{f}(\tilde{\boldsymbol{x}}_i) = \hat{f}(\hat{\boldsymbol{x}}_i) = f(\boldsymbol{x}_i)$—i.e., they are both sufficient to explain f—and $\#\tilde{I} < \#\hat{I}$, then for any finite and sufficiently large M*

$$\mathrm{E}\left[\delta\left(\tilde{I}\right)\right] < \mathrm{E}\left[\delta\left(\hat{I}\right)\right] .$$

Proof.

$$\begin{aligned}\mathrm{E}\left[\delta\left(\tilde{I}\right)\right] &= \frac{1}{2}\mathrm{E}\left[\left(Y_i - Y_{N_{\tilde{I}}(i)}\right)^2\right] \\ &= \frac{1}{2}\mathrm{E}\left[\left(f\left(X_i\right) + \varepsilon_i - f\left(X_{N_{\tilde{I}}(i)}\right) - \varepsilon_{N_{\tilde{I}}(i)}\right)^2\right]\end{aligned}$$

and further, as $f = \tilde{f}$ and the noise is independent,

$$= \sigma^2 + \frac{1}{2}\mathrm{E}\left[\left(\tilde{f}\left(\tilde{X}_i\right) - \tilde{f}\left(\tilde{X}_{N_{\tilde{I}}(i)}\right)\right)^2\right]$$

where the first term is obviously identical for $\tilde{I}$ and $\hat{I}$. According to [29] the second term is of order $M^{-2/\#\tilde{I}}$. So, for a sufficiently large M, this will be the dominating term, implying that a smaller selection produces a smaller Delta test estimate, proving the lemma. □

Theorem 1. *Assuming a finite but sufficiently large number of points, the expectation of the Delta test is minimised by the* smallest *subset of I which can fully explain f on C.*

Proof. Provided the number of points is sufficiently large, by Lemma 2 the minimising selection must be able to fully explain f, and by Lemma 3 it must be the smallest such selection. □

It is shown in [25] that the variance of the Delta test converges to 0 with increasing M. As the expectation of the Delta Test under the above assumptions is strictly minimised by the "best" selection, this means that the probability of the method choosing this selection generally increases by increasing the number of available samples.

4.3 On other noise estimators

On some level, it is intuitively sensible to optimise a model by "minimising the noise", but it is far from obvious whether the proposed scheme is justified beyond that. In

Section 4.2, it is shown to hold for the Delta test, but it is worth investigating which kinds of noise estimators can be used in this way.

As in the proof of Lemma 3 previously, the expectation of the Delta test can be expanded as

$$\mathrm{E}\left[\delta\left(\tilde{I}\right)\right] = \sigma^2 + \frac{1}{2}\,\mathrm{E}\left[\left(f\left(X_i\right) - f\left(X_{N_{\tilde{I}}(i)}\right)\right)^2\right].$$

Here it is clearly seen that unless f is constant, the Delta Test has a positive bias (for a finite M; the bias converges to 0 in the limit $M \to \infty$). This makes the estimator relatively poor for estimating noise variance compared to more optimised alternatives. However, the proposed method works for variable selection effectively by exploiting this bias. All noise estimators should be able to identify the important variables (since excluding one would inflate the noise estimate) but the Delta test has the unique ability to also prune unnecessary variables. This is because other—better—noise estimators are generally designed to be unbiased, so they do not have the property that the bias increases with the number of selected variables.

Chapter 5

Experiments

This chapter demonstrates the performance of the Delta test variable selection method on a variety of different data sets. After some practical details of implementing the method, a synthetic experiment is presented to show how the probability of choosing the best selection increases with the number of data points. Section 5.3 examines the behaviour of the method in certain corner cases by way of toy examples. Some real world data examples are studied in Section 5.4, where a regression model is trained with the selected variables, and the performance is compared to that of other variable selection methods.

5.1 Practical Considerations

When using the Delta test, it is important to normalise the data beforehand. In particular, the variances of the input variables need to be of the same order for the method to be effective. Otherwise, the variables with larger variance will have an artificially inflated significance in the selection. The standard normalisation process of whitening (scaling to unit variance and zero mean) the inputs is often a good idea, although changing the mean has no effect on the nearest neighbour search.

The naïve implementation of the Delta test is to separately find the nearest neighbour of each point, leading to a complexity of $\mathcal{O}(M^2)$ per evaluation. However, this can be improved to $\mathcal{O}(M \log M)$ by using k-d trees [30] to determine the nearest neighbours. Still, performing an exhaustive search over the space of all possible selections requires $2^d - 1$ evaluations. Our rule of thumb is that on a conventional, reasonably modern, desktop computer an exhaustive search takes 5–60 s for a data set with $M = 1000$ points and $d = 10$ variables, depending on the implementation and hardware. The exhaustive search is then practical in situations with up to 10

or 20 variables. However, many interesting problems are much larger than this.

Due to the nature of the noise variance estimator, it is often not necessary to find the selection providing the global minimum test value. Rather, a pragmatic approach is that, in general, reducing the estimate results in a better selection. Based on this notion it is generally beneficial to use different heuristics for searching the space of all possible selections:

- The *sequential forward selection* [31, 1] method, which starts from the empty selection and proceeds by sequentially adding that variable which results in the best improvement of the Delta test. Similarly, the *sequential backward elimination* (or *pruning*) method starts from the full selection and iteratively removes variables. Each method requires at most d evaluations of the Delta test.

- The *forward-backward* (or *stepwise*) search [1], which if started from an empty initialisation is like the forward search, but in addition to adding variables, it also considers the option of removing each of the previously selected variables, and makes the change which improves the target metric the most. This addition/removal of single variables is continued until convergence. The search can also be started from the full selection, or any number of random initialisation, to more extensively explore the search space. The method appears to converge in $\mathcal{O}(d)$ steps, requiring a total of $\mathcal{O}(d^2)$ evaluations, and has been found to often give good results.

- *Tabu* search [32], which is similar to the forward-backward search, but with additional conditions allowing it to efficiently get out of local minima, leading to better results. This search methodology has been successfully applied to optimising the Delta test for variable selection in [33, 34].

As the Delta test is an estimate of the residual noise, it represents the lowest generalisation error that a model is expected to be able to reach. Alternatively, it can be seen as the lowest possible training error without resorting to overfitting. In fact, as the Delta test has a bias which is always very slightly *positive*, it can be considered a safe choice to train a model until its training error matches the Delta test estimate. Consequently, it is often useful to store the final *value* of the Delta test in addition to the set of selected variables when using the method. Another interpretation of the Delta test is that it is half of the leave-one-out error of the 1-NN regression model. This provides another useful metric to compare to when performing model structure selection, since any sophisticated model should be able

to perform better than the simple 1-NN model. Essentially, if $\tilde{I}$ is the set of variables that are selected, and $\delta(\tilde{I})$ is the respective value of the Delta test, the leave-one-out or generalisation error of a good non-linear model using the variables $\tilde{I}$ should be between $\delta(\tilde{I})$ and $2\delta(\tilde{I})$, and preferably close to the lower limit.

5.2 Synthetic Experiment

To illustrate the effectiveness of the procedure, an artificial experiment is conducted to compare the Delta test to variable selection by mutual information. A synthetic experiment is appropriate as it allows repeatable instances of identical setups, which can be used to illustrate how the accuracy of the methods improves with increasing sample sizes. Here, and in the subsequent experiments, the Kraskov method [22] is used for estimating the mutual information. The method contains a parameter p which must be chosen, and in this case a medium value of $p = 6$ is used, in accordance with the author's general suggestions.

For this synthetic test, a very non-linear function is intentionally chosen:

$$f(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}, x^{(6)}) = \cos(2\pi x^{(1)}) \cos(4\pi x^{(2)}) \exp(x^{(2)}) \exp(2x^{(3)})$$

with $\boldsymbol{x}$ distributed uniformly on the unit cube $[0, 1]^6 \subset \mathbb{R}^6$. Obviously, the optimal selection of variables for modelling is $I = \{1, 2, 3\}$. To make the task challenging, the signal-to-noise ratio of the data is made to be 1:1 by choosing the variance of the noise to be equal to the variance of $f(\boldsymbol{x})$:

$$\mathrm{Var}\,[\varepsilon] = \mathrm{Var}\,[f(\boldsymbol{x})] = \frac{(8\pi^2 + 1)(e^2 - 1)(e^4 - 1)}{16(16\pi^2 + 1)} \approx 10.77\,.$$

The estimators are given all $2^6 - 1 = 63$ non-empty selections of variables, and evaluated on each of these. The subset which minimises the Delta test or maximises the mutual information estimate is returned as the result for the respective method. Comparing the selections of each method to the known answer gives an idea of the accuracy of each method. The results are presented in Figure 5.1, where the vertical axis represents the fraction of cases where the correct selection was chosen. The experiment was performed as a Monte Carlo simulation with 1000 repetitions for each value of the data set size M.

It is clear that with increasing data size, the Delta test is eventually able to *very reliably* choose the correct selection, as the curve tends towards 1. The necessary size of about 1000 points in this case might seem high, but recall that the situation

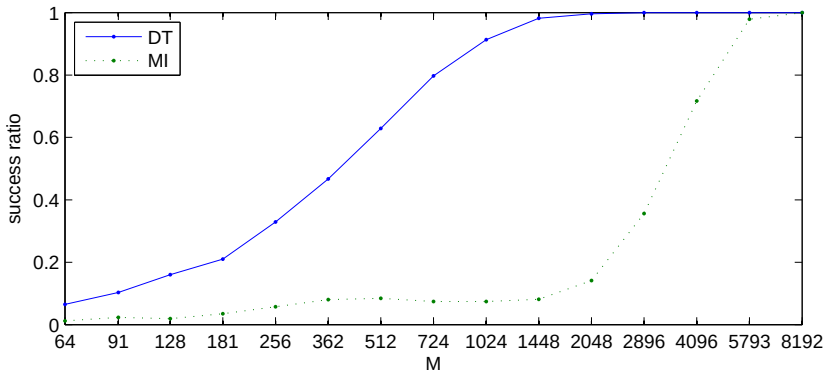

Figure 5.1: Comparison of the Delta test (DT) and mutual information (MI) on synthetic data. The vertical axis represents the ratio of cases where each method correctly identified the optimal selection from a total of 1000 tests for each point. Note the logarithmic scale for M.

was deliberately chosen to be problematic with the high amount of noise.

The mutual information method is less successful. Although the success rate does increase with M, the accuracy is much lower for smaller values of M when compared to the Delta test. The method also requires a significantly larger number of points to converge towards 1.

5.3 Toy Examples

This section presents how the Delta test behaves in corner cases where the input variables are correlated in inconvenient ways. Assume there is a latent variable x on $\mathbb{R}$, and the output is a direct function $y = f(x)$ with no additional noise. The inputs for regression are "measurements" of x.

It may appear that the theoretical framework of Chapter 4 does not apply here as some of the assumptions are not strictly fulfilled. However, the treatment still largely holds when interpreting the conditional expectation of the output

$$g(x_1, x_2) = \mathrm{E}\left[f(x) \mid X_1 = x_1, X_2 = x_2\right]$$

to be the function one is trying to model, and the Delta test as a measure of the residual error of this function.

5.3.1 Perfectly correlated variables

If $x_1 = x$ and $x_2 = 2x$ the Delta test will not discriminate between the different selections. The selections $\{x_1\}$, $\{x_2\}$, and $\{x_1, x_2\}$ all give the same nearest neighbours for every point, and so the noise estimates are also equal. The user is then free to select the smallest selection of these. In the presence of other variables, however, the different selections regarding perfectly correlated ones will effectively alter the weighting between the variables, and would affect the value of the Delta test in a manner which is not generally predictable.

5.3.2 Perfect vs noisy variable

If $x_1 = x$ and $x_2 = x + \varepsilon$ the Delta test will choose the desired result $\{x_1\}$, with high probability. To illustrate, a simple experiment was performed where $y = \sin\left(20\pi x^2\right)$ with $x \in [0, 1]$ distributed uniformly, $M = 100$ points and noise $\mathrm{Var}[\varepsilon] = 0.0001$. The simulations showed that the Delta test chooses $\{x_1\}$ in 99% of cases, even with this relatively low amount of noise to discriminate between x_1 and x_2.

5.3.3 Several noisy variables (measurements)

If $x_1 = x + \varepsilon_1$ and $x_2 = x + \varepsilon_2$ the optimal selection for regression is to choose both x_1 and x_2, as this allows one to minimise the effect of the noise by averaging.

An experiment is conducted similar to the above: $y = \sin\left(20\pi x^2\right)$, $x \in [0, 1]$ distributed uniformly with $M = 100$ points, and a noise variance of $\mathrm{Var}[\varepsilon_1] = \mathrm{Var}[\varepsilon_2] = 0.0001$. Now the Delta test chooses the desired selection $\{x_1, x_2\}$ in 80% of cases. When the number of points was increased to 1000, $\{x_1, x_2\}$ was chosen *every time* of 1000 repetitions.

5.4 Real World Data

In this section, the Delta test is benchmarked on some known datasets consisting of real-world measurements. The method is compared to variable selection by mutual information [20, 21], and variable ranking by least angle regression (LARS) [19]. The resulting selections are evaluated by training a least squares support vector machine (LS-SVM) [10] non-linear model and calculating the leave-one-out error. This provides a fair and unbiased method to compare the methods' performance for modelling.

The mutual information is, again, estimated by Kraskov's method [22] using a parameter value of $p = 6$. The Delta test and mutual information estimate are

optimised by exhaustively searching the selection space in all experiments except Section 5.4.4, where the number of inputs is prohibitively large, and the forward-backward scheme is used instead.

The LARS rankings are calculated using the implementation of [35]. As the method only gives a ranking, without any hint of how many variables to choose, the LS-SVM is here sequentially evaluated for each number of variables, successively choosing the top ranked ones until all variables are selected.

The LS-SVM models are obtained by the implementation [36]. The model has two parameters which need to be specified: the width σ of the Gaussian kernel and the regularisation parameter γ. Here these hyper-parameters for each model are obtained by running the toolbox function `tunelssvm` with initial values $\sigma^2 = 1$ and $\gamma = 1$. These initial values are sensible, as the datasets were all normalised as a preprocessing step. The function performs a sequence of two 10×10 grid searches to optimise a (random) 10-fold cross-validation in order to tune the parameters. The random component unfortunately introduces a certain degree of variability to the hyperparameters, and further, the model output. To eliminate discrepancies caused by this random effect, all the LS-SVM models were tuned and evaluated 12 times, and the mean LOO error is reported. The leave-one-out error of the LS-SVM is chosen as the performance criterion since the LS-SVM provides an efficient and exact method to calculate it, and it is a fair measure of the suitability of the selected variables for modelling.

5.4.1 Boston Housing

The Boston housing data set [37] is a set with 14 attributes for 506 objects, and the modelling task is to predict the (median) value of a house or apartment from the 13 other properties. The variables selected by the Delta test, mutual information, and LARS, as well as the LOO errors of the LS-SVM are all presented in Table 5.1. There are no obviously redundant variables in the data set, as is evidenced by the constantly decreasing error when successively choosing the variables ranked by LARS. The Delta test, also, selects all but three of the available variables. Observing the leave-one-out errors still reveals that the selection by the Delta test provides a solid improvement in the model accuracy compared to the mutual information or any of the LARS selections.

The final value of the Delta test is $\delta(\tilde{I}) = 0.0710$ here. The resulting mean LOO error of 0.0909 falls appropriately between $\delta(\tilde{I})$ and $2\delta(\tilde{I})$, while being close to the lower value.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	LS-SVM mean LOO
DT	•		•		•	•	•	•	•	•		•	•	**0.0909**
MI					•	•				•			•	0.1485
LARS													1	0.3248
						2							1	0.2333
						2					3		1	0.2054
						2					3	4	1	0.1920
				5		2					3	4	1	0.1798
	6			5		2					3	4	1	0.1553
	6			5		2		7			3	4	1	0.1456
	6			5	8	2		7			3	4	1	0.1347
	6	9		5	8	2		7			3	4	1	0.1379
	6	9	10	5	8	2		7			3	4	1	0.1301
	6	9	10	5	8	2		7	11		3	4	1	0.1177
	6	9	10	5	8	2		7	11	12	3	4	1	0.1058
	6	9	10	5	8	2	13	7	11	12	3	4	1	0.0944

Table 5.1: The variables selected by each method for predicting the median value in the Boston housing data set and the mean leave-one-out error of an LS-SVM model built using the specified variables.

5.4.2 Forest Fires

The variable selection methods are compared on a data set related to the spread of forest fires [37, 38] with 517 samples and 12 variables. The set includes some general variables (location as X-, Y-coordinates, the month and weekday of the fire), some physical measurements concerning the weather conditions (temperature, humidity, wind, rain), as well as a few variables called Fire Weather Index components (labelled as Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), and Initial Spread Index (ISI)) derived from the weather conditions. The originators of the data suggest building a model using only the actual weather measurements [38]. As the output is heavily biased towards small values, the logarithm transform $\hat{y} = \log(y + 1)$ is used instead of the original values, as also suggested in the paper.

Some of the variables here are integers, and others happen to be equal for several samples (apparently there had been several fires between the update intervals of the weather measurements). As this could cause issues in the nearest-neighbour

	X coord.	Y coord.	Month	Day of week	FFMC	DMC	DC	ISI	Temp	Rel. hum.	Wind	Rain	LS-SVM mean LOO
DT								•	•			•	0.9924
MI						•	•	•	•			•	1.0002
LARS			1										1.0108
			1								2		0.9964
	3		1								2		0.9939
	3		1							4	2		0.9937
	3		1			5				4	2		0.9911
	3		1			5		6		4	2		0.9939
	3	7	1			5		6		4	2		1.0016
	3	7	1			5		6		4	2	8	0.9898
	3	7	1		9	5		6		4	2	8	1.0007
	3	7	1		9	5	10	6		4	2	8	0.9892
	3	7	1		9	5	10	6	11	4	2	8	0.9899
	3	7	1	12	9	5	10	6	11	4	2	8	0.9923
From [38]									•	•	•	•	**0.9848**

Table 5.2: The variables selected by each method for predicting the burned forest area in the Forest fires data set and the resulting mean leave-one-out error of the LS-SVM.

search due to cases where several candidates are exactly at the same distance, a very slight perturbation is added in the variable selection phase to the inputs in order to randomise and even out the effect of this phenomenon.

As the month and weekday variables are cyclic in nature, directly using the Euclidean measure for these is not sensible. Instead, they are mapped to points on a circle, and the distance between points is measured *through* the circle. This may seem somewhat heuristic, but is still sufficiently accurate to be appropriate for a nearest-neighbour search. The mapping is done in this way to still allow the use of highly optimised functions for performing the nearest-neighbour search with the Euclidean metric.

The variables chosen by each method and resulting LOO errors can be seen in Table 5.2. As all of the LOO errors are close to the variance of the data, it is safe to say this is a very difficult problem; in fact, none of the methods perform satisfactorily. The set of variables suggested in [38] does provide the lowest error metric, but by a small margin. As the MSE is nearly as large as the variance of the noise, it may appear that the models are useless, but all hope is not lost. This phenomenon was

	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model year	American?	European?	Japanese?	LS-SVM mean LOO
DT	•	•	•	•		•		•		0.1183
MI		•		•		•				0.1331
LARS				1						0.2893
				1		2				0.1349
				1		2	3			0.1282
			4	1		2	3			0.1119
			4	1	5	2	3			0.1128
			4	1	5	2	3		6	0.1133
		7	4	1	5	2	3		6	0.1113
	8	7	4	1	5	2	3		6	0.1111
	8	7	4	1	5	2	3	9	6	**0.1109**

Table 5.3: The variables selected by each method for predicting the MPG usage in the auto MPG data set and the resulting mean leave-one-out error of the LS-SVM.

recognised in [38], and appears to be related to the extreme difficulty of accurately predicting the spread of large fires. The small fires still tend to be predicted with reasonable accuracy.

The final value of the Delta test is 0.7447, and the reported leave-one-out errors are squarely inside the 1–2 times bracket.

5.4.3 Auto MPG

Here the methods are compared on a data set for predicting the fuel consumption (miles per gallon) of a number of car models [37] with 398 samples. The 7 variables include other performance measures of the models, as well as the year and a discrete "origin" variable. The horsepower information is missing for 6 samples, and for this experiment these are replaced by the mean of the horsepowers of the other models. The origin variable is a class representing the source of the car with three possible values: American, European, or Japanese (all of the models included in the data fall into one of these categories). For this experiment, the information was divided into three binary variables, each representing whether a samples is included in the respective class. Again, as some of the variables are discrete, a slight perturbation is added to the inputs for the variable selection.

The results are shown in table 5.3. As the smallest error is obtained by the

full selection, it seems clear that all of the input variables are relevant. It can be seen that the Delta test fares better than the mutual information, but LARS does manage to get an even smaller LOO error by fewer selected variables, reaching 0.1119 by choosing the 4 top-ranked variables.

The optimal value of the Delta test is 0.0710, and while the leave-one-out error is between one and two times this, there is a significant difference between the noise estimate and the resulting LOO error. In this case, the LS-SVM model is not able to reach close to the smallest generalisation error as predicted by the Delta test. Hence, either the Delta test estimate is over-optimistic—which could be caused by the several discrete variables in the data—or the LS-SVM that is used is not able to capture all the information that could be extracted from the data.

5.4.4 AnthroKids

The AnthroKids data set consists of anthropological measurements of children conducted in the USA in 1977 [39]. The full original data included a total 122 measurements of 3900 individuals. As that data contains several missing values, it has been converted to a regression problem in [40] by assigning the weight to be the target, and retaining 53 variables and 1019 samples without missing values. See Table A.1 of Appendix A for the full list of variables which were retained in this pruning. In addition to physical attributes, the data contains general information about the individuals and the measurement event. It is clear that there are several entirely redundant variables, and variable selection should prove effective.

As 53 variables is far too many to perform an exhaustive search over the selection space, the forward-backward search (starting from the empty selection) method was used instead to optimise the Delta test as well as the mutual information estimator. The selected variables with resulting mean LOO errors are presented in Table 5.4. For LARS, only the results for the first 20 variables are included, as the addition of any further variables did not notably decrease the LOO error.

The Delta test chooses 9 out of the 53 variables, resulting in a better model than by any of the other selection methods. The value returned by the Delta test is 0.0096, which is a reasonably accurate estimate of the resultant mean leave-one-out error of 0.0109. The lowest mean LOO provided by the LARS selections is very close, 0.0115, which is obtained by using the 11 top ranked variables. However, as the method does not specify how many variables to use, actually reaching this level of accuracy would require the tuning, training, and evaluating of up to 53 models in order to consider the different possibilities. The Delta test, on the other hand, gives

	The selected variables:	LS-SVM mean LOO
DT	1 2 4 18 20 35 36 37 39	**0.0109**
MI	1 35 37 39	0.0130
LARS	35	0.0475
	35 39	0.0285
	35 39 21	0.0277
	35 39 21 37	0.0193
	35 39 21 37 17	0.0181
	35 39 21 37 17 2	0.0149
	35 39 21 37 17 2 3	0.0149
	35 39 21 37 17 2 3 36	0.0140
	35 39 21 37 17 2 3 36 19	0.0140
	35 39 21 37 17 2 3 36 19 33	0.0141
	35 39 21 37 17 2 3 36 19 33 20	0.0115
	35 39 21 37 17 2 3 36 19 33 20 48	0.0119
	35 39 21 37 17 2 3 36 19 33 20 48 44	0.0120
	35 39 21 37 17 2 3 36 19 33 20 48 44 49	0.0122
	35 39 21 37 17 2 3 36 19 33 20 48 44 49 51	0.0130
	35 39 21 37 17 2 3 36 19 33 20 48 44 49 51 53	0.0139
	35 39 21 37 17 2 3 36 19 33 20 48 44 49 51 53 52	0.0154
	35 39 21 37 17 2 3 36 19 33 20 48 44 49 51 53 52 46	0.0162
	35 39 21 37 17 2 3 36 19 33 20 48 44 49 51 53 52 46 16	0.0166
	35 39 21 37 17 2 3 36 19 33 20 48 44 49 51 53 52 46 16 40	0.0169
	⋮	
All	all 1–53	0.0355

Table 5.4: The variables selected by each method in the AnthroKids data set, and the resulting mean LOO errors for the LS-SVM. See Table A.1 for the descriptions of the variables.

a one-shot result, which is still slightly better in this case.

5.4.5 Time Series Prediction: Santa Fe A Laser Data

One interesting application where variable selection is often required is in auto-regressive time series prediction, and hence the methods are also tested on a well known time series problem selected from the Santa Fe Time Series Competition: the laser data known as Santa Fe A [41, 42] (Figure 5.2). The data consists of 1000 samples of intensity data of a Far-Infrared-Laser in a chaotic state, and it features

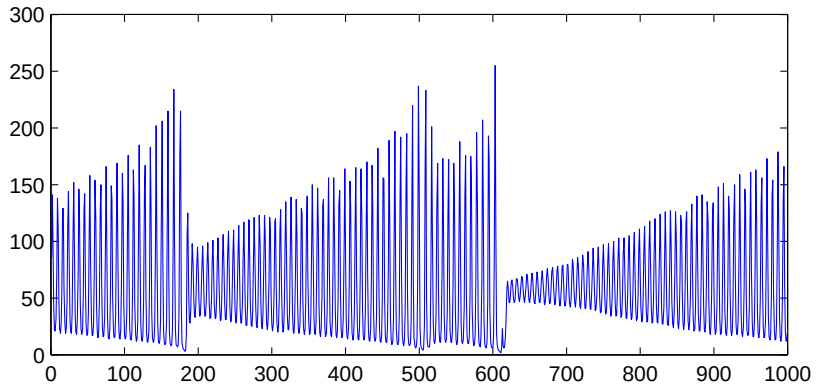

Figure 5.2: The Santa Fe A Laser time series data.

a clear periodicity. The amplitude of the oscillation varies by time, and appears to collapse at irregular intervals. The task is to perform one-step-ahead prediction, and it has been shown that a regressor size of 12 should suffice to train an efficient model. The variable selection then pertains to which of the delayed regressors (up to a delay of 12) should be used to build the model.

The results are shown in Table 5.5. The Delta test performs very well, leading to the best model by a significant margin, while choosing only three of the regressor variables.

The final value of the Delta test is 0.0165. As the LOO error is slightly smaller, this could suggest that overfitting might be occurring. However, it appears more likely that the discrepancy would be caused by behaviour around the rare events in the time series where the collapses occur.

	1	2	3	4	5	6	7	8	9	10	11	12	LS-SVM mean LOO
DT	•	•										•	**0.0144**
MI	•						•	•					0.0837
LARS								1					0.3774
							2	1					0.1329
			3				2	1					0.1123
	4		3				2	1					0.0279
	4	5	3				2	1					0.0206
	4	5	3	6			2	1					0.0205
	4	5	3	6	7		2	1					0.0200
	4	5	3	6	7	8	2	1					0.0209
	4	5	3	6	7	8	2	1	9				0.0250
	4	5	3	6	7	8	2	1	9	10			0.0332
	4	5	3	6	7	8	2	1	9	10		11	0.0341
	4	5	3	6	7	8	2	1	9	10	12	11	0.0361

Table 5.5: The regressor variables selected by each method for predicting the next value in the Santa Fe Data A Laser data, and the resulting mean LOO errors for the LS-SVM. The indexes represent the delay in the auto-regressive model relative to the sample to be predicted.

Chapter 6

Conclusions

The importance of variable selection procedures in non-linear regression analysis is becoming increasingly important as the size of data sets which can be gathered and handled continues to grow. In addition to reducing the size of the problem, variable selection can improve the performance of regression models by discarding noisy data. Additionally, variable selection provides valuable interpretability of the data by specifying which variables are more relevant than others. This thesis assesses some of the currently available state-of-the-art methods and presents the use of the "Delta test" noise variance estimator for input variable selection.

The theoretical claims presented in Section 4 show that, under reasonable assumptions, the expectation of the Delta test is minimised by the smallest input subset which can optimally explain the variation in the output. Minimising the expectation of the algorithm may seem insufficient considering that data often consists of a single realisation of some random process. However, as has been shown in [25], the variance becomes sufficiently small with a sufficient number of data points so this still implies that a near-minimal Delta test value corresponds to a near-optimal selection of variables.

The method is compared to two alternative methods on five real-world data sets, and the performances of the resulting models are summarised in Table 6.1. The Delta test beats the competition in three out of five cases, and provides comparable results in the remaining two. The method can consequently be considered a viable competitor among the current state-of-the-art.

One particular strength of the Delta test is that it is entirely non-parametric, i.e., it can output an optimal subset of variables without asking the user for parameter values or having to infer them from the data. This is particularly important in the field of machine learning, where entirely automatic processes are valued.

	Housing	Forest Fires	Auto MPG	AnthroKids	Santa Fe
Delta test	**0.0909** (10)	0.9924 (3)	0.1183 (6)	**0.0109** (9)	**0.0144** (3)
Mutual information	0.1485 (4)	1.0002 (5)	0.1331 (3)	0.0130 (4)	0.0837 (3)
LARS	—	**0.9892** (10)	—	0.0115 (11)	0.1123 (3)
No variable selection	0.0944 (13)	0.9923 (12)	**0.1109** (9)	0.0355 (53)	0.0361 (12)

Table 6.1: The best leave-one-out errors achieved by each method on each data set. The integers in parenthesis denote the number of selected variables for the achieved result. The dashes indicate situations where using any subset of top-ranked LARS variables resulted in worse performance than by selecting all the variables.

As the Delta test is based on a nearest-neighbour approach, the idea scales well to high-dimensional situations. However, for large problems, the computational cost of the method may become intractable with a naïve implementation. Hence, care should be taken to appropriately implement both the evaluation of the nearest-neighbour search as well as how to explore the search space efficiently.

In light of the experimental results, the Delta test is apparently most suitable for data where both the inputs and output are continuous. Both of the data sets where the method was outperformed by another method included discretised input variables. The possible extension from regression to classification problems is also interesting, but far from obvious. Another appealing future development is if and how the idea can be extended to encompass other forms of dimensionality reduction, such as scaling or linear projection.

The main limitation for the method's use is currently a practical one of size. An exhaustive survey of the search space is insurmountable if the number of variables is much beyond 20. Although there are some tried and true search methods, they appear sub-optimal, and research is ongoing on new, more efficient, algorithms.

The theoretical and experimental results support the notion that the method can provide desirable results in a wide variety of regression modelling problems. As the technique is both simple and robust it can be recommended as a suggested preprocessing step for nearly any regression task.

Bibliography

[1] Antti Sorjamaa, Jin Hao, Nima Reyhani, Yongnan Ji, and Amaury Lendasse. Methodology for long-term prediction of time series. *Neurocomputing*, 70(16-18):2861–2869, October 2007.

[2] Qi Yu, Eric Séverin, and Amaury Lendasse. A global methodology for variable selection: Application to financial modeling. In *Mashs 2007, Computational Methods for Modelling and learning in Social and Human Sciences, Brest (France)*, May 10-11 2007.

[3] Amir Navot, Lavi Shpigelman, Naftali Tishby, and Eilon Vaadia. Nearest neighbor based feature selection for regression and its application to neural activity. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 995–1002. MIT Press, Cambridge, MA, 2006.

[4] Antonia Jane Jones. New tools in non-linear modelling and prediction. *Computational Management Science*, 1(2):109–149, 2004.

[5] Dafydd Evans. *The Gamma Test: Data-derived estimates of noise for unknown smooth models using near neighbour asymptotics*. Ph.d. thesis, Cardiff University, 2002.

[6] Elia Liitiäinen, Francesco Corona, and Amaury Lendasse. On nonparametric residual variance estimation. *Neural Processing Letters*, 28(3):155–167, December 2008.

[7] Vladimir Spokoiny. Variance estimation for high-dimensional regression models. *Journal of Multivariate Analysis*, 82(1):111–133, 2002.

[8] Emil Eirola, Elia Liitiäinen, Amaury Lendasse, Francesco Corona, and Michel Verleysen. Using the Delta test for variable selection. In *ESANN 2008, European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 25–30, April 2008.

[9] Emil Eirola, Elia Liitiäinen, Amaury Lendasse, Francesco Corona, Olli Simula, and Michel Verleysen. Variable selection with the Delta test: Theory and practice. *IEEE Transaction on Neural Networks*, 2009. Submitted.

[10] Johan A. K. Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[11] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[12] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[13] Gavin C. Cawley and Nicola L. C. Talbot. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17(10):1467–1475, 2004.

[14] Richard E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, New Jersey, U.S.A., 1961.

[15] Michel Verleysen, Damien François, Geoffroy Simon, and Vincent Wertz. On the effects of dimensionality on data analysis with neural networks. In *Artificial Neural Nets Problem solving methods*, Lecture Notes in Computer Science 2687, pages II105–II112. Springer-Verlag, 2003.

[16] Damien François. *High-dimensional data analysis: from optimal metrics to feature selection*. VDM Verlag Dr. Muller, 2008.

[17] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lotfi A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[18] Mark A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, April 1999.

[19] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[20] Roberto Battiti. Using the mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

[21] Fabrice Rossi, Amaury Lendasse, Damien François, Vincent Wertz, and Michel Verleysen. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems / I Mathematical Background Chemometrics Intell Lab Syst*, 80:215–226, 2006.

[22] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6), 2004.

[23] Hong Pi and Carsten Peterson. Finding the embedding dimension and variable dependencies in time series. *Neural Computation*, 6(3):509–520, 1994.

[24] Aðalbjörn Stefànson, Nenad Koncar, and Antonia J. Jones. A note on the gamma test. *Neural Computing & Applications*, 5(3):131–133, 1997.

[25] Elia Liitiäinen, Michel Verleysen, Francesco Corona, and Amaury Lendasse. Residual variance estimation in machine learning. *Neurocomputing*, 72(16-18):3692–3703, 2009.

[26] Elia Liitiäinen, Francesco Corona, and Amaury Lendasse. Nearest neighbor distributions and noise variance estimation. In *ESANN 2007, European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 67–72, April 2007.

[27] Elia Liitiäinen and Amaury Lendasse. Variable scaling for time series prediction: Application to the ESTSP'07 and the NN3 forecasting competitions. In *IJCNN 2007, Orlando, FL, USA*, pages 2812–2816, August 2007.

[28] Elia Liitiäinen, Amaury Lendasse, and Francesco Corona. Bounds on the mean power-weighted nearest neighbour distance. *Proceedings of the Royal Society A*, 464(2097):2293–2301, September 2008.

[29] Mathew D. Penrose. Laws of large numbers in stochastic geometry with statistical applications. *Bernoulli*, 13(4):1124–1150, 2007.

[30] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, 1975.

[31] Christopher M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, Oxford, UK, 1996.

[32] Fred Glover and Manuel Laguna. *Tabu Search*. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

[33] Alberto Guillén, Dušan Sovilj, Fernando Mateo, Ignacio Rojas, and Amaury Lendasse. Minimizing the Delta test for variable selection in regression problems. *Int. J. High Performance Systems Architecture*, 1(4):269–281, 2008.

[34] Alberto Guillén, Dušan Sovilj, Fernando Mateo, Ignacio Rojas, and Amaury Lendasse. New methodologies based on delta test for variable selection in regression problems. In *Workshop on Parallel Architectures and Bioinspired Algorithms*, October 2008.

[35] Timo Similä and Jarkko Tikka. Multiresponse sparse regression with application to multidimensional scaling. In *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005, 15th International Conference, Warsaw, Poland, September 11-15, 2005, Proceedings, Part II*, volume 3697 of *Lecture Notes in Computer Science*, pages 97–102. Springer, 2005.

[36] Kristiaan Pelckmans, Johan A. K. Suykens, T. Van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor, and J. Vandewalle. LS-SVMlab: a Matlab/C toolbox for least squares support vector machines. Available: `http://www.esat.kuleuven.be/sista/lssvmlab/`.

[37] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[38] Paulo Cortez and Aníbal Morais. A data mining approach to predict forest fires using meteorological data. In J. Neves, M. F. Santos, and J. Machado, editors, *New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007—Portuguese Conference on Artificial Intelligence*, pages 512–523, December 2007.

[39] AnthroKids — Anthropometric data of children, 1977. Available: `http://ovrt.nist.gov/projects/anthrokids/`.

[40] Fernando Mateo and Amaury Lendasse. A variable selection approach based on the delta test for extreme learning machine models. In *Proceedings of the European Symposium on Time Series Prediction*, pages 57–66. d-side publ. (Evere, Belgium), September 2008.

[41] Andreas S. Weigend and Neil A. Gershenfeld, editors. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley, Reading, MA, 1994.

[42] The Santa Fe time series competition data, 1991. Available: `http://www-psych.stanford.edu/~andreas/Time-Series/SantaFe.html`.

Appendix A

AnthroKids Variables

This appendix presents the complete list of measured variables included from the AnthroKids data set after pruning some variables and samples with missing values, and the ranking by the LARS algorithm.

Index	Selected by DT	Selected by MI	LARS ranking	Description
1	•	•	50	Stature
2	•		6	Erect sitting height
3			7	Maximum hip breadth, sitting
4	•		47	Buttock-knee length
5			52	Head circumference
6			36	Head breadth
7			45	Bizygomatic breadth
8			33	Frontal breadth
9			26	Lower face height
10			23	Face height
11			49	Tragion to back of head
12			31	Tragion to top of head
13			34	Ear-sellion depth
14			22	Bitragion breadth
15			21	Mouth breadth
16			19	Nose length
17			5	Shoulder breadth
18	•		44	Shoulder-elbow length
19			9	Upper arm circumference
20	•		11	Elbow-hand length (lower arm length)
21			3	Forearm circumference
22			39	Hand length
23			37	Hand breadth
24			53	Minimum hand clearance
25			46	Thumb length
26			30	Thumb diameter
27			28	Index finger length
28			41	Index finger diameter
29			48	Middle finger length
30			42	Middle finger diameter
31			43	Middle finger-thumb grip length
32			38	Maximum fist circumference
33			10	Maximum fist breadth
34			40	Maximum fist depth
35	•	•	1	Chest circumference at axilla
36	•		8	Waist circumference
37	•	•	4	Hip circumference at buttocks
38			24	Upper thigh circumference
39	•	•	2	Calf circumference
40			20	Foot length
41			25	Foot breadth
42			27	Age in years
43			29	Sex
44			13	Location
45			51	Age in months
46			18	Birth date
47			35	Measurement date
48			12	Measurer number
49			14	Computer number
50			32	Race
51			15	Handedness
52			17	Twin
53			16	Birth order

Table A.1: The full list of AnthroKids variables.