

Fitting Generalized Additive Models for very large datasets with Apache Spark

Chapter Excerpt

Kai Thomas Brusch

24.12.2015

Summary This document contains three introductory chapters for my bachelor thesis titled "Fitting General Additive Models for very large datasets with Apache Spark".

Contact `kai.brusch@gmail.com`

Location Hochschule für Angewandte Wissenschaften Hamburg

Department Dept. Informatik

Examiner Prof. Dr. Michael Köhler-Bußmeier

Second examiner Dipl.-Math. Markus Schmaus

Contents

1	Linear models	3
1.1	Introduction to linear models	3
1.2	Example of a linear model	3
1.3	Ordinary least square estimation of β	3
1.4	Gauss-Markov Theorem	5
1.5	Estimating $\hat{\beta}$ with orthogonal decomposition	6
2	Generalized Linear Models	6
2.1	Introduction to Generalized Linear Models	6
2.2	Example of a Generalized Linear Models	6
2.3	Maximum likelihood estimation	7
2.4	Fitting Generalized Linear Model	9
3	Generalized Additive Models	10
3.1	Introduction to Generalized Additive Models	10
3.2	Generalized Additive Model example	10
3.3	Smoothing Functions	10
3.4	Regression Splines	11
3.5	Smoothing Parameter estimation	11
3.6	Fitting Generalized Additive Models	11
4	Matrix Algebra	11
4.1	Orthogonal Matrices	11
4.2	QR decomposition	11
5	References	12

1 Linear models

1.1 Introduction to linear models

Linear models are statistical models in which an univariate response is modeled as the sum of a ‘linear predictor’ and a zero mean random error term. The linear predictor depends on some predictor variables y , measured with the response variable x , and some unknown parameters β plus an error term ϵ , which must be estimated. This process is formally stated for a given row i of data as: [Wood, 2006] [Wood et al., 2015] [Zaharia et al., 2010]

$$y_i = \beta x_i + \epsilon_i \quad (1)$$

There are many choices for β and finding the best possible β stands at the heart of the following chapter. A key feature of linear models is that the linear predictor depends linearly on these parameters. Statistical inference with such models is usually based on the assumption that the response variable has a normal distribution. Linear models are used widely in most branches of science and an example is given in the next section.

1.2 Example of a linear model

Before describing the relevant theory I would like to give an example of a simple linear model. Let's say we have a data set called `mtcars` that describes cars with respect to miles per gallon (`mpg`) and horse power (`hp`). After careful thought and the examination of a scatter plot I believe that there is a linear relationship between the miles per gallon and the horse power of a car. I also believe that miles per gallon follow normal distribution. I would try to explain the relationship between `hp` and `mpg` as a linear model takes miles per gallon as the dependant variable and horse power as the independant variable. This model description written in R yields the following line of code:

```
model <- lm(data=mtcars, hp ~ mpg)

summary(model)
```

The summary gives me the estimated model and we can see the slope and the intercept of function explaining the relationship between dependant and independant variable:

```
Call:
lm(formula = hp ~ mpg, data = mtcars)

Coefficients:
(Intercept)          mpg
    297.688         -8.022
```

Our model describes `hp` power as $-8.022 * mpg + 297.688$. We can see that the more `hp` a car has the fewer `mpg` does it offer. The red line in the plot is our estimated function that illustrates the estimated model

The natural question that arises is: how do we estimate the line and the error of the chart above? The relevant method is called ordinary least squares and will be introduced in the next section.

1.3 Ordinary least square estimation of β

We are now looking at methods of finding β . Ideally we want to choose a β that produces a line through our data points with minimal distance between our points and our estimated line.

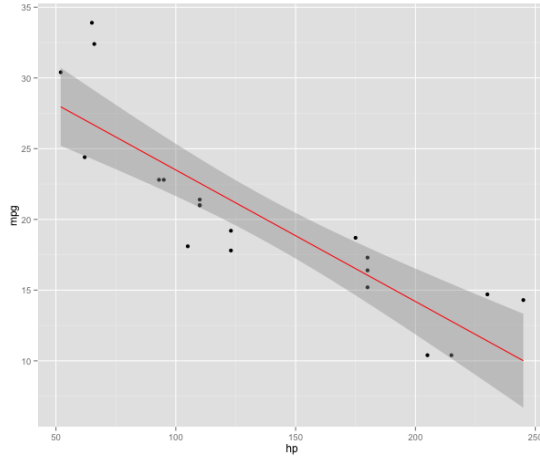


Figure 1: $-8.022 * mpg + 297.688$

More precicely we are looking to estimate a β that minimizes the squared distance between an estimated β times the given x and y . We are squaring the distance to normalize negative and positive differences. This distance formaly describes as S:

$$\mathbf{S} = \sum_{i=1}^n (y_i - x_i\beta)^2 \quad (2)$$

The close our S gets to 0 the better our line fits the data. The Markov-Gauss Theorem states that the minimization of S yields $\hat{\beta}$ which is the best possible estimation for β . This shall be discussed n the next chapter. 1 represents the univariate case where y is explained with only one variable. The process of minimizing S is commonly refered to as ordinary least squares (OLS).

There are two ways to think about estimating β , thinking off OLS in terms of calulus on functions allows us to compute $\hat{\beta}$ for the univariate case while computing $\hat{\beta}$ for multiple independant variables requires linear algebra. From the calculus perspective we can see OLS as function with two parameters: S and β . Minimizing S equates to taking the partial derivative of S with repsect to β . This is the most common approach and offers insight by stating S as the following equation:

$$\frac{\partial S}{\partial \beta} = - \sum_{i=1}^n 2x_i(y_i - x_i\beta) \quad (3)$$

Rewriting the partial dereviative to yiel $\hat{\beta}$ gives a very good idea of β

$$- \sum_{i=1}^n 2x_i(y_i - x_i\hat{\beta}) = 0 \quad (4)$$

$$- \sum_{i=1}^n x_i y_i - \hat{\beta} \sum_{i=1}^n x_i^2 = 0 \quad (5)$$

$$\hat{\beta} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2 \quad (6)$$

Minimizing S w.r.t. β to compute $\hat{\beta}$ is a reasonable approach when dealing with one independant variable. However, almost all relevant applications involve much more than one independant

variable and require linear algebra. To estimate $\hat{\beta}$ with OLS for multiple independent variables involves rephrasing the questions in terms of linear algebra. First we have to state the process of finding β as a linear combination problem which equates to asking what linear combination of column vectors of our model matrix X and our vector β of unknown coefficients yields the vector of y . Formally:

$$X\beta = y \quad (7)$$

Now that we have restated OLS as a matrix problem we can apply some linear algebra to find $\hat{\beta}$. First we can restate the problem of finding β as finding the unknown vector β equates to finding a linear combination of our column vectors of X with β that result in y . Finding this linear combination is highly dependant on the properties of X . Taking a look at relevant model matrices X will show that they almost exclusively consist of m rows and n columns with $m > n$. Being $m > n$ implies that the matrix is not symmetric and not invertible. Given that X has more rows than columns we can think of $X\beta = y$ as a system of equations with more equations than variables which causes this system of equations to have no solution. We have to stress the fact that the system of equations of 7 does not have a solution because there is no possible selection for β that lies in the column vector space of X . Whilst the nature of X makes finding β impossible we can find an estimation $\hat{\beta}$ of β by projecting it back into the column space of X . The solution becomes to multiply by A^T . The proper projection p is defined as $p = X\hat{\beta}$. The process of projecting of finding the projection involves multiplying by the transpose of the model matrix yielding:

$$X^T X \hat{\beta} = X^T y \quad (8)$$

This projection comes however at the cost of an error term:

$$y = X\hat{\beta} + \epsilon \quad (9)$$

$$\epsilon = y - X\hat{\beta} \quad (10)$$

The questions that now arises is: how to we make $\epsilon = y - X\hat{\beta}$ as small as possible? Using algebra we can split the vector β into two parts. One part in the column space is our projection p and the perpendicular part in the nullspace of A^T which the ϵ . It is essential to remember that the column space is always perpendicular to the nullspace of A^T . The solution to $X\beta = p$ leaves the least possible error ϵ , returning to the previously stated method of least squares, but this time in the world of linear algebra:

$$\|X\beta - y\|^2 = \|X\beta - p\|^2 + \|\epsilon\|^2 \quad (11)$$

This is the law $c^2 = a^2 + b^2$ for a right angle. The vector $X\beta - p$ in the column space is perpendicular to ϵ in the nullspace of A^T . We reduce $X\beta - p$ to zero choosing β to be $\hat{\beta}$, this leaves us with the smallest possible error vector ϵ . The projection leaves us with an invertible matrix that can be solved by usual elimination. The least squares solution $\hat{\beta}$ makes $\epsilon = X\beta$ as small as possible.

1.4 Gauss-Markov Theorem

So far we have introduced two different methods on how to estimate β . The calculus way for univariate data and the linear algebra way for multivariate data, both estimated β with an approximate *beta* which seeks to minimize the squared error of actual and estimate. While squared error seems like the an obvious choice we have no formal reason to prefer it to other measures. This section will lay out the formal proof that shows why $\hat{\beta}$ is the best possible estimator.[Wood, 2006] states that there exists no estimator with lower variance than least squares estimation. The

1.5 Estimating $\hat{\beta}$ with orthogonal decomposition

The previously suggested method lends a great tool to think about the univariate least squares technique. But the suggested method is rarely applied. Any m by n matrix X with independent columns can be factored into QR. The m by n matrix Q has orthonormal columns and the square matrix R is upper triangular with positive diagonal. $X^T X$ equals $R^T Q^T Q R = R^T R$ simplifying the least squares equation to $Rx = Q^T y$; allowing us to multiply by R^T instead of X^T for form the square matrix. This additional simplicity allows us to restate the problem of least squares for matrices using QR decomposition as:

$$R^T R \hat{\beta} = R^T Q^T y \text{ or } R \hat{\beta} = Q^T y \text{ or } \hat{\beta} = R^{-1} Q^T y \quad (12)$$

A very important property of orthogonal matrices is that their multiplication with a vector is independent with respect to the length of the vector.

Instead of multiplying with a fully formed model matrix X we only multiply with small subset, thus reducing the amount of multiplications and memory requirements drastically.

The QR decomposition is formally stated in the appendix and will appear as an essential part later in this thesis.

2 Generalized Linear Models

2.1 Introduction to Generalized Linear Models

[?] describes Generalized Linear Models (GLMs) as an extension of the general linear model with the ability to model more exponential family response distributions. While linear models with OLS estimation only allow for a normal distributed response, GLMs allow to model the response variable from any arbitrary exponential family. The exponential family of distributions contains many practical useful distributions. A formal description of its basis structure can be given as

$$g(\mu_i) = X_i \beta_i \quad (13)$$

Where $\mu_i \equiv E(Y_i)$ and Y_i is distributed according to some exponential family. Its members include but are not limited to Poisson, Binomial, Gamma and Normal Distributions. Every exponential family distribution has a link function $g()$. The idea here is that linear functions of the predictor variables are obtained by transforming the right side of the equation ($f(x)$) by a link function. The data are then fit in this transformed scale (using an iterative routine based on least squares), but the expected variance is calculated on the original scale of the predictor variables. X_i is the i^{th} row of a model matrix X and β is a vector of unknown parameters. The link function allows us to model an exponential function in terms of a linear link function. To account for exponential family distributions comes at a cost however: While OLS was sufficient for estimating β for normal distributed data now have to generalize this notion to account for an arbitrary amount of distribution parameters. The generalization of OLS is called maximum likelihood estimation (MLE) and involves an iterative method called iterative re-weighted least squares (IRLS). An important practical feature of generalized linear models is that they can all be fit to data using IRLS, independent of response variable distribution. A further layer of complexity is the error term now has varying variance, our estimations and inference needs weight the likelihood accordingly.

2.2 Example of a Generalized Linear Models

Examples for generalized linear models are found in many fields of science. Allowing the response to be distributed to Poisson or binomial distributions allow to model many practical applications. Let us consider an example of a Poisson distributed response. Poisson distributions are most

commonly used to model the number of occurrences of a given event. Lets say we want to discover the relationship between smoking tabaco und number of death for a given age group drawn from a random population. We are given a data set that describes a random sample in terms of age range, deaths and a categorial desription of the smoking habit. The age groups are non-overlapping five year spans starting at 40. The smoking habit is categorised as ("no", "cigarPipeOnly", "cigarettePlus", "cigaretteOnly") We can use the given data to produce a model that models the number of deaths as a function of smoking habit and age group. The model we are trying to build can be best described with the following R code snipped:

```
Call: glm(formula = dead ~ +0 + as.factor(smoke) + as.factor(age),
family = poisson(link = "log"), data = data)
```

Coefficients:

as.factor(smoke) cigarPipeOnly	as.factor(smoke) cigaretteOnly
3.5088	4.6131
as.factor(smoke) cigarettePlus	as.factor(smoke) no
4.8864	3.2702
as.factor(age)45–59	as.factor(age)50–54
0.1340	0.3163
as.factor(age)55–59	as.factor(age)60–64
1.3956	1.9252
as.factor(age)65–69	as.factor(age)70–74
1.8867	1.5894
as.factor(age)75–79	as.factor(age)80+
1.1377	0.7576

The estimated coefficients paint a clear picture: Less deaths occur in non-smoking groups. A little more for cigar pipe only smokers and many more for cigaretter smokers, indeferent to frequency. However, the coefficient interpretation should be secondary in the example and focus should on the model description. A Poisson distribution need only one parameter thus we only supply the number of death occured. The have to apply the link function in the right side of our equation, yielding in the following model description $Poisson(dead) = Constant + log(smoke) + log(age)$. We can see that used a GLM model with death as the reponse and specified the expotential family and its link. Taking a look at the variance of the

2.3 Maximum likelihood estimation

Parameter estimation stands at the heart of all statistical models. Random variables and their distribution allows us to explain a variables behavior based on their parameters. Once a model is specified with its parameters, and data have been collected, one is in a position to evaluate its goodness of fit, that is, how well it fits the observed data. Goodness of fit is assessed by finding parameter values of a model that best fits the data procedure called parameter estimation. The OLS linear model estimated the μ and σ for $N(\mu, \sigma^2)$ by minimizing S . [J. A. Nelder, 1972] introduces a method to account estimate parameters for any expotential family distributions members. Expotential family distributions have diffent number of parameters and parameter meaning, cf. $P(\lambda) G(K, \theta)$. Maximum likelihood estimation provides a single framework to allow parameter estimation for any of the expotential family distributions.

From a statistical analysis point of view the vector y of observed data is a random sample from an unknown population. The goal of maximum likelihood estimation is to find the parameters of the given distribution that most likely has produced this sample. This process is discribed with a probability density function (PDF) $f()$ of observed data y given a parameter w : $f(y|w)$. If individual observations, y_i 's, are statistically independent of one another, then according to the

theory of probability, the PDF for the data y given the parameter vector w can be expressed as a multiplication of PDFs for individual observations.

$$f(y|w) = f((y_1, y_2, \dots, y_n)|(w_1, w_2, \dots, w_n)) = \prod_{i=1}^n f_i(y_i|w_i) \quad (14)$$

Given a set of parameter values, the corresponding PDF will show that some data are more probable than other data. In reality the data is already given and we are searching for the parameters of the distribution that most likely produced the data. We thus inverse our function $f(y|w)$ to $L(w|y)$ to produce the likelihood of y given the parameters w . The principle of maximum likelihood estimation, originally developed by R.A. Fisher in the 1920s, states that the desired probability distribution is the one that makes the observed data “most likely,” which means that one must seek the value of the parameter vector that maximizes the likelihood function $L(w|y)$: The resulting parameter vector, which is sought by searching the multi-dimensional parameter space, is called the MLE estimate.

Theoretically MLE estimates need not exist nor be unique. But if they exist and are unique they can be estimated. For computational convenience, the MLE estimate is obtained by maximizing the log-likelihood function, $\ln(L(w|y))$: This is because the two functions, $\ln(L(w|y))$ and $L(w|y)$; are monotonically related to each other so the same MLE estimate is obtained by maximizing either one and the log-likelihood is preferred for obvious reasons. Assuming that the log-likelihood function, $\ln(L(w|y))$ is differentiable, if wMLE exists, it must satisfy the following partial differential equation known as the likelihood equation:

$$\frac{\partial \ln L(w|y)}{\partial w_i} = 0 \quad (15)$$

This is because the definition of maximum or minimum of a continuous differentiable function implies that its first derivatives vanish at such points. The likelihood equation represents a necessary condition for the existence of an MLE estimate. An additional condition must also be satisfied to ensure that $\ln(L(w|y))$ is a maximum and not a minimum, since the first derivative cannot reveal this. To be a maximum, the shape of the log-likelihood function should be convex (it must represent a peak, not a valley) in the neighborhood of wMLE: This can be checked by calculating the second derivatives of the log-likelihoods and showing whether they are all negative.

$$\frac{\partial^2 \ln L(w|y)}{\partial w_i^2} < 0 \quad (16)$$

In practice, however, it is usually not possible to obtain an analytic form solution for the MLE estimate, especially when the model involves many parameters and its PDF is highly non-linear. In such situations, the MLE estimate must be sought numerically using nonlinear optimization algorithms. The basic idea of nonlinear optimization is to quickly find optimal parameters that maximize the log-likelihood. This is done by searching much smaller sub-sets of the multi-dimensional parameter space rather than exhaustively searching the whole parameter space, which becomes intractable as the number of parameters increases. The “intelligent” search proceeds by trial and error over the course of a series of iterative steps. Specifically, on each iteration, by taking into account the results from the previous iteration, a new set of parameter values is obtained by adding small changes to the previous parameters in such a way that the new parameters are likely to lead to improved performance. Different optimization algorithms differ in how this updating routine is conducted. This process is called

2.4 Fitting Generalized Linear Model

Before looking at the fitting process of GLM it is important linear models assume an error with constant variance and no covariance, the combination of the two conditions is called homoscedastic. The assumption of homoscedastic is now dropped, in GLM the error term has a varying variance and error can have covariance. This situation is called heteroscedastic. Our fitting method thus needs to account for the varying error. As we have discussed previously, we obtain MLEs by setting the score vector equal to 0 Recall that for a GLM using the canonical link function, the score vector is:

$$u(\beta) = \phi^{-1} X^T (y - \mu) \quad (17)$$

In the above equation μ is a function of $\eta = X\beta$. However this is not required to be a linear function, if this is not the case lack a closed-formed solution for β . If we are given the canoical link function g we may use a Taylor series approach to obtain the following approxiamtion of the point $\hat{\beta}$:

$$\mu = \hat{\mu} + W(X\beta - X\hat{\beta}) \quad (18)$$

Where $\hat{\mu} = g^{-1}(X\hat{\beta})$ when the g is defined as

$$\frac{d}{d\eta} g^{-1}(\eta) = W(\eta) \quad (19)$$

Thus we can obtain the following linear approximation to the score for a given β :

$$\frac{\partial l}{\partial \beta} \approx \phi^{-1} X^T W(z - X\beta) \quad (20)$$

where $z = X\hat{\beta} + W^{-1}(y - \hat{\mu})$ is known as the adjusted response. Note that this approximation is based at $\hat{\beta}$ or, equivalently, $\hat{\mu}$, which are treated as constants in the above expression, thereby rendering the score equation linear in β after the approximation Again, recall that this approximation will be accurate near the fitted values $\hat{\mu}$ but not necessarily accurate far away from them. As we saw previously this gives the max likelih hood estimate:

$$\hat{\beta}^{(m)} = (X^T W X)^{-1} X^T W z \quad (21)$$

Note that W here plays the role of the weights in weighted least squares, and for that reason is often referred to as the weight matrix. Again, recall that for the canonical link, W is entirely determined by the mean-variance relationship, and that it plays a prominent role in the variability of $\hat{\beta}$ as well. Note that in the above equation, we require a superscript on $\hat{\beta}^{(m)}$ because this is a case of unknown weights, where W (and z) will change depending on $\hat{\beta}$ and vice versa

As we saw earlier, one way to address this problem is to iterate the process of reweight–estimate–reweight–estimate–. . . until convergence; this iteratively reweighted least squares (IRLS) algorithm is how generalized linear models are fit:

1. Choose an intial value $\hat{\beta}^{(0)}$
2. For $m = 0, 1, 2, \dots$
 - 2.1. Calculate z and W based on $\hat{\beta}^{(m)}$
 - 2.2. Solve for $\hat{\beta}^{(m+1)}$
 - 2.3. Check to see whether $\hat{\beta}^{(0)}$ has converged: if yes then stop

This concept is generalized with Newton-Raphson

3 Generalized Additive Models

3.1 Introduction to Generalized Additive Models

Generalized Additive Models (GAMs) extends the GLM by specifying the linear prediction in terms of the summation of smooth functions. This allows for a more flexible modeling of the influence for each explanatory variable. The gained flexibility comes at the cost of additional questions concerning the smooth function:

GAMs are fomally described by the following equation:

$$g(\mu_i) = \mathbf{X}_i\Theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i})\dots \quad (22)$$

22 explains y_i as the model matrix for this row and the smooth functions $f_j(x_{1j})$ of the x values for this row. X_i is a row of the model matrix with parametric component θ . Unlike the linear model we can now specify a smooth function for each explanatory variable. This proves to be way more flexible than only allowing for a constant influence per explanatory variable. The natural question that arises now are: How do I find proper smoothing functions? Finding the right smooth function stands at the heart of GAM fitting and can be best illustrated in the univariate case 23.

3.2 Generalized Additve Model example

WOODS HELP ME!!!

3.3 Smoothing Functions

$$y_i = f(x_i) + \epsilon_i \quad (23)$$

Smooth functions form a vector space, which can be approximated using a linear basis. Only allowing linear basis allow us to heavily leverage the theory already developed for linear models and S as the optimal model fit. For the sake of illustration we assume that 23 can be rewritten as the following equation if $b_i(x)$ is the i th basis function:

$$f(x) = \sum_{i=1}^q b_i(x)\beta_i \quad (24)$$

In 24 we already know $f()$ is linear in regard to 24. We now have to specify a basis function to represent b_i . We can choose from many basis functions for b_i , each with advantages and disadvantages. A common choice however is a fourth order polynomial basis function. 24 represented by a fourth order polynomial yields the following model:

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5 \quad (25)$$

Applying 25 to 23 we get the modeling of y_i as the sum of smoothing functions.

$$y_i = \beta_1 + x_i\beta_2 + x_i^2\beta_3 + x_i^3\beta_4 + x_i^4\beta_5 + \epsilon_i \quad (26)$$

3.4 Regression Splines

3.5 Smoothing Parameter estimation

3.6 Fitting Generalized Additive Models

4 Matrix Algebra

4.1 Orthogonal Matrices

We are concerned with matrices that do have orthogonal column vectors, severely concerned

4.2 QR decomposition

We want to decompose a matrix A in two parts, one orthonormal and one upper triangular, $X = QR$

5 References

- [Dobson and Barnett, 2008] Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models, Third Edition*. Texts in Statistical Science. Chapman & Hall/CRC Press, Boca Raton, FL. For more information about this book please refer to the publisher’s website (see link) or contact the author.
- [J. A. Nelder, 1972] J. A. Nelder, R. W. M. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.
- [Wood, 2006] Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
- [Wood et al., 2015] Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.
- [Zaharia et al., 2010] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud’10, pages 10–10, Berkeley, CA, USA. USENIX Association.