

Maximum likelihood estimation under inequality constraints in univariate linear models

Katarzyna Filipiak, Dietrich von Rosen, Martin Singull

Abstract This paper determines maximum likelihood estimates under general univariate linear models with a priori information related to maximum effects in the models. The negative log-likelihood functions and the constraints are convex functions, so convex optimization theory can be utilized to obtain relevant estimates. In particular, the complementary slackness condition, common in convex optimization, implies two alternative types of solutions, strongly dependent on the data and the restriction.

1 Introduction

Estimation techniques for estimating regression parameters are often based on penalized fitting functions with the ultimate goal to find estimators with desirable properties and also sometimes simultaneously include appropriate covariables. Several methods have been proposed; for example, [3] proposed bridge regression, [6] introduced LASSO, [2] came up with SCAD, [8] proposed the elastic net and [7] introduced the adaptive LASSO. However, if the variable selection is not of main concern, the ridge regression method [5] is also commonly used (see also Tikhonov regularization) and can be viewed as a penalized estimation method. In fact, the

Katarzyna Filipiak
Poznań University of Technology, Institute of Mathematics, Piotrowo 3A, PL-60965 Poznań,
Poland
e-mail: katarzyna.filipiak@put.poznan.pl

Dietrich von Rosen
Linköping University, Department of Mathematics, SE-581 83 Linköping, Sweden
e-mail: dietrich.von.rosen@liu.se

Martin Singull
Linköping University, Department of Mathematics, SE-581 83 Linköping, Sweden
e-mail: martin.singull@liu.se

above given references connect to ridge regression which appeared earlier. All the approaches modify the estimation function and the usual scenario is that the knowledge between data and the statistical model is vague.

In this work it will be utilized that there is specific knowledge (a priori information) about the maximum effects in a model which will be studied; for example, in physical and chemical processes one often knows the “bounds” of the processes. This knowledge is supposed to be quantifiable and then it can be built into the models via inequality constraints leading to a model closer to reality.

So far these ideas have not been implemented in statistics. In this article it will be shown how it can be carried out in linear models. We will stand on convex optimization theory. This means that the estimation functions (least squares function and negative log-likelihood function) should be convex functions with respect to the parameters in the model, which sometimes requires a model reparametrization. Moreover, the function which quantifies the prior information should also be convex. The approach in the present work also demands the functions to be differentiable which except ridge regression, usually is not the case with the above-mentioned variable selection methods.

In the article bold lower cases denote vectors whereas bold upper cases denote matrices. Other notations will be defined when they are introduced in the main body of the text.

2 Safety belt estimates and the general univariate linear model

Maximum likelihood estimation will be applied, mainly because estimation of the mean parameters and estimation of the variance can be handled simultaneously.

Let \mathbf{y} : $n \times 1$, be the response vector. We will not distinguish between the observation vector \mathbf{y} , which is thought of being a realization of a random vector \mathbf{y} , and the random vector \mathbf{y} itself. This should however not lead to any confusion. Moreover, let \mathbf{X} : $n \times k$, represent the usual known design matrix in a linear model (usually of full column rank, with the first column consisting of vector of ones) with corresponding unknown mean effects $\boldsymbol{\beta}$: $k \times 1$, and let $\boldsymbol{\varepsilon}$: $n \times 1$, be the vector of unobservable independently normally distributed variables with mean 0 and common variance σ^2 . These assumptions can be summarized as

$$\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \quad (1)$$

where \mathbf{I}_n stands for the $n \times n$ identity matrix. A crucial assumption is that there exist a priori information about $\boldsymbol{\beta}$ so that it makes sense to assume that $\boldsymbol{\beta}^\top \boldsymbol{\beta} \leq t$ for some given $t > 0$, “ \top ” denotes the transpose, or a bit more general $\boldsymbol{\beta}^\top \mathbf{H} \boldsymbol{\beta} \leq t$, for some positive semi-definite matrix \mathbf{H} . The matrix \mathbf{H} can be used to select components from $\boldsymbol{\beta}$, for example via $\mathbf{H} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, where $\mathbf{0}$ stands for a matrix with elements equal to 0. Obviously $\mathbf{H} = \mathbf{I}$ is another choice. However it can also be written that for $\mathbf{H} = (h_{ij})$ and $\boldsymbol{\beta} = (\beta_i)$, $\boldsymbol{\beta}^\top \mathbf{H} \boldsymbol{\beta} = \sum_{ij} h_{ij} \beta_i \beta_j$ indicating the effect of choice of elements

in \mathbf{H} . The use of an \mathbf{H} -matrix in ridge regression analysis has been considered in the literature (e.g., see [4, p.113]) and for example by choosing $\mathbf{H} = \mathbf{X}^\top \mathbf{X}$ one obtains the James-Stein estimator of β .

Since the a priori information is quantified via inequality constraints on the parameter space (squares of the parameters), we do not deal with linear subspaces and therefore cannot refer to the usual linear models theory. Moreover, in many applications $\beta^\top \mathbf{H} \beta \leq t$ makes sense. For example, if using linear regression for the study of a human body height over time we usually know pretty precisely what the maximum growth intensity will be. In fact, in most planned experiments one has the knowledge about maximum effects. In particular, measurement instruments have to be calibrated within certain ranges. However, in observational studies with many interacting variables, the upper bound t can be more difficult to set.

A general rule in statistics is to include a priori information when building a model as long as it is mathematically motivated and one does not lose interpretability. The next proposition balances inclusion of information and interpretability.

The proposition presents maximum likelihood estimates under inequality restrictions on $\beta^\top \mathbf{H} \beta$, where \mathbf{H} is supposed to be non-random known positive semi-definite matrix. Moreover, it is supposed that the design matrix \mathbf{X} is of full rank.

Proposition 1. *In model (1) with a priori information $\beta^\top \mathbf{H} \beta \leq t$, where $t > 0$ is a known constant and \mathbf{H} is a known non-random positive semi-definite matrix, maximum likelihood estimates $\hat{\beta}$ and $\hat{\sigma}^2$ of β and σ^2 , respectively, are presented:*

(i) if $\mathbf{y}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \leq t$ then

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad n\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta});$$

(ii) if $\mathbf{y}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} > t$ then

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X} + \hat{\lambda} \mathbf{H})^{-1} \mathbf{X}^\top \mathbf{y}, \quad n\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}), \quad (2)$$

where $\hat{\lambda}$ is the solution to the non-linear equation in λ ,

$$\mathbf{y}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{H})^{-1} \mathbf{H}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{H})^{-1} \mathbf{X}^\top \mathbf{y} = t.$$

Proof. First, we find an upper bound of a likelihood with respect to σ^2 and thereafter convex optimization with respect to β takes place. Thus, we put $\theta = \sigma^2$ and then the negative log-likelihood function equals

$$f(\beta, \theta) = \frac{n}{2} \log(2\pi) + \frac{n}{2} \log(\theta) + \frac{1}{2} \theta^{-1} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$

Note, that the above function is convex with respect to θ^{-1} with minimum for $\theta^{-1} = n / ((\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta))$. Therefore

$$\begin{aligned} f(\beta, \theta) &\geq f(\beta, (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) / n) \\ &= \frac{n}{2} \log(2\pi) + \frac{n}{2} \log((\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) / n) + \frac{n}{2} \end{aligned} \quad (3)$$

and

$$\min_{\beta^\top \mathbf{H}\beta \leq t, \theta} f(\beta, \theta) = \min_{\beta^\top \mathbf{H}\beta \leq t} f(\beta, (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)/n).$$

Finding a minimizer of (3) under the a priori information $\beta^\top \mathbf{H}\beta \leq t$ is equivalent to finding a minimizer of $(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$ under the condition $\beta^\top \mathbf{H}\beta \leq t$.

The mathematical problem can be introduced by estimating β , under a priori information on β , using a linear model and applying the least squares approach, i.e.,

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta), \quad \beta^\top \mathbf{H}\beta \leq t,$$

where $t > 0$ is given. To study this problem a well known approach is to set up the Lagrangian function:

$$L(\beta, \lambda) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda(\beta^\top \mathbf{H}\beta - t), \quad (4)$$

where $\lambda > 0$ is the Lagrangian multiplier.

Note that $(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$ and $\beta^\top \mathbf{H}\beta$ are convex functions in β which follows by differentiating the functions twice. Therefore the Karush-Kuhn-Tucker (KKT) conditions give necessary and sufficient conditions for solving (4) (e.g., see [1]). The KKT-conditions equal

$$\frac{dL(\beta, \lambda)}{d\beta} = \mathbf{0}, \quad \beta^\top \mathbf{H}\beta - t \leq 0, \quad \lambda \geq 0, \quad \lambda(\beta^\top \mathbf{H}\beta - t) = 0. \quad (5)$$

The derivative in (5) is a vector derivative.

Spelling out the derivative implies that the following system of equations has to be solved:

$$-2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta) + 2\lambda \mathbf{H}\beta = \mathbf{0}, \quad (6)$$

$$\beta^\top \mathbf{H}\beta - t \leq 0, \quad \lambda \geq 0, \quad (7)$$

$$\lambda(\beta^\top \mathbf{H}\beta - t) = 0. \quad (8)$$

Condition (8) is usually called the complementary slackness condition and is very important in our approach. The condition implies that either $\lambda = 0$ or $\beta^\top \mathbf{H}\beta - t = 0$. If $\lambda = 0$ then from (6) it appears that

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

which is the usual least squares estimator. However if $\lambda > 0$ then $\beta^\top \mathbf{H}\beta - t = 0$ and then the task is to choose β and λ so that (6)–(8) are true. Usually this problem has to be solved numerically. Hence the two results for $\hat{\beta}$ of the proposition have been established. Coming to $\theta = (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})/n$ and inequality (3) it follows that $\hat{\sigma}^2$ is obtained.

In our approach, which is mathematically optimal, if the least squares estimates are too far from what can be assumed through the a priori information, the estimates are shrunk via the Lagrangian multiplier. Thus, on top of the usual estimation approach, a “safety belt” has been put so that in particular a “bad” \mathbf{X} matrix can be handled, i.e., when the eigenvalues of $(\mathbf{X}^\top \mathbf{X})^{-1}$ become large, or if there exist influential observations. In these cases $\hat{\beta}$ and $\hat{\sigma}^2$ in (2) can be used.

The next proposition, which is presented, with sketch of a proof, includes a different type of restriction. It is the kind of restriction which will be possible to handle in a straightforward manner when finding maximum likelihood estimates in the general multivariate linear model (MANOVA model). In the univariate case this restriction is of the form $(\sigma^2)^{-1} \beta^\top \mathbf{H} \beta \leq t$, where as before t is a positive known constant and \mathbf{H} is a known positive semi-definite matrix. This kind of restriction has been discussed in, e.g., [4, Corollary 3.2.2], when comparing ridge regression estimates with least squares estimates. However, our motivation for considering $(\sigma^2)^{-1} \beta^\top \mathbf{H} \beta \leq t$ is different. In this article one of the main ideas is to rely on convex optimization when finding estimators. The approach in this paper is inspired by the fact that the normal distribution (univariate and multivariate) belongs to an exponential family.

Proposition 2. *In model (1) with a priori information $(\sigma^2)^{-1} \beta^\top \mathbf{H} \beta \leq t$, where $t > 0$ is a known constant and \mathbf{H} is a known non-random positive semi-definite matrix, maximum likelihood estimates $\hat{\beta}$ and $\hat{\sigma}^2$ of β and σ^2 , respectively, are presented. Put*

$$\tilde{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad n\tilde{\sigma}^2 = (\mathbf{y} - \mathbf{X}\tilde{\beta})^\top (\mathbf{y} - \mathbf{X}\tilde{\beta}).$$

- (i) *If $\tilde{\beta}^\top \mathbf{H} \tilde{\beta} \leq \tilde{\sigma}^2 t$, then maximum likelihood estimates are given by $\hat{\beta} = \tilde{\beta}$ and $\hat{\sigma}^2 = \tilde{\sigma}^2$.*
- (ii) *If $\tilde{\beta}^\top \mathbf{H} \tilde{\beta} > \tilde{\sigma}^2 t$ then maximum likelihood estimates of β and σ^2 satisfy*

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X} + \hat{\lambda} \mathbf{H})^{-1} \mathbf{X}^\top \mathbf{y}, \quad n\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) + \hat{\lambda} \hat{\beta}^\top \mathbf{H} \hat{\beta},$$

where $\hat{\lambda}$ is the solution to $\mathbf{y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{H})^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{H})^{-1} \mathbf{X}^\top \mathbf{y} = t \hat{\sigma}^2$.

Proof. (sketch of the proof) Let us introduce the canonical parameters $\theta_1 = (\sigma^2)^{-1} \beta$ and $\theta_2 = (\sigma^2)^{-1}$. Then, the negative log-likelihood function can be expressed as

$$f(\theta_1, \theta_2) = \frac{n}{2} \log(2\pi) - \frac{n}{2} \log \theta_2 + \frac{1}{2} \theta_2 \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X} \theta_1 + \frac{1}{2} \theta_2^{-1} \theta_1^\top \mathbf{X}^\top \mathbf{X} \theta_1,$$

which is a convex function of θ_1 and θ_2 . Moreover, since the condition $g(\beta, \sigma^2) = (\sigma^2)^{-1} \beta^\top \mathbf{H} \beta - t$ can be also expressed in terms of canonical variables, i.e., $g(\theta_1, \theta_2) = \theta_2^{-1} \theta_1^\top \mathbf{H} \theta_1 - t$, it can be checked that this is also a convex function of θ_1 and θ_2 . Therefore applying the KKT conditions for solving

$$L(\theta_1, \theta_2) = f(\theta_1, \theta_2) + \lambda g(\theta_1, \theta_2),$$

where $\lambda > 0$ is the Lagrangian multiplier, we obtain

$$\widehat{\theta_2^{-1}\theta_1} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{H})^{-1} \mathbf{X}^\top \mathbf{y}$$

and

$$n\widehat{\theta_2^{-1}} = \mathbf{y}^\top \mathbf{y} - \widehat{\theta_2^{-1}\theta_1}^\top \mathbf{X}^\top \mathbf{X} \widehat{\theta_2^{-1}\theta_1} - \lambda \widehat{\theta_2^{-1}\theta_1}^\top \mathbf{H} \widehat{\theta_2^{-1}\theta_1}.$$

Coming back to original notation we obtain the results.

Acknowledgements This work was supported by the Poznań University of Technology under Grant no. 0213/SBAD/0118 (K. Filipiak).

References

1. Boyd, S., Vandenberghe, L., *Convex Optimization*, Cambridge University Press, 2004.
2. Fan, J., Li, R., *Variable selection via nonconcave penalized likelihood and its oracle properties*, *Journal of the American Statistical Association*, 96, 1348–1360, 2001.
3. Frank, I.E. and Friedman, J.H., *A statistical view of some chemometrics regression tools*, *Technometrics*, 35, 109–135, 1993.
4. Gruber, M., *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*, CRC Press, New York, 1998.
5. Hoerl, A.E., Kennard, R.W., *Ridge regression: biased estimation for nonorthogonal problems*, *Technometrics*, 12, 55–67, 1970.
6. Tibshirani, R., *Regression shrinkage and selection via the Lasso*, *Journal of the Royal Statistical Society, Series B*, 58, 267–288, 1996.
7. Zou, H., *The adaptive Lasso and its oracle properties*, *Journal of the American Statistical Association*, 101, 1418–1429, 2006.
8. Zou, H., Hastie, T., *Regularization and variable selection via the elastic net*, *Journal of the Royal Statistical Society, Series B*, 67, 301–320, 2005.