Estimation under inequality constraints in univariate and multivariate linear models

Katarzyna Filipiak, Dietrich von Rosen, Martin Singull, Wojciech Rejchel

LiTH-MAT-R-2024/01
Estimation under inequality constraints
in univariate and multivariate linear models

Katarzyna Filipiak*, Dietrich von Rosen†, Martin Singull‡, Wojciech Rejchel§

Abstract

In this paper least squares and maximum likelihood estimates under univariate and
multivariate linear models with a priori information related to maximum effects in
the models are determined. Both loss functions (the least squares and negative log-
likelihood) and the constraints are convex, so the convex optimization theory can be
utilized to obtain estimates, which in this paper are called Safety belt estimates. In par-
ticular, the complementary slackness condition, common in convex optimization, implies
two alternative types of solutions, strongly dependent on the data and the restriction.

It is experimentally shown that, despite of the similarity to the ridge regression es-
timation under the univariate linear model, the Safety belt estimates behave usually
better than estimates obtained via ridge regression. Moreover, concerning the multivari-
ate model, the proposed technique represents a completely novel approach.

1 Introduction

Penalized estimation techniques are usually used to estimate regression parameters and simul-
taneously select appropriate covariables. There are several methods which have been proposed
by many authors; for example, Frank and Friedman (1993) proposed bridge regression (a term
Friedman, 1994 used in an interesting paper), Tibshirani (1996) introduced LASSO, Fan and
Li (2001) came up with SCAD, Zou and Hastie (2005) with elastic net and Zou (2006) pro-
posed adaptive LASSO. However, if the variable selection is not of main concern, the ridge

*Institute of Mathematics, Poznan University of Technology, Poznań, Poland
†Department of Energy and Technology, Swedish University of Agricultural Sciences and Department of
Mathematics, Linköping University, Sweden
‡Department of Mathematics, Linköping University, Sweden
§Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland
regression (Hoerl and Kennard, 1970) method is also commonly used (see also Tikhonov regularization) and can be viewed as a penalized estimation method. In fact, many of the above given references connect to ridge regression which appeared earlier than the other approaches where additionally variable selection was considered. All these 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 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 and multivariate linear models. We will stand on the 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 usually is not the case with the above-mentioned variable selection methods.

Section 2 is a continuation of the introduction where the reader will be acquainted with the ideas of the paper via a univariate linear model and least squares. In Section 3 we introduce our ”Safety belt” estimate for the univariate linear model and also include results for normally distributed data. Section 4 presents simulations which show that the Safety belt estimator often performs better than the ridge regression estimator and the classical maximum likelihood estimator. This shows that the Safety belt estimator makes sense. In Section 5 the main ideas and results concerning the multivariate linear model are presented and these are also the main results of the paper. Note that most penalized estimation approaches only consider univariate responses. In the section it is shown, for example, how the a priori information should be quantified. The paper is concluded with two appendices comprising some technical results.

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

2 Notation and basic ideas

Although the main results will be connected to multivariate linear models, this section is aimed to introduce the reader to the basic ideas of the paper via the general univariate linear
model. Therefore for a while it is focused on the least squares theory, but in subsequent sections the approach will be likelihood based. One advantage of using the likelihood function is that, for example, unlike the well known ridge regression approach (see Hoerl and Kennard, 1970, Gruber, 1998), both mean and dispersion parameters can be handled simultaneously.

Let \( \mathbf{y} : n \times 1 \), be a 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 design matrix in a linear model with corresponding unknown mean effects \( \mathbf{\beta} : k \times 1 \), and let \( \mathbf{\epsilon} : n \times 1 \) be the vector of unobservable uncorrelated errors following a symmetric distribution with mean \( \mathbf{0} \). The dispersion of \( \mathbf{\epsilon} \) will not be considered in the least squares approach. The aim will be to estimate \( \mathbf{\beta} \) under some a priori information which will be specified later.

The univariate linear model which will be considered includes the above vectors (matrix) and can be written

\[
\mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{\epsilon},
\]

where “\( = \)” stands for equality in distribution. To estimate \( \mathbf{\beta} \) the idea is to use the least squares theory, i.e.,

\[
\min_{\mathbf{\beta}} (\mathbf{y} - \mathbf{X}\mathbf{\beta})^\top (\mathbf{y} - \mathbf{X}\mathbf{\beta}),
\]

where \( ^\top \) denotes the transpose, which yields

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

where “\( - \)” denotes an arbitrary generalized inverse. If the rank of \( \mathbf{X} \), denoted \( \text{rank}(\mathbf{X}) \), equals \( k \), and \( n > k \), it follows that

\[
\hat{\mathbf{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.
\]

From now on \( \mathbf{X} \) will always be assumed to be of full rank. Once the starting point for finding alternative estimators to the least squares estimator was that columns in \( \mathbf{X} \) could almost be collinear (the multicollinearity problem), implying that \( \hat{\mathbf{\beta}} \) become unreliable (see Gruber, 1998). Moreover, the dispersion of \( \hat{\mathbf{\beta}} \) is also a function of \( (\mathbf{X}^\top\mathbf{X})^{-1} \) and it can be concluded that almost collinearity of columns in \( \mathbf{X} \) leads to a poor estimator. There exists a huge literature on handling multicollinearity, including the above-mentioned ridge regression.
method. Other regression approaches which are dealing with the multicollinearity problem
are, for example, principal component regression, (see e.g., Jolliffe, 1982) and partial least
squares (see e.g., Helland, 1990). For a comparison of ridge regression with these methods see
Frank and Friedman (1993).

In this research we do not focus on multicollinearity in $X$. Instead a crucial assumption is
that there exist a priori information about $\beta$ so that it makes sense to assume that $\beta^{\top} \beta \leq t$
for some given $t > 0$, or a bit more general $\beta^{\top} H \beta \leq t$, for some positive semi-definite matrix $H$. The matrix $H$ can be used to select components from $\beta$, for example via $H = (0 \quad 0 \quad I)$, where $I$ stands for the identity matrix. Obviously $H = I$ is another choice. However it can also be written that for $H = (h_{ij})$ and $\beta = (\beta_i)$, $\beta^{\top} H \beta = \sum_{ij} h_{ij} \beta_i \beta_j$ indicating the effect of choice of elements in $H$. The use of an $H$-matrix in ridge regression analysis has been considered in the literature (e.g., see Gruber, 1998, p. 113) and for example by choosing $H = X^{\top} X$ one obtains the James-Stein estimator of $\beta$.

The a priori information is quantified via inequality constrains on the parameter space
(squares of the parameters) which means that we do not deal with linear subspaces and
cannot refer to the usual linear models theory. Moreover, in many applications $\beta^{\top} H \beta \leq t$
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,
used measurements 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.

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

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

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

$$L(\beta, \lambda) = (y - X \beta)^{\top} (y - X \beta) + \lambda (\beta^{\top} H \beta - t),$$ (2)

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

Note that $(y - X \beta)^{\top} (y - X \beta)$ and $\beta^{\top} H \beta$ are convex functions in $\beta$ which follows by
differentiating the functions twice, and therefore the Karush-Kuhn-Tucker (KKT) conditions
give necessary and sufficient conditions for solving (2) (e.g., see Boyd and Vandenberghe, 2004). The KKT-conditions equal

\[
\frac{d L(\beta, \lambda)}{d \beta} = 0, \quad \beta^T H \beta - t \leq 0, \quad \lambda \geq 0, \quad \lambda(\beta^T H \beta - t) = 0. \tag{3}
\]

The derivative in (3) is a vector derivative, which also might be considered as a matrix derivative. Because such objects appear in many places in the paper, in the Appendix A we devote them a section to present the most basic results.

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

\[
\begin{align*}
-2X^T(y - X\beta) + 2\lambda H \beta &= 0, \tag{4} \\
\beta^T H \beta - t &\leq 0, \\
\lambda &\geq 0, \\
\lambda(\beta^T H \beta - t) &= 0. \tag{5}
\end{align*}
\]

Condition (5) is usually called the complementary slackness condition and is very important in our approach. The condition implies that either \( \lambda = 0 \) or \( \beta^T H \beta - t = 0 \). If \( \lambda = 0 \) then from (4) it appears that

\[
\tilde{\beta} = (X^T X)^{-1} X^T y, 
\]

which is the usual least squares estimator. However if \( \lambda > 0 \) then \( \beta^T H \beta - t = 0 \) and then the task is to choose \( \beta \) and \( \lambda \) so that the equation is true. Usually this problem has to be solved numerically.

**Proposition 2.1.** Let \( y \) be as in (1) and \( \beta^T H \beta \leq t \) for a given non-random positive semi-definite \( H \) and a known positive \( t \). Moreover, suppose the design matrix \( X \) to be of full rank. Put

\[
\tilde{\beta} = (X^T X)^{-1} X^T y. \tag{6}
\]

If \( \beta^T H \tilde{\beta} > t \) then the least squares estimate \( \tilde{\beta} \) satisfies

\[
\tilde{\beta} = (X^T X + \lambda H)^{-1} X^T y, \tag{7}
\]
where $\widehat{\lambda}$ is the solution to

$$\widehat{\beta}^\top H \widehat{\beta} = t. \hspace{1cm} (8)$$

If $\widehat{\beta}^\top H \widehat{\beta} \leq t$ then $\widehat{\beta} = \tilde{\beta}$.

Thus depending on if the $\tilde{\beta}^\top H \tilde{\beta}$ is larger or smaller than $t$ there are two alternative estimates. This means that the approach is data driven. Classical ridge regression estimation would be based on (7) but $\lambda$ is not estimated through (8) because there is no prior information $\beta^\top H \beta \leq t$ involved. Instead some statistical aspects guide how to estimate from the data a good value of $\lambda$. The literature on this topic is extensive. We just cite a few papers from the 70s of the previous century: Goldstein and Smith (1974), Hemmerle (1975), Hoerl et al. (1975), McDonald and Galarneau (1975), Hoerl and Kennard (1976), Lawless and Wang (1976), Dempster et al. (1977). Nowadays, using high performance CPUs, we can go another way: first we choose a sequence of $\lambda$’s, say $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Next, for each $\lambda_j, j \in \{1, \ldots, m\}$, we compute a minimizer

$$\hat{\beta}^{(j)} = \arg \min_{\beta \in \mathbb{R}^k} (y - X\beta)^\top (y - X\beta) + \lambda_j \beta^\top H \beta.$$ 

Then we choose the final estimate from a set $\{\hat{\beta}^{(j)} : j \in \{1, \ldots, m\}\}$ according to some criterion. One of the most popular method is the M-fold cross-validation (CV), which tries to minimize the out of sample prediction error. This approach will be used in the experimental section to compare properties of our proposal to ridge regression estimates. One can easily find drawbacks of the above mentioned CV procedure. First, a choice of $\lambda_j$’s is tricky, because it should depend on a current data set, in particular $n, k$, and the correlations between predictors, and the dispersion of the noise $\epsilon$. Notice that the latter two are unknown. The second shortcoming relates to the fact that the CV method needs a relatively large sample size to work well. Finally, CV can be time-consuming, because it has to calculate $m$ minimizers (over $\mathbb{R}^k$) for each of $M$ folds. We take $m = 200$ and $M = 5$ in our experiments, which is a common choice. Therefore, we have to compute 1000 minimizers over $\mathbb{R}^k$. On the other hand, to compute the estimate from Proposition 2.1 one needs at most to calculate (6) and (7), which correspond to two minimizers over $\mathbb{R}^k$. Besides this we have only to solve a one dimensional equation (8).

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 function (penalized least squares). Thus, on top of the least squares approach,
a "Safety belt" has been put so that in particular a "bad" $X$ matrix can be handled, i.e.,
when eigenvalues of $(X^TX)^{-1}$ become large, or if there exist influential information then also
$\hat{\beta}$ in (7) can be used.

In the subsequent the data driven estimates which were presented in Proposition 2.1 will
be referred to as "Safety belt" estimates. However, the presentation of the main results will
be based on the likelihood instead of the least squares loss function. Moreover, in particular
it will be shown how the classical MANOVA model can be treated.

3 Safety belt estimates and the general univariate linear
model

Consider the univariate linear model in (1) but now also assume that the response follows a
normal distribution:

$$y \sim \mathcal{N}_n(X\beta, \sigma^2I_n),$$

(9)

where $\sigma^2 > 0$ is the unknown variance parameter and $I_n$ stands for the $n \times n$ identity matrix.
The observations are supposed to be independent normally distributed. Note that $X$ is
supposed to be of full rank.

In the next proposition maximum likelihood estimates are presented under inequality re-
strictions on $\beta^T H \beta$, where $H$ is supposed to be non-random positive semi-definite and known.

**Proposition 3.1.** In model (9) with a priori information $\beta^T H \beta \leq t$, where $t > 0$ is a known
constant and $H$ is a known non-random positive semi-definite matrix, maximum likelihood
estimates $\hat{\beta}$ and $\hat{\sigma}^2$ of $\beta$ and $\sigma^2$, respectively, are presented:

(i) if $y^T X (X^TX)^{-1} H (X^TX)^{-1} X^Ty \leq t$ then

$$\hat{\beta} = (X^TX)^{-1}X^Ty,$$
$$n\hat{\sigma}^2 = (y - X\hat{\beta})^T(y - X\hat{\beta});$$

(ii) if $y^T X (X^TX)^{-1} H (X^TX)^{-1} X^Ty > t$ then

$$\hat{\beta} = (X^TX + \hat{\lambda}H)^{-1}X^Ty,$$
$$n\hat{\sigma}^2 = (y - X\hat{\beta})^T(y - X\hat{\beta});$$

7
where \( \hat{\lambda} \) is the solution to the non-linear equation in \( \lambda \),
\[
y^\top X (X^\top X + \lambda H)^{-1} H (X^\top X + \lambda H)^{-1} X^\top y = t.
\]

**Proof.** First, we find an upper bound of a likelihood with respect to \( \sigma^2 \) and thereafter convex optimization with respect to \( \beta \) takes place. Thus, we put \( \theta = \sigma^2 \) and then the negative log-likelihood equals
\[
f(\beta, \theta) = \frac{n}{2} \log(2\pi) + \frac{n}{2} \log(\theta) + \frac{1}{2} \theta^{-1} (y - X\beta)^\top (y - X\beta).
\]

For \( a > 0 \) and \( b \geq 0 \) the function \( h(\theta) = a \log(\theta) + b/\theta \) has the unique minimum at \( \theta = b/a \). Therefore, for each \( \beta \) and \( \theta \) we have
\[
f(\beta, \theta) \geq f(\beta, (y - X\beta)^\top (y - X\beta)/n)
\]
\[
= \frac{n}{2} \log(2\pi) + \frac{n}{2} \log((y - X\beta)^\top (y - X\beta)/n) + \frac{b}{2a}
\]  \hspace{1cm} (10)

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

Hence, finding a minimizer of (10) under the a priori information \( \beta^\top H\beta \leq t \) is equivalent to finding a minimizer of \( (y - X\beta)^\top (y - X\beta), \beta^\top H\beta \leq t \), which has been already considered in Proposition 2.1. \( \square \)

The next proposition 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 multivariate linear model in Section 5. In the univariate case this restriction is of the form \( (\sigma^2)^{-1} \beta^\top H\beta \leq t \), where as before \( t \) is a positive known constant and \( H \) is a known positive semi-definite matrix. This type of restriction has been discussed in, e.g., Gruber (1998, Corollary 3.2.2), when comparing ridge regression estimates with least squares estimates. However, our motivation for considering \( (\sigma^2)^{-1} \beta^\top H\beta \leq t \) is different. In this article one of the main ideas is to rely on convex optimization when finding estimators. The approach is inspired by the fact that the normal distribution (univariate and multivariate) belongs to the exponential family (log-linear model).

Put \( \theta_1 = (\sigma^2)^{-1} \beta \) and \( \theta_2 = (\sigma^2)^{-1} \) which are the canonical parameters. The negative logarithm of the density corresponding to the model in (9), using the canonical parameterization,
is given by
\[ \frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \theta_2 + \frac{1}{2} \theta_2 y^\top y + \frac{1}{2} \theta_2^{-1} \theta_1^\top X^\top X \theta_1 - \theta_1^\top X^\top y. \]

Expressions \(- \ln \theta_2, \theta_2 y^\top y\) and \(-\theta_1^\top X^\top y\) are obviously convex functions (differentiate the expressions twice) and now some attention is paid to the convexity of \(\theta_2^{-1} \theta_1^\top X^\top X \theta_1\). The Hessian can be constructed and it clearly appears that the Hessian is positive semi-definite, i.e.,
\[
\begin{pmatrix}
\frac{d^2}{d\theta_1^2} L(	heta_1, \theta_2, \lambda) & \frac{d}{d\theta_1} L(	heta_1, \theta_2, \lambda) \\
\frac{d}{d\theta_1} L(	heta_1, \theta_2, \lambda) & \frac{d^2}{d\theta_2^2} L(	heta_1, \theta_2, \lambda)
\end{pmatrix}
= 2 \begin{pmatrix}
\theta_2^{-1/2} X^\top & \theta_1^\top X^\top
\end{pmatrix}
\begin{pmatrix}
\theta_2^{-1/2} X^\top \\
-\theta_2^{-3/2} \theta_1^\top X^\top
\end{pmatrix}
\top.
\]

Thus it has been shown that the normal density is a convex function in the canonical parameters. It remains to consider the a priori information
\((\sigma^2)^{-1} \beta^\top H \beta \leq t\). Since \((\sigma^2)^{-1} \beta^\top H \beta = \theta_2^{-1} \theta_1^\top H \theta_1\), using the same differentiation methods as for obtaining the above Hessian matrix it follows that the a priori information is convex, and determination of the estimates comes down to solving a convex optimization problem.

As before the KKT conditions are necessary and sufficient conditions for finding the smallest value of the negative log-likelihood function (for notational convenience \(\lambda/2\) is used)
\[
L(\theta_1, \theta_2, \lambda) = \frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln(\theta_2) + \frac{1}{2} \theta_2 (y - \theta_2^{-1} X \theta_1)^\top (y - \theta_2^{-1} X \theta_1)
+ \frac{1}{2} \lambda \theta_2^{-1} \theta_1^\top H \theta_1 - t,
\]
and, according to (3), we have the following form:
\[
\frac{d}{d\theta_1} L(\theta_1, \theta_2, \lambda) = X^\top (\theta_2^{-1} X \theta_1 - y) + \lambda \theta_2^{-1} H \theta_1,
\]
\[
\frac{d}{d\theta_2} L(\theta_1, \theta_2, \lambda) = -\frac{n}{2} \theta_2^{-1} + \frac{1}{2} y^\top y - \frac{1}{2} \theta_2^{-2} \theta_1^\top X^\top X \theta_1 - \frac{1}{2} \lambda \theta_2^{-2} \theta_1^\top H \theta_1,
\]
\[
\theta_2^{-1} \theta_1^\top H \theta_1 - t \leq 0,
\]
\[
\lambda \geq 0,
\]
\[
\lambda (\theta_2^{-1} \theta_1^\top H \theta_1 - t) = 0.
\]

The condition \(\frac{d}{d\theta_1} L(\theta_1, \theta_2, \lambda) = 0\) gives
\[
\beta = \theta_2^{-1} \theta_1 = (X^\top X + \lambda H)^{-1} X^\top y,
\]
and
\[
\beta = \theta_2^{-1} \theta_1 = (X^\top X + \lambda H)^{-1} X^\top y.
\]
and the condition \( \frac{d}{d\theta_2} L(\theta_1, \theta_2, \lambda) = 0 \) implies

\[
n\theta_2^{-1} = y^\top y - \beta^\top X^\top X \beta - \lambda \beta^\top H \beta.
\]

Thus, the next result has been established.

**Proposition 3.2.** In model (9) with a priori information \((\sigma^2)^{-1} \beta^\top H \beta \leq t\), where \(t > 0\) is a known constant and \(H\) is a known non-random positive semi-definite matrix, maximum likelihood estimates \(\hat{\beta}\) and \(\hat{\sigma}^2\) of \(\beta\) and \(\sigma^2\), respectively, are presented. Put

\[
\begin{align*}
\hat{\beta} &= (X^\top X)^{-1} X^\top y, \\
n\hat{\sigma}^2 &= (y - X\hat{\beta})(y - X\hat{\beta}).
\end{align*}
\]

(i) \(\beta^\top H\beta \leq \sigma^2 t\), then maximum likelihood estimates are given by \(\hat{\beta} = \tilde{\beta}\) and \(\hat{\sigma}^2 = \tilde{\sigma}^2\).

(ii) \(\beta^\top H\beta > \sigma^2 t\) then maximum likelihood estimates of \(\beta\) and \(\sigma^2\) satisfy

\[
\begin{align*}
\hat{\beta} &= (X^\top X + \hat{\lambda} H)^{-1} X^\top y, \\
n\hat{\sigma}^2 &= y^\top y - \hat{\beta}^\top X^\top X \hat{\beta} - \hat{\lambda} \hat{\beta}^\top H \hat{\beta},
\end{align*}
\]

where \(\hat{\lambda}\) is the solution of

\[
t\hat{\sigma}^2 = y^\top X (X^\top X + \lambda H)^{-1} H (X^\top X + \lambda H)^{-1} X^\top y.
\]

**Remark 3.1.** Concerning computations we know that due to convexity there should be one unique solution to the equation in Proposition 3.2. Moreover, for a given \(\hat{\lambda}\), estimates \(\hat{\beta}\) and \(\hat{\sigma}^2\) are immediately obtained and thus calculations seem to be doable.

Some comparisons between Proposition 3.1 and Proposition 3.2 are appropriate to carry out. If the choice of \(t\) leads to the existence of the usual maximum likelihood estimates, we can observe the following properties. When replacing the parameters by the proposed estimates given in the propositions it follows that for Proposition 3.1 (i)

\[
\hat{\beta}^\top H\hat{\beta} = y^\top X (X^\top X)^{-1} H (X^\top X)^{-1} X^\top y,
\]

whereas for Proposition 3.2 (i)

\[
(\hat{\sigma}^2)^{-1} \hat{\beta}^\top H\hat{\beta} = \frac{y^\top X (X^\top X)^{-1} H (X^\top X)^{-1} X^\top y}{\frac{1}{n} y^\top (I - X (X^\top X)^{-1} X^\top) y}.
\]
The main difference between (11) and (12) is that the distribution of the statistic in (11) depends on $\sigma^2$ whereas this is not the case in (12). Moreover, when considering respective expressions from Proposition 3.1 (ii) and Proposition 3.2 (ii) with a fixed value of $\lambda$, that is

\[ y^\top X (X^\top X + \lambda H)^{-1} H (X^\top X + \lambda H)^{-1} X^\top y \]  

(13)

and

\[ y^\top X (X^\top X + \lambda H)^{-1} H (X^\top X + \lambda H)^{-1} X^\top y \]

\[ \frac{1}{n} y^\top (I - X (X^\top X + \lambda H)^{-1} X^\top) y \]  

(14)

it can be observed that again the distribution of the statistic in (13) depends on $\sigma^2$ which is not the case in (14).

To complete this section the results with a known variance in model (9) are briefly considered. The reason is that in the multivariate case it can be natural to present results according to if the dispersion is known or unknown. The proofs are very close to those which have earlier been presented and are therefore omitted. There are two cases to be considered, namely $\beta^\top H \beta \leq t$ and $(\sigma^2)^{-1} \beta^\top H \beta \leq t$. The next two propositions emerge.

Proposition 3.3. In model (9) with a priori information $\beta^\top H \beta \leq t$, where $t > 0$ is a known constant, $\sigma^2$ is known and $H$ is a known non-random positive semi-definite matrix, the maximum likelihood estimate $\hat{\beta}$ of $\beta$ is presented:

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

\[ \hat{\beta} = (X^\top X)^{-1} X^\top y; \]

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

\[ \hat{\beta} = (X^\top X + \hat{\lambda} H)^{-1} X^\top y, \]

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

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

Proposition 3.4. In model (9) with a priori information $(\sigma^2)^{-1} \beta^\top H \beta \leq t$, where $t > 0$ is a known constant, $\sigma^2$ is the known variance parameter, and $H$ is a known non-random positive
semi-definite matrix, the maximum likelihood estimate $\hat{\beta}$ of $\beta$ is presented. Put

$$\tilde{\beta} = (X^\top X)^{-1}X^\top y.$$  

(i) If $\tilde{\beta}^\top H\tilde{\beta} \leq \sigma^2 t$, then the maximum likelihood estimate is given by $\hat{\beta} = \tilde{\beta}$.  

(ii) If $\tilde{\beta}^\top H\tilde{\beta} > \sigma^2 t$ then the maximum likelihood estimate of $\beta$ satisfies

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

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

$$y^\top X (X^\top X + \lambda H)^{-1} H (X^\top X + \lambda H)^{-1} X^\top y = t\sigma^2.$$

4 Experimental results

In this section we investigate properties of the Safety belt estimator via several simulated data sets and put the estimate in relation to the maximum likelihood and ridge regression estimates. Eight versions of the univariate linear model (9) will be used:

- Model 1: $n = 15$, $k = 6$, $\sigma^2 = 3$, $\beta = (-1, 1\top_6)^\top$;
- Model 2: Model 1 with $\sigma^2 = 9$;
- Model 3: Model 1 with $n = 60$;
- Model 4: Model 3 with $\sigma^2 = 9$;
- Model 5: $n = 40$, $k = 11$, $\sigma^2 = 3$, $\beta = (-1, 1\top_{10})^\top$;
- Model 6: Model 5 with $\sigma^2 = 36$;
- Model 7: $n = 60$, $k = 41$, $\sigma^2 = 9$, $\beta = (-1, 2^{-1.5} \cdot 1\top_{40})^\top$;
- Model 8: Model 7 with $n = 160$.

In all these models the first column of the $n \times k$-matrix $X$ in (9) is a vector of ones, while the remaining elements are independently generated from the standard normal distribution. The error vector is generated from the normal distribution $N(0, \sigma^2)$.

Recall that the maximum likelihood estimator of $\beta$ under a normality assumption and without a priori information is unbiased and has the variance $\sigma^2 (X^\top X)^{-1}$, which is the
smallest among all unbiased estimators. However, the variance can still be inconveniently large, implying that models are difficult to handle; for example when $\sigma^2$ is large and/or the matrix $X^\top X$ is close to being singular (for instance, when $k$ is relatively close to $n$ or columns of $X$ are almost collinear). Among our considered models the most “difficult” ones to handle are Model 6 and Model 7.

In the experiments three estimation methods were studied:

(i) classical maximum likelihood estimation of $\beta$ and $\sigma^2$;

(ii) the new Safety belt estimation of $\beta$ and $\sigma^2$, under the constraint $\beta^\top H\beta \leq t$, where $H = \text{diag}(0, 1_{k-1})$ and $t > 0$, so the a priori information about $t$ does not bound the intercept. In the experiments we assume that this $t$ is 10% greater than the true value of $\beta^\top H\beta$. So, in Models 1-4 and 7-8 we have $t = 5.5$, while $t = 11$ in Models 5-6;

(iii) a least squares estimation of $\beta$ with the squared $\ell_2$-norm regularizer, i.e., the ridge regression estimator

$$\arg\min_{\beta \in \mathbb{R}^k} \left[ (y - X\beta)^\top (y - X\beta) + \lambda \beta^\top H\beta \right], \quad (15)$$

where $H$ is the diagonal matrix used in (ii). Note that the ridge regression estimate does not include any a priori information about the parameters (as $t$ in the Safety belt approach), but requires a value of the tuning parameter $\lambda$ in (15). To find this value, we applied 5-fold cross-validation and we choose the sequence of candidates for $\lambda$ as $2^{s_m}\sqrt{\log(k)/n}$, $m \in \{1, \ldots, 200\}$, where $\{s_m\}$ is the sequence of 200 equally spaced numbers over the interval $[-10, 15]$ (this choice worked satisfactory for all our models). The calculations were performed via the glmnet package (Friedman et al., 2010) in the R software (R Core Team, 2021).

In the simulations, for each model (Models 1-8), the maximum likelihood (ML), Safety belt (SB) and ridge regression (RR) estimates were calculated 200 times. Hence, there are 200 estimates of $\beta$ and $\sigma^2$ for the maximum likelihood and Safety belt estimation methods (denoted respectively, $\hat{\beta}_i^{(ML)}$, $\sigma_i^{2(ML)}$ and $\hat{\beta}_i^{(SB)}$, $\sigma_i^{2(SB)}$, $i \in \{1, \ldots, 200\}$), and 200 estimates of $\beta$ for the ridge regression method (denoted $\hat{\beta}_i^{(RR)}$, $i \in \{1, \ldots, 200\}$). The estimation of $\sigma^2$ on the basis of the RR estimate seems to be tricky. There exist some heuristic methods, but, to the best of our knowledge, none of them is thoroughly investigated and generally accepted. Therefore, we do not consider them.

Next, for each estimation method we computed:
• the squared Euclidean distance between \( \hat{\beta} \) and the true \( \beta \), i.e.,

\[
d^j_i = ||\hat{\beta}^j_i - \beta||^2, \quad i \in \{1, \ldots, 200\}, \quad j \in \{(ML), (SB), (RR)\};
\]

(16)

• the squared Euclidean norm, i.e.,

\[
g^j_i = \hat{\beta}^j_i \top H \hat{\beta}^j_i, \quad i \in \{1, \ldots, 200\}, \quad j \in \{(ML), (SB), (RR)\}
\]

(17)

(the expression does not include the estimated intercept);

• a distance between \( \hat{\sigma}^2 \) and true \( \sigma^2 \), i.e.,

\[
h^j_i = |\hat{\sigma}^2(j) - \sigma^2|, \quad i \in \{1, \ldots, 200\}, \quad j \in \{(ML), (SB)\}.
\]

(18)

Each distance/norm is averaged over 200 repetitions giving respectively \( \overline{d}^{(j)}, \overline{g}^{(j)}, j \in \{(ML), (SB), (RR)\} \) and \( \overline{h}^{(j)}, j \in \{(ML), (SB)\} \), and the results are presented in Table 1. Moreover, for each average value its standard deviation (\( \delta_{d}^{j}, \delta_{g}^{j}, \delta_{h}^{j}, j \in \{(ML), (SB), (RR)\} \)) is calculated. Note, that \( d^{(SB)} \) can be treated as an empirical mean square error of the estimator (MSE).

All results of the experiments are given in Table 1 and Figures 1 and 2. We can observe that for Model 1 the SB estimate has the smallest MSE (\( \overline{d} \)). The RR estimate finds its parameter \( \lambda \) via cross-validation, so it requires a relatively large sample size in order to work well, while in Model 1 the sample size is small (\( n = 16 \)). Moreover, for the SB estimate the norm (\( \overline{g} \)) is the closest to the true value (\( \beta \top H \beta = 5 \)) and the standard deviations of \( d^{(SB)} \) and \( g^{(SB)} \) are the smallest.

In Model 2 we increase \( \sigma^2 \), which implies that the MSE of all the estimates gets larger than in Model 1. The most significant change can be noted for the ML estimates. The SB estimates once again outperform its competitors.

For Model 3 it can be seen that all estimation approaches work fairly well, because the sample size has been increased.

In Model 4, which is simply Model 3 but with a higher variance, the performance of the estimates gets worse. When MSE of the estimates is taken into account, the SB estimate is slightly better than the other two. Moreover, comparing the norms (\( \overline{g} \)), SB is the closest to the true value (\( t = 5 \)) and is the most stable.

In Model 5 some more covariates than in Models 1-4 have been included. In this case the ML and SB estimates perform comparably, while the RR estimate is worse, because its norm
Table 1. The averages $\bar{d}$, $\bar{g}$ and $\bar{h}$ based on 200 simulations where $d_i^j$, $g_i^j$, $h_i^j$, $j \in \{\text{ML}, \text{SB}, \text{RR}\}$ were calculated according to (16), (17) and (18), respectively, with related standard deviation in parentheses.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>$\bar{d}$ ($\delta_d$)</th>
<th>$\bar{g}$ ($\delta_g$)</th>
<th>$\bar{h}$ ($\delta_h$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>SB</td>
<td>1.47 (1.20)</td>
<td>5.07 (0.78)</td>
<td>1.23 (0.64)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>2.07 (1.74)</td>
<td>3.78 (2.50)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>6.35 (4.57)</td>
<td>9.91 (4.68)</td>
<td>3.92 (1.98)</td>
</tr>
<tr>
<td>Model 2</td>
<td>SB</td>
<td>3.35 (2.60)</td>
<td>5.20 (0.75)</td>
<td>3.53 (1.91)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>4.78 (3.32)</td>
<td>3.21 (4.16)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.35 (0.23)</td>
<td>5.25 (1.04)</td>
<td>0.45 (0.32)</td>
</tr>
<tr>
<td>Model 3</td>
<td>SB</td>
<td>0.31 (0.20)</td>
<td>4.95 (0.63)</td>
<td>0.45 (0.32)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>0.36 (0.22)</td>
<td>4.64 (1.14)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>1.05 (0.69)</td>
<td>5.81 (1.90)</td>
<td>1.38 (0.96)</td>
</tr>
<tr>
<td>Model 4</td>
<td>SB</td>
<td>0.84 (0.54)</td>
<td>4.91 (0.84)</td>
<td>1.34 (0.96)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>1.04 (0.64)</td>
<td>4.14 (2.00)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>1.13 (0.59)</td>
<td>11.07 (2.86)</td>
<td>0.89 (0.49)</td>
</tr>
<tr>
<td>Model 5</td>
<td>SB</td>
<td>0.97 (0.53)</td>
<td>9.90 (1.39)</td>
<td>0.86 (0.48)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>1.50 (1.01)</td>
<td>8.40 (3.38)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>13.60 (7.10)</td>
<td>22.62 (12.14)</td>
<td>10.67 (5.85)</td>
</tr>
<tr>
<td>Model 6</td>
<td>SB</td>
<td>8.72 (4.29)</td>
<td>10.67 (1.05)</td>
<td>9.63 (5.65)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>9.82 (3.04)</td>
<td>3.80 (7.82)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>22.80 (9.27)</td>
<td>27.64 (9.38)</td>
<td>6.17 (0.92)</td>
</tr>
<tr>
<td>Model 7</td>
<td>SB</td>
<td>3.84 (0.98)</td>
<td>5.5 (0.01)</td>
<td>4.48 (1.11)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>3.15 (1.05)</td>
<td>2.87 (2.02)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>3.28 (0.84)</td>
<td>8.25 (1.44)</td>
<td>2.29 (0.85)</td>
</tr>
<tr>
<td>Model 8</td>
<td>SB</td>
<td>2.20 (0.56)</td>
<td>5.49 (0.06)</td>
<td>2.06 (0.86)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>1.94 (0.43)</td>
<td>3.04 (1.22)</td>
<td></td>
</tr>
</tbody>
</table>

($\bar{g}$) is shrunk too much. In Model 6 increasing the variance strongly weakens the properties of ML. Again SB outperforms its competitors.

Model 7 is a “difficult” case, because $k$ is large comparing to $n$ and the variance in the error term is significant. As one may expect, in this scenario the ML approach works poorly. We can also observe that the ridge regression estimate behaves better than the SB estimate, so the additional knowledge about the upper bound of the norm of $\beta$ for SB ($t = 5.5$) does not improve the properties of this estimator. In such a difficult setup the main goal of the estimation process is to control the variance of the estimator, and thus keeping the bias small.
becomes relatively less important. Ridge regression is focused on fulfilling this requirement, and thus RR has the smallest MSE among the considered estimates, as given in Table 1 and Figure 1. Indeed, the upper quartile of the MSE of RR is below the median of the MSE of SB. To control the variance (and MSE) ridge regression significantly shrinks coordinates of the estimate: for the true value of $\beta$ we have $\beta^\top H \beta = 5$, while for RR of $\beta$ this value is, on average, less than three (see Figure 2). It should also be pointed out that the SB algorithm executes only the second step of the procedure from Proposition 3.1, i.e., it minimizes the square loss over the sphere $\beta^\top H \beta = 5.5$. Therefore, it cannot shrink coordinates of $\beta$ as much as the ridge regression does. Obviously, the price which is paid for using RR is higher computational cost. Each time one has to compute 1000 minimizers over $\mathbb{R}^k$ instead of only two minimizers for SB (plus solving a one-dimensional equation).

In Model 8, in which the sample size increased, the MSE of SB is pretty close to the MSE of RR. However, again ridge regression shrinks the coordinates of $\beta$ more significantly than the Safety belt method.

Finally, let us take a short look onto the last column of Table 1. The SB estimator of $\sigma^2$ is, in average, comparable or better than the one based on ML. The advantage of SB is the most visible in Models 6 and 7 as can be observed from ratios between $\delta_h$ and $\bar{h}$.

Summing up, we can conclude that the SB estimates behave well in most of the scenarios. The SB estimate of $\beta$ loses compared to the RR estimate only in the case with a "difficult" model (Model 7), where there are many covariates (in comparison to $n$) and $\sigma^2$ is relatively large; however, even in this model the SB estimation method allows to get the estimate of $\sigma^2$ as well. Finally, notice that SB always satisfies the constraints.

## 5 Safety belt estimates in MANOVA

In this section the following multivariate linear model (MANOVA)

$$Y \sim \mathcal{N}_{n,p}(XB, I_n, \Sigma)$$

is considered, where $Y$ is a matrix of observations of size $n \times p$, $X$ is a known $n \times k$ design matrix of rank($X$) = $k$ which is the same matrix as in (9), $B$ is a $k \times p$ matrix of unknown parameters, and $\Sigma$ is an unknown symmetric positive definite covariance matrix of size $p \times p$.

Similarly as in previous sections, when considering a priori information, the restrictions will be formulated with the use of some non-random, symmetric positive semi-definite matrix $H$, which in the section presenting simulation studies excluded the intercept parameters from the
constraint. Before working with the likelihood function, a multivariate least squares approach is presented. Assuming $\Sigma = I_p$ in (19),

$$\text{tr}\{(Y - XB)^	op(Y - XB)\}$$

will be considered as a loss function.

**Proposition 5.1.** Assume model (19) without normality assumption and with $\Sigma = I_p$. Let the a priori information be quantified by $\text{tr}(B^\top HB) \leq t$, where $t > 0$ is a known constant, and $H$ is a known, non-random positive semi-definite matrix. According to (20) a multivariate least squares estimate, $\hat{B}$, of $B$ equals:
Figure 2. Boxplots of the squared norms given in (17) for ML, SB and RR estimation methods and eight considered models.

(i) if \( \text{tr} (Y^{\top} X (X^{\top} X)^{-1} H (X^{\top} X)^{-1} X^{\top} Y) \leq t \) then

\[
\hat{B} = (X^{\top} X)^{-1} X^{\top} Y;
\]

(ii) if \( \text{tr} (Y^{\top} X (X^{\top} X)^{-1} H (X^{\top} X)^{-1} X^{\top} Y) > t \) then

\[
\hat{B} = (X^{\top} X + \hat{\lambda} H)^{-1} X^{\top} Y,
\]

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

\[
\text{tr} (Y^{\top} X (X^{\top} X + \lambda H)^{-1} H (X^{\top} X + \lambda H)^{-1} X^{\top} Y) = t.
\]
Proof. Both \( \text{tr}\{ (Y - XB)^\top (Y - XB) \} \) and \( \text{tr}(B^\top HB) \) are convex differentiable functions in \( B \). Thus convex optimization takes place. The Lagrangian function is given by

\[
L(B, \lambda) = \text{tr}\{ (Y - XB)^\top (Y - XB) + \lambda(B^\top HB - t) \},
\]

where \( \lambda \) is the Lagrange multiplier, implying that the KKT conditions equal

\[
\frac{d}{dB} L(B, \lambda) = 0 \tag{21}
\]
\[
\text{tr}(B^\top HB) - t \leq 0,
\]
\[
\lambda \geq 0,
\]
\[
\lambda \left( \text{tr}(B^\top HB) - t \right) = 0. \tag{22}
\]

The \text{vec}(\bullet) operator will be utilized, which stands for the vector of columns of a matrix which are put underneath starting with the first one. Using the chain rule (A.1) and Lemma 6.1 (ii) - (iv)

\[
\frac{d}{dB} L(B, \lambda) = -2 \text{vec}^\top(X^\top(Y - XB)) + 2 \lambda \text{vec}^\top(HB),
\]

implying that condition (21) is satisfied by

\[
\tilde{B} = (X^\top X + \lambda H)^{-1}X^\top Y,
\]

which is a necessary relation for obtaining the least squares estimates. Thus \( \lambda \) has to be determined and the complementary slackness condition (22) will guide us. Condition (22) implies that either \( \lambda = 0 \) or \( \text{tr}(\tilde{B}^\top HB) - t = 0 \). If

\[
\text{tr} \left( Y^\top X (X^\top X)^{-1}H (X^\top X)^{-1}X^\top Y \right) < t,
\]

then it is necessary that \( \lambda = 0 \) and \( \tilde{B} = \hat{B} \), where \( \hat{B} \) is the estimate given in statement (i). If

\[
\text{tr} \left( Y^\top X (X^\top X)^{-1}H (X^\top X)^{-1}X^\top Y \right) = t
\]

then, if the KKT conditions should be satisfied, once again it follows that \( \lambda = 0 \). Thus statement (i) is proven.

If

\[
\text{tr} \left( Y^\top X (X^\top X)^{-1}H (X^\top X)^{-1}X^\top Y \right) > t,
\]

19
then, according to (22), \( \lambda > 0 \), and it follows that \( \lambda \) has to be determined via

\[
\text{tr} \left( Y^T X (X^T X + \lambda H)^{-1} H (X^T X + \lambda H)^{-1} X^T Y \right) \leq t.
\]

In this case the KKT conditions are satisfied which establishes statement (ii).

Remark 5.1. Once again the estimator obtained in Proposition 5.1 can be called a Safety belt estimator since it takes care of the a priori information in an optimal way.

Similarly to Section 3 the focus is on maximum likelihood estimation, and here we are interested in the maximum likelihood estimation of the unknown parameters, \( B \) and \( \Sigma \), under some restrictions. However, before we prove the main results of this paper it will be assumed that \( \Sigma \) is known. The reason is the possibility to screen for difficulties with the model formulation, i.e., a difficult result with a known \( \Sigma \) will also become difficult to obtain and interpret when \( \Sigma \) has to be estimated. The following propositions are generalizations of Propositions 3.3 and 3.4 to the MANOVA model.

Proposition 5.2. Assume model (19) with a known \( \Sigma \), and a priori information quantified by \( \text{tr}(B^T H B) \leq t \), where \( t > 0 \) is a known constant, and \( H \) is a known, non-random positive semi-definite matrix. The maximum likelihood estimator, \( \hat{B} \), of \( B \) equals:

(i) if \( \text{tr}\{Y^T X (X^T X)^{-1} H (X^T X)^{-1} X^T Y\} \leq t \), then

\[
\hat{B} = (X^T X)^{-1} X^T Y;
\]

(ii) if \( \text{tr}\{Y^T X (X^T X)^{-1} H (X^T X)^{-1} X^T Y\} > t \), then

\[
\text{vec} \hat{B} = (I_p \otimes X^T X + \hat{\Sigma} \otimes \hat{\lambda} H)^{-1} (I_p \otimes X^T) \text{vec} Y,
\]

where \( \hat{\lambda} \) is the solution to

\[
\text{tr}\{Y^T X (X^T X + \lambda H)^{-1} H (X^T X + \lambda H)^{-1} X^T Y\} = t.
\]

Proof. Since \( \Sigma \) is known, model (19) can be post-multiplied by \( \Sigma^{-1/2} \), where without loss of generality \( \Sigma^{-1/2} \) is chosen to be symmetric, and obtain

\[
Z \sim N_{n,p}(X \Gamma, I_n, I_p),
\]
with $Z = Y \Sigma^{-1/2}$, $\Gamma = B \Sigma^{-1/2}$. The constraint is expressed as $\text{tr}(\Sigma \Gamma^T H \Gamma) \leq t$, which is a convex function in $\Gamma$. Moreover, the negative log-likelihood for the model in (19) is also a convex function with respect to $\Gamma$. The Lagrangian function based on the negative log-likelihood equals (for technical reasons $\frac{\lambda}{2}$ is used)

$$L(B, \lambda) = \frac{1}{2} \ln(2\pi) + \frac{1}{2} \text{tr}\{(Z - X \Gamma)(Z - X \Gamma)^T\} + \frac{\lambda}{2}(\text{tr}\{\Sigma \Gamma^T H \Gamma\} - t).$$

Thus the KKT conditions are given by

\begin{align}
- \text{vec}^T(X^T(Z - X \Gamma)) + \lambda \text{vec}^T(H \Gamma \Sigma) &= 0, \\
\text{tr}(\Sigma \Gamma^T H \Gamma) &\leq t, \\
\lambda &\geq 0, \\
\lambda(\text{tr}\{\Sigma \Gamma^T H \Gamma\} - t) &= 0,
\end{align}

(23)

where in particular the expression in (23) is obtained via application of the results for the matrix derivatives presented in Appendix A (the chain rule in (A.1) and Lemma 6.1 (ii)–(iv)). Using (23) we obtain

$$\text{vec}(\Gamma) = (I_p \otimes X^T X + \lambda \Sigma \otimes H)^{-1}(I_p \otimes X^T) \text{vec} Z,$$

which expressed in terms of the original matrices is identical to

$$\text{vec}(B) = (I_p \otimes X^T X + \lambda \Sigma \otimes H)^{-1}(I_p \otimes X^T) \text{vec} Y.$$

(25)

If $\lambda = 0$ then (25) yields

$$B = (X^T X)^{-1} X^T Y,$$

which is identical to $\hat{B}$ in statement (i).

Now the complementary slackness condition (24) expressed in terms of the original matrices equals $\lambda(\text{tr}\{B^T H B\} - t) = 0$ and then the remaining proof follows by copying the proof of Proposition 5.1. When

$$\text{tr}\{Y^T X (X^T X)^{-1} H (X^T X)^{-1} X^T Y\} \leq t,$$
then $\hat{\mathbf{B}}$ with $\lambda = 0$ satisfy the KKT conditions. On the other hand, if we have
\[
\text{tr}\{\mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}\} > t,
\]
then $\lambda$ has to be determined via $\text{tr}(\hat{\mathbf{B}}^\top \mathbf{H} \hat{\mathbf{B}}) = t$, as in statement (ii). Thus the proposition is established. \(\square\)

In the aforementioned proposition the Safety belt estimator of $\mathbf{B}$ in statement (ii) is given in a vectorized form. In the next proposition the known covariance matrix is included in the constraint, and it is shown that this formulation allows to express the Safety belt estimator of $\mathbf{B}$ in matrix form.

**Proposition 5.3.** Assume model (19) with a known $\Sigma$ and with a priori information quantified by $\text{tr}(\Sigma^{-1} \mathbf{B}^\top \mathbf{H} \mathbf{B}) \leq t$, where $t > 0$ is a known constant, and $\mathbf{H}$ is a known, non-random positive semi-definite matrix. The maximum likelihood estimate, $\hat{\mathbf{B}}$, of $\mathbf{B}$ equals:

(i) if $\text{tr}\left[\Sigma^{-1} \mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}\right] \leq t$ then
\[
\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y};
\]

(ii) if $\text{tr}\left[\Sigma^{-1} \mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}\right] > t$ then
\[
\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X} + \hat{\lambda} \mathbf{H})^{-1} \mathbf{X}^\top \mathbf{Y},
\]
where $\hat{\lambda}$ is the solution to
\[
\text{tr}\left[\Sigma^{-1} \mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{H})^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{H})^{-1} \mathbf{X}^\top \mathbf{Y}\right] = t.
\]

**Proof.** Since $\Sigma$ is known, (19) can be post-multiplied by $\Sigma^{-1/2}$ and then
\[
\mathbf{Z} \sim \mathcal{N}_{n,p}(\mathbf{X} \Gamma, \mathbf{I}_n, \mathbf{I}_p),
\]
with $\mathbf{Z} = \mathbf{Y} \Sigma^{-1/2}$, $\Gamma = \mathbf{B} \Sigma^{-1/2}$. The constraint is expressed as $\text{tr}(\Gamma^\top \mathbf{H} \Gamma) \leq t$, which obviously is a convex function in $\Gamma$. Furthermore, the Lagrangian function, based on the negative log-likelihood, equals
\[
L(\mathbf{B}, \lambda) = \frac{1}{2} \ln(2\pi) + \frac{1}{2} \text{tr}\{(\mathbf{Z} - \mathbf{X} \Gamma)(\mathbf{Z} - \mathbf{X} \Gamma)^\top\} + \frac{\lambda}{2} (\text{tr}(\Gamma^\top \mathbf{H} \Gamma) - t).
\]
Therefore the KKT conditions can be written as

\[- \text{vec}^\top (X^\top (Z - X\Gamma)) + \lambda \text{vec}^\top (H\Gamma) = 0, \]
\[\text{tr}(\Gamma^\top H\Gamma) \leq t,\]
\[\lambda \geq 0,\]
\[\lambda (\text{tr}\{\Gamma^\top H\Gamma\} - t) = 0,\]  

(26)

where (26) is obtained via application of the results for matrix derivatives presented in Appendix A, i.e., the chain rule (A.1) and the derivatives in Lemma 6.1 (ii)–(iv). When solving (26) yields

\[B = (X^\top X + \lambda H)^{-1} X^\top Y.\]  

(28)

If \(\lambda = 0\) then (28) is identical to

\[B = (X^\top X)^{-1} X^\top Y,\]

which is simply the same as \(\hat{B}\) in statement (i). Now the complementary slackness condition in (27), expressed in terms of the original matrices, equals

\[\lambda (\text{tr}\{\Sigma^{-1} B^\top HB\} - t) = 0.\]

(27)

The remaining part of the proof of the proposition follows by copying the proof of Proposition 5.2.

Note that if \(p = 1\), Propositions 5.2 and 5.3 reduce to Propositions 3.3 and 3.4, respectively.

The most natural generalization of the results presented in Proposition 5.2 is when \(\Sigma\) is supposed to be unknown. In the univariate case, estimators of the mean and variance were presented in Proposition 3.1. However, the proof of Proposition 3.1 cannot be extended to the multivariate case, as the convexity of \(\ln |(Y - XB)^\top (Y - XB)|\) with respect to \(B\) does not hold; see Appendix B for the proof.

In addition, observe that for \(p = 1\) the aforementioned function is concave. This fact did not play significant role in the proof of Proposition 3.1, as it was enough that the function under the determinant is convex. Note however, that for \(p > 1\) this property is not valid anymore, as the determinant of p.d. function is not convex nor concave; again see Appendix B for a proof. The question that may arise is whether the composition of convex function, \((Y - XB)^\top (Y - XB)\), and the determinant, implies convexity of \(|(Y - XB)^\top (Y - XB)|\).
The answer is no, as many counterexamples for indefiniteness of respective Hessian matrix (see Appendix B) can be found.

Another approach that would allow to state the multivariate version of Proposition 3.1 is a vectorization of the original model (19); however, this transformation reduces the model to the univariate case, model (9), only for $\Sigma = \sigma^2 I_p$.

To prove the main results of this section it will be utilized that estimates can be derived via convex optimization. The negative log-likelihood function

$$f(B, \Sigma) = \frac{n}{2} \ln(2\pi) + \frac{n}{2} \ln |\Sigma| + \frac{1}{2} \text{tr} \left[ \Sigma^{-1} (Y - XB)^\top (Y - XB) \right]$$

(29)

is neither convex nor concave with respect to the original parameters $B$ and $\Sigma$. However, the multivariate normal distribution, as the univariate one, belongs to the exponential family. For the MANOVA model the canonical parameters are given by

$$\Theta_1 = B\Sigma^{-1} : k \times p, \quad \Theta = \text{vech} \Theta_2 : \frac{1}{2}p(p+1) \times 1,$$

(30)

where $\Theta_2 = \Sigma^{-1}: p \times p$ and vech($\bullet$) denotes the vectorized lower triangle of a square matrix. This implies the important fact that the negative log-likelihood function for $Y$ in (19) is a convex function when the canonical parameters are used as parameters in the model. Note that vec$\Theta_2 = D_p \Theta$, where $D_p$ is the well known duplication matrix, see e.g., Magnus and Neudecker (1986).

Using the canonical parameters the expression in (29) equals

$$f(\Theta_1, \Theta) = \frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln |\Theta_2| + \frac{1}{2} \text{tr}(\Theta_2 Y^\top Y) - \text{tr}(Y^\top X \Theta_1)$$

$$+ \frac{1}{2} \text{tr}(\Theta_2^{-1} \Theta_1^\top X^\top X \Theta_1),$$

(31)

which is a convex function in $\Theta_1$ and $\Theta$. One way of showing the convexity is to establish the positive semi-definiteness of the corresponding Hessian. For this purpose the matrix differentiation rules presented in Appendix A are used and the proof of convexity of (31) is given in details in Appendix C.

In the last result we present the Safety belt estimates of $B$ and $\Sigma$ in model (19) and the constraint $\text{tr}(\Sigma^{-1} B^\top HB) \leq t$, for unknown $\Sigma$ and known positive scalar $t$. Note, that this is a multivariate generalization of Proposition 3.2.

**Proposition 5.4.** Assume model (19) with a priori information quantified by $\text{tr}(\Sigma^{-1} B^\top HB) \leq t$, where $t > 0$ is a known constant, and $H$ is a known, non-random positive semi-definite
matrix. The maximum likelihood estimates, \( \hat{B} \) and \( \hat{\Sigma} \), of \( B \) and \( \Sigma \) will be presented. Put

\[
\hat{B} = (X^\top X)^{-1}X^\top Y, \\
n\hat{\Sigma} = (Y - X\hat{B})^\top(Y - X\hat{B}).
\]

Then,

(i) if \( \text{tr}(\hat{\Sigma}^{-1}\hat{B}^\top H\hat{B}) \leq t \), the maximum likelihood estimates are given by \( \hat{B} = \tilde{B} \) and \( \hat{\Sigma} = \tilde{\Sigma} \);

(ii) if \( \text{tr}(\hat{\Sigma}^{-1}\hat{B}^\top H\hat{B}) > t \), the maximum likelihood estimates of \( B \) and \( \Sigma \) satisfy

\[
\hat{B} = (X^\top X + \hat{\lambda}H)^{-1}X^\top Y \\
n\hat{\Sigma} = (Y - X\hat{B})^\top(Y - X\hat{B}) + \hat{\lambda}\hat{B}^\top H\hat{B},
\]

where \( \hat{\lambda} \) is the unique solution to

\[
\text{tr}\left[\hat{\Sigma}^{-1}Y^\top X(X^\top X + \lambda H)^{-1}H(X^\top X + \lambda H)^{-1}X^\top Y\right] = t.
\]

**Proof.** Note that the restriction \( \text{tr}(\Sigma^{-1}B^\top HB) \leq t \) can be expressed in terms of the canonical parameters \( \Theta_1 \) and \( \Theta \), which are defined as in (30), as \( \text{tr}(\Theta_1^\top H\Theta_1\Theta_2^{-1}) \leq t \). Since both, \( f(\Theta_1, \Theta) \) given in (31) and \( g(\Theta_1, \Theta) = \text{tr}(\Theta_1^\top H\Theta_1\Theta_2^{-1}) - t \) are convex in \( \Theta_1 \) and \( \Theta \), it is enough to consider the convex optimization problem via the Lagrangian function

\[
L(\Theta_1, \Theta, \lambda) = f(\Theta_1, \Theta) + \lambda g(\Theta_1, \Theta).
\]

Since

\[
\frac{dL(\Theta_1, \Theta, \lambda)}{d\Theta} = -\text{vec}^\top \left[ X^\top Y - (X^\top X + \lambda H)\Theta_1\Theta_2^{-1} \right], \\
\frac{dL(\Theta_1, \Theta, \lambda)}{d\Theta_1} = -\frac{1}{2} \text{vec}^\top \left[ n\Theta_2^{-1} - Y^\top Y + \Theta_2^{-1}\Theta_1^\top(X^\top X + \lambda H)\Theta_1\Theta_2^{-1} \right]D_p,
\]

the KKT conditions are equivalent to

\[
X^\top Y = (X^\top X + \lambda H)\Theta_1\Theta_2^{-1}, \quad (32) \\
n\Theta_2^{-1} = Y^\top Y - \Theta_2^{-1}\Theta_1^\top(X^\top X + \lambda H)\Theta_1\Theta_2^{-1}, \quad (33) \\
\text{tr}(\Theta_1^\top H\Theta_1\Theta_2^{-1}) - t \leq 0, \\
\lambda \geq 0, \\
\lambda(\text{tr}(\Theta_1^\top H\Theta_1\Theta_2^{-1}) - t) = 0.
\]

25
From (32) we obtain
\[ \Theta_1 \Theta_2^{-1} = (X^\top X + \lambda H)^{-1} X^\top Y \]
and plugging the above into (33) implies
\[ n \Theta_2^{-1} = Y^\top Y - \Theta_2^{-1} \Theta_1^\top X \Theta_1 \Theta_2^{-1} X^\top X - \lambda \Theta_2^{-1} \Theta_1^\top H \Theta_1 \Theta_2^{-1} \).

Since \( \Theta_2 = \Sigma^{-1} \) and \( \Theta_1 = B \Sigma^{-1} \), we obtain
\[ \hat{B} = (X^\top X + \lambda H)^{-1} X^\top Y, \]
\[ n \hat{\Sigma} = Y^\top Y - \hat{B}^\top X^\top X \hat{B} - \lambda \hat{B}^\top H \hat{B}. \]

Using \( X^\top Y = (X^\top X + \lambda H) \hat{B} \) we can rewrite the last estimate as
\[ n \hat{\Sigma} = (Y - X \hat{B})^\top (Y - X \hat{B}) + \lambda \hat{B}^\top H \hat{B}, \]
which is exactly the same expression as in statement (ii).

As several times before the complimentary slackness condition implies that if
\[ \text{tr} \left[ \tilde{\Sigma}^{-1} Y^\top X (X^\top X)^{-1} H (X^\top X)^{-1} X^\top Y \right] \leq t, \]
then \( \lambda = 0 \), and \( \hat{B} \) and \( \hat{\Sigma} \) with \( \lambda = 0 \) satisfy the KKT conditions. If we have
\[ \text{tr} \left[ \tilde{\Sigma}^{-1} Y^\top X (X^\top X)^{-1} H (X^\top X)^{-1} X^\top Y \right] > t, \]
then \( \lambda \) has to be determined via \( \text{tr}(\tilde{\Sigma}^{-1} \hat{B}^\top H \hat{B}) = t \).

Summing up it is worth noting, that the for \( p = 1 \) the above result reduces to Proposition 3.2.

6 Final remarks

In this paper we considered a Safety belt estimate, based on maximum likelihood estimation method with a given a priori information related to the maximum effects in the model. Under various conditions we showed similarities and differences between Safety belt estimation and ridge regression.

Let us recall the problem with Safety belt estimation of \( B \) and \( \Sigma \) in model (19), with a
priori information \( \text{tr}(B^\top HB) \leq t \), where \( t > 0 \) is a known constant and \( H \) is some known symmetric p.d. matrix. One can imagine, that these estimates might have a similar form as in Proposition 3.1, with \( \lambda \) replaced by some \( p \times p \) matrix \( \Lambda \), i.e.,

\[
\text{vec} \hat{B} = (I_p \otimes X^\top X + \Lambda \otimes H)^{-1}(I_p \otimes X^\top)y. \tag{34}
\]

Similar approach, i.e., the estimate given in (34) with \( H \) replaced by identity matrix, has been proposed by, e.g., Brown and Zidek (1980, 1982), Haitovsky (1987), to address generalization of ridge-type estimation to the multivariate case. Nevertheless, this estimator is rather heuristic, even if motivated by a Bayesian approach. Another problem that arises here is the method of determination of \( \Lambda \). Since in the considered case the a priori information is precisely stated (\( t \) is known), \( \Lambda \) should satisfy

\[
y^\top (I_p \otimes X)(I_p \otimes X^\top X + \Lambda \otimes H)^{-1}(I_p \otimes H)(I_p \otimes X^\top X + \Lambda \otimes H)^{-1}(I_p \otimes X^\top)y \leq t.
\]

However, even if the equality in the above condition would be considered, there is just one equation with \( p^2 \) unknown parameters (or \( p(p+1)/2 \) if \( \Lambda \) would be assumed to be symmetric). Summing up, the form of Safety belt estimates of \( B \) and \( \Sigma \) under the considered condition is worth investigating, however, it will be the topic of future research.

Acknowledgments

The authors thank to the International Banach Center of the Polish Academy of Sciences for providing the opportunity and support for this paper.

The research of K. Filipiak was funded by the Poznan University of Technology under Grant no. 0213/SBAD/0118.

The research of W. Rejchel was funded by the Polish National Science Center grant: NCN UMO-2018/31/B/ST1/00253.

Appendix A: Matrix derivatives

Let \( F(X) \) be an \( r \times s \) differentiable real matrix function of \( X \in \mathbb{R}^{p \times q} \) with non-constant and functionally independent elements. Following Magnus and Neudecker (1988) the Jacobian
matrix or the derivative of $F(X)$ by $X$ is a $rp \times sq$ matrix defined as
\[
\frac{d F(X)}{d^\top X} = \frac{d \text{vec} F(X)}{d \text{vec}^\top X}.
\]
In particular, let us assume that $F(X)$ depends on $X$ through $Y(X)$, that is we are interested in the derivative of $F(Y(X))$ with respect to $X$. According to the chain rule defined by Magnus and Neudecker (1988) we have
\[
\frac{d F(Y(X))}{d^\top X} = \frac{d F(Y)}{d^\top Y} \frac{d Y}{d^\top X}.
\]
(A.1)

For a $p \times p$ symmetric matrix $X$ the derivative is defined with respect to the vectorized lower triangle of $X$, i.e., $\frac{d Y}{d^\top X} = \frac{d \text{vec} Y}{d \text{vech}^\top X}$. Using the above presented chain rule for $X = X^\top$ yields,
\[
\frac{d Y}{d^\top X} = \frac{d \text{vec} Y}{d \text{vec}^\top X} \cdot \frac{d \text{vec} X}{d \text{vech}^\top X} = \frac{d \text{vec} Y}{d \text{vec}^\top X} \cdot D_p.
\]

Following Magnus and Neudecker (1988), the second derivative is defined as
\[
\frac{d^2 Y}{d X d^\top X} = \frac{d^2 Y}{d \text{vec} X d \text{vec}^\top X}.
\]
(A.2)

The differentiation formulas presented in the next lemma were established by Magnus and Neudecker (1988). However, both (i) and (v) of the lemma are not used explicitly in the main body of the presentation but are needed when the Hessian for the negative log-density of the normal distribution in Appendix C is derived.

Let $K_{m,q}$ be the well known $mq \times mq$ commutation matrix, see e.g., Magnus and Neudecker (1986).

Lemma 6.1.

(i) Let $X \in \mathbb{R}^{p \times p}$. Then
\[
\frac{d \ln |X|}{d^\top X} = \text{vec}^\top ((X^{-1})^\top).
\]
(ii) Let $A \in \mathbb{R}^{p \times q}$ be a matrix of constants and $X \in \mathbb{R}^{q \times p}$. Then
\[
\frac{d \text{tr}(AX)}{d^\top X} = \text{vec}^\top (A^\top).
\]
Let $A \in \mathbb{R}^{r \times q}$ and $B \in \mathbb{R}^{p \times s}$ be matrices of constants and $X \in \mathbb{R}^{q \times p}$. Then

$$\frac{d AXB}{d^\top X} = B^\top \otimes A.$$  

(iv) Let $X \in \mathbb{R}^{p \times q}$. Then

$$\frac{d X^\top X}{d^\top X} = (I_p^2 + K_{p,p})(I_p \otimes X^\top).$$

(v) Let $X \in \mathbb{R}^{p \times p}$. Then

$$\frac{d X^{-1}}{d^\top X} = -(X^{-1})^\top \otimes X^{-1}.$$ 

Appendix B: Non-convexity of log-determinant and the determinant of the product of two matrices

The log-determinant function of some p.d. matrix is concave (cf. Boyd and Vandenberghe, 2004); however, the composition of log-determinant function with the product $(Y - XB)^\top (Y - XB)$ (which is obviously convex function of $B$) is not convex nor concave. Indeed, if

$$F_1(B) = \ln |(Y - XB)^\top (Y - XB)|$$

then

$$\frac{d F_1(B)}{d B} = - \text{vec}^\top [(Y - XB)^\top (Y - XB)]^{-1} (I_p^2 + K_{p,p})(I_p \otimes (Y - XB)^\top X$$

$$= -2 \text{vec}^\top \{X^\top (Y - XB)[(Y - XB)^\top (Y - XB)]^{-1}\}.$$ 

Denoting $A = Y - XB$ we can write Hessian matrix as

$$2 \{ (A^\top A)^{-1} \otimes X^\top Q_A X - [(A^\top A)^{-1} A^\top A^\top A (A^\top A)^{-1} A X X^\top A (A^\top A)^{-1} K_{k,p}] \}.$$ 

It is easy to state a counterexample, that this Hessian matrix is indefinite (neither p.d., nor n.d.).

Let $F_2(W) = |W|$, with an $p \times p$ symmetric p.d. matrix $W$. The first derivative of $F_2(W)$
equals \[
\frac{d F_2(W)}{d W} = |W| \text{vec}^\top (W^{-1}) D_p,
\]
and the Hessian can be presented as
\[
D_p^\top \left[ \text{vec}(W^{-1})|W| \text{vec}^\top (W^{-1}) - |W|(W^{-1} \otimes W^{-1}) \right] D_p
\]
which is equivalent to
\[
D_p^\top |W|(W^{-1/2} \otimes W^{-1/2}) \left( \text{vec} I_p \text{vec}^\top I_p - I_p^2 \right) (W^{-1/2} \otimes W^{-1/2}) D_p.
\]
Since the matrix in the middle is indefinite, the whole Hessian is indefinite as well.

Let now \( F_3(B) = |(Y - XB)^\top (Y - XB)| \). Denoting \( Y - XB \) by \( A \) and using the same rules as before, the Hessian has the form
\[
2|A^\top A| \left\{ (A^\top A)^{-1} \otimes X^\top Q_A X - [(A^\top A)^{-1} A^\top X \otimes X^\top A(A^\top A)^{-1}] K_{k,p} 
+ 2 \text{vec}(X^\top A(A^\top A)^{-1}) \text{vec}^\top (X^\top A(A^\top A)^{-1}) \right\}.
\]
The examples for indefiniteness of the above can be easily found.

**Appendix C: Convexity of negative log-likelihood function and the a priori information given in Proposition 5.4 with respect to canonical parameters**

Differentiation of \( f(\Theta_1, \Theta) \) given in (31) with respect to \( \Theta_1, \Theta = \text{vech} \Theta_2 \), gives
\[
\frac{d f(\Theta_1, \Theta)}{d \Theta_1} = - \text{vec}^\top (X^\top Y) + \frac{1}{2} \text{vec}^\top (\Theta_2^{-1})(I_p^2 + K_{p,p})(I_p \otimes \Theta_1^\top X^\top)(I_p \otimes X)
\]
\[
= - \text{vec}^\top (X^\top Y - X^\top X \Theta_1 \Theta_2^{-1}),
\]
\[
\frac{d f(\Theta_1, \Theta)}{d \Theta} = [-\frac{n}{2} \text{vec}^\top \Theta_2^{-1} + \frac{1}{2} \text{vec}^\top (Y^\top Y) - \frac{1}{2} \text{vec}^\top (\Theta_1^\top X^\top X \Theta_1)(\Theta_2^{-1} \otimes \Theta_2^{-1})] D_p
\]
\[
= -\frac{1}{2} \text{vec}^\top (n\Theta_2^{-1} - Y^\top Y + \theta_2^{-1} \Theta_1^\top X^\top X \Theta_1 \theta_2^{-1}) D_p.
\]
Note, that in the derivative with respect to \( \Theta \) the equality \( K_{p,p} \text{vec} \Theta_2^{-1} = \text{vec} \Theta_2^{-1} \) holds because of symmetry of \( \Theta_2 \).

To prove convexity of \( f(\Theta_1, \Theta) \) we compute now second derivatives of \( f(\Theta_1, \Theta) \), which
can be calculated as
\[
\frac{d^2 f}{d \Theta_1 d \Theta_1} = \Theta_2^{-1} \otimes X^\top X, \\
\frac{d^2 f}{d \Theta_1 d \Theta} = -\frac{1}{2} D_p (I_p^2 + K_{p,p})(I_p \otimes \Theta_2^{-1} \Theta_1^\top X^\top)(\Theta_2^{-1} \otimes X), \\
\frac{d^2 f}{d \Theta d \Theta} = \frac{1}{2} D_p^\top \left[ n(\Theta_2^{-1} \otimes \Theta_2^{-1})(I_p^2 + K_{p,p})(I_p \otimes \Theta_2^{-1} \Theta_1^\top X^\top)(I_p \otimes X \Theta_1)(\Theta_2^{-1} \otimes \Theta_2^{-1}) \right] D_p
\]
since \(K_{p,p} D_p = D_p\). Therefore, the upper triangle of the Hessian matrix equals
\[
\begin{pmatrix}
\Theta_2^{-1} \otimes X^\top X & -(\Theta_2^{-1} \otimes X^\top X \Theta_1 \Theta_2^{-1}) D_p \\
D_p^\top (\frac{n}{2} \Theta_2^{-1} \otimes \Theta_2^{-1} + \Theta_2^{-1} \otimes \Theta_2^{-1} \Theta_1^\top X^\top X \Theta_1 \Theta_2^{-1}) D_p
\end{pmatrix}.
\]
The above given Hessian matrix can be represented as the sum of two positive semi-definite matrices:
\[
\frac{n}{2}
\begin{pmatrix}
0_{k p \times p^2} \\
\Theta_2^{-1} \otimes \Theta_2^{-1}
\end{pmatrix}
(0_{p^2 \times k p} : D_p)
\]
and
\[
\begin{pmatrix}
-I_p \otimes I_k \\
D_p^\top (I_p \otimes \Theta_2^{-1} \Theta_1^\top)
\end{pmatrix}
(\Theta_2^{-1} \otimes X^\top X)
\begin{pmatrix}
-I_p \otimes I_k : (I_p \otimes \Theta_1 \Theta_2^{-1}) D_p
\end{pmatrix},
\]
and hence it is also positive semi-definite, which implies convexity of \(f(\Theta_1, \Theta)\).

Let
\[
g(\Theta_1, \Theta) = \frac{1}{2} \left( \text{tr} \left( \Theta_1^\top H \Theta_1 \Theta_2^{-1} \right) - t \right)
\]
with \(\Theta = \text{vech} \Theta_2\). Applying the chain rule given in (A.1) together with Lemma 6.1 (ii)–(iv) yields
\[
\frac{d g(\Theta_1, \Theta)}{d \Theta_1} = \frac{1}{2} \text{vec}^\top \Theta_2^{-1}(I_p + K_{p,p})(I_p \otimes \Theta_1^\top H) = \text{vec}^\top (H \Theta_1 \Theta_2^{-1}), \\
\frac{d g(\Theta_1, \Theta)}{d \Theta} = -\frac{1}{2} \text{vec}^\top (\Theta_1^\top H \Theta_1)(\Theta_2^{-1} \otimes \Theta_2^{-1}) D_p = -\frac{1}{2} \text{vec}^\top (\Theta_2^{-1} \Theta_1^\top H \Theta_1 \Theta_2^{-1}) D_p,
\]
where \(D_p\) as before is the duplication matrix.
The second derivatives which build up the Hessian are given by

\[
\frac{d^2 g(\Theta_1, \Theta)}{d \Theta_1 d^\top \Theta_1} = \Theta_2^{-1} \otimes H,
\]

\[
\frac{d^2 g(\Theta_1, \Theta)}{d \Theta_1 d^\top \Theta} = -\frac{1}{2} D_p^\top (I_p^2 + K_{p,p})(I_p \otimes \Theta_2^{-1} \Theta_1^\top)\Theta_2^{-1} \otimes H
\]

\[
= -D_p^\top (\Theta_2^{-1} \otimes \Theta_2^{-1} \Theta_1^\top H),
\]

\[
\frac{d^2 g(\Theta_1, \Theta)}{d \Theta d^\top \Theta} = \frac{1}{2} D_p^\top (I_p^2 + K_{p,p})(I_p \otimes \Theta_2^{-1} \Theta_1^\top)(I_p \otimes H \Theta_1)(\Theta_2^{-1} \otimes \Theta_2^{-1})D_p
\]

\[
= D_p^\top (\Theta_2^{-1} \otimes \Theta_2^{-1} \Theta_1^\top H \Theta_1 \Theta_2^{-1})D_p,
\]

since \( K_{p,p} D_p = D_p \). The Hessian matrix equals

\[
\left(\begin{array}{cc}
\frac{d^2 g(\Theta_1, \Theta)}{d \Theta_1 d^\top \Theta_1} & \frac{d^2 g(\Theta_1, \Theta)}{d \Theta_1 d^\top \Theta} \\
\frac{d^2 g(\Theta_1, \Theta)}{d \Theta d^\top \Theta_1} & \frac{d^2 g(\Theta_1, \Theta)}{d \Theta d^\top \Theta}
\end{array}\right)
\]

\[
= \left(\begin{array}{c}
-I_p \otimes I_u \\
D_p^\top (I_p \otimes \Theta_2^{-1} \Theta_1^\top)
\end{array}\right) \left(\Theta_2^{-1} \otimes H\right) \left(\begin{array}{c}
-I_p \otimes I_k : (I_p \otimes \Theta_1 \Theta_2^{-1})D_p
\end{array}\right),
\]

which is positive semi-definite. Thus \( g(\Theta_1, \Theta) \) is a convex function in \( \Theta_1 \) and \( \Theta \).

**References**


