Failure to fit a heteroscedastic model due to an unbounded likelihood

M.J.P. Nijmeijer

nijmeijer@posthuma-partners.nl

Posthuma Partners
Zuidelijk Halfrond 1
2801 DD Gouda
the Netherlands

March 16, 2020

Abstract

We look at a heteroscedastic model which is linear in the expected value and log-linear in the variance and show that a particular structure of the models for expected value and variance, leads to an unbounded likelihood and possibly failing model fits. We show that a bounded likelihood is obtained when certain degrees of freedom are removed from the model for the variance. We develop an algorithm to identify these. Our ideas are illustrated with a numerical example.

Keywords: heteroschedastic models, unbounded likelihood, log-linear variance

1 Introduction

Modeling a data set sometimes calls for heteroscedasticity [6]. For example, at Posthuma Partners we have experienced that heteroscedastic models can significantly improve the description of insurance-claims data over their homoscedastic counterparts [7, 9].

One of the simplest heteroscedastic models is a Gaussian model which is linear for the expectation value and log-linear for the variance [1, 17, 16]. We call this the LMVAR model and give a detailed definition in the next section.

It is our experience that fitting a LMVAR model is not as straightforward as fitting a classical linear model. Although a successful LMVAR fit can yield better results than a classical linear fit, attempts to fit a data set to an LMVAR model fail quite frequently, with the software returning an error or warning message. The software we use is R [12], in particular the public domain package 'lmvar' [11] which was written by the current author. However, the phenomenon also occurs with another R-package capable of treating the LMVAR model: 'dglm' [4].

Fit-routines typically search for the maximum likelihood. In this paper, we describe that the maximum likelihood of an LMVAR model is undefined if the models for the expected value and for the variance have a specific structure. We show that the characteristics of this structure match what we see in an example of a failing fit. We present a way to restore the existence of a maximum likelihood and show that this stabilizes the previously failing fit. Presumably, a similar mechanism as for LMVAR exists for other heteroscedastic models as well.

All our numerical calculations were carried out on a laptop with an Intel Core i5-6200U processor and 16GB RAM.

In the next section we describe the LMVAR model. In Section 3, we give a numerical example of a model that does not fit. We describe the conditions under which the likelihood is not bounded from above in Section 4. The mathematical background for an algorithm is given in Section 6. In Section 7, the algorithm is developed, which is demonstrated in Section 8 for the numerical example introduced in Section 3. Section 9 discusses some aspects of the algorithms efficiency and its complexity class. Alternative approaches are discussed in Section 10. Conclusions are given in Section 11.

2 The LMVAR Model

The LMVAR model is defined by

$$Y \sim \mathcal{N}_n(\mu, \Sigma), \tag{1}$$

[1] with $Y \in \mathbb{R}^n$ the response variable, $\mathcal{N}_n$ the multivariate Gaussian distribution, $\mu \in \mathbb{R}^n$ the expected value of Y and $\Sigma \in \mathbb{R}^{n,n}$ the diagonal covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}. \tag{2}$$

The model for μ is linear

$$\mu = X_\mu \beta_\mu \tag{3}$$

with $X_\mu \in \mathbb{R}^{n,p_\mu}$ a matrix of explanatory variables and $\beta_\mu \in \mathbb{R}^{p_\mu}$ the parameter vector of betas for μ, while the σ_i have a log-linear model

$$\begin{pmatrix} \log \sigma_1 \\ \vdots \\ \log \sigma_n \end{pmatrix} = X_\sigma \beta_\sigma \tag{4}$$

with $X_\sigma \in \mathbb{R}^{n,p_\sigma}$ and $\beta_\sigma \in \mathbb{R}^{p_\sigma}$.

The two matrices are called the model matrices or design matrices. We assume both are full-rank. We also assume that the parameter vectors β_μ and β_σ are estimated as maximum-likelihood estimators when fitting a data set to the LMVAR model.

3 Numerical Example, Part 1

As a numerical example, we take a design matrix X and a response vector y originating from the study of an insurance portfolio. The vector has 34,511 observations, the matrix has 188 columns and includes an intercept term (that is, a column in which each matrix element is equal to 1). Except for the intercept term and six other columns, all columns represent factor-levels: the matrix elements in those columns are either 0 or 1. The value 1 appears at least 21 times and at most 34,127 in any of those columns.

We fit the response vector to an LMVAR model with $X_\mu = X$ and $X_\sigma = X$. Using the 'lmvar' package, the fit is returned together with two warnings: one warning which indicates that the iterative process ran into trouble and one warning that the final log-likelihood is not at a local or global maximum. The 'dglm' package returns an error and no fit.

If we take the fit obtained with the 'lmvar' package, we notice some unusual things. First, Table 1 lists the four smallest and the four largest elements of the result for β_σ. There are two values that are unlikely large in absolute value compared to the other values. Second, Figure 1 shows the fitted values of σ for all observations. Observation 32,072 stands out with an incredibly small value for σ.

4 An Unbounded Likelihood

We argue that the convergence problems of the previous section are caused by a likelihood which diverges to infinity for ever more extreme values of β_σ.

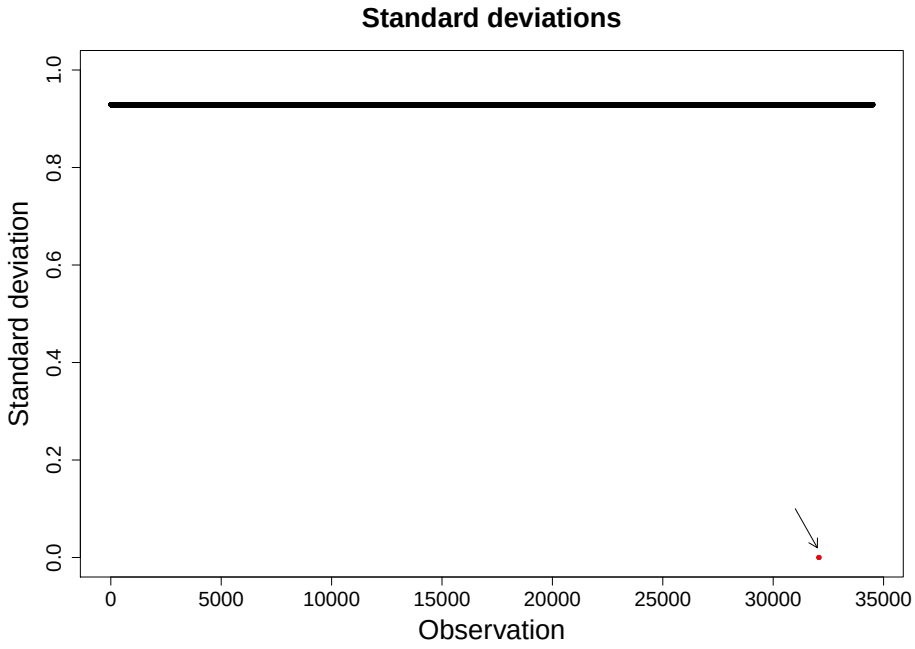

Figure 1: The fitted standard deviations for all observations. The value for observations 32,072 (indicated by the arrow and shown in red) is much smaller than for any other observation.

element	value
94	-9.10536
1	-0.07228
141	-0.00071
38	-0.00010
⋮	⋮
93	0.00202
88	0.00203
72	0.00205
87	9.10698

Table 1: The four smallest and the four largest elements of the maximum-likelihood estimate of β_σ. The first column shows the index of the vector element, the second column the element-value. Element 1 corresponds to the intercept term in X_σ.

The log-likelihood is

$$\log \mathcal{L} = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \log \sigma_i - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma_i^2}. \tag{5}$$

It will diverge to plus infinity if one or more of the σ_i go to zero while the corresponding residuals $y_i - \mu_i$ are zero (or go to zero quickly enough). This is the only way the divergence can come about. It means there must be a (non-empty) subset S_1 of all n observations and a β'_μ such that $y_i = \mu'_i$ for all $i \in S_1$ (with $\mu' = X_\mu \beta'_\mu$). The log-likelihood as a function of this β'_μ and some β'_σ takes the form

$$\log \mathcal{L}(\beta'_\mu, \beta'_\sigma) = -\frac{n}{2}\log(2\pi) - \sum_{i \in S_1} \left(X_\sigma \beta'_\sigma\right)_i - \sum_{i \in S_2} \left(X_\sigma \beta'_\sigma\right)_i - \sum_{i \in S_2} \frac{(y_i - \mu'_i)^2}{2} \exp\left(-2\left(X_\sigma \beta'_\sigma\right)_i\right). \tag{6}$$

The most straightforward way to establish the divergence is to construct a sequence $\{\beta'_{\sigma j}\}_{j=1}^{\infty}$ such that

$$\sum_{i \in S_1} \left(X_\sigma \beta'_{\sigma j}\right)_i + \sum_{i \in S_2} \left(X_\sigma \beta'_{\sigma j}\right)_i \xrightarrow{j \to \infty} -\infty \tag{7}$$

while

$$\left(X_\sigma \beta'_{\sigma j}\right)_i > B \text{ for all } i \in S_2 \text{ and all } j \tag{8}$$

for some, possibly negative, bound B. Condition (7) requires that at least one of the components of $\beta_{\sigma j}$ diverges with increasing j whereas condition (8) requires that $X_\sigma \beta_{\sigma j}$ remains bounded from below for all elements in S_2.

Such a sequence exists if the sub-matrix of X_σ obtained by removing all rows in S_1, is rank-deficient. It can be seen as follows. To ease the notation, take S_1 to be the first n_1 rows in the vector of observations Y and the model matrices X_μ and X_σ. The remaining n_2 rows (with $n_1 + n_2 = n$) form the set S_2. Split the vector of observations and the model matrices accordingly:

$$Y = \begin{pmatrix} y_{S_1} \\ y_{S_2} \end{pmatrix}, \quad X_\mu = \begin{pmatrix} X_{\mu 1} \\ X_{\mu 2} \end{pmatrix}, \quad X_\sigma = \begin{pmatrix} X_{\sigma 1} \\ X_{\sigma 2} \end{pmatrix} \tag{9}$$

with $y_{S_1} \in \mathbb{R}^{n_1}$, $y_{S_2} \in \mathbb{R}^{n_2}$, $X_{\mu 1} \in \mathbb{R}^{n_1, p_\mu}$, etc.

As discussed, we assume there exists a β'_μ such that $y_i = \mu'_i$ for all $i \in S_1$. Because $X_{\sigma 2}$ is rank-deficient, there exists a β'_σ with $\beta'_\sigma \neq 0$ such that $X_{\sigma 2} \beta'_\sigma = 0$. In terms of our matrices:

$$X_\mu \beta'_\mu = \begin{pmatrix} X_{\mu 1} \beta'_\mu \\ X_{\mu 2} \beta'_\mu \end{pmatrix} = \begin{pmatrix} y_{S_1} \\ \mu'_{S_2} \end{pmatrix} \text{ and } X_\sigma \beta'_\sigma = \begin{pmatrix} X_{\sigma 1} \beta'_\sigma \\ X_{\sigma 2} \beta'_\sigma \end{pmatrix} = \begin{pmatrix} v \\ 0 \end{pmatrix}. \tag{10}$$

Because X_σ is full-rank, $v \neq 0$. Now, $\sum_{i=1}^{n} \left(X_\sigma \beta'_\sigma\right)_i = \sum_{i=1}^{n_1} v_i \equiv s$. Unless it happens to be that $s = 0$, we can always choose the sign of β'_σ such that $s < 0$. The residuals $y_{S_2} - \mu'_{S_2}$ play no further role.

Now

$$\log \mathcal{L}(\beta'_\mu, L\beta'_\sigma) = -\frac{n}{2} \log(2\pi) - Ls - \sum_{i \in S_2} \frac{(y_i - \mu'_i)^2}{2} \xrightarrow{L \to \infty} \infty. \tag{11}$$

Besides the sequence $\{L\beta'_\sigma\}$ with $L \to \infty$, any sequence $\{L\beta'_\sigma + \hat{\beta}_\sigma\}$ with an arbitrary $\hat{\beta}_\sigma \in \mathbb{R}^{p_\sigma}$ will also give the desired divergence of the log-likelihood.

If the set of observations can not be split in a subset S_1 and a subset S_2 such that the residuals can be made to vanish in S_1 and $X_{\sigma 2}$ is rank-deficient, one can sill obtain a log-likelihood which is not bounded from above under slightly relaxed conditions for X_{σ_2}. We still need $y_{S_1} = \mu'_{S_1}$. But it is sufficient that there exists a β'_σ such that

$$\sum_{i \in S_1} \left(X_\sigma \beta'_\sigma\right)_i + \sum_{i \in S_2} \left(X_\sigma \beta'_\sigma\right)_i < 0 \tag{12}$$

and

$$\left(X_\sigma \beta'_\sigma\right)_i > 0 \text{ for at least one } i \in S_2 \tag{13}$$

while

$$\left(X_\sigma \beta'_\sigma\right)_i = 0 \text{ for the remaining } i \in S_2. \tag{14}$$

Note that if $(X_\sigma \beta'_\sigma)_i = 0$ for all $i \in S_2$, $X_{\sigma 2}$ is rank-deficient and we are in the scenario discussed before.

The condition $y_i = \mu'_i$ for all $i \in S_1$, implies that the number of observations in S_1 can be at most of the order p_μ. In the common situation that $n \gg p_\mu$, S_2 will have many more observations than S_1. If there is a sizable fraction of elements $i \in S_2$ for which $(X_\sigma \beta'_\sigma)_i > 0$, it is hard to satisfy (12).

Numerical routines calculating maximum-likelihood estimators for β_μ and β_σ typically use an iterative algorithm in which the estimators are improved with each iteration. One would expect that in the situation described in this section, the components of the estimator for β_σ diverge further with each iteration while the residuals in S_1 shrink to zero and the standard deviations in S_1 either shrink to zero or diverge, depending on the sign of v_i. At some point, numerical instabilities will thwart the iterative process, presumably leading to an abnormal termination.

5 Restoring the Bound on the Maximum Likelihood

The situation described in the previous section occurs when the removal of no more than about p_μ rows from X_σ, results in a matrix that is rank-deficient. The observations that are removed form the set S_1 and we assume that the set of equations $X_{\mu 1} \beta'_\mu = y_{S_1}$ has a solution for β'_μ. A strategy one can follow is to remove degrees of freedom (columns, that is) from X_σ such that the removal of the rows no longer leads to a rank-deficient matrix. This assumes there is no compelling reason to stay with those degrees of freedom for σ. An operational procedure to do this is as follows.

- Identify the rows which removal makes X_σ rank-deficient. Check that there exists a β'_μ such that the residuals of the observations corresponding to the rows are zero.
- Remove the rows from X_σ (which is now rank-deficient) and make it full-rank again by removing appropriately chosen columns.

- Reinsert the rows in X_σ.

Given that we know how to make a rank-deficient matrix full-rank, that leaves us with the task of identifying rows which make X_σ rank-deficient when they are removed. The author is not aware of any practical method to accomplish this. A brute-force search for such rows will only be feasible for models with very few observations. For example, for our numerical example with 34,511 observations, looking for combinations of merely two rows would bring one to inspect nearly 6×10^8 combinations of rows. Clearly, a more sophisticated approach is needed.

In the next section, we establish mathematical results that help us to develop such an approach.

6 Mathematical Background

We employ the following definitions and notations. A matrix $X \in \mathbb{R}^{n,p}$ is called left-orthogonal if $X^T X = I$ where $I \in \mathbb{R}^{p,p}$ is the identity matrix. The columns of a left-orthogonal matrix form an orthonormal set. The matrix is therefore full-rank and $p \leq n$.

Let $X \in \mathbb{R}^{n,p}$. With $\hat{R}_{i_1,\ldots,i_m} X$ we denote the matrix X with its rows $i_1, \ldots i_m$ removed. $\hat{R}_{i_1,\ldots,i_m} X \in \mathbb{R}^{n-m,p}$ and, obviously, the row-numbers i_1, $\ldots$,i_m must be all different and refer to existing rows in X.

The range (also called the 'column space' or 'image') of a matrix X is denoted as $\mathrm{Im}(X)$. The kernel (also called the 'null-space') of a matrix X is denoted as Ker(X). The rank of a matrix is denoted as $\mathrm{rank}(X)$.

The vector e_i is the i-th standard Euclidean basis vector of $\mathbb{R}^n$. Its vector elements are

$$(e_i)_j = \begin{cases} 0 & j \neq i \\ 1 & j = i \end{cases}.$$

The norm $||v||$ of a vector $v \in \mathbb{R}^n$ is the Euclidean norm

$$||v|| = \sqrt{v_1^2 + \cdots + v_n^2}.$$

We also recall the following result [8]. Let $X \in \mathbb{R}^{n,p}$ with $p \leq n$ be a full-rank matrix. X can be decomposed as $X = QR$ with $Q \in \mathbb{R}^{n,p}$ a left-orthogonal matrix and $R \in \mathbb{R}^{p,p}$ a full-rank upper-triangular matrix. This is called the QR-decomposition of X. Each column vector in Q is defined uniquely up to a sign. Given the signs, R is defined uniquely.

We now develop the mathematical results we need. First, we develop a criterion for the rank-deficiency of a matrix with rows removed

Lemma 1. *Let $X \in \mathbb{R}^{n,p}$ with $p \leq n - m$ be a full-rank matrix. Then, $\hat{R}_{i_1,\ldots,i_m}X$ is rank-deficient iff there exists a linear combination $l = \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}$ with $\lambda_i \in \mathbb{R}$ such that $l \neq 0$ and $l \in \mathrm{Im}(X)$.*

Proof. Let $\hat{R}_{i_1,\ldots,i_m}X$ be rank-deficient. There exists a $v \in \mathbb{R}^p$ with $v \neq 0$ such that $(\hat{R}_{i_1,\ldots,i_m}X)v = 0$. Therefore $Xv = \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}$ with $Xv \neq 0$ because X is full-rank and $v \neq 0$.

Conversely, let $l = \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}$ with $l \neq 0$ and $l \in \mathrm{Im}(X)$. $Xv = l$ for some $v \in \mathbb{R}^p$ with $v \neq 0$. Then $(\hat{R}_{i_1,\ldots,i_m}X)v = \hat{R}_{i_1,\ldots,i_m}(Xv) = 0$ which shows that $\hat{R}_{i_1,\ldots,i_m}X$ is rank-deficient. □

Next, observe that rank-deficiency of $\hat{R}_{i_1,\ldots,i_m}X$ is equivalent to rank-deficiency of $\hat{R}_{i_1,\ldots,i_m}Q$.

Lemma 2. *Let $X \in \mathbb{R}^{n,p}$ with $p \leq n-m$ be a full-rank matrix. Let $X = QR$ be a QR-decomposition of X. Then, $\hat{R}_{i_1,\ldots,i_m}X$ is rank-deficient iff $\hat{R}_{i_1,\ldots,i_m}Q$ is rank-deficient.*

Proof. Let $\hat{R}_{i_1,\ldots,i_m}X$ be rank-deficient. According to Lemma 1, there exists a linear combination $l = \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}$ with $l \neq 0$ and $l \in \mathrm{Im}(X)$. That is, there exists a $v \in \mathbb{R}^p$ such that $Xv = QRv = l$. This shows $l \in \mathrm{Im}(Q)$ and hence $\hat{R}_{i_1,\ldots,i_m}Q$ is rank-deficient.

Conversely, let $\hat{R}_{i_1,\ldots,i_m}Q$ be rank-deficient. There exists a linear combination $l = \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}$ with $l \neq 0$ and $l \in \mathrm{Im}(Q)$. Note that R is full-rank and therefore invertible. Hence $Qv = XR^{-1}v = l$ for some $v \in \mathbb{R}^p$. This shows $l \in \mathrm{Im}(X)$ and hence $\hat{R}_{i_1,\ldots,i_m}X$ is rank-deficient. □

We now develop a criterion to identify the matrix-rows which reduce the matrix rank upon their removal.

Lemma 3. *Let $Q \in \mathbb{R}^{n,p}$ with $p \leq n - m$ be a left-orthogonal matrix. Let $\{i_1, \ldots, i_m\}$ be a set of rows of Q. Let $\hat{R}_{j_1,\ldots,j_k}Q$ be full-rank for any subset $\{j_1, \ldots, j_k\} \subsetneq \{i_1, \ldots, i_m\}$. Then $\hat{R}_{i_1,\ldots,i_m}Q$ is rank-deficient iff the matrix*

$$\begin{pmatrix} q_{i_1}^T q_{i_1} & \cdots & q_{i_1}^T q_{i_m} \\ \vdots & \ddots & \vdots \\ q_{i_m}^T q_{i_1} & \cdots & q_{i_m}^T q_{i_m} \end{pmatrix}$$

has an eigenvalue 1, where $q_i \in \mathbb{R}^p$ is the i-th row-vector of Q.

Proof. Let $\hat{R}_{i_1,\ldots,i_m}Q$ be rank-deficient. According to Lemma 1, there exists a linear combination $l = \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}$ with $l \in \mathrm{Im}(Q)$. That is,

there exists a $v \in \mathbb{R}^p$ such that $Qv = l$. According to the conditions of the lemma, none of the λ_i can be zero. Suppose, as an example, $\lambda_1 = 0$, then $\hat{R}_{i_2,\ldots,i_m} Qv = \lambda_1 e_{i_1} = 0$, showing $\hat{R}_{i_2,\ldots,i_m} Q$ would be rank-deficient, which is not allowed. From $Qv = l$, we know $q_{i_j}^T v = \lambda_j$. Because Q is left-orthogonal, $v = Q^T l = \lambda_1 q_{i_1} + \cdots + \lambda_m q_{i_m}$. Combining results we have $\lambda_j = \lambda_1 q_{i_j}^T q_{i_1} + \cdots + \lambda_m q_{i_j}^T q_{i_m}$. Hence the matrix of the lemma has an eigenvalue 1 with eigenvector $(\lambda_1, \ldots, \lambda_m)^T$.

Conversely, let the matrix of the lemma have an eigenvalue 1. There is an eigenvector $v \in \mathbb{R}^m$ with $v \neq 0$. Let $w = v_1 q_{i_1} + \cdots + v_m q_{i_m}$. Then

$$\begin{aligned} Qw =& \sum_{j=1}^{n} \left(q_j^T (v_1 q_{i_1} + \cdots + v_m q_{i_m}) \right) e_j \\ =& v_1 e_{i_1} + \cdots + v_m e_{i_m} + \sum_{\substack{j=1 \\ j \notin \{i_1,\ldots,i_m\}}}^{n} \left(q_j^T (v_1 q_{i_1} + \cdots + v_m q_{i_m}) \right) e_j. \end{aligned}$$

We define

$$\begin{aligned} w_1 =& v_1 e_{i_1} + \cdots + v_m e_{i_m} \\ w_2 =& \sum_{\substack{j=1 \\ j \notin \{i_1,\ldots,i_m\}}}^{n} \left(q_j^T (v_1 q_{i_1} + \cdots + v_m q_{i_m}) \right) e_j \end{aligned}$$

such that $Qw = w_1 + w_2$. Note that

$$\begin{aligned} w =& Q^T Q w \\ =& Q^T w_1 + Q^T w_2 \\ =& v_1 q_{i_1} + \cdots + v_m q_{i_m} + Q^T w_2 \\ =& w + Q^T w_2. \end{aligned}$$

In other words

$$Q^T w_2 = 0.$$

Because $\mathrm{Ker(Q^T)} = \perp \mathrm{Im}(Q)$ (where $\perp \mathrm{Im}(Q)$ is the set of vectors perpendicular to all vectors in $\mathrm{Im}(Q)$), $w_2 \in \perp \mathrm{Im}(Q)$. Because $(w_1 + w_2) \in \mathrm{Im}(Q)$, $w_1 + w_2$ is perpendicular to w_2, that is $w_1^T w_2 + w_2^T w_2 = 0$. From their definitions, $w_1 \perp w_2$ and therefore we conclude that $w_2^T w_2 = 0$ which means that $w_2 = 0$.

We conclude that $w_2 = 0$ and $Qw = w_1 = v_1 e_{i_1} + \cdots + v_m e_{i_m}$. Because $||w_1|| = ||v|| \neq 0$, $w_1 \neq 0$. Lemma 1 now asserts that $\hat{R}_{i_1,\ldots,i_m} Q$ is rank-deficient. □

For the case of 1 row, the lemma states that the removal of row vector q_i lowers the rank of Q iff $||q_i|| = 1$. This can be understood if one realizes that $q_i^T q_i$ is a diagonal element of the symmetric, idempotent matrix QQ^T. If a symmetric, idempotent matrix has a diagonal element equal to 1, the other elements on the same row (or column) are equal to zero. Hence $q_i^T q_j = 0$ for all $j = 1, \ldots, n$ with $j \neq i$ if $q_i^T q_i = 1$. In other words: the row vector q_i is perpendicular to all other row vectors if $||q_i|| = 1$.

For the sake of completeness, we emphasize that the matrix rank reduces by 1 under the conditions of the lemma.

Lemma 4. *Let $Q \in \mathbb{R}^{n,p}$ with $p \leq n - m$ be a left-orthogonal matrix. If $\hat{R}_{i_1,\ldots,i_m} Q$ is rank-deficient for a set of rows $\{i_1, \ldots i_m\}$ while $\hat{R}_{j_1,\ldots,j_k} Q$ is full-rank for any subset $\{j_1, \ldots, j_k\} \subsetneqq \{i_1, \ldots, i_m\}$, then* $\operatorname{rank}(\hat{R}_{i_1,\ldots,i_m} Q) = p - 1$.

Proof. Remove all row vectors $q_{i_1}, \ldots, q_{i_m}$ from the set $\{q_1, \ldots q_n\}$ except q_{i_m}. According to the conditions of the lemma, the remaining set still contains p linear independent row vectors. Removal of q_{i_m} can reduce the number of linear independent row vectors by at most 1. □

It is noteworthy hat the eigenvalue 1 that appears in Lemma 3, is an extreme value. This notion is developed in the following two lemmas.

Lemma 5. *Let $Q \in \mathbb{R}^{n,p}$ be a left-orthogonal matrix. Then*

$$\sup_{\substack{v \in \mathbb{R}^n \\ ||v||=1}} ||Q^T v|| = 1.$$

The supremum is obtained iff $v \in \operatorname{Im}(QQ^T)$ with $||v|| = 1$.

Proof. Calculate the supremum of $||Q^T v||^2 = v^T Q Q^T v$ under the constraint $||v||^2 = v^T v = 1$ using a Lagrange multiplier

$$\mathcal{L}(v, \lambda) = v^T Q Q^T v + \lambda(1 - v^T v).$$

The supremum is taken by a stationary point of $\mathcal{L}$. Its stationary points are given by

$$\begin{aligned} \nabla_v \mathcal{L} &= 2QQ^T v - 2\lambda v = 0 \\ \nabla_\lambda \mathcal{L} &= 1 - v^T v = 0 \end{aligned}$$

hence the stationary points (v_i, λ_i) are given by all eigenvalues λ_i of QQ^T with v_i an element of the eigenspace of λ_i and $||v_i|| = 1$. At a stationary

point (v_i, λ_i), $||Q^T v_i||^2 = v_i^T Q Q^T v_i = \lambda_i^2 v_i^T v_i = \lambda_i^2$. Hence the supremum is taken by the eigenvalue of QQ^T which has the largest absolute value. Call this eigenvalue $\lambda_{\max}$. Because QQ^T is idempotent, its only possible eigenvalues are 0 and 1, hence $\lambda_{\max} = 1$ unless the corresponding eigenspace is equal to $\{0\}$ [8]. Because the eigenspace corresponding to the eigenvalue 1 of an idempotent matrix is the range of the matrix, the eigenspace is $\mathrm{Im}(QQ^T)$. $\mathrm{Im}(QQ^T)$ would be equal to $\{0\}$ iff all column-vectors of QQ^T would be equal to 0. That would require that $q_i^T q_i = 0$, that is $q_i = 0$, for all row vectors q_i. Hence it would require that $Q = 0$. This can obviously not be the case, which completes the proof. □

Lemma 6. *Let $Q \in \mathbb{R}^{n,p}$ be a left-orthogonal matrix. Let $\{q_{i_1}, \ldots, q_{i_m}\}$ be an arbitrary subset of the set of row vectors $\{q_1, \ldots, q_n\}$ of Q. If λ is an eigenvalue of the matrix*

$$\begin{pmatrix} q_{i_1}^T q_{i_1} & \cdots & q_{i_1}^T q_{i_m} \\ \vdots & \ddots & \vdots \\ q_{i_m}^T q_{i_1} & \cdots & q_{i_m}^T q_{i_m} \end{pmatrix}$$

then $\lambda \leq 1$.

Proof. Let A be the matrix defined in the lemma. If $B \in \mathbb{R}^{m,p}$ is the matrix

$$B = \begin{pmatrix} q_{i_1}^T \\ \vdots \\ q_{i_m}^T \end{pmatrix},$$

$A = BB^T$. $Av = \lambda v$ for an eigenvector v in the eigenspace corresponding to λ. Hence

$$\lambda = \frac{v^T A v}{v^T v} = \frac{||B^T v||^2}{||v||^2}.$$

This shows

$$\lambda \leq \sup_{\substack{v \in \mathbb{R}^m \\ ||v|| = 1}} ||B^T v||^2.$$

Lemma 5 shows that the norm of any linear combination of row vectors of Q (under the constraint that the coefficients v_i satisfy $\sum_i v_i^2 = 1$) is at most 1. Therefore the norm of the linear combination $B^T v$ is also bounded by 1. □

7 Algorithm

We now return to the task of finding a way to identify the rows which lower the rank of X_σ when removed. To make use of lemma 3, the first step is to calculate the QR-decomposition $X_\sigma = QR$. The removal of m rows from Q lowers the rank of Q (and hence of X_σ, c.f. lemma 2) iff the matrix defined in lemma 3 has an eigenvalue 1. Let's call this matrix the 'criterion matrix'. The matrix is positive semi-definite and therefore all eigenvalues λ_i are bounded from below as $\lambda_i \geq 0$. Hence $\sum_i \lambda_i \geq 1$ if the removal of the rows lowers the rank. The sum of the eigenvalues is equal to the trace of the matrix. Hence the trace must be at least 1: $\sum_{j=1}^{m} q_{i_j}^T q_{i_j} \geq 1$. We can limit our search for a combination of m rows to those combinations of rows which satisfy this criterion. Because $0 \leq q_i^T q_i \leq 1$ for all row vectors q_i and $\sum_{i=1}^{n} q_i^T q_i = p$, this will be a useful criterion in many cases.

The algorithm is now as follows.

1. Calculate the QR-decomposition of X_σ

2. Calculate the norm of each row vector of Q and sort the row vectors in order of decreasing norm. Let $\{q_{(1)}, \ldots, q_{(n)}\}$ be the sorted set, that is, $||q_{(1)}|| \geq ||q_{(2)}|| \geq \cdots \geq ||q_{(n)}||$.

3. Look for a single row which lowers the rank of Q upon removal, while there exists a β'_μ such that the residual of the corresponding observation is zero (such a β'_μ normally always exists). If you find one, remove an appropriately chosen column from X_σ such that removing the row does not make X_σ rank-deficient and return to step 1. If you do not find a single row, go to the next step.

4. Look for a combination of two rows which lower the rank of Q upon removal, while there exists a β'_μ such that the residuals of the corresponding observations are zero. If you find such a combination, remove an appropriately chosen column from X_σ such that removing the rows does not make X_σ rank-deficient and return to step 1. If you do not find a combination of two rows, go to the next step.

5. Continue with combinations of three, four, five, etc. rows as long as no combination is found.

6. Once you can not find any combination anymore, the X_σ you have obtained can not lead to a diverging log-likelihood in the way described in section 4.

Looking for a single row in step 3 is easy: if $||q_{(1)}|| < 1$, there is no such row. If $||q_{(1)}|| = 1$, removal of the row-vector will lower the rank of Q and it remains to be checked if there exists a β'_μ such that the residual of the corresponding observation is zero. That will be the case if the corresponding row in X_μ has at least one element unequal to zero.

Looking for a combination of m rows with $m = 2$ in step 4 and $m = 3, 4, 5, \ldots$ in step 5, works as follows. Start with combination $(1, 2, 3, \ldots, m)$ as the current combination in the iteration described in the next paragraph.

Suppose the current combination is $(j_1, j_2, \ldots, j_m)$. If it happens to be that $(j_1, j_2, \ldots, j_m) = (k+1, \ldots, k+m)$ for some k and $||q_{(j_1)}||^2 + \cdots + ||q_{(j_m)}||^2 < 1$, there are no combinations of m rows which lower the rank of Q upon removal. Assume $(j_1, j_2, \ldots, j_m) \neq (k+1, \ldots, k+m)$ for all k. Check if $||q_{(j_1)}||^2 + \cdots + ||q_{(j_m)}||^2 < 1$. If so, continue with the next combination. If not, check if the criterion matrix has an eigenvalue 1 and if there exists a β'_μ such that the residuals of the observations in the combination are zero. If it meets these two requirements, a combination of rows has been identified. If it does not, continue with the next combination.

If the current combination is $(j_1, j_2, \ldots, j_m)$, the next combination is obtained by increasing the counter at the largest possible position with 1, staying within the requirement $1 \leq j_1 < j_2 < \cdots < j_m \leq n$. All following counters, if any, must be reset to a value one larger than the preceding counter. As an example for $m = 2$, starting with $(1, 2)$, the next combination is $(1, 3)$, then $(1, 4)$ etc. until $(1, n)$, after which follow $(2, 3)$, $(2, 4)$ etc. until $(2, n)$, after which follow $(3, 4)$, $(3, 5)$ etc. As another example, for $m = 3$, if the current combination is $(1, 3, 5)$, the next ones will be $(1, 3, 6)$, $(1, 3, 7)$, ... until $(1, 3, n)$ after which come $(1, 4, 5)$, $(1, 4, 6)$, etc. After $(1, n-1, n)$, comes $(2, 3, 4)$.

The iteration for a given m stops when one encounters a combination matching the criteria, or one arrives at a combination $(k+1, k+2, \ldots, k+m)$ for some k with $||q_{(k+1)}||^2 + \cdots + ||q_{(k+m)}||^2 < 1$, or one has reached the combination $(n-m+1, n-m+2, \ldots, n)$.

8 Numerical example, Part 2

We return to the numerical example of section 3. The observations we made, carry some of the traits of the situation described in section 4: components 87 and 94 of β_σ that become unusually large, and a standard deviation that becomes very small for a well-defined set of observations (in this case a set of only one observation, observation 32072). Because components 87 and 94

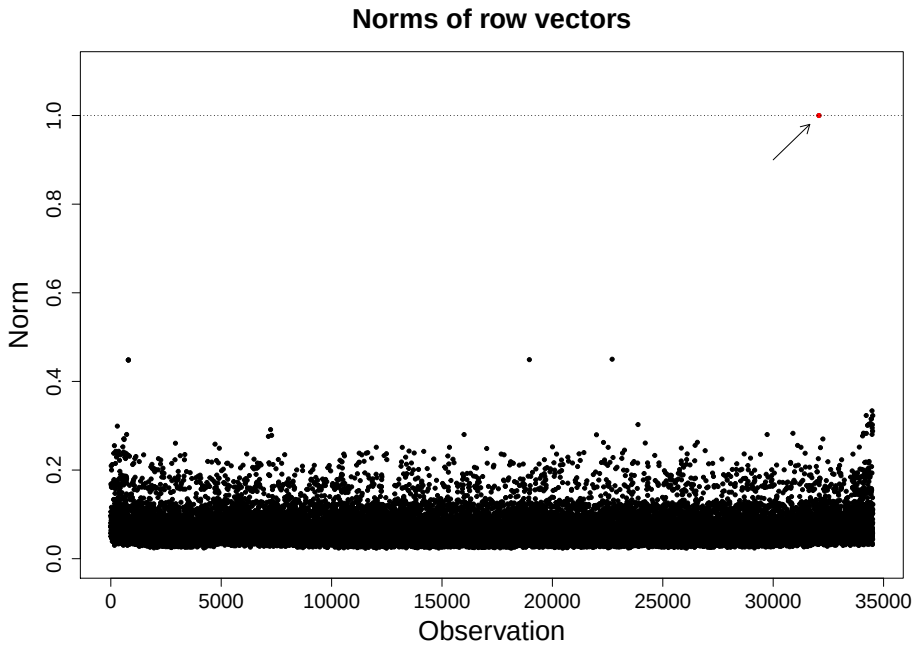

Figure 2: The norms of all the row vectors $||q_i||$ of Q of the QR-decomposition of X. Observations 32,072 (indicated by the arrow and shown in red) is the only one for which the norm is 1.

are nearly of equal size but of opposite sign, this suggests that the vector $e_{87} - e_{94}$ is an element of $\mathrm{Ker}(\hat{\mathrm{R}}_{32072}\mathrm{X})$. That is indeed the case, as can be checked by a direct calculation.

Carrying out the QR decomposition of X and plotting the norm of the row-vectors q_i of Q, shows that $||q_{32072}|| = 1$, see Fig. 2.

Following the algorithm, we remove a degree of freedom from X such that $\hat{R}_{32072}X$ is no longer rank-deficient. This degree of freedom is a matrix column in which 157 matrix elements have a value 1 and the other elements are 0. Call the resulting matrix $X_1 \in \mathbb{R}^{34511,187}$.

Continuing the algorithm, we calculate the QR-decomposition of X_1. This time, $||q_{(1)}||^2 + \cdots + ||q_{(4)}||^2 < 1$ hence there is no single row and no combination of 2, 3 or 4 rows which makes X_1 rank-deficient if removed. However, $||q_{(1)}||^2 + \cdots + ||q_{(5)}||^2 > 1$ and the criterion matrix constructed from the corresponding observations has an eigenvalue 1 while there is also a β'_μ such that the residuals of those observations are zero. Hence we found a combination of 5 rows meeting both criteria. They are pointed out in Fig. 3

The algorithm removes one degree of freedom from X_1 to make sure that taking out the 5 rows doesn't give a rank-deficient matrix. This degree of freedom is a column in which 34,022 matrix elements are equal to 1 and the other ones are 0. Call the resulting matrix $X_2 \in \mathbb{R}^{34511,186}$. Calculating the QR-decomposition of this matrix, we now find that $||q_{(1)}||^2 + \cdots + ||q_{(10)}||^2 < 1$ which means that up to and including combinations of 10 rows, there are no combinations which lower the rank of Q if they are removed. $||q_{(1)}||^2 + \cdots + ||q_{(11)}||^2 > 1$ and the algorithm checks 7,122,025 combinations of 11 rows before finding that $||q_{(6)}||^2 + ||q_{(7)}||^2 + \cdots + ||q_{(16)}||^2 < 1$ and hence also no combination of 11 rows qualifies. Note that the 7,122,025 combinations are a tiny fraction of the order 10^{-36} of all possible combinations of 11 rows from the 34,511 matrix rows. However, the observations corresponding to $q_{(1)}, \ldots, q_{(12)}$ qualify: the corresponding criterion matrix has an eigenvalue 1 and there exists a β'_μ such that the residuals are zero. The 12 observations are pointed out in Fig. 4. They lead to $X_3 \in \mathbb{R}^{34511,185}$.

Continuing the algorithm, checking combinations of 13 rows took longer than we were willing to wait. Keeping in mind that it took many hours to check combinations of 11 rows, we terminated the search for a combination of 13 rows after about 3 hours. This shows that looking for combinations with many rows can still be too time-consuming, even with the filtering of combinations that we have developed. For example, we did not find the combination of 21 rows which, upon removal, lowers the rank of X (X has a column with 21 elements equal to 1 and all other elements equal to 0).

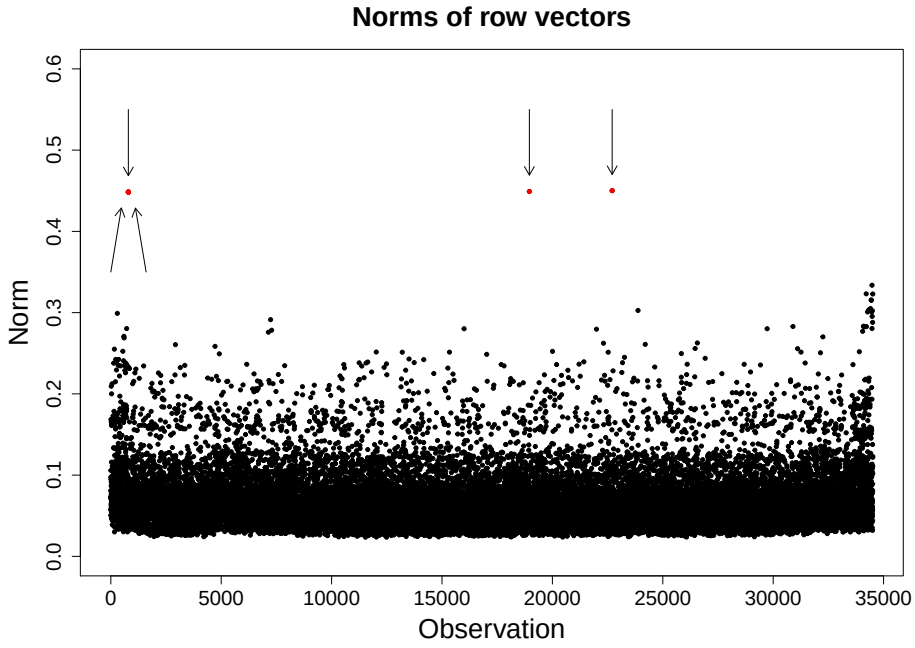

Figure 3: The norms of all the row vectors $||q_i||$ of Q of the QR-decomposition of X_1. The 5 row vectors which lower the rank of Q when removed, are indicated by arrows and shown in red. The leftmost 3 of these 5 are so close together that they are indistinguishable at the scale of the figure.

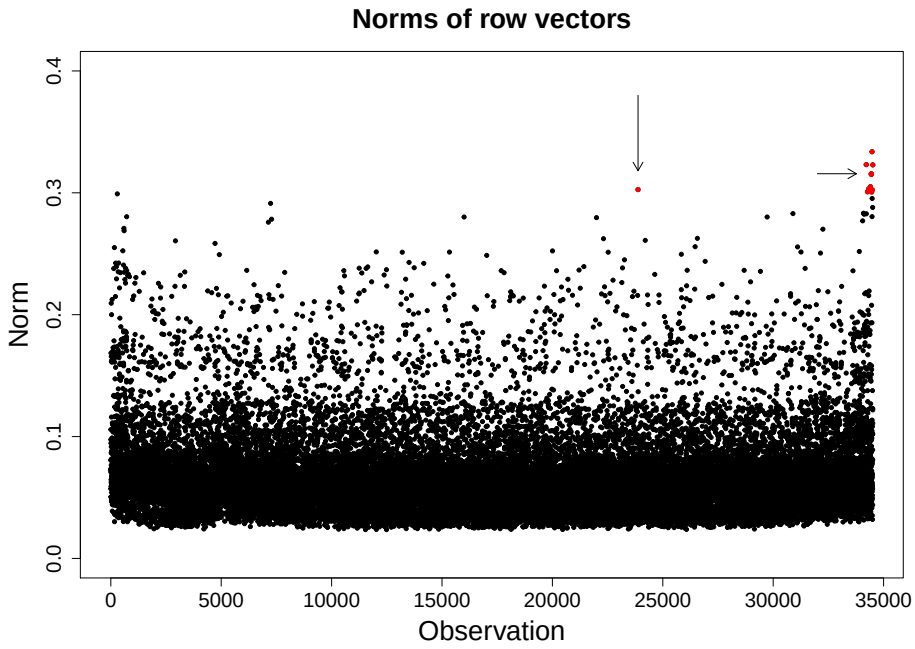

Figure 4: The norms of all the row vectors $||q_i||$ of Q of the QR-decomposition of X_2. The 12 row vectors which lower the rank of Q when removed, are indicated by arrows and shown in red. The right arrow points to a cluster of 11 points.

Nevertheless, the algorithm is a vast improvement compared to a brute-force search.

If we once again try to fit the model, now taking $X_\mu = X$ and $X_\sigma = X_1$ both the 'lmvar' and the 'dglm' package return a fit without errors or warnings. The two fits yield nearly identical estimates for β_μ and β_σ (apart from a factor 2 in the latter, because β_σ is defined as $\log \sigma^2 = X_\sigma \beta_\sigma$ in the 'dglm' package). The relative difference is at most 0.8% for the elements in β_μ and at most 4% for the elements in β_σ. There are no betas that stand out as remarkably large in absolute value. The 'lmvar' fit gives $\sigma_{32072} = 0.92$, very much in line with the other observations.

It is a little surprising that both packages carry out the fit without problems for $X_\sigma = X_1$. The log-likelihood is still unbounded for the triplet y, $X_\mu = X$ and $X_\sigma = X_1$ and ideally, an attempt to calculate the maximimum-likelihood estimators for β_μ and β_σ must fail. Apparently, the iterative methods used by the packages converge to a local maximum of the log-likelihood. It is intuitively clear why it can be difficult to discover a diverging log-likelihood: as $L \to \infty$ in (11), the residuals $y_i - \mu_i$ in the set S_1 must be kept at the order of σ_i, otherwise the log-likelihood will not be large. Because at least one of the σ_i will go to zero, the region in β_μ-space in which the divergence can be observed, shrinks to zero for some of the components of β_μ as L becomes larger.

9 Computational Complexity

Our algorithm addresses the problem: given a full-rank matrix of size $n \times p$ with $n \geq p$, find a set of rows which is as small as possible and lowers the rank of the matrix when the rows are removed. The numerical example shows that our algorithm is good at finding a set of m rows when m is small, but not so when m is large. Given that the average $||q_i||^2$ of the norms in Figure 2 is 0.005, one expects that any set of about 200 rows has a good change of meeting the criterion that the sum of the norms squared must be at least 1. Indeed, taking random samples of sets of rows shows that if the set has 250 rows or more, the probability of meeting the criterion is virtually 1. For combinations of that many rows, nearly each combination meets the criterion and the algorithm is not more efficient than a brute-force search through all possible combinations. In practice, we already hit our computational limits at a set size of 13 rows. Although the probability that a set of 13 rows meets the criterion is only 0.0004 (calculated by taking random samples of

13 rows), the total number of combinations $\binom{34511}{13} = 2 \times 10^{49}$ is such that the number of combinations meeting the criterion is prohibitively large.

A problem of a similar nature, determining the spark of a matrix, has been proven to be NP-hard [14, 18] and the same may be true for our problem. Our algorithm is not much better than checking all possible combinations of size m when m is large enough. This appears to be a trait of NP-hard problems. A formal treatment of the complexity-class of our problem is non-trivial and beyond our scope.

10 Alternative Solutions

The problem of an unbounded likelihood happens in other systems as well. Examples are models with a three-parameter probability-density function (p.d.f.) in which one of the parameters is the boundary of the half-space on which the p.d.f. is defined, and models of Gaussian mixtures, as discussed by [10]. They mention two alternatives to maximum-likelihood estimation: the maximum product of spacings (MPS) method [13, 2, 3, 19] and an approach in which the likelihood is replaced by a 'discretized' version.

In the context of the LMVAR model, the MPS method works as follows: given a β_μ and β_σ, calculate the z-scores $z_i = (y_i - \mu_i)/\sigma_i$ and order them in increasing magnitude $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$. Calculate the $n+1$ distances D_i, defined as $D_i = \Phi(z_{(i)}) - \Phi(z_{(i-1)})$ with Φ the cumulative distribution function of the standard normal distribution, $z_{(0)} = -\infty$ and $z_{(n+1)} = \infty$. Estimate β_μ and β_σ as the values which maximize the logarithm of the geometric mean of the spacings $D_1, \ldots, D_{n+1}$.

We tried the method on a generated data set of 5000 observations with few degrees of freedom. It works reasonably well with $p_\sigma = 1$, a classical linear model. However, taking for example $p_\mu = 2$ and $p_\sigma = 2$ we found it hard to obtain acceptable estimates for especially β_σ. Using various optimization algorithms available in the base-R function 'optim()', we obtain estimates $\beta_\sigma = (-0.77, 0.12)$, $(-0.56, -0.09)$ and $(-0.75, 0.34)$ where the exact value is $\beta_\sigma = (-1.1, 0.8)$. The maximum-likelihood estimator is $\beta_\sigma = (-1.11, 0.79)$. This experiment convinced us that the MPS method is not a straightforward alternative to likelihood maximization for an LMVAR model. It might be useful in the case of a diverging likelihood, but it does require careful study.

The second alternative mentioned in [10], is to replace the likelihood $\mathcal{L}$

by

$$\mathcal{L}'(\beta_\mu, \beta_\sigma; y, \Delta) = \int_{y_1-\frac{1}{2}\Delta_1}^{y_1+\frac{1}{2}\Delta_1} \mathrm{d}y'_1 \cdots \int_{y_n-\frac{1}{2}\Delta_n}^{y_n+\frac{1}{2}\Delta_n} \mathrm{d}y'_n \; \mathcal{L}(\beta_\mu, \beta_\sigma; y').$$

$\mathcal{L}'$ no longer represents a probability density but rather the probability that each observation i is in the range $(y_i - \frac{1}{2}\Delta_i, y_i + \frac{1}{2}\Delta_i)$. It will therefore never diverge. The choice of Δ influences the resulting maximum-likelihood estimates. For example, once most of the probability mass for observation i is in the interval $(y_i - \frac{1}{2}\Delta_i, y_i + \frac{1}{2}\Delta_i)$, the contribution of this observation to $\log \mathcal{L}'$ can not increase much anymore and it becomes uninteresting to improve the model for this observation. Hence the effect of the Δ_i on the resulting estimators must be studied in some detail.

As a third alternative, one could add a penalty term to the log-likelihood. This term must prevent the likelihood to grow unbounded. In our case, one can imagine a term like $-\lambda \sum_{i=1}^{n} (1/\sigma_i)$ with $\lambda > 0$. An extreme case of such a penalty term is to calculate the maximum-likelihood estimates under the constraint that all σ_i must have a value larger than a minimum value.

This can be tested with the 'lmvar' package which allows the application of this constraint. We constrained $\sigma_i > \sigma_{\min}$ for all observations i and three values of $\sigma_{\min}$. With $\sigma_{\min} = 0.001$, 'lmvar' exits with a warning that the iteration limit has been exceeded and that the likelihood is not at a (local or global) maximum. The first warning means that the iteration did not converge to a solution within 200 steps. If we increase the iteration limit to 400 steps, the iteration converges but the resulting solution is still not at a maximum of the likelihood. If we accept this (because we know that the maximum is unbounded), we find that the fit gives $\sigma_{32072} = 0.001001$, which suggests that the fit makes σ_{32072} as small as possible while converging as close as possible to the unbounded likelihood situation.

Taking $\sigma_{\min} = 0.1$, again makes 'lmvar' exit with the warning that the likelihood is not at a maximum. The fit gives $\sigma_{32072} = 0.101$. Taking $\sigma_{\min} = 0.5$ and setting the iteration limit to 400 steps, 'lmvar' exists with the same warning. The fit now gives $\sigma_{32072} = 0.50005$.

We conclude that setting a lower bound on the σ_i, results in a solution that is as close to the unbounded likelihood situation as allowed by the bound. The choice of the bound has a significant influence on the resulting fit.

Closely related is a Baysian approach in which a conditional probability $P(\beta_\mu, \beta_\sigma|Y)$ is constructed and maximized over β_μ and β_σ:

$$P(\beta_\mu, \beta_\sigma|Y) = \frac{P(Y|\beta_\mu, \beta_\sigma)P(\beta_\mu, \beta_\sigma)}{P(Y)}.$$

with $P(Y|\beta_\mu, \beta_\sigma)$ the multivariate Gaussian distribution (1) and the prior $P(\beta_\mu, \beta_\sigma)$ chosen suitably. Taking logarithms, the relation with the maximization of the log-likelihood with a penalty term, can be seen:

$$\log P(\beta_\mu, \beta_\sigma|Y) = \log \mathcal{L} + \log P(\beta_\mu, \beta_\sigma) - \log P(Y)$$

with $\log P(\beta_\mu, \beta_\sigma)$ a penalty on improbable values of the parameter vectors β_μ and β_σ and $\log P(Y)$ independent of the parameter vectors.

Finally, shrinkage methods such as ridge regression or lasso, are yet another form of maximizing the log-likelihood under a penalty term. They can suppress the unbounded growth of the log-likelihood and are a potential alternative to the approach taken in this paper. As far as we know, there are no R-packages supporting a shrinkage method for the LMVAR model. Because an efficient implementation of shrinkage methods is non-trivial [5, 20], such methods are not readily available for the this model.

11 Conclusions

We have shown that for particular model matrices X_μ and X_σ, the log-likelihood of an LMVAR model is unbounded, which in turn can result in a failing model-fit. The case we studied, is when the removal of not-too-many observations from X_σ makes the matrix rank-deficient, while the residuals of the removed observations can be made zero and the product of the estimated standard deviations of the removed observations is not equal to 1. In this case, the bound on the log-likelihood can be restored by removing specific degrees of freedom from X_σ. We have developed an algorithm to identify these degrees of freedom. These degrees are related to the combinations of matrix-rows that make X_σ rank-deficient when removed. Our algorithm efficiently finds these combinations as long as they contain relatively few rows. Looking for combinations with many rows, our algorithm is not more efficient than a brute-force search.

Removing degrees of freedom associated with combinations of a few rows, might already produce a successful fit, even though not all troublesome combinations have been found and the likelihood is still unbounded. Our numerical example shows this. Such a fit must be to a local maximum of the likelihood, as there is no global maximum. It is a question whether the properties of this fit are satisfactory, and whether there are other local maxima that can give alternative estimators.

We did not discuss whether an unbounded likelihood must be considered a finite-size effect, which in some sense vanishes when the number of obser-

vations grows. Adding rows to X_μ that are identical to one of the existing rows, and likewise for X_σ, as a way of increasing the number of observations will not help. The added rows increase the size of set S_2 but do not change the null-space of the matrix $X_{\sigma 2}$. Also, a vector β'_μ which makes the residuals in the set S_1 vanish, will continue to do so.

Even if an unbounded likelihood is in some sense a finite-size effect, it can still hamper the fit of an LMVAR model in not-so-small models. In our numerical example, it was one observation of the 34,511 that prevented a fit.

Another question is whether heteroscedastic models such as the heteroscedastic genereralizations of GLIM models [17, 16, 15] exhibit a similar mechanism.

Acknowledgments

The author thanks E. Cator and E.W. van Zwet for discussions and encouragement. Many thanks to M. Dijkstra and S. Draijer for proofreading the manuscript.

References

[1] Murray Aitkin. "Modelling Variance Heterogeneity in Normal Regression Using GLIM". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 36.3 (1987), pp. 332–339. ISSN: 00359254, 14679876.

[2] Stanislav Anatolyev and Grigory Kosenok. "AN ALTERNATIVE TO MAXIMUM LIKELIHOOD BASED ON SPACINGS". In: *Econometric Theory* 21.2 (2005), pp. 472–476. DOI: 10.1017/S0266466605050255.

[3] R. C. H. Cheng and N. A. K. Amin. "Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 45.3 (1983), pp. 394–403. ISSN: 00359246. URL: http://www.jstor.org/stable/2345411.

[4] Peter K Dunn and Gordon K Smyth. *dglm: Double Generalized Linear Models*. R package version 1.8.3. 2016. URL: https://CRAN.R-project.org/package=dglm.

[5] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software, Articles* 33.1 (2010), pp. 1–22. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/v033/i01.

[6] Harvey Goldstein. *Heteroscedasticity and Complex Variation*. John Wiley & Sons, Ltd, 2014. ISBN: 9781118445112. DOI: 10.1002/9781118445112.stat06249. URL: http://dx.doi.org/10.1002/9781118445112.stat06249.

[7] G. Z. Heller et al. "Mean and Dispersion Modeling for Policy Claims Costs". In: *Scandinavian Actuarial Journal* 2007 (Dec. 2007), pp. 281–292. DOI: 10.1080/03461230701553983.

[8] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2018. ISBN: 978-0-521-83940-2.

[9] Piet de Jong and Gillian Z. Heller. *Generalized Linear Models for Insurance Data*. International Series on Actuarial Science. Cambridge University Press, 2008. ISBN: 978-0-521-87914-9.

[10] Shiyao Liu, Huaiqing Wu, and William Q. Meeker. "Understanding and Addressing the Unbounded "Likelihood" Problem". In: *The American Statistician* 69.3 (2015), pp. 191–200. DOI: 10.1080/00031305.2014.1003968. eprint: https://doi.org/10.1080/00031305.2014.1003968. URL: https://doi.org/10.1080/00031305.2014.1003968.

[11] Posthuma Partners. *lmvar: Linear Regression with Non-Constant Variances*. Posthuma Partners. Gouda, the Netherlands, 2018. URL: https://CRAN.R-project.org/package=lmvar.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[13] Bo Ranneby. "The Maximum Spacing Method. An Estimation Method Related to the Maximum Likelihood Method". In: *Scandinavian Journal of Statistics* 11.2 (1984), pp. 93–112. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4615946.

[14] Michael Sipser. *Introduction to the theory of computation*. Cengage Learning, 2012. ISBN: 978-1-133-18779-0.

[15] Gordon K. Smyth. "Generalized Linear Models with Varying Dispersion". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 51.1 (1989), pp. 47–60. ISSN: 00359246.

[16] Gordon K. Smyth and Arūnas P. Verbyla. "Adjusted Likelihood Methods for Modelling Dispersion in Generalized Linear Models". In: *Environmetrics* 10 (1999), pp. 695–709.

[17] Gordon K. Smyth and Arūnas P. Verbyla. "Double Generalized Linear Models: Approximate REML and Diagnostics". In: *Proceedings of the 14th International Workshop on Statistical Modelling* (July 19–23, 1999). Ed. by H. Friedl, A. Berghold, and G. Kauermann. Technical University, Graz, Austria. Graz, Austria, 1999, pp. 66–80.

[18] A. M. Tillmann and M. E. Pfetsch. "The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing". In: *IEEE Transactions on Information Theory* 60.2 (Feb. 2014), pp. 1248–1259. ISSN: 0018-9448. DOI: 10.1109/TIT.2013.2290112.

[19] T. S. T. Wong and W. K. Li. "A note on the estimation of extreme value distributions using maximum product of spacings". In: *Time Series and Related Topics*. Ed. by Hwai-Chung Ho, Ching-Kang Ing, and Tze Leung Lai. Vol. 52. Lecture Notes–Monograph Series. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2006, pp. 272–283. DOI: 10.1214/074921706000001102. URL: https://doi.org/10.1214/074921706000001102.

[20] Y. Zheng and P. Breheny. *The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R*. Jan. 20, 2017. URL: https://arxiv.org/abs/1701.05936.