ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Chair for Mathematical Information Science
Prof. Dr. H. Bölcskei

Master Thesis in Information Technology and Electrical Engineering

Spring Semester 2022

Genc Kqiku

On the Existence of Maximum Likelihood Estimate in High-dimensional Misspecified Logistic Regression

Supervisor: Weigutian Ou

March 2022

Abstract

High-dimensional generalized linear models have attracted growing attention, especially in the determination of the existence of maximum likelihood estimate (MLE). A paper from Candès and Sur studies this question for logistic regression with Gaussian covariates, and shows that the asymptotic probability of existence of the MLE goes from 1 to 0 as the limit of the parametrization ratio crosses a certain value. We present a proof of this phenomenon, based on convex geometry and linear separability, and we extend this result to misspecified models with latent variable, by imposing assumptions on the correlation between the observable and unobservable variables, and on the magnitude of the unknown parameters.

Acknowledgments

I would like to thank Prof. Dr. H. Bölcskei for his supervision, as well as Weigutian Ou for his precious guidance all along this project. I am grateful for his availability, his willingness to help, and for all the insightful meetings that we had.

Contents

Notation

- We assume a classic probability space $(\Omega, \mathcal{F}, P)$, with probability measure P.
- For any vector $\mathbf{v} \in \mathbb{R}^n$, $\|\cdot\|$ will denote the Euclidean norm, i.e. $\|\mathbf{v}\| = \sqrt{\sum_{i=1}^{n}(\mathbf{v}_i)^2}$.
- For any matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, $\|\cdot\|$ will denote the Frobenius norm, i.e. $\|\mathbf{M}\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(\mathbf{M}_{ij})^2}$.

Chapter 1

Introduction

When dealing with a classification task, generalized linear model (GLM), and in particular the logistic model, has established itself as a very popular option. The binary response is modeled as

$$P[Y_i = 1|\boldsymbol{X}_i] = 1 - P[Y_i = -1|\boldsymbol{X}_i] = \sigma(\boldsymbol{X}_i^T\beta)$$

for response variables $\{Y_i\}_{i=1}^n$ taking value in $\{-1, 1\}$, covariates $\{\boldsymbol{X}_i\}_{i=1}^n$ with values in $\mathbb{R}^p$, and $\beta \in \mathbb{R}^p$ the unknown parameter. The link function σ is the sigmoid

$$\sigma : \mathbb{R} \to [0, 1], \quad t \mapsto \frac{1}{1 + e^{-t}}.$$

The ratio of the number of parameters p over the sample size n and how it affects the estimator and its performance is of big interest in machine learning. Recent works have showed that some models can increase their performance when one increases the complexity of the model even after reaching interpolation, i.e. zero training error. Indeed, it can happen that increasing the number of parameters makes the test error decrease also in an overparametrized regime, in which $p > n$. The test error curve then displays a double descent, and this phenomenon has been studied notably for neural networks [1, 2], linear models [3–5], GLMs [6, 7].

In order to prove results such as a double descent curve, one needs to have closed-form solution for the estimator, that we choose as the empirical risk minimizer (ERM), i.e.

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \, \mathbf{x}_i^T \beta), \tag{1.1}$$

with ℓ the logistic loss function, defined by $\ell(t) = \log(1+e^{-t})$, and $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ our data set, realizations of the $(\boldsymbol{X}_i, Y_i)$. It is direct to see that β defined as in (1.1) corresponds to the maximum likelihood estimator (MLE) of the problem, i.e.

$$\hat{\beta} = \arg\max_{\beta} \; \log\left(\prod_{i=1}^{n} P_\beta[Y_i = y_i]\right).$$

However, the MLE does not always exist. And it is in the center of our interest throughout this work to determine when the MLE does exist and when it does not.

A well known fact by statisticians, that is crucial for us, is the following. The MLE does not exist if and only if the data set $(\mathbf{x}_i, y_i)$ is linearly separable, i.e. there exist $b \neq 0$ and $\mathbf{w} \in \mathbb{R}^p$ a non-zero vector with

$$y_i \left(b + \mathbf{x}_i^T \mathbf{w}\right) \geq 0 \quad \text{for all } 1 \leq i \leq n.$$

The latter is equivalent to saying that the hyperplane with direction $\mathbf{w}$ and offset b does classify correctly all the n samples of the data set $(\mathbf{x}_i, y_i)$, meaning that all the points $\mathbf{x}_i$ with $y_i = 1$ lie on one side of it, and all the points $\mathbf{x}_i$ with $y_i = -1$ on the other. Note that our definition of separability does allow for equality $y_i \left(b + \mathbf{x}_i^T \mathbf{w}\right) = 0$, i.e. for points $\mathbf{x}_i$ lying on the separating hyperplane.

Of course, this equivalence for the existence of the MLE does not say it all by itself. Indeed, one can say if the data set is separable or not only when one has observed realisations of the random variables. However, one may want to be able to say something about the existence of the MLE by only knowing the distribution of the data. A couple of papers have shown that there is a phase transition in the existence of the MLE for the logistic regression, controlled by the ratio p/n. Indeed, it has been shown that when n and p grow to infinity at similar speeds, with $p/n \to \kappa$ a constant, there exists a threshold h such that

$$\begin{aligned} P[\text{MLE exists}] &\to 1 \text{ if } \kappa < h, \text{ and} \\ P[\text{MLE exists}] &\to 0 \text{ if } \kappa > h. \end{aligned}$$

This has been proven in [8] in the setup above, when the explanatory variables are independent and identically distributed (i.i.d.) copies of a random vector with centered Gaussian distribution. We add a layer of complexity to this case, by considering a misspecified model. The phase transition has also been studied in other settings of the logistic regression, notably in [7, 9].

1.1 Model considered

Let

$$\boldsymbol{Y} \sim \mathrm{Rad}(\sigma(\gamma + \boldsymbol{X}\beta + \boldsymbol{Z}\xi)), \tag{1.2}$$

where Rad denotes the Rademacher distribution, i.e.

$$P[Y_i = 1] = 1 - P[Y_i = -1] = \sigma(\gamma + \boldsymbol{X}_i^T \beta + \boldsymbol{Z}_i^T \xi) \quad \forall\, 1 \leq i \leq n$$

where $\boldsymbol{X}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, $i = 1, \ldots, n$ are observed covariates, $\boldsymbol{Z}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $i = 1, \ldots, n$ are unobserved covariates, $\beta \in \mathbb{R}^p$, $\xi \in \mathbb{R}^d, \gamma \in \mathbb{R}^n$ the model parameters, and the sigmoid function σ is applied to vectors component-wise. The $\boldsymbol{Z}_i$'s being unobserved means that they affect Y_i, but we do not get realizations $\mathbf{z}_i$'s of them, so they cannot be used to build the estimator. The covariance matrix Σ of $\boldsymbol{X}$ is invertible, otherwise arbitrary. It has been shown in [7], using an extension of the Convex Gaussian Min-Max Theorem, that there is a phase transition in that misspecified model if $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent. We will present in Chapter 2 another proof of this case, based on [8].

This will serve us as a base to show phase transition for a couple a specific cases where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are **not** independent.

We will assume the following for the entirety of this work.

Assumption 1.1.1. *Throughout this work we will be interested in the asymptotic framework where $n, p, d \to \infty$, with $\frac{d}{n} \to \delta \geq 1$, $\frac{p}{n} \to \kappa$, where δ, κ are constants. That way we get a sequence of data inputs and response, with diverging dimensions. For each n, p, d fixed, $\boldsymbol{X}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\boldsymbol{U}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ for each $i = 1, \ldots, n$, and $\boldsymbol{X}$ and $\boldsymbol{U}$ are independent. Moreover, we assume*

$$\mathbb{V}\mathrm{ar}(\mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}_i^T\mathbf{B}^T\xi) \to \nu^2$$

and

$$\sqrt{\mathbb{E}[(\gamma + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}_i^T\mathbf{B}^T\xi)^2]} \to \sqrt{\gamma^2 + \nu^2}$$

That way we are ensured that $\gamma + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}_i^T\mathbf{B}^T\xi$ does not explode and hence does not yield trivial probabilities for Y_i being equal to 1 or -1. Also, a good estimator (β, ξ) yields now to a good estimator of a non-trivial probability of Y_i being equal to 1.

Chapter 2 presents a proof of the existence of a phase transition in the case where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent. The proof is inspired by [8]. It describes the separability of the data set as a non-trivial intersection of a vector space and a convex cone, and uses convex geometry to describe the asymptotic behaviour of the probability of this non-trivial intersection happening. We then consider a more complex model in which $\mathbf{Z}$ is a latent variable. We present three specific cases in which our phase transition result still holds. To our knowledge, the phase transition in those three cases has not been covered in past literature.

Chapter 2

Case of independent features

2.1 Main result

We assume the set-up above and treat first the case where the observable variable $\boldsymbol{X}$ and the unobservable variable $\boldsymbol{Z}$ are independent. We show that there is a phase transition in the existence of the MLE and characterize its threshold. We will show that the following result holds with different assumptions throughout this work.

Theorem 2.1.1. *Let $\boldsymbol{Y}$ distributed as in (1.2) and assume that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent. Let $Q \sim \mathcal{N}(0,1)$, independent from $\boldsymbol{Y}$ and $\boldsymbol{V}$. Define*

$$h = \min_{t_1,t_2 \in \mathbb{R}} \mathbb{E}\left[(t_1 Y_1 + t_2 V_1 - Q)_+^2\right] \tag{2.1}$$

Then, we have

$$\begin{aligned} &\text{if } \kappa > h, \text{ then } \lim_{n,p,d\to\infty} P[\text{MLE exists}] = 0 \\ &\text{if } \kappa < h, \text{ then } \lim_{n,p,d\to\infty} P[\text{MLE exists}] = 1. \end{aligned}$$

Remark 2.1.1. *One can see that the value of h depends on ν^2, which expresses the strength of the signal in $\boldsymbol{X}$ and $\boldsymbol{Z}$.*

We present here a proof following the steps of [8], which is based on convex geometry theory. Hence this whole chapter is inspired by the latter paper.

2.2 Some preliminary results

Let us first show two results which simplify our set-up and allow us to have stronger assumptions. Indeed, we show that we can in fact assume that the observable features are independent, i.e. $\Sigma = \mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

Lemma 2.2.1. *Let Y distributed as in (1.2). Recall that $\boldsymbol{X}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, $i = 1, \dots, n$, with Σ invertible. Then, there exists a separating hyperplane based on the features $\boldsymbol{X}$ if and only if there exists a separating hyperplane based on features $\tilde{\boldsymbol{X}}$, with $\tilde{\boldsymbol{X}}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $i = 1, \dots, n$.*

Proof. We can write for $i \in \{1, \dots n\}$

$$\boldsymbol{X}_i = \Sigma^{\frac{1}{2}} \tilde{\boldsymbol{X}}_i$$

where $\Sigma^{\frac{1}{2}}$ is such that $\Sigma^{\frac{1}{2}} \cdot \Sigma^{\frac{1}{2}} = \Sigma$. Hence, there exists a separating hyperplane with probability p if and only if there exist $b \neq 0$ and $w \in \mathbb{R}^p$, $w \neq \mathbf{0}$ such that

$$P\left[Y_i \left(b + \tilde{\boldsymbol{X}}_i^T \Sigma^{\frac{1}{2}} \mathbf{w}\right) \geq 0 \text{ for all } i\right] = p$$

It is easy to see that the the latter is equivalent to the existence of $\tilde{b} \neq 0$ and $\tilde{\mathbf{w}} \in \mathbb{R}^p$, $\tilde{\mathbf{w}} \neq \mathbf{0}$ such that

$$P\left[Y_i \left(\tilde{b} + \tilde{\boldsymbol{X}}_i^T \Sigma^{\frac{1}{2}} \tilde{\mathbf{w}}\right) \geq 0 \text{ for all } i\right] = p$$

as wanted. Indeed, we can simply multiply the directing vector by $\Sigma^{\frac{1}{2}}$ (resp. $(\Sigma^{\frac{1}{2}})^{-1}$) to go from one set of features to another, and keep the same offset $b = \tilde{b}$. □

Now, we show that in our case of independent features with Gaussian distribution, the model can be reduced to a model with one observable feature.

Lemma 2.2.2. *Let Y distributed as in (1.2), with $\boldsymbol{X}$ and $\boldsymbol{Z}$ independent. Denote $\boldsymbol{X}_1$ the first column of $\boldsymbol{X}$. Then there exists $\tilde{\beta}, \tilde{\xi} \in \mathbb{R}$ and standard Gaussian covariates $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1, \dots, \tilde{\boldsymbol{X}}_n)^T$, $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{Z}}_1, \dots, \tilde{\boldsymbol{Z}}_n)^T$ such that for*

$$\tilde{\boldsymbol{Y}} \sim \mathrm{Rad}(\tilde{\gamma} + \tilde{\beta}\tilde{\boldsymbol{X}}_1 + \tilde{\xi}\tilde{\boldsymbol{Z}}_1)$$

we have

$$\boldsymbol{Y} \overset{d}{=} \tilde{\boldsymbol{Y}}$$

and $(\boldsymbol{Y}, \boldsymbol{X})$ is linearly separable if and only if $(\boldsymbol{Y}, \tilde{\boldsymbol{X}})$ is linearly separable.

Proof. Let $i \in \{1, \dots, n\}$. We write

$$\beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \xi_1 Z_{i,1} + \cdots + \xi_d Z_{i,d} = \\ \beta_1 X_1 + \cdots + \beta_p X_p + \xi_1 Z_1 + \cdots + \xi_d Z_d$$

for clarity. Hence we have, for observation i

$$Y \sim \mathrm{Rad}(\gamma + \beta_1 X_1 + \cdots + \beta_p X_p + \xi_1 Z_1 + \cdots + \xi_d Z_d)$$

Now, since $X_1, \dots, X_p$ are independent, $(X_1, \dots, X_p)$ is a Gaussian random vector. The Gaussian distribution is rotationally invariant, meaning that

$$\mathbf{R}(X_1, \dots, X_p) \overset{d}{=} (\tilde{X}_1, \dots, \tilde{X}_p)$$

for any orthogonal matrix $\mathbf{R}$ and $\tilde{X}_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. Let

$$v = \left(\beta_1^2 + \cdots + \beta_p^2\right)^{\frac{1}{2}}.$$

We construct a matrix $\mathbf{R}_v$ with first row

$$r_1 = \frac{1}{v}\,(\beta_1, \cdots, \beta_p,)$$

and with the $p-1$ other rows chosen such that $\mathbf{R}_v^T\mathbf{R}_v = I_p$. The rotational invariance implies then $\mathbf{R}_v(X_1,\ldots,X_p)^T \stackrel{d}{=} (\tilde{X}_1,\ldots,\tilde{X}_p)^T$, which gives for in the first row

$$\frac{1}{v}(\beta_1 X_1 + \cdots + \beta_p X_p) \stackrel{d}{=} \tilde{X}_1,$$

or

$$\beta_1 X_1 + \cdots + \beta_p X_p \stackrel{d}{=} v\,\tilde{X}_1.$$

Using the rotational invariance of $(Z_1,\ldots,Z_p)$, we obtain by using the same argument that

$$\xi_1 Z_1 + \cdots + \xi_p Z_d \stackrel{d}{=} w\,\tilde{Z}_1$$

for $w = (\xi_1^2 + \cdots + \xi_d^2)^{\frac{1}{2}}$ and $\tilde{Z}_i \overset{iid}{\sim} \mathcal{N}(0,1)$. Since i was arbitrary we indeed found that $\boldsymbol{Y} \stackrel{d}{=} \tilde{\boldsymbol{Y}}$, if we define

$$\tilde{\boldsymbol{Y}} \sim \mathrm{Rad}(\gamma + v\tilde{\boldsymbol{X}}_1 + w\tilde{\boldsymbol{Z}}_1).$$

Finally, the fact that $(\boldsymbol{Y},\boldsymbol{X})$ is linearly separable if and only if $(\boldsymbol{Y},\tilde{\boldsymbol{X}})$ is linearly separable follows from Lemma 2.2.1, since $\tilde{\boldsymbol{X}}$ is an invertible linear transformation of $\boldsymbol{X}$. □

Hence we can assume that the signal holds in the first feature. Based on this observation, it will turn out to be useful to define the random variables

$$V_i = Y_i\,X_{i,1} \text{ for } i = 1,\ldots,n, \quad \boldsymbol{V} = (V_1,\ldots,V_n)^T. \tag{2.2}$$

We now derive a equivalent condition for the separability of the data set in terms of intersection of subsets of $\mathbb{R}^n$.

Proposition 2.2.1. *If $p < n-1$, then*

$$P[\textit{MLE does not exist}] = P[\mathrm{span}(\boldsymbol{Y},\boldsymbol{V},\boldsymbol{X}_2,\ldots,\boldsymbol{X}_p) \cap \mathbb{R}^n_+ \neq \{\mathbf{0}\}], \tag{2.3}$$

where $\mathbb{R}^n_+ = \{(v_1,\ldots,v_n)^T \in \mathbb{R}^n : v_i \geq 0\ \forall i\}$. Note that $\mathrm{span}(\boldsymbol{Y},\boldsymbol{V},\boldsymbol{X}_2,\ldots,\boldsymbol{X}_p)$ is a random subspace of $\mathbb{R}^n$.

Proof. We know that the MLE does not exist if and only if the dataset is separable, i.e. there exist $b \neq 0$ and $\mathbf{w} = (w_1,\ldots,w_p) \in \mathbb{R}^p$, $\mathbf{w} \neq \mathbf{0}$ such that

$$Y_i\,(b + w_1 X_{i,1} + \cdots + w_p X_{i,p}) \geq 0 \text{ for all } i = 1,\ldots,n$$

that is, if and only if there exists $\mathbf{u} \in \mathbb{R}^n_+$ such that

$$\begin{aligned} &\boldsymbol{Y}(b + w_1\boldsymbol{X}_1 + \cdots + w_p\boldsymbol{X}_p) = \mathbf{u} \\ \Leftrightarrow\ &b\boldsymbol{Y} + w_1\boldsymbol{Y}\boldsymbol{X}_1 + w_2\boldsymbol{Y}\boldsymbol{X}_2 \cdots + w_p\boldsymbol{Y}\boldsymbol{X}_p = \mathbf{u} \end{aligned} \tag{2.4}$$

where the products are performed component-wise. With the notation introduced earlier, we have $\boldsymbol{Y}\boldsymbol{X}_1 = \boldsymbol{V}$. Also, since the entries of $\boldsymbol{Y}$ take values in $\{-1, 1\}$ and the $\boldsymbol{X}_i$'s are assumed to be standard Gaussians (hence with symmetric distribution), we have $\boldsymbol{Y}\boldsymbol{X}_i \overset{d}{=} \boldsymbol{X}_i$ for all $2 \leq i \leq n$. Statement (2.4) is hence equivalent to the existence of $b, \mathbf{w}, \mathbf{u}$ as above such that

$$b\boldsymbol{Y} + w_1\boldsymbol{V} + w_2\boldsymbol{X}_2 + \dots w_p\boldsymbol{X}_p = \mathbf{u}$$

so $\mathbf{u} \in \mathbb{R}^n_+$ and $\mathbf{u} \in \text{span}(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p)$ concurrently and the proposition is proved. □

The right hand-side of (2.3) is not yet the expression we will work with. We want that probability to boil down to the probability of a subspace of $\mathbb{R}^n$ intersecting a convex cone. Let us first introduce the latter concept.

Definition 2.2.1. *A **cone** is a set $C \in \mathbb{R}^n$ such that*

$$\lambda\,\mathbf{x} \in C \quad \forall\lambda \geq 0,\, \forall \mathbf{x} \in C$$

*Let $E \subset \mathbb{R}^n$ a subspace. The **cone generated by** $\mathbf{E}$ is defined as*

$$\mathcal{C}(E) = \{\mathbf{e} + \mathbf{u} : \mathbf{e} \in E, \mathbf{u} \in \mathbb{R}^n_+\}$$

It is easy to see that this construction is indeed a cone, and that it is moreover a closed and convex set, which is crucial in order to be able to use known results from convex geometry.

For the rest of this work we write

$$\mathcal{L} = \text{span}(\boldsymbol{X}_2, \dots, \boldsymbol{X}_p) \text{ and } \mathcal{W} = span(\boldsymbol{Y}, \boldsymbol{V}) \tag{2.5}$$

Proposition 2.2.2. *Let $\{$no MLE in univariate$\}$ be the event that the dataset is separable using the intercept and the first feature $\boldsymbol{X}_1$ only. We have*

$$P[\text{MLE does not exist}] = P[\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\mathbf{0}\} \text{ and } \{\text{no MLE in univariate}\}^c] + P[\{\text{no MLE in univariate}\}]$$

Proof. We partition

$$P[\text{MLE does not exist}] = P[\{\text{MLE does not exist}\} \cap \{\text{no MLE in univariate}\}] + P[\{\text{MLE does not exist}\} \cap \{\text{no MLE in univariate}\}^c]$$

The first term of the left-hand side simplifies to

$$P[\{\text{MLE does not exist}\} \cap \{\text{no MLE in univariate}\}] = P[\{\text{no MLE in univariate}\}]$$

because if the dataset is separable using only the intercept and the first feature, it obviously also is using the intercept and all the features. For the second term, assume that {no MLE in univariate} does not occur. Then, we know from the proof of Proposition 2.2.1 that the MLE does not exist if and only if there is $b \neq 0$, a non zero vector $(w_1, \dots, w_p)$ and a vector $\mathbf{u} \in \mathbb{R}^n_+$ such that

$$b\boldsymbol{Y} + w_1\boldsymbol{V} + w_2\boldsymbol{X}_2 + \dots w_p\boldsymbol{X}_p = \mathbf{u}$$

Note that $(\boldsymbol{Y}, \boldsymbol{V})$ and $(\boldsymbol{X}_2, \dots, \boldsymbol{X}_p)$ being independent implies that $\mathbf{u} \neq \mathbf{0}$ almost surely. Now, $b\boldsymbol{Y} + w_1\boldsymbol{V} = \mathbf{u}$ is impossible since this would imply $\boldsymbol{Y}(b + w_1\boldsymbol{X}_1) \geq \mathbf{0}$ which cannot hold under $\{\text{no MLE in univariate}\}^c$. Hence $w_2\boldsymbol{X}_2 + \dots w_p\boldsymbol{X}_p$ is a non-zero vector. Also,

$$w_2\boldsymbol{X}_2 + \dots w_p\boldsymbol{X}_p = \mathbf{u} - b\boldsymbol{Y} + w_1\boldsymbol{V} = \mathbf{u} + \mathbf{e}$$

with $\mathbf{u} \geq \mathbf{0}$ and $\mathbf{e} = -b\boldsymbol{Y} + w_1\boldsymbol{V} \in \mathcal{W}$. So $w_2\boldsymbol{X}_2 + \dots w_p\boldsymbol{X}_p$ is a non-zero element of $\mathcal{C}(\mathcal{W})$ and we have found

$$P[\{\text{MLE does not exist}\} \cap \{\text{no MLE in univariate}\}^c] = \\ P[\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\mathbf{0}\} \text{ and } \{\text{no MLE in univariate}\}^c]$$

which concludes the proof. □

Corollary 2.2.1. *Using the same notations as in Proposition 2.2.2, we have*

$$P[\textit{MLE does not exist}] = P[\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\mathbf{0}\}] + \mathcal{O}_P(n^{-\frac{1}{2}})$$

Proof. First, note that Proposition 2.2.2 directly implies

$$0 \leq P[\text{MLE does not exist}] \leq P[\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\mathbf{0}\}] + P[\text{no MLE in univariate}]$$

and thus

$$0 \leq P[\text{MLE does not exist}] - P[\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\mathbf{0}\}] \leq P[\text{no MLE in univariate}]$$

so it suffices to show that $P[\text{no MLE in univariate}] = \mathcal{O}_P(n^{-\frac{1}{2}})$. We have a way stronger statement, since actually the latter probability decays exponentially to zero as n, p, d grow. Indeed, in the bivariate case $P(Y_i = 1|X_i, Z_i) = \sigma(\gamma + \beta X_i + \xi Z_i)$ (by Lemma 2.2.2) and for any $t \in \mathbb{R}$ (hyperplane in $\mathbb{R}$), it is easy to see that the probability that t separates the set $(X_i)_{i=1,\dots,n}$ indexed by $(Y_i)_{i=1,\dots,n}$ goes to zero exponentially because of the Gaussian distribution of the X_i's and Z_i's. □

2.3 The approximate kinematic formula

With Corollary 2.2.1 the problem boils down to the study of $P[\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \mathbf{0}]$. Intuitively, for this probability to be big, the cone $\mathcal{C}(\mathcal{W})$ needs to be "rich" enough. We will measure how "rich" a convex cone is by using the notion of statistical dimension. For a convex cone $C \in \mathbb{R}^n$, its statistical dimension is defined as

$$\delta(C) = \sum_{k=0}^{n} k\, \nu_k(C),$$

where the ν_k's form a probability distribution on $\{0, 1, \dots, n\}$ and are called the intrisic volumes of C. The definition of the intrisic volumes is not explicit and they offer no workable expression. We will thus use the following metric characterization, proposed by [10], as a definition of the statistical dimension.

Definition 2.3.1. *Let C be a closed convex cone in $\mathbb{R}^n$. The **statistical dimension** of C is*

$$\delta(C) = \mathbb{E}\left[\|\Pi_C(Q)\|^2\right],$$

where $\mathbf{Q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and Π_C is the Euclidean projection onto the cone C, defined as

$$\Pi_C(\mathbf{x}) = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{c}\| : \mathbf{c} \in C\}, \quad \textit{for } \mathbf{x} \in \mathbb{R}^n.$$

We now present the approximate kinematic formula and how it applies to our case.

Theorem 2.3.1. *Let $\epsilon \in (0, 1)$ and $a_\epsilon = \sqrt{8 \log 4/\epsilon}$. Let C and K be convex cones in $\mathbb{R}^n$ and $\mathbf{Q} \in \mathbb{R}^{n \times n}$ a random orthogonal basis. Then*

$$\begin{aligned}
\delta(C) + \delta(K) \leq n - a_\epsilon \sqrt{n} &\Rightarrow P[C \cap \mathbf{Q}\, K \neq \{\mathbf{0}\}] \leq \epsilon, \\
\delta(C) + \delta(K) \geq n + a_\epsilon \sqrt{n} &\Rightarrow P[C \cap \mathbf{Q}\, K \neq \{\mathbf{0}\}] \geq 1 - \epsilon
\end{aligned}$$

The interested reader can have a look at the proof of this result in [10], Theorem I. We are facing a special case of this theorem.

Corollary 2.3.1. *Let $\mathcal{L}$ as in (2.5) and a subspace $E \subset \mathbb{R}^n$. Let also $\epsilon \in (0, 1)$ and $a_\epsilon = \sqrt{8 \log 4/\epsilon}$. We have*

$$\begin{aligned}
p - 1 + \delta(\mathcal{C}(E)) \leq n - a_\epsilon \sqrt{n} &\Rightarrow P[\mathcal{L} \cap \mathcal{C}(E) \neq \{\mathbf{0}\}] \leq \epsilon, \\
p - 1 + \delta(\mathcal{C}(E)) \geq n + a_\epsilon \sqrt{n} &\Rightarrow P[\mathcal{L} \cap \mathcal{C}(E) \neq \{\mathbf{0}\}] \geq 1 - \epsilon
\end{aligned}$$

Proof. The result is a simple application of Theorem 2.3.1 with $Q\, K = \mathcal{L}$. We also use the fact that the statistical dimension agrees with the usual dimension on vector spaces (see [10], Proposition 3.1). Thus, $\delta(\mathcal{L}) = \dim(\mathcal{L}) = p - 1$. □

The cone that is of our interest is a cone generated by a subset of $\mathbb{R}^n$. We derive a handy expression for the statistical expression of such a cone.

Lemma 2.3.1. *For $E \subset \mathbb{R}^n$, we have*

$$\delta(\mathcal{C}(E)) = n - \mathbb{E}\left[\min_{\mathbf{v} \in E} \|(\mathbf{v} - \mathbf{Q})_+\|^2\right]$$

with $\mathbf{Q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

Proof. We expand

$$\begin{aligned}
\mathbb{E}\|\mathbf{Q} - \Pi_{\mathcal{C}(E)}(\mathbf{Q})\|^2 &= \mathbb{E}\|\mathbf{Q}\|^2 - 2\mathbb{E}[\langle \mathbf{Q},\, \Pi_{\mathcal{C}(E)}(\mathbf{Q})\rangle] + \mathbb{E}\|\Pi_{\mathcal{C}(E)}(\mathbf{Q})\|^2 \\
&= n - 2\mathbb{E}[\langle \Pi_{\mathcal{C}(E)}(\mathbf{Q}),\, \Pi_{\mathcal{C}(E)}(\mathbf{Q})\rangle] + \mathbb{E}\|\Pi_{\mathcal{C}(E)}(\mathbf{Q})\|^2 = n - \mathbb{E}\|\Pi_{\mathcal{C}(E)}(\mathbf{Q})\|^2 \\
&= n - \delta(\mathcal{C}(E))
\end{aligned}$$

where we used in the second equality that $\Pi_{\mathcal{C}(E)}$ is an orthogonal projection. Thus,

$$\delta(\mathcal{C}(E)) = n - \mathbb{E}\|\mathbf{Q} - \Pi_{\mathcal{C}(E)}(\mathbf{Q})\|^2 \tag{2.6}$$

We now derive an equivalent expression for $\|\mathbf{Q} - \Pi_{\mathcal{C}(E)}(\mathbf{Q})\|^2$. For a fixed $\mathbf{q} \in \mathbb{R}^n$, $\|\mathbf{q} - \Pi_{\mathcal{C}(E)}(\mathbf{Q})\|^2 = \text{dist}^2(\mathbf{q}, \mathcal{C}(E))$ is the optimal value of the following quadratic program

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{q} - (\mathbf{v} + \mathbf{u})\|^2 \\ &\text{under} \quad \mathbf{v} \in E,\, \mathbf{u} \geq \mathbf{0} \end{aligned}$$

For a fixed $\mathbf{v} \in E$, the optimal $\mathbf{u} \geq \mathbf{0}$ is $\mathbf{u}^* = (\mathbf{q} - \mathbf{v})_+$. Indeed, one can check that $\mathbf{u}^*$ minimizes each term of the sum $\|\mathbf{q} - (\mathbf{v} + \mathbf{u})\|^2 = \sum_i (q_i - v_i - u_i)^2$. Thus, we can rewrite the objective value of the program as

$$\min_{\mathbf{v} \in E} \|\mathbf{q} - (\mathbf{v} + (\mathbf{q} - \mathbf{v})_+)\|^2 \;=\; \min_{\mathbf{v} \in E} \|\mathbf{q} - \mathbf{v} - (\mathbf{q} - \mathbf{v})_+\|^2 \;=\; \min_{\mathbf{v} \in E} \|(\mathbf{v} - \mathbf{q})_+\|^2$$

and plugging it in (2.6) yields the result. □

Remark 2.3.1. *Observe that $\delta(\mathcal{C}(E))$ is large whenever $\mathbb{E}\left[\min_{\mathbf{v} \in E} \|(\mathbf{v} - \mathbf{Q})_+\|^2\right]$ is small. The latter is smaller whenever E contains more directions in $\mathbb{R}^n$, so that for each realization of $\mathbf{Q}$ one can find $\mathbf{v} \in E$ with small distance to $\mathbf{Q}$. Again, this reinforces the idea that the statistical dimension is a natural stochastic extension of the usual dimension.*

The last big piece of the puzzle is the following convergence result.

Theorem 2.3.2. *Let $\boldsymbol{Y}$ and $\boldsymbol{V}$ as in (1.2) and (2.2), $\mathbf{Q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and define the sequence (R_n) by*

$$R_n = \min_{t_1, t_2 \in \mathbb{R}} \frac{1}{n} \|(t_1 \boldsymbol{Y} + t_2 \boldsymbol{V} - \mathbf{Q})_+\|^2.$$

Define the constant

$$h = \min_{t_1, t_2 \in \mathbb{R}} \mathbb{E}\left[(t_1 Y_1 + t_2 V_1 - Q_1)_+^2\right]$$

as in (2.1). Then,

$$R_n \to h \text{ in probability, as } n, p, d \to \infty.$$

We delay the proof of this theorem and claim that we now have all the ingredients needed to prove Theorem 2.1.1.

Proof of Theorem 2.1.1: Let $\mathcal{F}$ the σ-algebra generated by $\boldsymbol{Y}$ and $\boldsymbol{V}$. Let $\alpha > 0$, $\epsilon_n = n^{-\alpha}$, $b_n = \sqrt{8\alpha \log(4n)}$ and define sequences of events by

$$A_n = \left\{ \frac{p-1}{n} \geq \mathbb{E}[R_n | \mathcal{F}] + b_n n^{-\frac{1}{2}} \right\}, \quad J_n = \{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\mathbf{0}\}\}.$$

Assume now that $\kappa > h$. We want to show $P[\text{MLE exists}] \to 0$, i.e. $P[J_n] \to 1$. Fix $n \geq 1$. Recall Corollary 2.3.1. For $a_\epsilon = \sqrt{8 \log(4/\epsilon_n)}$, we know that if

$$B_n = \{p - 1 + \delta(\mathcal{C}(E)) \geq n + a_\epsilon \sqrt{n}\}$$

occurs, then $P[J_n] \geq 1 - \epsilon$. Now, assume that A_n occurs. Observe that

$$\begin{aligned} \mathbb{E}[R_n | \mathcal{F}] = \mathbb{E}\left[\min_{t_1, t_2 \in \mathbb{R}} \frac{1}{n} \|(t_1 \boldsymbol{Y} + t_2 \boldsymbol{V} - \mathbf{Q})_+\|^2\right] &= \mathbb{E}\left[\min_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} \|(\mathbf{w} - \mathbf{Q})_+\|^2\right] \\ &= \frac{1}{n}(n - \delta(\mathcal{C}(\mathcal{W})) = 1 - \frac{1}{n}\delta(\mathcal{C}(\mathcal{W})) \end{aligned}$$

and $b_n = a_{\epsilon_n}$. Hence the occurrence of A_n can be written as

$$\frac{p-1}{n} > 1 - \frac{1}{n}\delta(\mathcal{C}(\mathcal{W})) + a_{\epsilon_n} n^{-\frac{1}{2}},$$

or equivalently

$$p - 1 + \delta(\mathcal{C}(\mathcal{W})) \geq n + a_{\epsilon_n}\sqrt{n}$$

i.e. $A_n \Rightarrow B_n$, thus $A_n \Rightarrow P[J_n] \geq 1 - \epsilon_n$. This implies

$$\mathbb{1}\{A_n\} \leq \mathbb{1}\{P[J_n|\mathcal{F}] \geq 1 - \epsilon_n\} \leq P[J_n|\mathcal{F}] + \epsilon_n$$

and

$$P[J_n] \geq P[A_n] - \epsilon_n \tag{2.7}$$

by taking expectations. Now, since $R_n - h$ is the minimum of an average of i.i.d. sub-exponential variables, $(R_n - h)$ is uniformly integrable. We have

$$|\mathbb{E}[R_n|\mathcal{F}] - h| \leq E[|R_n - h||\mathcal{F}]$$

and the right-hand side goes to zero because the convergence in probability of Theorem 2.3.2 implies convergence in mean by uniform integrability. Taking expectations, yields the convergence of $\mathbb{E}[R_n|\mathcal{F}]$ to h in mean, and hence in probability, by uniform integrability. Thus,

$$P[A_n] = P\left[\frac{p}{n} \geq E[R_n|\mathcal{F}] + \frac{a_{\epsilon_n}}{\sqrt{n}} + \frac{1}{n}\right]$$

Recall that $\frac{p}{n} \to \kappa$. Under the assumption $\kappa > h$, we thus find $P[A_n] \to 1$ and plugging this into (2.7) concludes the proof. The case $\kappa < h$ is treated in the same way and is let to the reader. □

It now remains to show that $R_n \xrightarrow{P} h$.

Proof of Theorem 2.3.2: We first rephrase what we want to show. Define the function $G : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \|x_+\|^2$. Let $\mathbf{D} = (\boldsymbol{Y}, \boldsymbol{V}) \in \mathbb{R}^{n\times 2}$, $\mathbf{Q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ as before and define

$$F : \mathbb{R}^2 \to \mathbb{R}, \quad \lambda \mapsto \frac{1}{n}G(\mathbf{D}\lambda - \mathbf{Q})$$
$$f : \mathbb{R}^2 \to \mathbb{R}, \quad \lambda \mapsto \mathbb{E}F(\lambda)$$

Since G is convex, F is convex as well and one can show that f is strictly convex. Hence, we can take an almost sure minimizer λ_* of F, and let λ_0 be the unique minimizer of f. We will show a stronger statement that Therorem 2.3.2, namely that $R_n = h + \mathcal{O}_P(n^{-\frac{1}{2}})$. Equivalently, we show

$$F(\lambda_*) = f(\lambda_0) + \mathcal{O}_P(n^{-\frac{1}{2}}). \tag{2.8}$$

The first step consists of using concentration inequalities to show that for a fixed $\lambda \in \mathbb{R}^2$, the random variable $F(\lambda)$ does not deviate a lot from its expectation, namely $f(\lambda)$.

Observe that for $\lambda = (\lambda_1, \lambda_2)^T \in \mathbb{R}^2$

$$F(\lambda) = \frac{1}{n} G(\mathbf{D}\lambda - \mathbf{Q}) = \frac{1}{n} \|(\mathbf{D}\lambda - \mathbf{Q})_+\|^2$$
$$= \frac{1}{n} \left\| \left[\begin{pmatrix} \lambda_1 Y_1 + \lambda_2 V_1 \\ \vdots \\ \lambda_1 Y_n + \lambda_2 V_n \end{pmatrix} - \begin{pmatrix} Q_1 \\ \vdots \\ Q_n \end{pmatrix} \right]_+ \right\|^2 = \frac{1}{n} \sum_{i=1}^{n} (\lambda_1 Y_i + \lambda_2 V_i - Q_i)_+^2 \quad (2.9)$$

so $F(\lambda)$ is an average of n i.i.d. copies of $(\lambda_1 Y_1 + \lambda_2 V_1 - Q_1)_+^2$. Observe that

$$(\lambda_1 Y_1 + \lambda_2 V_1 - Q_1)_+^2 \leq (\lambda_1 Y_1 + \lambda_2 V_1 - Q_1)^2$$
$$\leq 3\,(\lambda_1^2 Y_1^2 + \lambda_2^2 V_1^2 + Q_1^2) = 3\,(\lambda_1^2 + \lambda_2^2 X_{1,1}^2 + Q_1^2)$$

where we used the basic inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and $Y_1^2 = 1$. We know that squared Gaussian RVs are sub-exponential, and that a sum of sub-exponential RVs is itself a sub-exponential RV. Thus, $(\lambda_1 Y_1 + \lambda_2 V_1 - Q_1)_+^2$ is a sub-exponential random variable. Note that the expectation of the right-hand side $\mathbb{E}[3\,(\lambda_1^2 + \lambda_2^2 X_{1,1}^2 + Q_1^2)] = 3\,(1 + \|\lambda\|^2)$, so the parameters of the sub-exponential distribution mentioned above depend on the argument λ, via $1 + \|\lambda\|^2$. Combining the sub-exponentiality with (2.9), we can directly apply Corollary 5.17. of [11] and we get

$$P[|F(\lambda) - f(\lambda)| \geq t] \leq 2 \exp\left[-c_0\, n \min\left(\frac{t^2}{c_1^2 (1 + \|\lambda\|^2)^2}, \frac{t}{c_1 (1 + \|\lambda\|^2)} \right) \right] \quad (2.10)$$

for some constants c_0, c_1. We now work on the gradient of F. First observe that

$$(\nabla G(\mathbf{x}))_j = 2\, x_+^j \frac{d\, x_+}{dx} = \begin{cases} 0, \text{ if } x^j \leq 0 \\ 2\, x_+^j, \text{ otherwise} \end{cases} = 2\, x_+^j$$

so $\nabla G(\mathbf{x}) = 2\,\mathbf{x}_+$ and it is easy to see that $\nabla F(\lambda)$ is an average of i.i.d. copies of a sub-exponential random variable in the same manner as $F(\lambda)$. Thus, we have

$$P[|\nabla F(\lambda) - \nabla f(\lambda)| \geq t] \leq 2 \exp\left[-c_2\, n \min\left(\frac{t^2}{c_3^2 (1 + \|\lambda\|^2)^2}, \frac{t}{c_3 (1 + \|\lambda\|^2)} \right) \right] \quad (2.11)$$

for some constants c_2, c_3. Moreover, F is convex, so it lies above its tangents. In particular,

$$F(\lambda_*) \geq F(\lambda_0) + \langle \nabla F(\lambda_0), \lambda_* - \lambda_0 \rangle \geq F(\lambda_0) - \|\nabla F(\lambda_0)\|\, \|\lambda_* - \lambda_0\| \quad (2.12)$$

using also the Cauchy-Schwarz inequality. Also,

$$\begin{aligned} |F(\lambda_*) - F(\lambda_0)| &= n^{-1} |G(\mathbf{D}\lambda_* - \mathbf{Q}) - G(\mathbf{D}\lambda_0 - \mathbf{Q})| \\ &= n^{-1} |\|(\mathbf{D}\lambda_* - \mathbf{Q})_+\|^2 - \|(\mathbf{D}\lambda_0 - \mathbf{Q})_+\|^2| \\ &\leq n^{-1} \|(\mathbf{D}\lambda_* - \mathbf{Q})_+ - (\mathbf{D}\lambda_0 - \mathbf{Q})_+\|^2 \leq n^{-1} \|\mathbf{D}\lambda_* - \mathbf{Q} - (\mathbf{D}\lambda_0 - \mathbf{Q})\|^2 \\ &= n^{-1} \|\mathbf{D}(\lambda_* - \lambda_0)\|^2 \leq n^{-1} \|\mathbf{D}\|^2 \|\lambda_* - \lambda_0\|^2 \end{aligned}$$

where the last inequality is a direct consequence of the Cauchy-Schwarz inequality. We have

$$n^{-1}\|\mathbf{D}\|^2 = n^{-1}(\|\boldsymbol{Y}\|^2 + \|\boldsymbol{V}\|^2) = 1 + n^{-1}\|\boldsymbol{V}\|^2$$

so we obtain

$$F(\lambda_*) \leq F(\lambda_0) + (1 + n^{-1}\|\boldsymbol{V}\|^2)\,\|\lambda_* - \lambda_0\|^2. \tag{2.13}$$

Combining (2.12) and (2.13) yields

$$F(\lambda_0) - \|\nabla F(\lambda_0)\|\,\|\lambda_* - \lambda_0\| \leq F(\lambda_*) \leq F(\lambda_0) + (1 + n^{-1}\|\boldsymbol{V}\|^2)\,\|\lambda_* - \lambda_0\|^2.$$

Yet, (2.10) implies

$$F(\lambda_0) = f(\lambda_0) + \mathcal{O}_P(n^{-\frac{1}{2}})$$

so we can rewrite the latter bounds on $F(\lambda_*)$ as

$$f(\lambda_0) + \mathcal{O}_P(n^{-\frac{1}{2}}) - \|\nabla F(\lambda_0)\|\,\|\lambda_* - \lambda_0\| \leq F(\lambda_*) \leq \\ f(\lambda_0) + \mathcal{O}_P(n^{-\frac{1}{2}}) + (1 + n^{-1}\|\boldsymbol{V}\|^2)\,\|\lambda_* - \lambda_0\|^2.$$

Thus, showing

$$\|\nabla F(\lambda_0)\|\,\|\lambda_* - \lambda_0\| = \mathcal{O}_P(n^{-\frac{1}{2}}), \text{ and} \tag{2.14}$$

$$(1 + n^{-1}\|\boldsymbol{V}\|^2)\,\|\lambda_* - \lambda_0\|^2 = \mathcal{O}_P(n^{-\frac{1}{2}}) \tag{2.15}$$

would mean reaching the goal (2.8). First, (2.11) implies $\|\nabla F(\lambda_0)\| = \|\nabla f(\lambda_0)\| + \mathcal{O}_P(n^{-\frac{1}{2}})$, and by definition of λ_0 we have $\nabla f(\lambda_0) = \mathbf{0}$ so we get

$$\|\nabla F(\lambda_0)\| = \mathcal{O}_P(n^{-\frac{1}{2}})$$

and (2.14) holds if $\|\lambda_* - \lambda_0\| = \mathcal{O}_P(1)$. Moreover, the law of large numbers yields $n^{-1}\|\boldsymbol{V}\|^2 \xrightarrow{P} \mathbb{E}\boldsymbol{V}^2 < \infty$. Thus, both (2.14) and (2.15) hold (and thus (2.8)) if we show the following

Lemma 2.3.2. $\|\lambda_* - \lambda_0\| = \mathcal{O}_P(n^{-\frac{1}{4}})$.

Proof. We will use second-order Taylor expansions, so we compute the Hessian of F and f, which are twice differentiable. Computing the Hessian of F we get

$$\nabla^2 F(\lambda_1, \lambda_2) = n^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i^2\,\psi_i(\lambda_1, \lambda_2) & \sum_{i=1}^n Y_i\,V_i\,\psi_i(\lambda_1, \lambda_2) \\ \sum_{i=1}^n Y_i\,V_i\,\psi_i(\lambda_1, \lambda_2) & \sum_{i=1}^n V_i^2\,\psi_i(\lambda_1, \lambda_2) \end{pmatrix}$$

where $\psi_i(\lambda_1, \lambda_2) = \mathbb{1}\{\lambda_1 Y_i + \lambda_2 V_i - Q_i \geq 0\}$. It can be rewritten as

$$\nabla^2 F(\lambda) = n^{-1}\mathbf{D}^T\mathbf{L}\mathbf{D}, \quad \mathbf{L} = \operatorname{diag}(\mathbb{1}\{\mathbf{D}\lambda - \mathbf{Q} \geq \mathbf{0}\})$$

so that

$$\nabla^2 f(\lambda) = n^{-1}\,\mathbb{E}[\mathbf{D}^T\mathbf{L}\mathbf{D}], \quad \mathbf{L} = \operatorname{diag}(\mathbb{1}\{\mathbf{D}\lambda - \mathbf{Q} \geq \mathbf{0}\}).$$

Now, expand

$$(\nabla^2 f(\lambda))_{1,1} = (n^{-1}\,\mathbb{E}[\mathbf{D}^T\mathbf{L}\mathbf{D}])_{1,1} = n^{-1}\sum_{i=1}^{n}\mathbb{E}[Y_i^2\mathbb{1}\{\lambda_1 Y_i + \lambda_2 V_i - Q_i \geq 0\}]$$
$$= n^{-1}\sum_{i=1}^{n}\mathbb{E}[Y_i^2]P[\lambda_1 Y_i + \lambda_2 V_i - Q_i \geq 0|\boldsymbol{Y},\boldsymbol{V}] = \sum_{i=1}^{n}\mathbb{E}[Y_i^2]P[Q_i \leq \lambda_1 Y_i + \lambda_2 V_i]$$
$$= \mathbb{E}[Y_1^2\,\phi(\lambda_1 Y_1 + \lambda_2 V_1)]$$

with ϕ the density function of the standard normal distribution. Proceeding the same way for the other entries we get

$$\nabla^2 f(\lambda) = \begin{pmatrix} \mathbb{E}[Y_1^2\,\phi(\lambda_1 Y_1 + \lambda_2 V_1)] & \mathbb{E}[Y_1\,V_1\,\phi(\lambda_1 Y_1 + \lambda_2 V1)] \\ \mathbb{E}[Y_1\,V_1\,\phi(\lambda_1 Y_1 + \lambda_2 V_1)] & \mathbb{E}[V_1^2\,\phi(\lambda_1 Y_1 + \lambda_2 V_1)] \end{pmatrix}$$

and one can show that there exist $\alpha_0, \alpha_1 > 0$ such that

$$\alpha_0\,\mathbf{I}_2 \preceq \nabla^2 f(\lambda) \preceq \alpha_1\,\mathbf{I}_2 \tag{2.16}$$

uniformly over λ, with $\mathbf{M} \succeq \mathbf{0}$ meaning that $\mathbf{M}$ is positive semi-definite. Thus, also recalling that $\nabla f(\lambda_0) = \mathbf{0}$, the Taylor expansion of f around its minimizer gives for $\lambda \in \mathbb{R}^2$

$$f(\lambda) = f(\lambda_0) + \frac{1}{2}(\lambda - \lambda_0)^T\,\nabla^2 f(\lambda)\,(\lambda - \lambda_0) \overset{(2.16)}{\geq} f(\lambda_0) + \frac{\alpha_0}{2}\|\lambda - \lambda_0\|^2. \tag{2.17}$$

Now, for $x \geq 1$, define the 2-dimensional circle of radius $xn^{-1/4}$ centered at λ_0 $C(x) = \{\lambda \in \mathbb{R}^2 : \|\lambda - \lambda_0\| = xn^{-1/4}\}$. For any $\lambda \in C(x)$, we have using (2.17)

$$f(\lambda) \geq f(\lambda_0) + \frac{\alpha_0}{2}\|\lambda - \lambda_0\|^2 = f(\lambda_0) + \frac{\alpha_0}{2}x^2 n^{-\frac{1}{2}} = f(\lambda_0) + 3\,y_n \tag{2.18}$$

for $y_n = \frac{\alpha_0 x^2}{6\sqrt{n}}$. We write $z = f(\lambda_0) + 3\,y_n$, we fix $x \geq 1$ and define the event

$$E = \{F(\lambda_0) < z \quad \text{and} \quad \inf_{\lambda \in C(x)} F(\lambda) > z\}.$$

Under E, the value of F at the center of the circle is strictly less than everywhere on the circle. The convexity of F implies that the minimizer λ_* must then lie inside the circle, so

$$E \Rightarrow \|\lambda_* - \lambda_0\| \leq xn^{-\frac{1}{4}}.$$

Thus it remains to show that $P[E]$ is large enough.

To achieve this, we pick d equidistant points $\{\lambda_i\}_{i=1}^d \subset C(x)$. We define the event

$$B = \left\{\max_{1 \leq i \leq d}\|\nabla F(\lambda_i) - \nabla f(\lambda_i)\| \leq xn^{-\frac{1}{2}}\right\}$$

and want to study how the occurrence of B affects the one of E. First, (2.11) directly implies, using a union bound argument, that

$$P[B^c] \leq 2\,d\,\exp\left[-c_2\,\min\left(\frac{x^2}{c_3^2(1 + \max_i\|\lambda_i\|^2)^2}, \frac{\sqrt{n}\,x}{c_3(1 + \max_i\|\lambda_i\|^2)}\right)\right]. \tag{2.19}$$

Fix $\lambda \in C(x)$ and take $i = \arg\min_{1\leq i\leq d} \|\lambda - \lambda_i\|$. Note that we then have

$$\|\lambda - \lambda_i\| \leq \frac{1}{2}\frac{2\pi\, x\, n^{-\frac{1}{4}}}{d} = \frac{\pi\, x\, n^{-\frac{1}{4}}}{d} \tag{2.20}$$

In a similar fashion than for (2.12) we use the convexity of F to get

$$F(\lambda) \geq F(\lambda_i) + \langle \nabla F(\lambda_i),\, \lambda - \lambda_i \rangle \geq F(\lambda_i) - \|\nabla F(\lambda_i)\|\, \|\lambda - \lambda_i\|. \tag{2.21}$$

We start by finding an upper-bound for the second expression of the right-hand side. A first-order Taylor expansion of ∇f around λ_0 gives (recall that $\nabla f(\lambda_0) = \mathbf{0}$)

$$\|\nabla f(\lambda_i)\| \leq \|\nabla^2 f(\lambda_0)\|\, \|\lambda_i - \lambda_0\| \overset{(2.16)}{\leq} \alpha_1 \|\lambda_i - \lambda_0\| = \alpha_1 x\, n^{-\frac{1}{4}} \tag{2.22}$$

Now, when B occurs we have

$$\begin{aligned}
\|\nabla F(\lambda_i)\|\, \|\lambda - \lambda_i\| &\overset{(2.20)}{\leq} \|\nabla F(\lambda_i)\| \frac{\pi\, x\, n^{-\frac{1}{4}}}{d} \\
&\leq \left(\|\nabla f(\lambda_i)\| + \|\nabla F(\lambda_i) - \nabla f(\lambda_i)\|\right) \frac{\pi\, x\, n^{-\frac{1}{4}}}{d} \\
&\overset{(2.22),\, \text{def. } B}{\leq} \left(\alpha_1 x\, n^{-\frac{1}{4}} + x\, n^{-\frac{1}{2}}\right) \frac{1}{d}\, \pi\, x\, n^{-\frac{1}{4}} \leq C\frac{y_n}{d}
\end{aligned}$$

for n large enough and some constant C. Thus, on B, we have, using (2.21)

$$\inf_{\lambda\in C(x)} F(\lambda) \geq \min_{1\leq i\leq d} F(\lambda_i) - C\frac{y_n}{d} \geq \min_{1\leq i\leq d} F(\lambda_i) - y_n \tag{2.23}$$

for d chosen such that $d \geq C$. For the expression on the right-hand side, observe that if $F(\lambda_i) > f(\lambda_i) - y_n \forall i$, then $F(\lambda_i) - y_n > f(\lambda_0) + y_n = z$ because $f(\lambda_i) \geq f(\lambda_0) + 3\, y_n$ by (2.18). In particular, applying this to (2.23) yields,

$$B \cap \left(\bigcap_{i=1}^{d} \{F(\lambda_i) > f(\lambda_i) - y_n\}\right) \Rightarrow \left\{\inf_{\lambda\in C(x)} F(\lambda) > z\right\}$$

So, looking at the definition of the event E, we have obtained

$$B \cap \left(\bigcap_{i=1}^{d} \{F(\lambda_i) > f(\lambda_i) - y_n\}\right) \cap \{F(\lambda_0) < f(\lambda_0) + y_n\} \Rightarrow E$$

so that

$$P[E^c] \geq P[B^c] + \sum_{i=1}^{d} P[F(\lambda_i) \leq f(\lambda_i) - y_n] + P[F(\lambda_0) \geq f(\lambda_0) + y_n]$$

and applying (2.19) to the first term and (2.10) to the last two terms yields

$$P[E^c] = \mathcal{O}(n^{-\frac{1}{2}})$$

and in particular

$$P[E] \to 1$$

which concludes the proof of the lemma and hence the proof of Theorem 2.3.2. □

Remark 2.3.2. *We have throughout this chapter actually proved a stronger statement that Theorem 2.1.1, namely that the phase transition happens in an interval with length decreasing with rate $n^{-1/2}$, i.e. for any sequence $d_n \to \infty$,*

$$p/n > h + d_n n^{-1/2} \Rightarrow P[\text{MLE exists}] \to 0,$$
$$p/n < h - d_n n^{-1/2} \Rightarrow P[\text{MLE exists}] \to 1.$$

2.4 Empirical results

We run simulations to illustrate what we have derived theoretically. Given data points $(\mathbf{x}_i, y_i)$, one can check if the set is separable by considering the following linear program

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n} y_i(w_0 + \mathbf{x}_i^T \mathbf{w}) \\ \text{under} \quad & y_i(w_0 + \mathbf{x}_i^T \mathbf{w}) \geq 0, \ i = 1, \dots, n \\ & -1 \leq w_0 \leq 1, \ -\mathbf{1} \leq \mathbf{w} \leq \mathbf{1} \end{aligned} \tag{2.24}$$

If there exists a feasible non-zero solution, then the set is separable. If on the contrary there is no feasible solution, or the optimal solution is $w_0 = 0, \mathbf{w} = \mathbf{0}$, then the set is not separable and the MLE does exist. We will use the "pulp" package in python to do so. For a given setting, one can thus generate realizations $(\mathbf{x}_i, y_i)$ multiple times and get an estimate of P[MLE does not exist]. We do this for a couple of values of κ in order to visualize the phase transition phenomenon.

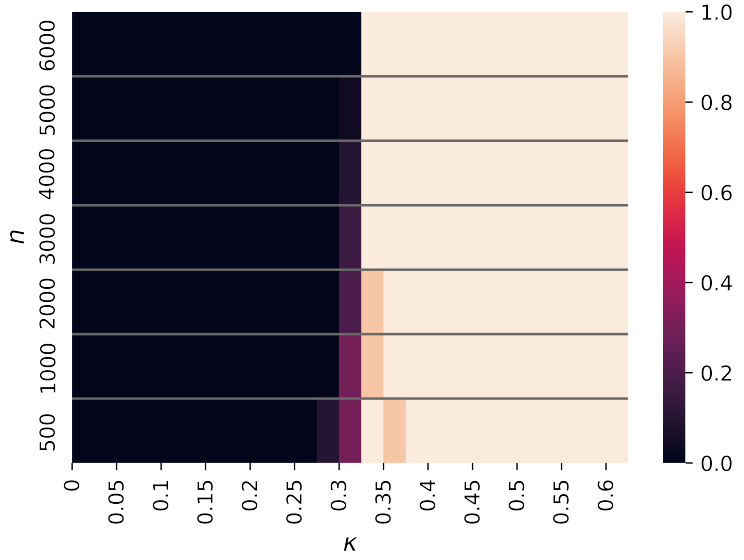


Figure 2.1: P[MLE does not exist] for well specified model, varying sample size n

We consider a sequence of κ equidistant between 0 and 0.6. For a fixed n, we fix $p = \kappa n$ and we generate 50 times variables $\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{I}_p)$, $i = 1, \dots, n$ and the

according y_i under the model (1.2). For simplicity, we do not include an intercept. Since the threshold h depends on ν^2, we fix $\nu^2 = 2.5$ and select coefficients β so that $\mathbb{V}\text{ar}(\mathbf{x}_i^T\beta) = \|\beta\|^2 = \nu^2$, because the value of ν^2 affects the value of the threshold h. For simplicity, we assume that there is no unobserved variable, so we are in a special case of the setup of Chapter 2. Figure 2.1 shows the results of the simulations for 7 different values of n (see y-axis). The colors represent the estimated probability that the MLE does **not** exist for some value of n and of κ (the lighter the color, the higher the probability). Our estimate of this probability is simply the average number of times that the data set is separable, over the 50 simulations. One can clearly observe the phase transition around the value $\kappa \approx 0.31$, and we can see that the transition is stronger as n increases, like we expect from Theorem 2.1.1. The interested reader can verify that the threshold $\kappa \approx 0.31$ is close to the theoretical value h, which is displayed in [8], Figure 2.

Chapter 3

Latent variable model

We now assume that the unobserved variable is a latent variable, which takes the form of linear transformation of the observed one plus another variable which is independent from it, i.e.

$$\boldsymbol{Z} = \boldsymbol{X}\mathbf{A}^T + \boldsymbol{U}\mathbf{B}^T$$

where $\boldsymbol{U}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $i = 1, \dots, n$, $\boldsymbol{X}, \boldsymbol{U}$ are independent, and $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times p}$. Hence our model (1.2) can be written as

$$\boldsymbol{Y} \sim \mathrm{Rad}(\sigma(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi) + \boldsymbol{U}\mathbf{B}^T\xi)). \tag{3.1}$$

3.1 Strongly correlated features

In this section we take a look at the case where

$$\|\mathbf{B}^T\xi\| \to 0 \text{ as } n, p, d \to \infty$$

More specifically, we assume the following

Assumption 3.1.1. *The random variable $\boldsymbol{U}$ is in L^2,*

$$\|\mathbf{B}\| = o\left(n^{-\frac{3}{2}}d^{-\frac{1}{2}}\right) \text{ as } n, p, d \to \infty \tag{3.2}$$

and

$$\sup_{d\geq 1}\sup_{j}|\xi_j| < \infty$$

where n and p grow at a similar speed under Assumption 1.1.1.

Theorem 3.1.1. *Let $\boldsymbol{Y}$ distributed as in (3.1), $\boldsymbol{V}$ as in (2.2) and $Q \sim \mathcal{N}(0, 1)$, independent from $\boldsymbol{Y}$ and $\boldsymbol{V}$. Define*

$$h = \min_{t_0,t_1\in\mathbb{R}} \mathbb{E}\left[(t_0Y_1 + t_1V_1 - Q)_+^2\right] \tag{3.3}$$

Then, under Assumption 1.1.1 and Assumption 3.1.1,

$$\text{if } \kappa > h, \text{ then } \lim_{n,p,d\to\infty} P[\text{MLE exists}] = 0$$
$$\text{if } \kappa < h, \text{ then } \lim_{n,p,d\to\infty} P[\text{MLE exists}] = 1.$$

Proof. First, let

$$Y_s^i = Y_s^i(W) = \begin{cases} 1, \text{ if } W \leq \sigma(\gamma + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi)) \\ -1, \text{ otherwise} \end{cases} \quad i = 1, \dots, n \tag{3.4}$$

and

$$Y^i = Y^i(W) = \begin{cases} 1, \text{ if } W \leq \sigma(\gamma + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}_i^T\mathbf{B}^T\xi) \\ -1, \text{ otherwise} \end{cases} \quad i = 1, \dots, n \tag{3.5}$$

where W is a RV with distribution $\mathrm{Unif}(0,1)$. Then $\boldsymbol{Y}_s \sim \mathrm{Rad}(\sigma(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi)))$ and $\boldsymbol{Y}$ is distributed as in (3.1). Let us write $p = \sigma(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi) + \boldsymbol{U}\mathbf{B}^T\xi)$ and $\tilde{p} = \sigma(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi))$. We compute, for $1 \leq i \leq n$:

$$\begin{aligned} P[Y^i \neq Y_s^i] &= P[W \in (\mathbb{E}(p_i), \mathbb{E}(\tilde{p}_i)]] + P[W \in (\mathbb{E}(\tilde{p}), \mathbb{E}(p_i)]] = |\mathbb{E}(p_i) - \mathbb{E}(\tilde{p}_i)| \\ &= |\mathbb{E}\left(\sigma(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}_i^T\mathbf{B}^T\xi) - \sigma(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi))\right)| \\ &\leq \mathbb{E}\left|\sigma(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}_i^T\mathbf{B}^T\xi) - \sigma(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi))\right| \\ &\leq K\mathbb{E}(|\mathbf{u}_i^T\mathbf{B}^T\xi|) \end{aligned} \tag{3.6}$$

where we used in the last inequality that σ is K-Lipschitz, because bounded. So we have

$$\begin{aligned} P[\boldsymbol{Y} \neq \boldsymbol{Y}_s] = P[\{\exists\, i \text{ s.t. } Y^i \neq Y_s^i\}] &= P\left[\bigcup_{i=1}^n \{Y^i \neq Y_s^i\}\right] \\ &\leq \sum_{i=1}^n P[Y^i \neq Y_s^i] \leq K\sum_{i=1}^n E|\mathbf{u}_i^T\mathbf{B}^T\xi| = K\,n\,E|\mathbf{u}^T\mathbf{B}^T\xi| \end{aligned}$$

where we let $\mathbf{u}$ be a copy of $\mathbf{u}_1$ and we used in the last equality that $\mathbf{u}_1, \dots, \mathbf{u}_n$ are independent, and (3.6) in the last inequality. We now want the term $\mathbb{E}(|\mathbf{u}^T\mathbf{B}^T\xi|)$ to decrease fast enough in order to have $P[Y \neq Y_s] \to 0$. We bound using the Cauchy-Schwarz inequality:

$$K\,n\,\mathbb{E}|\mathbf{u}^T\mathbf{B}^T\xi| \leq K\,n\,\|\mathbf{B}^T\xi\|\,\mathbb{E}\|\mathbf{u}\|$$

Now, since $\boldsymbol{U}$ is L^2 by Assumption 3.1.1, we have

$$\mathbb{E}\|\mathbf{u}\| = \mathcal{O}(\sqrt{p}) \text{ as } p \to \infty$$

so $K\,n\,\|\mathbf{B}^T\xi\|\,\mathbb{E}\|\mathbf{u}\| \sim K\,n\,\sqrt{p}\,\|\mathbf{B}^T\xi\|$ as $n, p, d \to \infty$ and thus

$$\|\mathbf{B}^T\xi\| = o\left(n^{-1}p^{-\frac{1}{2}}\right) \Rightarrow P[\boldsymbol{Y} \neq \boldsymbol{Y}_s] \to 0$$

or equivalently

$$\|\mathbf{B}^T\xi\| = o\left(n^{-\frac{3}{2}}\right) \Rightarrow P[\boldsymbol{Y} \neq \boldsymbol{Y}_s] \to 0.$$

We now show that the condition $\|\mathbf{B}^T\xi\| = o\left(n^{-\frac{3}{2}}\right)$ is implied by (3.2). Assume $\|\mathbf{B}\| = o\left(n^{-3}d^{-\frac{1}{2}}\right)$ and $\sup_{d\geq 1}\sup_j |\xi_j| < \infty$. Note that the latter is equivalent to

$$\|\xi\|^2 = \mathcal{O}(d). \tag{3.7}$$

We denote $\mathbf{b}_k$ the k^{th} column of $\mathbf{B}$. Using the Cauchy-Schwarz inequality we get

$$\|\mathbf{B}^T\xi\|^2 = \sum_{j=1}^{p}(\mathbf{b}_j^T\xi)^2 \leq \|\xi\|^2 \sum_{j=1}^{p}\|\mathbf{b}_j^T\|^2 = \|\xi\|^2\|\mathbf{B}\|^2$$

so, using (3.7) and Assumption 3.1.1, we have as wanted

$$\|\mathbf{B}^T\xi\| = o\left(n^{-\frac{3}{2}}\right)$$

and thus

$$P[Y \neq Y_s] \to 0 \tag{3.8}$$

We now show that (3.8) implies that the phase transition for $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$ boils down to the one for $(\boldsymbol{Y}_s, \boldsymbol{X})$. Indeed, Let $\boldsymbol{Y}_s, \boldsymbol{Y}$ as in (3.4) and (3.5) respectively, and let $\boldsymbol{V}_s = \boldsymbol{Y}_s\boldsymbol{X}$, $\boldsymbol{V}$ as in (2.2), $\mathbf{Q} \sim N(0, \mathbf{I}_n)$. Let

$$\tilde{R}_n = \min_{t_0,t_1\in\mathbb{R}} \frac{1}{n}\|(t_0\boldsymbol{Y}_s + t_1\boldsymbol{V}_s - \mathbf{Q})_+\|^2$$

and

$$R_n = \min_{t_0,t_1\in\mathbb{R}} \frac{1}{n}\|(t_0\boldsymbol{Y} + t_1\boldsymbol{V} - \mathbf{Q})_+\|^2$$

The case where Z and X are independent obviously contains the case of no unobservable variable. Thus, $\boldsymbol{Y}_s$ and $\boldsymbol{V}_s$ satisfy the condition of Theorem 2.3.2, which gives

$$\tilde{R}_n \xrightarrow{P} h = \min_{t_0,t_1\in\mathbb{R}} \mathbb{E}(t_0Y_s^{(1)} + t_1V_s^{(1)} - Q_1)_+^2 \text{ as } n, p, d \to \infty \tag{3.9}$$

We show that $R_n \xrightarrow{P} h$ as well. First note that

$$P[\boldsymbol{Y} \neq \boldsymbol{Y}_s] \to 0 \quad \Rightarrow \quad P[R_n \neq \tilde{R}_n] \to 0 \tag{3.10}$$

Now, let $\epsilon > 0$.

$$\begin{aligned} P[|R_n - h| > \epsilon] &= P[|R_n - h| > \epsilon,\ R_n = \tilde{R}_n] + P[|R_n - h| > \epsilon,\ R_n \neq \tilde{R}_n] \\ &\leq P[|\tilde{R}_n - h| > \epsilon] + P[R_n \neq \tilde{R}_n] \end{aligned}$$

and both terms of the last expression tend to zero by (3.9) and (3.10) respectively. Hence $P[|R_n - h| > \epsilon] \to 0$, so $R_n \xrightarrow{P} h$ as wanted. Thus, we have an equivalent of Theorem 2.3.2 for this case and now the rest of the proof is identical to the proof of Theorem 2.1.1 in Chapter 2. □

Remark 3.1.1. *Assumption 3.1.1 restrains ξ, i.e. the parameter vector of the unobservable variable Z. We restrain its norm to grow slower than linearly in the dimension d. If we impose a stronger assumption on ξ, a slower rate of decay for $\|B\|$ is sufficient for Theorem 3.1.1 to hold. Indeed, it is easy to show that the following assumption is equivalent to Assumption 3.1.1:*

$$\|\mathbf{B}\| = o\left(n^{-\frac{3}{2}}\right) \text{ as } n, p, d \to \infty$$

and

$$\|\xi\|^2 \to 0 \text{ as } n, p, d \to \infty$$

where n and p grow at a similar speed under Assumption 1.1.1. Here we assume that the effect of the unobservable variable Z on the response Y vanishes as the dimensions grow.

We now present a condition on the correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ which is equivalent to (3.2). We fix an observation index $i \in \{1, \dots, n\}$. We know that

$$\boldsymbol{Z}_i = \mathbf{A}\boldsymbol{X}_i + \mathbf{B}\mathbf{U}_i.$$

Let us have a look at the covariance matrix of our input. We write $\mathbf{D} = (\boldsymbol{X}, \boldsymbol{Z})$ a $n \times 2p$ random matrix. Its covariance matrix is of the form

$$\Sigma_{\mathbf{D}} = \mathbb{C}\mathrm{ov}(\mathbf{D}) = \begin{pmatrix} \Sigma_{\boldsymbol{X}} & \Sigma_{\boldsymbol{X},\boldsymbol{Z}} \\ \Sigma_{\boldsymbol{Z},\boldsymbol{X}} & \Sigma_{\boldsymbol{Z}} \end{pmatrix}$$

where $\Sigma_{\boldsymbol{X}} = (\mathbb{C}\mathrm{ov}(X_{i,k}, X_{i,\ell}))_{k,\ell}$, $\Sigma_{\boldsymbol{Z}} = (\mathbb{C}\mathrm{ov}(Z_{i,k}, Z_{i,\ell}))_{k,\ell}$, $\Sigma_{\boldsymbol{X},\boldsymbol{Z}} = (\mathbb{C}\mathrm{ov}(X_{i,k}, Z_{i,\ell}))_{k,\ell}$ and $\Sigma_{\boldsymbol{Z},\boldsymbol{X}} = (\mathbb{C}\mathrm{ov}(Z_{i,k}, X_{i,\ell}))_{k,\ell} = \Sigma_{\boldsymbol{X},\boldsymbol{Z}}^T$. We want to determine A and B in the expression of $\boldsymbol{Z}$ as functions of the blocks of Σ_W. Since $\boldsymbol{X}$ and $\boldsymbol{U}$ are independent, we know that

$$\mathbb{C}\mathrm{ov}(\boldsymbol{X}_i,\, \mathbf{B}\mathbf{U}_i) = \mathbf{0}$$

Rewriting $B\mathbf{U}_i$, we get

$$\mathbb{C}\mathrm{ov}(\boldsymbol{X}_i, \mathbf{B}\mathbf{U}_i) = \mathbb{C}\mathrm{ov}(\boldsymbol{X}_i,\, \boldsymbol{Z}_i - \mathbf{A}\boldsymbol{X}_i) = \Sigma_{\boldsymbol{X},\boldsymbol{Z}} - \Sigma_{\boldsymbol{X}}\mathbf{A}^T$$

Combining both equalities yields

$$\Sigma_{\boldsymbol{X}}\mathbf{A}^T = \Sigma_{\boldsymbol{X},\boldsymbol{Z}}$$

so we have

$$\mathbf{A}^T = \Sigma_{\boldsymbol{X}}^{-1}\Sigma_{\boldsymbol{X},\boldsymbol{Z}} = \Sigma_{\boldsymbol{X},\boldsymbol{Z}} \tag{3.11}$$

since $\Sigma_{\boldsymbol{X}} = I_p$ by assumption on the distribution of $\boldsymbol{X}_i$. We use a similar approach for $\mathbf{B}$, which is our interest. We compute

$$\mathbb{C}\mathrm{ov}(\boldsymbol{Z}_i,\, \mathbf{B}\mathbf{U}_i) = \mathbb{C}\mathrm{ov}(\mathbf{A}\boldsymbol{X}_i + \mathbf{B}\mathbf{U}_i,\, \mathbf{B}\mathbf{U}_i) = \mathbf{0} + \mathbf{B}\Sigma_{\mathbf{U}}\mathbf{B}^T$$

and

$$\mathbb{C}\mathrm{ov}(\boldsymbol{Z}_i,\, \mathbf{B}\mathbf{U}_i) = \mathbb{C}\mathrm{ov}(\boldsymbol{Z}_i,\, \boldsymbol{Z}_i - \mathbf{A}\boldsymbol{X}_i) = \Sigma_{\boldsymbol{Z}} - \Sigma_{\boldsymbol{Z},\boldsymbol{X}}\mathbf{A}^T$$

Combining both equalities yields

$$\mathbf{B}\mathbf{B}^T = \mathbf{I}_d - \Sigma_{\boldsymbol{Z},\boldsymbol{X}}\mathbf{A}^T = \mathbf{I}_d - \Sigma_{\boldsymbol{Z},\boldsymbol{X}}\Sigma_{\boldsymbol{X},\boldsymbol{Z}} \tag{3.12}$$

where we used the assumption on the distributions of $\mathbf{U}$ and $\boldsymbol{Z}$ to write $\Sigma_{\mathbf{U}} = \mathbf{I}_p$, $\Sigma_{\boldsymbol{Z}} = \mathbf{I}_d$. Now, $\|\mathbf{B}\| \to 0$ if and only if $\|\mathbf{B}\mathbf{B}^T\| \to 0$, so

$$\|\mathbf{B}\| \to 0 \quad \Leftrightarrow \quad \|\Sigma_{\boldsymbol{Z},\boldsymbol{X}}\Sigma_{\boldsymbol{X},\boldsymbol{Z}} - \mathbf{I}_d\| \to 0 \text{ as } n, p, d \to \infty$$

Equivalently,

$$\|\mathbf{B}\| \to 0 \quad \Leftrightarrow \quad \Sigma_{\boldsymbol{Z},\boldsymbol{X}}\Sigma_{\boldsymbol{X},\boldsymbol{Z}} \to \mathbf{I}_d \text{ entry-wise as } n, p, d \to \infty$$

And we have actually proved the stronger statement

Proposition 3.1.1.

$$\|\mathbf{B}\| = o\left(n^{-\frac{3}{2}}d^{-\frac{1}{2}}\right) \quad \Leftrightarrow \quad \|\Sigma_{\boldsymbol{Z},\boldsymbol{X}}\Sigma_{\boldsymbol{X},\boldsymbol{Z}} - \mathbf{I}_d\| = o\left(n^{-\frac{3}{2}}d^{-\frac{1}{2}}\right) \text{ as } n,p,d \to \infty.$$

Intuitively the correlation of the features has to grow with rate n^{-2} for Assumption 3.1.1 to hold.

Remark 3.1.2. *Expanding the condition $\Sigma_{\boldsymbol{Z},\boldsymbol{X}}\Sigma_{\boldsymbol{X},\boldsymbol{Z}} \to \mathbf{I}_d$, we get*

$$\begin{cases} \sum_{k=1}^{p} \mathbb{C}\mathrm{ov}(\boldsymbol{Z}_i, \boldsymbol{X}_k)^2 \to 1 \quad \textit{for all } i = 1,\dots,d \\ \sum_{k=1}^{p} \mathbb{C}\mathrm{ov}(\boldsymbol{Z}_i, \boldsymbol{X}_k)\,\mathbb{C}\mathrm{ov}(\boldsymbol{Z}_j, \boldsymbol{X}_k) \to 0 \quad \textit{for all } 1 \le i \ne j \le d \end{cases}$$

Note that we can take the transposed and get as a condition $\Sigma_{\boldsymbol{X},\boldsymbol{Z}}\Sigma_{\boldsymbol{Z},P} \to \mathbf{I}_p$, i.e.

$$\begin{cases} \sum_{k=1}^{d} \mathbb{C}\mathrm{ov}(\boldsymbol{X}_i, \boldsymbol{Z}_k)^2 \to 1 \quad \textit{for all } i = 1,\dots,p \\ \sum_{k=1}^{d} \mathbb{C}\mathrm{ov}(\boldsymbol{X}_i, \boldsymbol{Z}_k)\,\mathbb{C}\mathrm{ov}(\boldsymbol{X}_j, \boldsymbol{Z}_k) \to 0 \quad \textit{for all } 1 \le i \ne j \le p \end{cases}$$

Remark 3.1.3. *Doing the same development for the one-dimensional case $Z = aX + bU$ where $Z, X, U \sim \mathcal{N}(0,1)$, $a, b \in \mathbb{R}$, we obtain that*

$$b \to 0 \quad \Leftrightarrow \quad \left| \frac{\mathbb{C}\mathrm{ov}(X,Z)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\,\mathbb{V}\mathrm{ar}(Z)}} \right| \to 1$$

i.e.

$$b \to 0 \quad \Leftrightarrow \quad |\mathrm{Corr}(X,Z)| \to 1$$

so in this case the intuition is clear, because we know that if Z is a linear transformation of X (hence $\mathrm{Corr}(X,Z) = 1$ and have $b = 0$).

3.2 Weakly correlated features

We have seen that when the correlation between $\mathbf{X}$ and $\boldsymbol{Z}$ tends to be stronger and stronger, the response variable tends to behave like

$$\boldsymbol{Y}_s \sim \mathrm{Rad}(\sigma(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi))).$$

We can legitimately ask ourselves what happens when $\boldsymbol{X}$ and $\boldsymbol{Z}$ tend to be less and less correlated. We directly get the answer by recalling (3.11):

$$\mathbf{A}^T = \Sigma_{X,\boldsymbol{Z}} \quad \Leftrightarrow \quad \mathbf{A} = \Sigma_{\boldsymbol{Z},\boldsymbol{X}}$$

so we obtain

$$\|\Sigma_{\boldsymbol{Z},\boldsymbol{X}}\| \to 0 \quad \Leftrightarrow \quad \|\mathbf{A}\| \to 0 \text{ as } n,p,d \to \infty$$

Remark 3.2.1. *In the one-dimensional case $Z = aX + bU$ where $Z, X, U \sim \mathcal{N}(0,1)$, $a, b \in \mathbb{R}$, we get*

$$a \to 0 \quad \Leftrightarrow \quad \mathbb{C}\mathrm{ov}(X,Z) \to 0 \textit{ as } n,p,d \to \infty$$

Now, assuming this weak correlation set-up with $\mathbf{A}^T\xi \to \mathbf{0}$, we can show that for

$$\boldsymbol{Y}_w \sim \mathrm{Rad}(\sigma(\gamma + \boldsymbol{X}\beta + \boldsymbol{U}\mathbf{B}^T\xi))$$

we have

$$P[\boldsymbol{Y} \neq \boldsymbol{Y}_w] \to 0 \text{ as } n, p, d \to \infty$$

if we assume the following

Assumption 3.2.1. *The random variable X is in L^2,*

$$\|\mathbf{A}\| = o\left(n^{-\frac{3}{2}}d^{-\frac{1}{2}}\right) \text{ as } n, p, d \to \infty \tag{3.13}$$

and

$$\sup_{d\geq 1}\sup_{j}|\xi_j| < \infty$$

where n and p grow at a similar speed under Assumption 1.1.1.

This will allow us to prove

Theorem 3.2.1. *Let $\boldsymbol{Y}$ distributed as in (3.1), $\boldsymbol{V}$ as in (2.2) and $Q \sim \mathcal{N}(0,1)$, independent from $\boldsymbol{Y}$ and $\boldsymbol{V}$. Define*

$$h = \min_{t_0,t_1\in\mathbb{R}} \mathbb{E}\left[(t_0Y_1 + t_1V_1 - Q)_+^2\right] \tag{3.14}$$

Then, under Assumption 1.1.1 and Assumption 3.2.1,

$$\text{if } \kappa > h, \text{ then } \lim_{n,p,d\to\infty} P[\text{MLE exists}] = 0$$
$$\text{if } \kappa < h, \text{ then } \lim_{n,p,d\to\infty} P[\text{MLE exists}] = 1.$$

Proof. We follow exactly the same recipe as for proving Theorem 3.1.1, hence some details will be skipped here. Let

$$Y_w^i = Y_w^i(W) = \begin{cases} 1, \text{ if } W \leq \sigma(\gamma + \mathbf{x}_i^T\beta + \mathbf{u}_i^T\mathbf{B}^T\xi) \\ -1, \text{ otherwise} \end{cases} \quad i = 1, \ldots, n \tag{3.15}$$

and

$$Y^i = Y^i(W) = \begin{cases} 1, \text{ if } W \leq \sigma(\gamma + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}_i^T\mathbf{B}^T\xi) \\ -1, \text{ otherwise} \end{cases} \quad i = 1, \ldots, n \tag{3.16}$$

where W is a RV with distribution $\mathrm{Unif}(0,1)$. Then $\boldsymbol{Y}_w \sim \mathrm{Rad}(\sigma(\gamma + \boldsymbol{X}\beta + \boldsymbol{X}\mathbf{B}^T\xi))$ and $\boldsymbol{Y}$ is distributed as in (3.1). Let us write $p = \sigma(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi) + \boldsymbol{U}\mathbf{B}^T\xi)$ and $\tilde{p} = \sigma(\gamma + \boldsymbol{X}\beta + \boldsymbol{U}\mathbf{B}^T\xi)$. With the same development as in (3.6), we find for $1 \leq i \leq n$:

$$P[Y^i \neq Y_w^i] \leq K\mathbb{E}(|\mathbf{x}_i^T\mathbf{A}^T\xi|)$$

Now in order to have $P[\boldsymbol{Y} \neq \boldsymbol{Y}_w] \to 0$ as $n, p, d \to \infty$, we find using the same technique as in the proof of Theorem 3.1.1 that the following rate of decrease of the norm of A is sufficient:

$$\|\mathbf{A}\|^2 = o\left(n^{-3}d^{-1}\right), \text{ equivalently } \|A\| = o\left(n^{-\frac{3}{2}}d^{-\frac{1}{2}}\right)$$

With $P[\boldsymbol{Y} \neq \boldsymbol{Y}_w] \to 0$, we show with the exact same coupling argument that the phase transition for $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$ boils down to the one for $(\boldsymbol{Y}_w, \boldsymbol{X}, \boldsymbol{U})$. Since $\boldsymbol{X}$ and $\boldsymbol{U}$ are independent, we are exactly in the set-up of Chapter 2, where Theorem 2.3.2 holds. We thus have an equivalent of this theorem, and now the rest of the proof is identical to the proof of Theorem 2.1.1 in Chapter 2. □

3.3 Stable unobservable effect

We have seen that if the effect of the unobservable variable vanishes quickly enough, we still have a phase transition in the existence of the MLE. We now show that it is in fact sufficient to control this effect, by assuming $\mathbf{B}^T\xi \to \mathbf{c} \in \mathbb{R}^p, \mathbf{c} \neq \mathbf{0}$.

Theorem 3.3.1. *Let $\boldsymbol{Y}$ distributed as in (3.1), $\boldsymbol{V}$ as in (2.2) and $Q \sim \mathcal{N}(0,1)$, independent from $\boldsymbol{Y}$ and $\boldsymbol{V}$. Define*

$$h = \min_{t_0,t_1 \in \mathbb{R}} \mathbb{E}\left[(t_0 Y_1 + t_1 V_1 - Q)_+^2\right] \tag{3.17}$$

Then, under Assumption 1.1.1 and if

$$\mathbf{B}^T\xi \to \mathbf{c} \in \mathbb{R}^p, \mathbf{c} \neq \mathbf{0}, \tag{3.18}$$

$$\begin{aligned} &\text{if } \kappa > h, \text{ then } \lim_{n,p,d\to\infty} P[\text{MLE exists}] = 0 \\ &\text{if } \kappa < h, \text{ then } \lim_{n,p,d\to\infty} P[\text{MLE exists}] = 1. \end{aligned}$$

Proof. Recall

$$\boldsymbol{Y} \sim \mathrm{Rad}(\sigma(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi) + \boldsymbol{U}\mathbf{B}^T\xi)$$

Focusing on observation $i \in \{1, \dots, n\}$ we get

$$y_i \sim \mathrm{Rad}(\sigma(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}_i^T\mathbf{B}^T\xi))$$

Dropping the subscript i for readability, we have

$$P[y = 1|\mathbf{x}, \mathbf{u}] = \sigma(\gamma + \mathbf{x}^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}^T\mathbf{B}^T\xi)$$

Recall that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $u \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, and that $\mathbf{x}$ and $\mathbf{u}$ are independent. In the following, ϕ denotes the density function of the standard normal distribution. Taking expectations, and with a_k, b_k denoting the k^{th} column of A, B respectively, we get

$$\begin{aligned} P[y = 1] &= \mathbb{E}[\sigma(\gamma + \mathbf{x}^T(\beta + \mathbf{A}^T\xi) + \mathbf{u}^T\mathbf{B}^T\xi)] \\ &= \mathbb{E}\left[\sigma\left(\gamma + \sum_{j=1}^p (x_j(\beta_j + \mathbf{a}_j^T\xi) + u_j\mathbf{b}_j^T\xi)\right)\right] \\ &= \int_{\mathbb{R}^{2p}} \sigma\left(\gamma + \sum_{j=1}^p (\tilde{x}_j(\beta_j + \mathbf{a}_j^T\xi) + \tilde{u}_j\mathbf{b}_j^T\xi)\right) \cdot \\ &\qquad \phi(\tilde{x}_1)\dots\phi(\tilde{x}_1)\,\phi(\tilde{u}_1)\dots\phi(\tilde{u}_p)\,d\tilde{x}_1\dots d\tilde{x}_1\,d\tilde{u}_1\dots d\tilde{u}_p \\ &= \int_{\mathbb{R}^p} g\left(\gamma + \sum_{j=1}^p \tilde{x}_j(\beta_j + \mathbf{a}_j^T\xi)\right)\phi(\tilde{x}_1)\dots\phi(\tilde{x}_1)\,d\tilde{x}_1\dots d\tilde{x}_1 \\ &= \mathbb{E}\left[g\left(\gamma + \sum_{j=1}^p x_j(\beta_j + \mathbf{a}_j^T\xi)\right)\right] = \mathbb{E}\left[g\left(\gamma + \mathbf{x}^T(\beta + \mathbf{A}^T\xi)\right)\right] \end{aligned} \tag{3.19}$$

with, writing $\mathbf{v} = \mathbf{B}^T\xi$,

$$g(z) = \int_{\mathbb{R}^p} \sigma\left(z + \sum_{j=1}^{p} \tilde{u}_j \mathbf{b}_j^T \xi\right) \phi(\tilde{u}_1) \dots \phi(\tilde{u}_p)\, d\tilde{u}_1 \dots d\tilde{u}_p = \mathbb{E}\left[\sigma\left(z + \sum_{j=1}^{p} u_j v_j\right)\right]$$

Note that for each $1 \leq j \leq p$, $u_j v_j$ has distribution $\mathcal{N}(0, v_j^2)$ so that $\sum u_j v_j \sim \mathcal{N}(0, \sum v_j^2)$. Thus, we can write

$$g_{\|\mathbf{B}^T\xi\|}(z) = g(z) = \mathbb{E}[\sigma(z + W\|\mathbf{v}\|)] = \mathbb{E}[\sigma(z + W\,\|\mathbf{B}^T\xi\|)]$$

with $W \sim \mathcal{N}(0,1)$ and where we use the subscript to highlight the dependence of g on $\|\mathbf{B}^T\xi\|$ and hence also to n, p, d that grow to ∞.

Using (3.19), we can rewrite the distribution of $\boldsymbol{Y}$ as

$$\boldsymbol{Y} \sim \mathrm{Rad}(g_{\|\mathbf{B}^T\xi\|}(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi)))$$

where $g_{\|\mathbf{B}^T\xi\|}$ is applied component-wise.

We managed to transport the effect of Z into the link function. We now want to find a RV with fixed link function (that does not depend on n) that is close enough to Y. Recall that $\mathbf{c}$ is the limit of $\mathbf{B}^T\xi$. We show that $g_{\|\mathbf{B}^T\xi\|}$ converges to $g_{\|\mathbf{c}\|} : \mathbb{R}^d \to \mathbb{R}^d$, $z \mapsto \mathbb{E}[\sigma(z + W\,\|\mathbf{c}\|)]$ in L^∞ as $n, p, d \to \infty$. Since the link functions are applied component-wise, we can focus on the one-dimensional case. Let $z \in \mathbb{R}$. We observe, again using that σ is K-Lipschitz:

$$\begin{aligned}
|g_{\|\mathbf{B}^T\xi\|}(z) - g_{\|\mathbf{c}\|}(z)| &= \left|\mathbb{E}\left(\sigma(z + W\,\|\mathbf{B}^T\xi\|) - \sigma(z + W\,\|\mathbf{c}\|)\right)\right| \\
&\leq \mathbb{E}\left|\sigma(z + W\,\|\mathbf{B}^T\xi\|) - \sigma(z + W\,\|\mathbf{c}\|)\right| \leq K\left(\|\mathbf{B}^T\xi\| - \|\mathbf{c}\|\right)\mathbb{E}\,|W|
\end{aligned}$$

The last expression tends to 0 as $n, p, d \to \infty$ because $\|\mathbf{B}^T\xi\| \to \|\mathbf{c}\|$, and does not depend on z. Hence

$$\|g_{\|\mathbf{B}^T\xi\|} - g_{\|\mathbf{c}\|}\|_\infty \to 0 \text{ as } n, p, d \to \infty \tag{3.20}$$

Let us analyze the consequence of this convergence on the response variable. Let

$$\boldsymbol{Y}_g \sim \mathrm{Rad}(g_{\|\mathbf{c}\|}(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi))).$$

We have

$$\begin{aligned}
P[Y_i \neq Y_g^i] &= \left|\mathbb{E}[g_{\|\mathbf{B}^T\xi\|}(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) - g_{\|\mathbf{c}\|}(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi)]\right| \\
&\leq \mathbb{E}\left|g_{\|\mathbf{B}^T\xi\|}(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi) - g_{\|\mathbf{c}\|}(\gamma_i + \mathbf{x}_i^T(\beta + \mathbf{A}^T\xi)\right| \leq \|g_{\|\mathbf{B}^T\xi\|} - g_{\|\mathbf{c}\|}\|_\infty
\end{aligned}$$

Now, from (3.20), and since i is arbitrary, we get

$$P[\boldsymbol{Y} \neq \boldsymbol{Y}_g] \to 0 \text{ as } n, p, d \to \infty$$

Hence, we can use the same coupling argument as in Section 3.1 and we are done if the result holds for $\boldsymbol{Y}_g$. That is the case because we managed to get rid of the unobserved variable and hence fall into the case of Chapter 2, with a different link function than the sigmoid σ. However, one can check that the whole development of Chapter 2 holds for any link function. □

As for the case $\|\mathbf{B}^T\xi\| \to 0$, let us derive a condition that guarantees $\mathbf{B}^T\xi \to \mathbf{c} \in \mathbb{R}^p, \mathbf{c} \neq \mathbf{0}$. It is easy to see that there exists $\mathbf{c} \in \mathbb{R}^p, \mathbf{c} \neq \mathbf{0}$ such that $\mathbf{B}^T\xi \to \mathbf{c}$ if and only if there exists $k \in \mathbb{R}^p, k \neq \mathbf{0}$ such that $\mathbf{B}^T \to k$. The latter is itself equivalent to

$$\exists \mathbf{M} \in \mathbb{R}^{d\times d}, \mathbf{M} \neq \mathbf{0} \text{ such that } \mathbf{B}\mathbf{B}^T \to \mathbf{M} \tag{3.21}$$

Now, recalling (3.12), (3.21) is equivalent to

$$\exists \mathbf{H} \in \mathbb{R}^{d\times d}, \mathbf{H} \neq \mathbf{I}_d \text{ such that } \Sigma_{Z,X}\Sigma_{X,Z} \to \mathbf{H} \tag{3.22}$$

Since $\Sigma_{Z,X} = \Sigma_{X,Z}^T$, we obtain

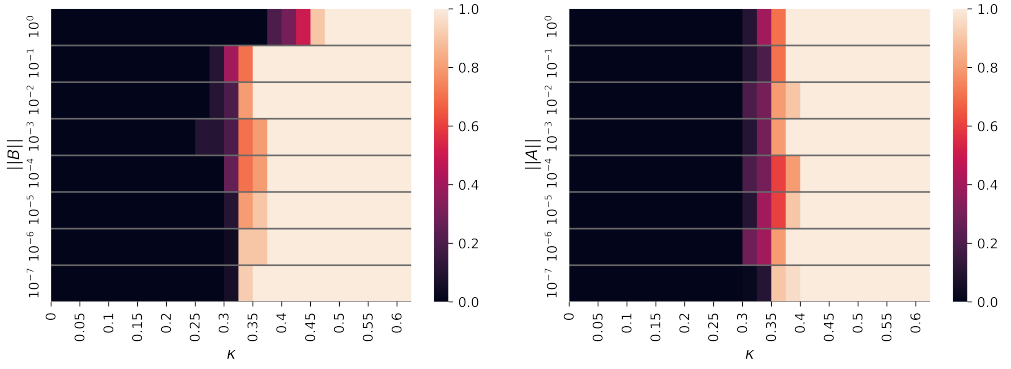
$$\mathbf{B}^T\xi \to \mathbf{c} \in \mathbb{R}^p, \mathbf{c} \neq \mathbf{0} \;\Leftrightarrow\; \Sigma_{X,Z} \text{ converges to a non-orthogonal matrix.}$$

3.4 Empirical results

We use the same procedure as in Section 2.4 to illustrate the theoretical results of this chapter. Here, we add unobservable features, so we generate $(\mathbf{x}_i, y_i)$ according to (3.1), that we recall here

$$\boldsymbol{Y} \sim \text{Rad}(\sigma(\gamma + \boldsymbol{X}(\beta + \mathbf{A}^T\xi) + \boldsymbol{U}\mathbf{B}^T\xi)).$$

$\boldsymbol{Z} = \boldsymbol{X}\mathbf{A}^T + \boldsymbol{U}\mathbf{B}^T$ being unobservable, we still use the same linear program (2.24) to determine the existence of the MLE. We want to visualize the impact of $\mathbf{B}$ and $\mathbf{A}$ that we have assessed. Thus, we fix $n = d = 500$ and we estimate the probability that the MLE does not exist for different values of $\|\mathbf{B}\|$ and $\|\mathbf{A}\|$.



(a) Strong correlation **(b)** Weak correlation

Figure 3.1: P[MLE does not exist] for latent model, $n = 500$

We first consider the set-up of Assumption 3.1.1 and Theorem 3.1.1. We run our simulations for eight different values of $\|\mathbf{B}\|$. Each time we fix the norms of the other parameters so that $\nu^2 = 2.5$. We observe on Figure 3.1a that the phase transition becomes sharper as $\|\mathbf{B}\|$ decreases, i.e. as the correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ becomes larger, as expected. Moreover, the threshold appears to be close to the theoretical

value $h \approx 0.31$ for $\nu^2 = 2.5$ (see [8] for theoretical curve) only from $\|\mathbf{B}\| = 10^{-6}$ and below. This seems to confirm the necessity of Assumption 3.1.1, since the latter requires $\|\mathbf{B}\| = o\left(n^{-\frac{3}{2}} d^{-\frac{1}{2}}\right) \overset{d=n}{=} o\left(n^{-2}\right)$ and we have for $n = 500$, $n^{-2} = 4 \times 10^{-6}$. The conclusions of Figure 3.1b are similar: the phase transition is stronger and close to the theoretical threshold for small values of $\|\mathbf{A}\|$ (weak correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$) and tend to validate the rate of decay that we assume in Assumption 3.2.1.

Chapter 4

Conclusion

We presented a proof of the existence of the phase transition and used it to extend the result to other settings, in a misspecified model with latent variable. We managed to do so by transposing the problem back to a model for which the phase transition is known. Indeed, if we have a response variable $\boldsymbol{Y}_1$ for which there is a phase transition, we proved that it is sufficient for our response variable to be close enough to $\boldsymbol{Y}_1$ in each of the observation with high probability, while the dimension and thus the number of observations explodes. It is likely that this approach can be used to show phase transition in other settings of the logistic regression. Future research could also be directed towards showing that there is a double descent phenomenon besides the phase transition in the cases that we presented.

Bibliography

[1] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and the double descent curve," *Communications on Pure and Applied Mathematics*, 06 2021.

[2] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, dec 2021.

[3] T. Hastie, A. Montanari, S. Rosset, and R. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," 03 2019, preprint on webpage at https://arxiv.org/abs/1903.08560.

[4] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1907378117

[5] P. Nakkiran, "More Data Can Hurt for Linear Regression: Sample-wise Double Descent," *arXiv e-prints*, p. arXiv:1912.07242, Dec. 2019.

[6] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Fundamental limits of ridge-regularized empirical risk minimization in high dimensions," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 2773–2781. [Online]. Available: https://proceedings.mlr.press/v130/taheri21a.html

[7] Z. Deng, A. Kammoun, and C. Thrampoulidis, "A model of double descent for high-dimensional binary linear classification," *Information and Inference: A Journal of the IMA*, 04 2021.

[8] E. Candès and P. Sur, "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression," *The Annals of Statistics*, vol. 48, pp. 27–42, 02 2020.

[9] W. Tang and Y. Ye, "The existence of maximum likelihood estimate in high-dimensional binary response generalized linear models," *Electronic Journal of Statistics*, vol. 14, no. 2, pp. 4028 – 4053, 2020. [Online]. Available: https://doi.org/10.1214/20-EJS1766

[10] D. Amelunxen, M. Lotz, M. McCoy, and J. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference*, vol. 3, 03 2013.

[11] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012, p. 210–268.