High-dimensional analysis of ridge regression for non-identically distributed data with a variance profile

Jérémie Bigot & Issa-Mbenard Dabo & Camille Male

Institut de mathématiques de Bordeaux & CNRS (UMR 5251)
Université de Bordeaux, France
jeremie.bigot@math.u-bordeaux.fr; issa-mbenard.dabo@math.u-bordeaux.fr;
camille.male@math.u-bordeaux.fr

Résumé. Les modèles de régression linéaire sont souvent étudiés dans le contexte de données indépendantes et identiquement distribuées. Nous proposons d'étudier ces modèles pour des données indépendantes mais non-identiquement distribuées. Dans ce but, nous supposons que l'ensemble des prédicteurs observés constitue une matrice aléatoire avec un profil de variance. En supposant un modèle à effets aléatoires, nous étudions le risque de prédiction de l'estimateur ridge pour la régression linéaire avec un tel profil de variance. Nous montrons un équivalent déterministe de ce risque, qui s'avère être un bon estimateur en grande dimension. Nous proposons également un équivalent déterministe du degré de liberté de l'estimateur ridge dans ce cadre. Nos travaux mettent aussi en évidence l'émergence du phénomène de double descente pour l'estimateur des moindres carrés de norme minimale lorsque le paramètre de régularisation ridge tend vers zéro. Les preuves de nos résultats s'appuient sur des outils issus de la théorie des matrices aléatoires en présence d'un profil de variance qui n'ont pas été considérés jusqu'à présent pour étudier les modèles de régression. Des expériences numériques sont fournies pour montrer l'exactitude de ces équivalents déterministes sur le calcul du risque de prédiction pour la régression ridge et des données non-identiquement distribuées.

Mots-clés. Régression ridge; Degrés de liberté; Double descente; Profil de variance; Hétéroscédasticité; Matrices aléatoires; Equivalents déterministes.

Abstract. Linear regression models are often studied in the context of independent and identically distributed data. We propose to investigate these models for independent but non-identically distributed data. To this end, we suppose that the set of observed predictors (or features) is a random matrix with a variance profile. Assuming a random effect model, we study the prediction risk of the ridge estimator for linear regression with such a variance profile. We provide a deterministic equivalent of this risk, which proves to be a good estimator in high-dimension. We also propose a deterministic equivalent of the degree of freedom of the ridge estimator in this setting. Our work also highlights the emergence of the double descent phenomenon for the minimum norm least-squares estimator when the ridge regularization parameter goes to zero. The proofs of our results are based on tools from random matrix theory in the presence of a variance profile that have not been considered so far to study regression models. Numerical experiments are provided to show the accuracy of the aforementioned deterministic equivalents on the computation of the prediction risk of ridge regression for non-identically distributed data.

Keywords. Linear ridge regression; Degrees of freedom; Double descent; Variance profile; Heteroscedasticity; Random Matrices; Deterministic equivalents.

1 Introduction

High dimensionality is a subject of interest in the field of statistics, especially in regression problems, driven by the advent of big data. This context gives rise to unexpected phenomena and contradictions with established statistical heuristics when the dimension p of the predictors is fixed and the number n of observations tends to infinity. These phenomena appear particularly in the context of linear regression. Indeed, as the sample size and dimension of acquired data increase, the study of this model is different from the classical framework. Indeed, in the asymptotic regime where $\min(n,p) \to +\infty$ and $\frac{p}{n} \to c > 0$, one can notably mention the obsolescence of certain estimators, the occurrence of double descent, or examples of overfitting. In this asymptotic setting, using tools from random matrix theory (RMT), many authors have therefore focused on the consequences of high-dimensionality on linear regression, see e.g. [DW18, Bac23, HMRT22, LC18]. In this paper, we focus on the linear regression model

$$Y_n = X_n \theta_* + \varepsilon_n, \tag{1.1}$$

where X_n is $n \times p$ matrix of random predictors, $\varepsilon_n \in \mathbb{R}^n$ is a noise vector independent of X_n with $\mathbb{E}[\varepsilon_n] = 0$ and $\mathbb{E}[\varepsilon_n \varepsilon_n^T] = \sigma^2 I_n$, $\theta_* \in \mathbb{R}^p$ is a vector of unknown parameters, and $Y_n \in \mathbb{R}$ is the vector of observed responses.

Classically, the predictors are assumed to be independent and identically distributed (iid) data, meaning that the rows of the matrix X_n are independent vectors sampled from the same probability distribution. In this paper, we propose to depart from this assumption by considering the setting where the rows of X_n are independent but non-identically distributed. To this end, we suppose that X_n is expressed in the following form

$$X_n = \Upsilon_n \circ Z_n,$$

where $\circ$ denotes the Hadamard (entry-wise) product between two matrices, $Z_n = (Z_{ij})$ has iid centered entries with variance one, and $\Upsilon_n = (\gamma_{ij})$ is a deterministic matrix. The matrix $(\gamma_{ij}^2) \in \mathbb{R}^{n \times p}$ governs the variance of the entries of X_n, and it is called a variance profile. The motivation for studying linear regression using such a variance profile is to consider the setting where one has n independent pairs of observations $(Y_i, X_i)_{1 \le i \le n}$ (with $X_i = (X_{ij})_{1 \le j \le p}$) that are not necessarily identically distributed. Note that in the standard setting of iid data, one that

$$\gamma_{ij} = \gamma_j \quad \text{for all} \quad 1 \le i \le n, \text{ and } 1 \le j \le p.$$

The main goal of this paper is then to understand how assuming such a variance profile for X_n influences the statistical properties of ridge regression in the linear model (1.1) when compared to the standard assumption of iid observations. In this setting, our approach also allows to analyze the performances of the minimum norm least-squares estimator when the ridge regularization parameter goes to zero.

We consider the possibly high-dimensional context (with $p \geq n$) when the least squares estimator is not uniquely defined. Thus, we focus our analysis on the ridge regression estimator that is the minimizer of the following loss function

$$\hat{\theta}_\lambda = \underset{\theta \in \mathbb{R}^p}{\arg\min} \frac{1}{n}\|Y_n - X_n\theta\|^2 + \lambda\|\theta\|^2,$$

for some regularization parameter $\lambda > 0$. Regardless of the ratio between n and p, this estimator has the following explicit expression

$$\hat{\theta}_\lambda = (X_n^T X_n + n\lambda I_p)^{-1} X_n^T Y_n = X_n^T (X_n X_n^T + n\lambda I_n)^{-1} Y_n.$$

Our analysis also includes the study of the minimum norm least-squares estimator defined as

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\arg\min} \left\{ \|\theta\| \ : \ \theta \text{ minimizes } \frac{1}{n}\|Y_n - X_n\theta\|^2 \right\}$$

to which the ridge regression estimator converges when λ tends to zero. Note that the minimum norm least-squares estimator is also known to be the limit of gradient descent when started at zero [GLSS18].

To study the statistical performances of the ridge regression estimator, we analyse its predictive risk defined as

$$\hat{r}_\lambda(X_n) = \mathbb{E}[(\tilde{Y} - \tilde{X}^T\hat{\theta}_\lambda)^2 | X_n], \tag{1.2}$$

where $(\tilde{Y}, \tilde{X}) \in \mathbb{R} \times \mathbb{R}^p$ is independent from (Y_n, X_n) and satisfies

$$\tilde{Y} = \tilde{X}^T\theta_* + \tilde{\varepsilon}, \text{ with } \mathbb{E}[\tilde{\varepsilon}] = 0, \ \mathbb{E}[\tilde{\varepsilon}^2] = \sigma^2.$$

In the above formula, $\tilde{X} = \tilde{\Gamma}_p^{1/2}\tilde{Z}$ with $\tilde{Z} \in \mathbb{R}^p$ a random vector with iid centered entries and variance one and $\tilde{\Gamma}_p = \mathbb{E}[\tilde{X}\tilde{X}^T] = \mathrm{diag}(\tilde{\gamma}_1^2, \ldots, \tilde{\gamma}_p^2)$ denotes the variance profile of $\tilde{X}$. Note that the risk $\hat{r}_\lambda(X_n)$ is conditioned on the predictors X_n, and it is thus a random variable.

Following [DW18], we focus on a random-effect hypothesis that assumes that the components of the vector θ_* are drawn independently at random. As argued in [DW18], this corresponds to an average case analysis over a set of dense regression coefficients as opposed to the "sparsity hypothesis" [HTW15] or the "manifold hypothesis" [LHT23] that are other popular assumptions in high-dimensional linear regression. More precisely, the following random coefficients assumption is made throughout the paper.

Assumption 1.1 *The vector θ_* of regression coefficients is random, independent from X_n, $\tilde{X}$, ε_n and $\tilde{\varepsilon}$, with $\mathbb{E}[\theta_*] = 0$ and*

$$\mathbb{E}[\theta_*\theta_*^T] = \frac{\alpha^2}{p} I_p.$$

The above coefficient $\alpha > 0$ represents the average amount of signal strength in model (1.1).

1.1 Main contributions

Recall that the prediction of Y_n by ridge regression is

$$\hat{Y}_\lambda = X_n \hat{\theta}_\lambda = A_\lambda Y_n, \quad \text{where} \quad A_\lambda = X_n (X_n^T X_n + n\lambda I_p)^{-1} X_n^T.$$

Then, the so-called degrees of freedom (DOF) of the estimator $\hat{\theta}_\lambda$, that is defined as

$$\widehat{df}_1(\lambda) = \mathrm{Tr}[A_\lambda] = \mathrm{Tr}[\widehat{\Sigma}_n(\widehat{\Sigma}_n + \lambda I_p)^{-1}], \text{ where } \widehat{\Sigma}_n = \frac{1}{n} X_n^T X_n,$$

represents the so-called effective dimension of the linear estimator $\hat{Y}_\lambda$. Inspired by recent results from [Bac23] in the setting of iid data, a first contribution of this work is to prove the following deterministic equivalence of the DOF

$$\widehat{df}_1(\lambda) \sim df_1(\lambda), \quad \text{where} \quad df_1(\lambda) = \mathrm{Tr}[\Gamma_n(\Gamma_n + \kappa(\lambda))^{-1}], \tag{1.3}$$

with

$$\Gamma_n = \mathbb{E}[\widehat{\Sigma}_n] = \frac{1}{n} \mathrm{diag}\left(\sum_{i=1}^n \gamma_{i1}^2, \ldots, \sum_{i=1}^n \gamma_{ip}^2 \right),$$

and $\kappa(\lambda)$ is a diagonal matrix that depends upon the regularization parameter λ and the variance profile matrix.

Hence, the equivalence relation (1.3), to be understood as

$$\lim_{n\to\infty,\, p/n\to c} |\widehat{df}_1(\lambda) - df_1(\lambda)| = 0, \text{ almost surely},$$

indicates that the DOF of the ridge regression estimator for the empirical covariance matrix $\widehat{\Sigma}_n$ corresponds to the DOF computed with the population covariance matrix Γ_n and another additive regularization structure than λI_p that is given by the diagonal matrix $\kappa(\lambda)$ whose explicit expression is given in Section 3.

Then, the second and main contribution of the paper is to derive a deterministic equivalent of the predictive risk $\hat{r}_\lambda(X_n)$ in the case where the number of samples n and the dimension p tend to infinity at a proportional rate. This deterministic equivalent allows us to understand the influence of the ratio c on the predictive risk and to also analyze the effect of the signal strength α. We also study the convergence of the predictive risk as λ tends 0 to analyze the statistical properties of the minimum norm least square estimator. In this setting, it appears a phenomenon arising from the curse of dimensionality that is commonly known as double descent for iid data. This phenomenon contradicts the consensus heuristic that, when a model becomes over-parameterized, then the predictive risk increases due to overfitting of the training data and the model is no longer capable of generalizing. This double descent has been thoroughly studied in the case of high-dimensional linear regression using tool from RMT, see e.g. [HMRT22, Bac23, BHX20a] and references therein. In this paper, we show that it also occurs for non iid data with a variance profile.

As a third contribution, using synthetic data and illustrative examples of variance profile, we conduct various numerical experiments to verify the accuracy of our deterministic equivalents of the DOF and the predictive risk using finite samples. We also investigate the similarities and differences that exist between the standard setting of iid data and the one of non-identically distributed data with a variance profile.

2 Related works

2.1 High-dimensional linear regression from the random matrix perspective

In the setting where the sample size is comparable to the dimensionality of the observations, recent advances in random matrix theory (RMT) have been successfully applied to various inference problems in high-dimensional multivariate statistics, see e.g. [NPW21] for a recent overview. Many works have considered the high-dimensional analysis of the linear model using tools from RMT for iid data with a general covariance structure $\Sigma \in \mathbb{R}^{p\times p}$ (assumed to be a positive semi-definite matrix) that is for

$$X_n = Z_n \Sigma^{1/2}, \text{ for an } n \times p \text{ matrix } Z_n \text{ with iid centered entries having variance one.}$$

In particular, for such data, the study of the minimum norm least-squares estimator and the double descent behavior of the predictive risk has been considered in [HMRT22, Bac23, BHX20b, RMR21]. The analysis of the predictive risk of ridge regression using iid data with a general covariance structure has been studied in [DW18, Bac23], while previous works on the statistical analysis of ridge regression from the RMT perspective include [EK18, Dic16] and [CD11, TV04] for applications in wireless communication.

2.2 Linear regression for independent but non-identically distributed data

While the statistical analysis of linear regression for iid data with a general covariance structure is very well understood, the literature on the study of the linear model for non-identically distributed predictors appears to be rather scarce. A first analysis of maximum likelihood estimation in standard models (including linear regression) for independent but non-identically distributed data dates back to [Ber82]. More recent works [BBK+19, KBB+20] on statistical inference in linear regression in the so-called model-free framework allow to consider the setting on non-identically distributed predictors. However, to the best of our knowledge, the high-dimensional analysis of the linear model using non-identically distributed data has not considered so far.

In this paper, we build upon results from [HLN07] to construct deterministic equivalents of the Stieltjes transforms

$$g_{\hat{\mu}_n}(z) = \frac{1}{p}\operatorname{Tr}[(\widehat{\Sigma}_n - zI_p)^{-1}] \text{ and } g_{\tilde{\mu}_n}(z) = \frac{1}{n}\operatorname{Tr}[(\widetilde{\Sigma}_n - zI_n)^{-1}] \quad \text{for} \quad z \in \mathbb{C}\setminus\mathbb{R}^+,$$

of the empirical eigenvalue distribution $\hat{\mu}_n$ of $\widehat{\Sigma}_n$, and the empirical eigenvalue distribution $\tilde{\mu}_n$ of $\widetilde{\Sigma}_n = \frac{1}{n}X_nX_n^T$ respectively, when $X_n = \Upsilon_n \circ Z_n$ has a variance profile (γ_{ij}^2).

Proposition 2.1 *The Stieltjes transforms $g_{\hat{\mu}_n}(z)$ and $g_{\tilde{\mu}_n}(z)$ satisfy*

$$\lim_{n\to\infty,\, p/n\to c}\left(g_{\hat{\mu}_n}(z) - \frac{1}{p}\operatorname{Tr}[T(z)]\right) = 0, \quad \textit{for all} \quad z \in \mathbb{C}\setminus\mathbb{R}^+,$$

$$\lim_{n\to\infty,\, p/n\to c} \left(g_{\tilde{\mu}_n}(z) - \frac{1}{n}\operatorname{Tr}[\tilde{T}(z)] \right) = 0, \quad \textit{for all} \quad z \in \mathbb{C}\setminus\mathbb{R}^+,$$

where

$$T(z) = \operatorname{diag}(T_1(z),\ldots,T_p(z)) \quad \textit{and} \quad \tilde{T}(z) = \operatorname{diag}(\tilde{T}_1(z),\ldots,\tilde{T}_n(z)),$$

are diagonal matrices of size $p\times p$ and $n\times n$ respectively, whose diagonal elements are the unique solutions of the deterministic system of $p+n$ equations

$$T_j(z) = \frac{-1}{z\left(1+(1/n)\operatorname{Tr}[\tilde{D}_j\tilde{T}(z)]\right)} \quad \textit{for} \quad 1\le j\le p, \tag{2.1}$$

$$\tilde{T}_i(z) = \frac{-1}{z\left(1+(1/n)\operatorname{Tr}[D_iT(z)]\right)} \quad \textit{for} \quad 1\le i\le n, \tag{2.2}$$

where

$$\tilde{D}_j = \operatorname{diag}(\gamma_{1j}^2,\ldots,\gamma_{nj}^2) \quad \textit{and} \quad D_i = \operatorname{diag}(\gamma_{i1}^2,\ldots,\gamma_{ip}^2).$$

Moreover, $\frac{1}{p}\operatorname{Tr}[T(z)]$ and $\frac{1}{n}\operatorname{Tr}[\tilde{T}(z)]$ are the Stieltjes transforms of probability measures denoted as ν_n and $\tilde{\nu}_n$ that are the deterministic equivalents of $\hat{\mu}_n$ and $\tilde{\mu}_n$ respectively.

3 Main results

In this section, we derive deterministic equivalents for the DOF and the predictive risk of ridge regression. We also obtain a deterministic equivalent of the predictive risk of minimum norm least square estimation when the ridge regularization parameter tends to zero. We compare these results to those that are already known in the standard setting of iid data, and we highlight the emergence of the double descent phenomenon for non-iid data.

Assumption 3.1 *There exists $\delta>0$ such that, $\mathbb{E}[|Z_{ij}|^{4+\delta}]<+\infty$.*

Assumption 3.2 *There exists $\gamma_{\max}>0$ such that, $\sup_{n\ge 1}\max_{\substack{1\le i\le n\\ 1\le j\le n}}|\gamma_{ij}|<\gamma_{\max}$.*

Assumption 3.3 *There exists $\gamma_{\min}>0$ such that, $\forall n\ge 1$, $\min_{\substack{1\le i\le n\\ 1\le j\le n}}|\gamma_{ij}|\ge\gamma_{\min}$.*

3.1 Predictive risk

Theorem 3.1 *Consider the linear model (1.1). Then, under Assumptions (3.1) and (3.2), a deterministic equivalent of the predictive risk is*

$$r_\lambda(\Upsilon_n) = \sigma^2 + \frac{\sigma^2}{n}\operatorname{Tr}[\tilde{\Gamma}_pT(-\lambda)] + \lambda\left(\frac{\lambda\alpha^2}{p}-\frac{\sigma^2}{n}\right)\operatorname{Tr}[\tilde{\Gamma}_pT'(-\lambda)],$$

in the sense that it satisfies

$$\lim_{n\to\infty,\, p/n\to c} \hat{r}_\lambda(X_n) - r_\lambda(\Upsilon_n) = 0 \quad \text{a.s.}$$

Moreover, for $\lambda > 0$ *and* $\lambda_* = \frac{\sigma^2 p}{\alpha^2 n}$, $r_{\lambda_*}(\Upsilon_n) \leq r_\lambda(\Upsilon_n)$ a.s.

Note that our deterministic equivalent $r_\lambda(\Upsilon_n)$ of the predictive risk allows to derive an optimal choice λ_* for the regularization parameter that corresponds to the one obtained for iid data in [DW18], and that is independent of the variance profile. Theorem 3.1 also allows us to understand the behavior of $r_\lambda(\Upsilon_n)$ when $\lambda \to 0$ and $\lambda \to +\infty$ through the following corollary.

Corollary 3.1 *Under Assumptions (3.1), (3.2) and (3.3), the limit of the deterministic equivalent* $r_\lambda(\Upsilon_n)$ *for* $\lambda \to +\infty$ *and* $\lambda \to 0$ *are as follows*

$$\lim_{\lambda\to+\infty} r_\lambda(\Upsilon_n) = \frac{\alpha^2}{p}\mathrm{Tr}[\tilde{\Gamma}_p] + \sigma^2.$$

If $\frac{p}{n} < 1$, *then*

$$\lim_{\lambda\to 0} r_\lambda(\Upsilon_n) = \frac{\sigma^2}{n}\mathrm{Tr}[\tilde{\Gamma}_p T(0)] + \sigma^2.$$

If $\frac{p}{n} > 1$, *then*

$$\lim_{\lambda\to 0} r_\lambda(\Upsilon_n) = \frac{\alpha^2}{p}\mathrm{Tr}[\tilde{\Gamma}_p \kappa(0)(\tilde{\Gamma}_p + \kappa(0))^{-1}] + \frac{\sigma^2}{n}\mathrm{Tr}[\kappa'(0)\tilde{\Gamma}_p \Delta_n (\tilde{\Gamma}_p + \kappa(0))^{-2}] + \sigma^2,$$

where for $\lambda \in \mathbb{R}^+$

$$\kappa(\lambda) = \operatorname*{diag}_{1\leq j\leq p}\left(\frac{\mathrm{Tr}[\tilde{D}_j]}{\mathrm{Tr}[\tilde{D}_j \tilde{T}(-\lambda)]}\right) \quad \text{and} \quad \Delta_n = \frac{1}{n}\operatorname{diag}(\mathrm{Tr}[\tilde{D}_1], \ldots, \mathrm{Tr}[\tilde{D}_p])$$

4 Numerical experiments

In this section, we illustrate our results with numerical experiments. Figure 1 compares the predictive risk and its deterministic equivalent. The red curve represent the deterministic equivalent $r_\lambda(\Upsilon_n)$ whereas the black one represents the actual predictive risk $\hat{r}_\lambda(\Upsilon_n)$. On the other had, the dotted line depicts $(\tilde{Y} - \tilde{X}^T\hat{\theta}_\lambda)^2$, which is a realisation of the predictive risk. Figure 1 confirms that $r_\lambda(\Upsilon_n)$ is a good estimator of $\hat{r}_\lambda(\Upsilon_n)$ in high dimension as the black and the red curves coincide. The dotted line stays around the two other curves as they estimate $\hat{r}_\lambda(\Upsilon_n)$ which is the expectation of $(\tilde{Y} - \tilde{X}^T\hat{\theta}_\lambda)^2$.

Figure 2 represents the predictive risk with for different values of the ratio n/p, the solid lines depicts the case with $\lambda = 0$ and the dashed lines depicts the case with $\lambda = \lambda_* = \frac{\sigma^2 p}{\alpha^2 n}$. The double descent phenomenon is illustrated by the solid lines since the curves are increasing for $p/n < 1$ and decreasing for $p/n > 1$. This is related to Corollary 3.1 which states that the expression of $\lim_{\lambda\to 0} r_\lambda(\Upsilon_n)$ depends upon the value of the ratio p/n with respect to one. The performances of $\hat{\theta}_0$ and $\hat{\theta}_*$ seem similar when p/n is large whereas the performances obtained with $\hat{\theta}_*$ are much better than with those using $\hat{\theta}_0$.

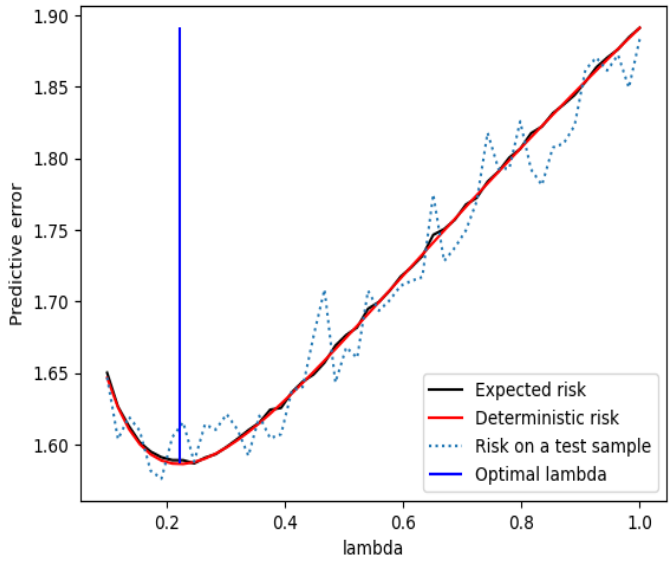

Figure 1: Comparison of $r_\lambda(\Upsilon_n)$ and $\hat{r}_\lambda(\Upsilon_n)$ with a doubly-stochastic variance profile (rows and columns sum up to the same constant value), $\alpha = 1.5$, $\sigma = 1$, $n = 800$ and $p = 500$,

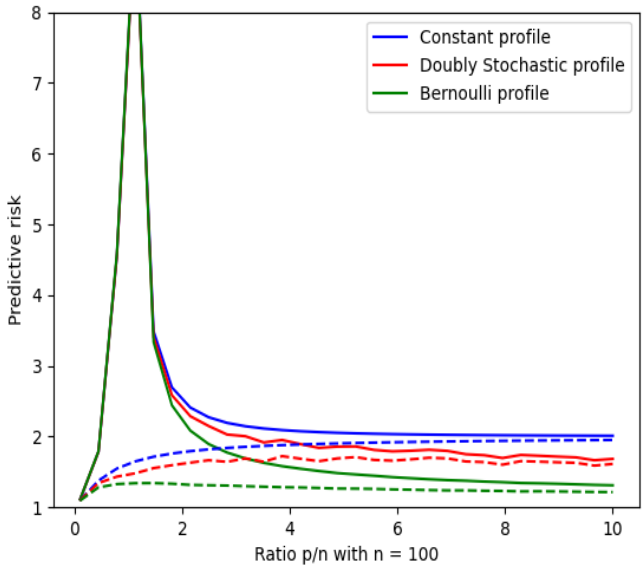

Figure 2: Double descent phenomenon for different variance profile with $\alpha = \sigma = 1$, $n = 100$ and p varying from 10 to 500. The Bernoulli variance profile corresponds to variance γ_{ij} randomly sampled from a Bernoulli distribution. The full lines correspond to $\lambda = 0$ and the dashed lines correspond to $\lambda = \lambda_*$.

5 Conclusion

In this paper, we have derived a deterministic equivalent of the DOF and the predictive risk of ridge (less) regression in a high-dimensional framework with a variance profile to handle the setting of non-iid data. The numerical experiments that we have conducted confirm that this deterministic equivalent accurately estimates the predictive risk in high-dimension. Our results also allow to understand how assuming such a variance profile for the data influences the statistical properties of ridge regression when compared to the standard assumption of iid observations. We hope that our approach on the use of variance profiles may lead to further research works on the statistical analysis of other estimators than ridge regression in more complex models with non-iid data.

References

[Bac23] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections, 2023.

[BBK+19] Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, and Linda Zhao. Models as Approximations II: A Model-Free Theory of Parametric Regression. *Statistical Science*, 34(4):545 – 565, 2019.

[Ber82] Rudolf Beran. Robust Estimation in Models for Independent Non-Identically Distributed Data. *The Annals of Statistics*, 10(2):415 – 428, 1982.

[BHX20a] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[BHX20b] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[CD11] Romain Couillet and Mérouane Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[Dic16] Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1 – 37, 2016.

[DW18] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018.

[EK18] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170:95–175, 02 2018.

[GLSS18] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas

Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841. PMLR, 10–15 Jul 2018.

[HLN07] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875 – 930, 2007.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[KBB+20] Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, Junhui Cai, Edward I. George, and Linda H. Zhao. Valid post-selection inference in model-free linear regression. *The Annals of Statistics*, 48(5):2953 – 2981, 2020.

[LC18] Zhenyu Liao and Romain Couillet. The dynamics of learning: A random matrix approach. In *International Conference on Machine Learning*, pages 3072–3081. PMLR, 2018.

[LHT23] Liangchen Liu, Juncai He, and Richard Tsai. Linear regression on manifold structured data: the impact of extrinsic geometry on solutions, 2023.

[NPW21] Jamshid Namdari, Debashis Paul, and Lili Wang. High-dimensional linear models: A random matrix perspective. *Sankhya A: The Indian Journal of Statistics*, 83(2):645–695, 2021.

[RMR21] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge(less) regression under general source condition. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 3889–3897. PMLR, 2021.

[TV04] Antonia Maria Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Found. Trends Commun. Inf. Theory*, 1(1), 2004.