Learning Linear Operator: Causality and Double Descent

TIANLE YANG

MARCH 2024

Learning Linear Operator: Causality and Double Descent

A dissertation submitted to
THE GRADUATE SCHOOL OF ENGINEERING SCIENCE
OSAKA UNIVERSITY
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN SCIENCE

By

TIANLE YANG

MARCH 2024

Abstract

This thesis delves into the intricacies of linear regression in machine learning, specifically focusing on its applications in diverse dimensions, causal discovery, and managing model complexity as follows:

We first study how dropout layers in neural networks can mitigate the double descent phenomenon. We demonstrate theoretically and empirically that optimal dropout in linear models can prevent this phenomenon. We estimate the true coefficients using a generalized ridge-type estimator and show that optimal dropout leads to a monotonic test error curve, even in nonlinear neural networks. This finding suggests the effectiveness of dropout regularization in managing risk curves and explains the absence of double descent in models employing similar regularization techniques.

Then we extend the Linear Non-Gaussian Acyclic Model (LiNGAM) to Functional LiNGAM (Func-LiNGAM), capable of handling infinite-dimensional data, such as fMRI and EEG datasets. This development addresses the limitations of the original LiNGAM in processing such complex datasets. We theoretically validate the identifiability of causal relationships in these high-dimensional spaces and employ functional principal component analysis to manage data sparsity. The effectiveness of Func-LiNGAM is demonstrated through synthetic and real fMRI data analysis.

In conclusion, this thesis presents a comprehensive exploration of linear regression in machine learning, contributing to both theoretical understanding and practical methodologies. By focusing on two interconnected areas—the application of Func-LiNGAM in infinite-dimensional spaces for causal discovery and the effective use of dropout regularization to address the double descent phenomenon—we have advanced the knowledge of linear regression's versatility. These studies offer valuable insights into the field and provide practical tools for dealing with complex, high-dimensional datasets. They showcase the potential of linear operators in advancing machine learning research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Causal Discovery with Functional Data Analysis: It's crucial to emphasize the significance of applying functional data analysis to neuroimaging data, particularly in EEG causality studies. Functional data, inherently infinite-dimensional, necessitates specialized approaches like Func-LiNGAM. Conventional models may lead to incorrect causal inferences, as standard methods may overlook the intricate structure of such data. Functional data analysis, through techniques like smoothing, reveals hidden patterns and connections, crucial for accurate causal discovery in EEG data. This approach, already impactful in various fields, offers a robust framework for unraveling complex causal relationships in brain activity, making it pivotal for advancements in neuroimaging and related research areas.

Double Descent: Dropout is a well-established regularization technique for training deep neural networks. Its primary objective is to prevent "co-adaptation" among neurons by randomly excluding them during training (Hinton et al., 2012). Dropout's effectiveness extends across various machine learning tasks, ranging from classification (Srivastava et al., 2014) to regression (Toshev and Szegedy, 2014). Notably, dropout played a vital role in the design of AlexNet (Krizhevsky et al., 2012), significantly outperforming its competitors in the 2012 ImageNet challenge. Given dropout's proven efficiency in mitigating overfitting (Srivastava et al., 2014) and its wide applicability, we propose that it may significantly mitigate the double descent phenomenon. This leads us to the following question:

Under what conditions and how does dropout mitigate the double descent phenomenon?

We acknowledge that the double descent phenomenon exists under both sample-wise and model-wise

conditions. This paper investigates its occurrence in both linear and nonlinear models to enhance test performance while avoiding unexpected non-monotonic responses. Eliminating the double descent phenomenon has indeed become a prominent research topic. For instance, ridge regularization can alleviate double descent (Nakkiran et al., 2021b), as can early stopping (Heckel and Yilmaz, 2021).

1.2 Organization and Contribution

The thesis is organized as follows: In Chapter 1, we briefly introduce the motivation of this thesis and present a summary of contributions. In Chapter 2, we propose an infinite-dimensional causal discovery framework for functional data. In particular,

Contribution of Chapter 2:

- We establish a framework for discovering causal orders for random vectors and functions, moving beyond the traditional focus on random variables.
- We theoretically prove that it is possible to identify the causal order under non-Gaussianity for random vectors (Theorem 5).
- We further demonstrate the identifiability of the causal order for non-Gaussian processes in infinite-dimensional Hilbert spaces (Theorem 8).
- To verify the validity of our method, we performed extensive experiments with simulated data as Table 2.1. Empirical results demonstrate the identifiability. The results show that it performs worse as the number of functions increases, which is reasonable. But as the sample size increases, it performs better. We need more data for larger dimensions, but the required amounts are still reasonable.

In Chapter 3, we theoretically prove that dropout can drop double descent. In particular,

Contribution of Chapter 3:

- **Eliminating the Sample-Wise Double Descent.** We empirically validate the monotonicity of the test error as the sample size increases (see Figure 3.1) and theoretically prove the monotonicity of the second-order Neumann series test error. We plan to detail the exact solution in future work.
- **Eliminating the Model-Wise Double Descent.** We empirically demonstrate the monotonicity of the test error as the model size increases.

- **Multi-layer CNN.** We provide empirical evidence showing that dropout can alleviate the double descent in multi-layer CNNs.

In Chapter 4, we summarize the proposed methods and provide a comprehensive discussion of their strengths and limitations. We also provide a discussion about possible directions for the follow-up work.

Chapter 2

Functional LiNGAM

2.1 Introduction

Numerous empirical sciences strive to uncover and comprehend causal mechanisms that underlie a wide range of natural phenomena and human social behavior. Causal discovery has a wide range of applications, including biology (Sachs et al., 2005), climate studies (Ebert-Uphoff and Deng, 2012), and healthcare (Lucas et al., 2004). When determining the cause-and-effect relationship between variables, such as X_1 and X_2, detecting their dependence alone is insufficient for determining the causal direction, i.e., whether $X_1 \to X_2$ or $X_2 \to X_1$.

Causal analysis based on the LiNGAM, proposed by Shimizu et al. (2006), addresses this challenge by identifying the causal directions in linear relationships. Specifically, supposing there is no latent common cause for X_1 and X_2, it figures out the causal direction between them by checking which of the following two models holds: $X_2 = aX_1 + \epsilon$ and $X_1 = a'X_2 + \epsilon'$, where $X_1 \perp\!\!\!\perp \epsilon$ and $X_2 \perp\!\!\!\perp \epsilon'$ and $a, a' \in \mathbb{R}$.[1] The sufficient and necessary condition of the identifiability is that LiNGAM requires at most one of the noise terms (including the root causes) to be non-Gaussian to make it possible to identify unique causal directions. Notably, zero correlation is synonymous with independence in Gaussian variables, making it impossible to distinguish between the two causal models when X_1 and X_2 are Gaussian.

In this linear, Gaussian case, one can only end up with the so-called Markov equivalence class (all members of the equivalence class have the same conditional independence relations), even when adhering to faithfulness assumption (Spirtes et al., 2000; Pearl, 2000). For instance, the three Directed Acyclic Graphs (DAGs) connecting three variables, such as X_1, X_2, X_3, in Fig. 2.1 are Markov equivalent be-

[1] $X_1 \perp\!\!\!\perp X_2$ denotes the independence of X and Y.

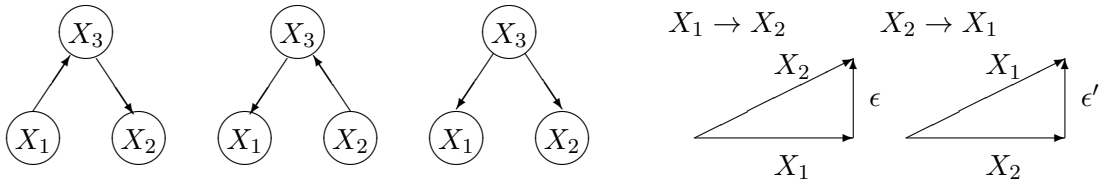

Figure 2.1: *Structure learning methods like the PC algorithm cannot distinguish these causal graphs that have the identical probability distribution $P(X_1X_2)P(X_2X_3)/P(X_2)$ (Left). But LiNGAM can differentiate them via the non-Gaussian assumption (Right).*

cause they have the same distribution in the Gaussian case. Here faithfulness refers to the property where any independence relations observed in the data can be explained by the causal relationships represented in the graphical model. However, this is not the case anymore in non-Gaussian cases. Due to this significant advancement, LiNGAM can uniquely determine the causal ordering among variables solely based on observational data, even without assuming faithfulness.

For the converse, the Darmois-Skitovich theorem (D-S) is employed to prove the identifiability of causal direction. From D-S, if at least one of the variables X_1 and X_2 are non-Gaussian, then only one unique direction of $X_1 \to X_2$ and $X_2 \to X_1$ exists. The Darmois-Skitovich (D-S) theorem originally focused on one-dimensional Gaussian random variables. Interestingly, Ghurye and Olkin (1962) expanded its application to random vectors, while Myronyuk (2008) generalized it to Banach spaces. In our paper, random elements that take values in a Banach space are called random functions.

This paper establishes a novel functional framework for modeling the causal structure of multivariate functional data, which is the realization of random functions. It is important to note that functional data is inherently infinite-dimensional. If we apply conventional models such as PC or LiNGAM directly, we might incorrectly identify causal relationships, as shown in Fig. 2.2. To demonstrate the benefits of functional data analysis (Ramsay and Silverman, 2005), we provide an example in Fig. 2.3, illustrating how smoothing the discrete points enables us to capture missing information. Functional data analysis has gained prominence in diverse fields, including neuroimaging (Wainwright, 2019), finance (Tsay and Pourahmadi, 2017), and genetics (Wu and Xu, 2020). Exploring causal relationships among random functions presents a significant challenge in multivariate functional data analysis.

This research is motivated by brain-effective connectivity (Advani et al., 2020), which explores the directional effects between neural systems. Learning brain-effective connectivity networks from electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), and electrocorticographic imaging (ECoG) records is crucial for understanding brain activities and neuron responses. Modeling these multivariate processes and accurately estimating effective connections between brain areas pose significant

challenges due to the continuous nature of the data and the need to treat the data as functions, considering the small time intervals between adjacent sample points. Previous studies, such as Qiao et al. (2019), has explored the functional aspects of the Gaussian graphical model by estimating the inverse covariance matrix. Nakkiran (2019) introduced the functional directional relationships under Gaussian assumption, enabling the determination of a directed acyclic graph (DAG) up to its equivalence class. The previous version of this paper Yang and Suzuki (2022) discussed the identifiability without considering one important point for functional data: the covariance operator's non-invertibility. Moreover, the previous algorithm for functional data is not accurate because it only tests the independence of every principal component rather than the whole random vector. Zhou et al. (2022) developed a novel Bayesian network model for multivariate functional data. Roy et al. (2023) considers the directed cyclic model for functional data. In contrast to previous works, our approach differs in that we first establish the identifiability of random vectors. Subsequently, we demonstrate the identifiability of random functions considering the non-invertibility and extend it into multivariate scenarios. **Our contributions are as follows**:

- We establish a framework for discovering causal orders for random vectors and functions, moving beyond the traditional focus on random variables.
- We theoretically prove that it is possible to identify the causal order under non-Gaussianity for random vectors (Theorem 5).
- We further demonstrate the identifiability of the causal order for non-Gaussian processes in infinite-dimensional Hilbert spaces (Theorem 8).
- To verify the validity of our method, we performed extensive experiments with simulated data as Table 2.1. Empirical results demonstrate the identifiability. The results show that it performs worse as the number of functions increases, which is reasonable. But as the sample size increases, it performs better. We need more data for larger dimensions, but the required amounts are still reasonable.

The structure of the paper is as follows. Section 2.2 provides the necessary background information to comprehend this paper. This includes introducing the LiNGAM, infinite-dimensional Hilbert spaces, and random elements (random functions). Section 2.3 and 2.4 present the main theoretical results extending the LiNGAM and outlines the corresponding procedure. Section 2.5 and 2.6 present the experimental results. Section 2.7 summarizes the key points.

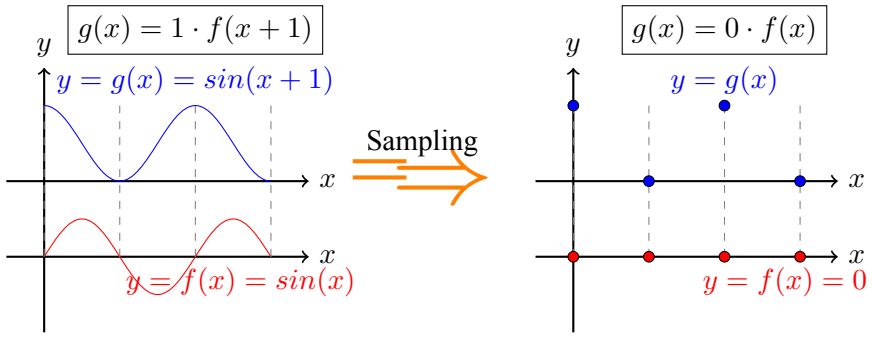

Figure 2.2: ***Illustration of why the original LiNGAM does not work***. *The Left Graph: original two stochastic processes with their causal relationships; The Right Graph: a possible situation where we sample the time series but miss the causal relationship.*

2.2 Background

2.2.1 Linear Non-Gaussian Acyclic Model (LiNGAM)

This section introduces the concept of the LiNGAM for inferring the causal relationships among random variables.

Suppose two random variables $X_1, X_2 \in \mathbb{R}$, we want to identify the causal directions of either $X_1 \to X_2$ or $X_2 \to X_1$. More specifically, our analysis assumes that X_1 and X_2 are linearly related and have zero means. Such as

$$X_1 = e_1, \quad X_2 = aX_1 + e_2 \,, \tag{2.1}$$

$$X_2 = e_1', \quad X_1 = a'X_2 + e_2' \tag{2.2}$$

with $a, a' \in \mathbb{R}$ and $\mathbb{E}[\epsilon] = \mathbb{E}[\epsilon'] = 0$. To be simple, we let

$$a \neq 0 \,, \text{ or } a' \neq 0 \,, \tag{2.3}$$

to avoid $X_1 \perp\!\!\!\perp X_2$. Specifically, in the context of LiNGAM, under the assumption of the noise terms, denoted as ϵ and ϵ', are independent of their respective covariates, X_1 and X_2 in (2.1) and (2.2). Therefore, based on the condition of $X_1 \perp\!\!\!\perp e_2$ or $X_2 \perp\!\!\!\perp e_2'$, we determine the true causal model to be either (2.1) or (2.2). It may initially appear that distinguishing between (2.1) and (2.2) is not possible, in other words, X_1 and X_2 could satisfy both equations for certain values of a, a', e_2, and e_2', where $X_1 \perp\!\!\!\perp e_2$ and $X_2 \perp\!\!\!\perp e_2'$. LiNGAM claims that this inconvenience occurs if and only if X_1 and X_2 are Gaussian. In other words, we can identify (2.1) and (2.2) if and only if at least one of X_1 and X_2 are non-Gaussian.

For the sufficient part, we show that if variables are both Gaussian, then causal order is unidentifiable. Suppose X_1, X_2 both are normally distributed, and the model (2.1) with $X_1 \perp\!\!\!\perp e_2$ is true for certain a

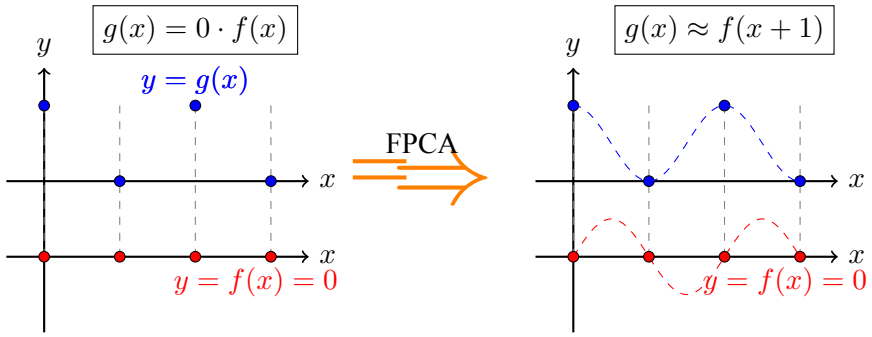

Figure 2.3: ***Illustration of Why Func-LiNGAM Work****. (Smoothing: Functional data analysis) The Left Graph: with the worst situation when we sample the time series and miss the causal relationship, where we get g and f have no causal relationship. The Right Graph: we can complete the discrete points into smooth curves with the Functional data analysis, capturing extra information when choosing suitable bases.*

and ϵ. Let σ_1^2, σ_2^2 be the variances of e_1 and e_2. Then, from $\mathbb{E}[e_1 e_2] = 0$, we have

$$e_1' = ae_1 + e_2 \tag{2.4}$$

$$e_2' = e_1 - a'e_1' = e_1 - a'(ae_1 + e_2) = (1 - a'a)e_1 - a'e_2 \,, \tag{2.5}$$

and $\mathbb{E}[e_1' e_2'] = (1 - a'a)\sigma_1^2 - a'\sigma_2^2$, which means that choosing

$$a' = \frac{a\sigma_1^2}{a^2\sigma_1^2 + \sigma_2^2} \tag{2.6}$$

will make the $\mathbb{E}[e_1' e_2'] = 0$ too. We call W and Z jointly Gaussian if the two random variables can be represented as $\begin{bmatrix} Z \\ W \end{bmatrix} = A \begin{bmatrix} U \\ V \end{bmatrix}$ where $A \in \mathbb{R}^{2\times 2}$ and U, V are independent Gaussian.

The well-known property states that independence is equivalent to zero correlation for jointly Gaussian variables[2]. By checking e_1' and e_2' belonging to joint Gaussian distribution, we can conclude that e_1' is independent of e_2'. Consequently,(2.2) holds with $X_2 \perp\!\!\!\perp \epsilon'$ for the corresponding a', ϵ'.

For the necessary part, assume that $X \perp\!\!\!\perp \epsilon$ for (2.1) and $Y \perp\!\!\!\perp \epsilon'$ for (2.2) both hold simultaneously for certain $a, a', \epsilon, \epsilon'$, where a' satisfies (2.6). Therefore, this means that $a, a' \neq 0$ due to (2.3) and (2.6). Now note the statement as follows:

Proposition 1 (Skitivic (1953); Darmois (1953))**.** Let $m \geq 2$ and independent random variables $\xi_1, \ldots, \xi_m \in \mathbb{R}$. Let two linear form $L_1 = \sum_{i=1}^m \alpha_i \xi_i$ and $L_2 = \sum_{i=1}^m \beta_i \xi_i$, if $L_1 \perp\!\!\!\perp L_2$, for $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m \in \mathbb{R}$. Then the random variable ξ_i such that $\alpha_i \beta_i \neq 0$ belongs to Gaussian for $i = 1, \ldots, m$.

[2]Suppose Z and W be binary taking ± 1 equiprobably and zero-mean Gaussian. Then, ZW and Z are not jointly Gaussian. Even though $\mathbb{E}[ZW \cdot Z] = \mathbb{E}[W] \cdot \mathbb{E}[Z^2] = 0$ but they are not independent.

Following (2.4)(2.5)(2.6) and the Proposition 1, then

$$\begin{aligned}(e_1, e_2, a, 1, 1 - aa', -a') &= (\xi_1, \xi_2, \alpha_1, \alpha_2, \beta_1, \beta_2) \\ &= (X, \epsilon, a, 1, \frac{\sigma_2^2}{a^2\sigma_1^2 + \sigma_2^2}, -\frac{a\sigma_1^2}{a^2\sigma_1^2 + \sigma_2^2})\end{aligned}$$

By combining (2.3), X_1, ϵ belong to Gaussian, then X_2 is also Gaussian-distributed.

Proposition 2 (Shimizu et al. (2011))**.** Assuming (2.3), we can identify the causal order using LiNGAM if at least one of two random variables belongs to non-Gaussian.

We can also identify the causal orders among multiple random variables. Suppose there are three linearly related random variables X_1, X_2, X_3 with zero means. Then, six potential causal orders exist, for instance, $X_2 \to X_1 \to X_3$, and $X_3 \to X_2 \to X_1$. First, we determine the top of them. Assuming X_1 is independent of $\{X_2 - aX_1, X_3 - a'X_1\}$ for $a, a' \in \mathbb{R}$, which means X_1 is the top variable. Furthermore, suppose that $X_2 - aX_1$ is independent of $X_3 - a'X_1 - a''(X_2 - aX_1)$ for some $a'' \in \mathbb{R}$, then regarding the X_2 as the middle and X_3 as the bottom. We obtain the causal order $X_1 \to X_2 \to X_3$. Following the steps, we can identify the causal order for X_1, X_2, X_3. Furthermore, we can estimate the causal order for an arbitrary number of random variables like

$$X_i = \sum_{j=1}^{i-1} b_{i,j} X_j + e_i$$

where $b_{i,j} \in \mathbb{R}$ and noise e_i is non-Gaussian for p random variables $X_1, \ldots, X_p$.

2.2.2 Hilbert Spaces

A Banach space is a complete normed vector space where completeness ensures that all Cauchy sequences converge within the space. It combines linearity, completeness, and the norm to provide a framework for studying mathematical structures and functions. More precisely, in our context, we consider the set of functions as a Hilbert space, denoted by $\mathscr{H}$. A Hilbert space is a Banach space equipped with an inner product that induces the norm, ensuring completeness.

We define a linear operator $T_{21} : \mathscr{H}_1 \to \mathscr{H}_2$ over $\mathbb{R}$ as a mapping that satisfies the linearity property: $T_{21}(\alpha f + \beta g) = \alpha T_{21} f + \beta T_{21} g$ for $f, g \in \mathscr{H}_1$ and $\alpha, \beta \in \mathbb{R}$. Furthermore, T_{21} is said to be bounded if there exists a positive constant C such that $\|T_{21} f\|_2 \leq C\|f\|_1$ holds for all $f \in \mathscr{H}_1$. Here, $\|\cdot\|_1, \|\cdot\|_2$ denote the norms within $\mathscr{H}_1, \mathscr{H}_2$, respectively.

For any bounded operator $T_{21} : \mathscr{H}_1 \to \mathscr{H}_2$, there exists its adjoint operator or dual operator, a unique bounded linear operator $T_{21}^* : \mathscr{H}_2 \to \mathscr{H}_1$ such that the following equality holds: $\langle T_{21} f_1, f_2 \rangle_2 =$

$\langle f_1, T_{21}^* f_2 \rangle_1$ for $f_1 \in \mathscr{H}_1$ and $f_2 \in \mathscr{H}_2$. The operator T_{21}^* is the adjoint operator of T_{21}. If $T_{21} = T_{21}^*$, we say that T_{21} is self-adjoint. Moreover, if the dimension of $\mathscr{H}$ is finite, the self-adjoint operator T_{21} is symmetric.

2.2.3 Random functions

Functional data analysis involves considering each individual element of data as a random function. These functions are defined over a continuous physical continuum, which is typically time but can also be spatial location, wavelength, probability, or other dimensions. Functionally, these data are infinite-dimensional. Random functions can be interpreted as random elements that take values in a Hilbert space or as stochastic processes. The former approach provides mathematical convenience, while the latter is more suitable for practical applications. These two perspectives align when the random functions are continuous and satisfy a mean-squared continuity condition.

Formally speaking, if a mapping $X : \Omega \to \mathbb{R}$ is measurable from a probability space $(\Omega, \mathcal{F}, \mu)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then it is a random variable:

$$B \in \mathcal{B}(\mathbb{R}) \Longrightarrow \{\omega \in \Omega | X(\omega) \in B\} \in \mathcal{F} \ ,$$

with the Borel sets $\mathcal{B}(\mathbb{R})$. Similarly, if $\chi : \Omega \to \mathscr{H}$ is measurable from $(\Omega, \mathcal{F}, \mu)$ to $(\mathscr{H}, \mathcal{B}(\mathscr{H}))$, then it is a random function (or random element) in a Hilbert space $\mathscr{H}$:

$$B \in \mathcal{B}(\mathscr{H}) \Longrightarrow \{\omega \in \Omega | X(\omega) \in B\} \in \mathcal{F} \ ,$$

with the Borel sets $\mathcal{B}(\mathscr{H})$ w.r.t. the norm of $\mathscr{H}$. Let E be one set, we suppose that every entry f of $\mathscr{H}$ is a function $f : E \ni x \mapsto f(x) \in \mathbb{R}$.

The mean of the random function χ is defined using the Bochner integral[3] as $\int_\Omega \chi d\mu$, under the condition that the expectation of $\|\chi\|$ is bounded. Moreover, if the means of χ_1, χ_2 in $\mathscr{H}$ are m, we give the definition of the covariance operator $\mathscr{K} : \mathscr{H} \to \mathscr{H}$ of random functions χ_1, χ_2 when $\mathscr{H} := \mathscr{H}_1 = \mathscr{H}_2$:

$$\langle \mathscr{K}\, g_1, g_2 \rangle = \langle \int_\Omega \langle \chi_1 - m, g_1 \rangle (\chi_2 - m) \rangle d\mu, g_2 \rangle = \int_\Omega \langle \chi_1 - m, g_1 \rangle \langle \chi_2 - m, g_2 \rangle d\mu \ ,$$

for $g_1, g_2 \in \mathscr{H}$. By using orthonormal bases $\{e_i\}$ in $\mathscr{H}$, we can compute the covariance values $\langle \mathscr{K}\, e_i, e_j \rangle$ for all pairs of indices i and j. Generally, if $\chi_1 \perp\!\!\!\perp \chi_2$, then we get $\langle \mathscr{K}\ g_1, g_2 \rangle = 0$ for $g_1, g_2 \in \mathscr{H}$.

In the context where each element in $\mathscr{H}$ is a mapping from E to $\mathbb{R}$, a random function $\chi : \Omega \to \mathscr{H}$ takes values $\chi(\omega, x) \in \mathbb{R}$ for each $\omega \in \Omega$ and $x \in E$. Furthermore, if we fix $\omega \in \Omega$, $\chi(\omega, \cdot)$ represents a

[3] See the definition of the Bochner integral in HSING and EUBANK (2015).

random function from E to $\mathbb{R}$. Henceforth, we adopt the notation $\chi(\cdot)$ to represent the random function $\chi(\omega, \cdot)$. This convention is analogous to the simplification employed for random variables, where $X(\omega)$ is denoted as X. Note that a mean m random function χ is referred as a Gaussian process if for any $n \geq 1$, the random vector $[\chi(x_1), \ldots, \chi(x_n)]$ of length n follows a Gaussian distribution with mean $[m(x_1), \ldots, m(x_n)]$, $x_1, \ldots, x_n \in E$.

When the Hilbert space $\mathscr{H}$ has a finite dimension d, the covariance operator can be represented by a covariance matrix, denoted as $\Sigma \in \mathbb{R}^{d \times d}$. This matrix is positive definite. Consequently, we can define the eigenvalues $\{\lambda_i\}$ and eigenvectors $\{\phi_i\}$ of Σ. Each vector in $\mathscr{H}$ can be expressed as a linear combination of the eigenvectors, specifically as $\sum_{i=1}^{d} \langle X, \phi_i \rangle \phi_i$. Moreover, for $\langle X, \phi_i \rangle$, the variance is given by λ_i. Then, for random function χ, if $\mathscr{H}$ is an infinite-dimensional function space,

Proposition 3 (HSING and EUBANK (2015)). *Let $\{\lambda_i\}$ and $\{\phi_i\}$ denote the eigenvalues and eigenfunctions obtained from the eigenvalue problem $\mathscr{K}\phi_i = \lambda_i \phi_i$, $i = 1, 2, \ldots$. With probability one, χ can be represented as:*

$$\chi = \sum_{i=1}^{\infty} \langle \chi, \phi_i \rangle_{\mathscr{H}} \phi_i,$$

where $\langle \chi, \phi_i \rangle_{\mathscr{H}}$ denotes the inner product between χ and ϕ_i in $\mathscr{H}$. Additionally, mean of χ is zero, and for $\langle \chi, \phi_i \rangle_{\mathscr{H}}$, the variance is equal to λ_i.

It is important to note the close relationship between stochastic processes and random functions. A set of random variables $\{X(t)\}_{t \in E}$ can be considered a stochastic process if the function $X : \Omega \times E \to \mathbb{R}$ is measurable with respect to the probability space $(\Omega, \mathcal{F}, \mu)$ and the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for each $t \in E$. It is worth mentioning that certain stochastic processes can also be regarded as random functions (HSING and EUBANK, 2015).

2.3 Extension to Functional Data

In this section, we generalize the concept of LiNGAM from random variables to encompass both random vectors and random functions.

Previous works have extended the D-S to encompass various scenarios. These extensions include incorporating random vectors (Ghurye and Olkin, 1962) and random functions in a Banach space (Myronyuk, 2008) as substitutes for random variables.

2.3.1 LiNGAM for Random Vectors

As shown by Shimizu et al. (2011), the identifiability of non-Gaussian random variables is outlined in Proposition 2. However, this proposition does not extend to the case of random vectors or random functions. This section provides proof of identifiability for non-Gaussian random vectors.

Proposition 4 (Multivariate Darmois-Skitovich (Ghurye and Olkin, 1962))**.** *Let* $L_1 = \sum_{i=1}^{m} A_i \xi^i$ *and* $L_2 = \sum_{i=1}^{m} B_i \xi^i$ *with mutually independent* k*-dimensional random vectors* ξ^i *and invertible matrices* A_i, B_i *for* $i = 1, \ldots, m$*. If* L_1 *and* L_2 *are mutually independent, then all* ξ^i *are Gaussian.*

Now we consider the identifiability of the following model when $x, y \in \mathbb{R}^m$ and invertible matrix $A \in \mathbb{R}^{m \times m}$, $e_1 \perp\!\!\!\perp e_2$ and zero means,

$$\begin{aligned} x &= \epsilon_1, & y &= \epsilon_1', \\ y &= Ax + \epsilon_2, & x &= A'y + \epsilon_2', \\ \epsilon_1' &= A\epsilon_1 + \epsilon_2, & \epsilon_2' &= (I - A'A)\epsilon_1 - A'\epsilon_2. \end{aligned} \tag{2.7}$$

We assume

$$A \text{ or } A' \text{ is invertible.} \tag{2.8}$$

Then, we have the following theorem.

Theorem 5. Assuming (2.8), which extends (2.3), we can identify the causal order between random vectors $X_1, X_2 : \Omega \to \mathbb{R}^m$ of dimension $m \in [1, \infty)$ if and only if at least one of them is non-Gaussian.

Proof. Since $\epsilon_1 \perp\!\!\!\perp \epsilon_2, E\epsilon_1' = E\epsilon_2' = 0$, and they are Gaussian random vectors with covariance matrix Σ_1, Σ_2, respectively. Then the correlated coefficient $\rho = 0 \iff Cov(\epsilon_1', \epsilon_2') = A\Sigma_1 \left(I - A^T A'^T\right) - \Sigma_2 A'^T = 0 \iff \epsilon_1' \perp\!\!\!\perp \epsilon_2'$, that is, when $A' = \Sigma_1 A^T (A\Sigma_1 A^T + \Sigma_2)^{-1}$, the causal relation between x, y is unable to be identified. This also satisfies the condition of $\epsilon_1' \perp\!\!\!\perp \epsilon_2'$ is that they follow the Gaussian distribution from the Proposition 4. □

2.3.2 LiNGAM for Random Functions

In this subsection, we present results that demonstrate identifiability can be achieved in non-Gaussian scenarios in infinite-dimensional Hilbert spaces as Theorem 8. In extending our approach to multivariate scenarios, we adopted methodologies from Lemmas 1 and 2 of DirectLiNGAM (refer to (Shimizu et al., 2011)). This involved identifying the exogenous function (see Appendix) and using residuals for causal

ordering, paralleling the process in Direct-LiNGAM. Owing to the procedural similarities with multivariate functions, we omitted a detailed proof in our main text, choosing to apply these established principles to our context. We have included the preliminary proof in Section 2.3.3 for clarity.

Let $\mathscr{H}_1, \mathscr{H}_2$ be Hilbert spaces. Assume that there are two causal models for $f_1 \in \mathscr{H}_1$ and $f_2 \in \mathscr{H}_2$,

$$\begin{aligned} f_1 &= h_1, \qquad f_2 = T_{21} f_1 + h_2, \\ f_2 &= h_1', \qquad f_1 = T_{12} f_2 + h_2'. \end{aligned} \tag{2.9}$$

where random functions $\{h_1, h_2'\} \in \mathscr{H}_1$ and $\{h_1', h_2\} \in \mathscr{H}_2$. We also assume the covariance operator $\mathscr{K}_{11}$ of h_1, $\mathscr{K}_{22}$ of h_2 have positive eigenvalues (> 0). The $T_{12} : \mathscr{H}_2 \to \mathscr{H}_1$, $T_{21} : \mathscr{H}_1 \to \mathscr{H}_2$ are linear bounded operators between $\mathscr{H}_1, \mathscr{H}_2$, and we identify the order by examining whether $h_2 \perp\!\!\!\perp f_1$ or $h_1 \perp\!\!\!\perp f_2$.

A bounded linear operator $T : \mathscr{H}_1 \to \mathscr{H}_2$ is considered continuous if the set $\{T(f) | f \in U\} \subseteq \mathscr{H}_2$ is open for any subset $U \subseteq \mathscr{H}_1$. Similarly, the inverse image U is also open. Furthermore, an operator $T : \mathscr{H}_1 \to \mathscr{H}_2$ is said to be invertible if it is both one-to-one (injective) and onto (surjective).

Let's confirm the statements before proceeding with our discussion:

- Proposition 6: There is an equivalence between independence and non-correlation for jointly Gaussian random functions. In other words, if χ_1 and χ_2 are jointly Gaussian random functions, they are independent if and only if they are uncorrelated.
- Proposition 7: The Darmois-Skitovich (D-S) theorem can be extended to random functions in Banach spaces.

The following Proposition 6 establishes the equivalence between independence and non-correlation for random functions in Banach spaces, which also includes Hilbert spaces as a special case.

Proposition 6 (van Neerven (2020))**.** *Suppose χ, χ' are joint Gaussian in Banach spaces. Then, $\chi \perp\!\!\!\perp \chi'$ if and only if they are uncorrelated.*

Proposition 7 (Darmois-Skitovich in Banach Space(Myronyuk, 2008))**.** *Suppose that $n \geq 2$, and random functions $\xi_1, \ldots, \xi_n$ are in a Banach space. Let $L_1 = \sum_{i=1}^m A_i \xi_i$, $L_2 = \sum_{i=1}^m B_i \xi_i$ with some continuous linear bounded operators $A_1, \ldots, A_m$, and $B_1, \ldots, B_m$. If $L_1 \perp\!\!\!\perp L_2$, then ξ_i is a Gaussian process for $i = 1, \ldots, m$ with invertible A_i, B_i.*

Theorem 8 (Causal Identifiability). *If either T_{12} or T_{21} is invertible, the causal order between random functions in infinite-dimensional Hilbert spaces can be identified if and only if at least one of them is a non-Gaussian process.*

Proof. For the sufficiency, from (2.9), we first assume the $f_1 = h_1$, $f_2 = T_{21}f_1 + h_2$, and represent the noise functions h'_1, h'_2 with h_1, h_2:

$$\begin{aligned} h'_1 &= f_2 = T_{21}h_1 + h_2 \\ h'_2 &= f_1 - T_{12}f_2 = h_1 - T_{12}(T_{21}h_1 + h_2) = (I - T_{12}T_{21})h_1 - T_{12}h_2 \,. \end{aligned} \tag{2.10}$$

Because h'_1, h'_2 are formed as the linear combinations of two independent Gaussian random functions h_1, h_2, we can conclude that h'_1 and h'_2 are jointly Gaussian (van Neerven, 2020). Then from Proposition 6, the zero-correlation implies independence. Since $h_1 \perp\!\!\!\perp h_2$ and $h_1 \in \mathscr{H}_1$, $h_2 \in \mathscr{H}_2$, the cross-covariance operator $\mathscr{K}_{12}$ is zero:

$$\langle \mathscr{K}_{12} g_1, g_2 \rangle_{H_2} = \int_\Omega \langle h_1, g_1 \rangle_{\mathscr{H}_1} \langle h_2, g_2 \rangle_{\mathscr{H}_2} = 0$$

for any $g_1 \in \mathscr{H}_1$,$g_2 \in \mathscr{H}_2$. Then, the cross-covariance operator $\mathscr{K}'_{12}$ between h'_1 and h'_2 is

$$\begin{aligned} &\langle \mathscr{K}'_{12} g_1, g_2 \rangle_{\mathscr{H}_2} = \int_\Omega \langle (I - T_{12}T_{21})h_1 - T_{12}h_2, g_1 \rangle_{\mathscr{H}_1} \langle T_{21}h_1 + h_2, g_2 \rangle_{\mathscr{H}_2} d\mu \\ &= \int_\Omega \langle (I - T_{12}T_{21})h_1, g_1 \rangle_{\mathscr{H}_1} \langle T_{21}h_1, g_2 \rangle_{\mathscr{H}_2} d\mu + \int_\Omega \langle -T_{12}h_2, g_1 \rangle_{\mathscr{H}_1} \langle h_2, g_2 \rangle_{\mathscr{H}_2} d\mu \\ &= \int_\Omega \langle h_1, (I - T_{12}T_{21})^* g_1 \rangle_{\mathscr{H}_1} \langle h_1, T_{21}^* g_2 \rangle_{\mathscr{H}_1} d\mu - \int_\Omega \langle h_2, T_{12}^* g_1 \rangle_{\mathscr{H}_2} \langle h_2, g_2 \rangle_{\mathscr{H}_2} d\mu \\ &= \langle \mathscr{K}_{11}(I - T_{12}T_{21})^* g_1, T_{21}^* g_2 \rangle_{\mathscr{H}_1} - \langle \mathscr{K}_{22} T_{12}^* g_1, g_2 \rangle_{\mathscr{H}_2} \\ &= \langle T_{21}\mathscr{K}_{11}(I - T_{21}^* T_{12}^*) g_1, g_2 \rangle_{\mathscr{H}_2} - \langle \mathscr{K}_{22} T_{12}^* g_1, g_2 \rangle_{\mathscr{H}_2} \end{aligned} \tag{2.11}$$

for any $g_1 \in \mathscr{H}_1$,$g_2 \in \mathscr{H}_2$, where $\mathscr{K}_{11}, \mathscr{K}_{22}$ are the covariance operators of h_1, h_2, respectively. We assume that $\mathscr{K}_{11}, \mathscr{K}_{22}$ are not zero. If $\mathscr{K}'_{12} = 0$, then we require

$$\mathscr{K}_{11} T_{21}^* = T_{12}\{T_{21}\mathscr{K}_{11}T_{21}^* + \mathscr{K}_{22}\} \,. \tag{2.12}$$

We have

$$\begin{aligned} (2.11) = 0 &\Leftrightarrow T_{21}\mathscr{K}_{11}(I - T_{21}^* T_{12}^*) = \mathscr{K}_{22} T_{12}^* \\ &\Leftrightarrow T_{21}\mathscr{K}_{11} = (T_{21}\mathscr{K}_{11}T_{21}^* + \mathscr{K}_{22})T_{12}^* \Leftrightarrow (2.12) \end{aligned}$$

However, the covariance operator K_{11} and K_{22} are not invertible because of they are compact operator:

- A covariance operator is trace-class operator (Theorem 7.2.5 in HSING and EUBANK (2015));
- A trace-class operator is Hibert-Schmidt operator (Theorem 4.5.2 in HSING and EUBANK (2015));
- An Hilbert–Schmidt operator is compact (Theorem 4.4.3 in HSING and EUBANK (2015));
- A compact operator is not invertible (Theorem 4.1.4 in HSING and EUBANK (2015)).

Then we know covariance operators are not invertible. But here, we need to notice that we can always define a Moore-Penrose inverse to make the equation (2.12) hold if

$$\text{Im}\left(\mathscr{K}_{11}T_{21}^{*}\right) \subseteq \text{Im}\left(\{T_{21}\mathscr{K}_{11}T_{21}^{*} + \mathscr{K}_{22}\}\right) \tag{2.13}$$

and the following is bounded (Li, 2018):

$$\{T_{21}\mathscr{K}_{11}T_{21}^{*} + \mathscr{K}_{22}\}^{\dagger}\mathscr{K}_{11}T_{21}^{*} \, . \tag{2.14}$$

Then the problem becomes determining the Images and boundness.

Next we prove $\text{Im}(A) \subseteq \text{Im}(A+B)$. Note that if A is positive semidefinite and $\langle Au, u\rangle = 0$, then $Au = 0$. To see why, let $v_1, \ldots, v_n$ be an orthonormal basis of eigenvectors of A (so $A\, v_i = \lambda_i v_i$) and write $u = \sum_{i=1}^{n} \langle u, v_i\rangle\, v_i$. Then

$$\langle Au, u\rangle = \sum_{i=1}^{n} \langle u, v_i\rangle^2 \lambda_i = 0$$

together with $\lambda_i \geq 0$ implies that $\langle u, v_i\rangle = 0$ if $\lambda_i > 0$ so $u \in \ker(A)$. To prove that $\text{Im}(A) \subseteq \text{Im}(A+B)$, it is enough to prove that

$$\ker(A+B) = \text{Im}(A+B)^{\perp} \subseteq \text{Im}(A)^{\perp} = \ker(A) \tag{2.15}$$

let $u \in \ker(A+B)$. Then $0 = \langle (A+B)u, u\rangle = \langle Au, u\rangle + \langle Bu, u\rangle$ which implies that $\langle Au, u\rangle = 0$, so $u \in \ker(A)$. Then (2.13) satisfys. Now we consider the boundness. As we know, the eigenvalue of $A+B$ (positive semidefinite) is bigger than A or B, which means the inverse eigenvalue of $A+B$ will be smaller than the inverse eigenvalue of A or B. Moreover, the smallest eigenvalue of the covariance operator tends to 0, then $(A+B)^{\dagger}A$ is bounded. Then we say the equation (2.12) holds. We can check more details in the Appendix.

Conversely, we first let $h_1 \perp\!\!\!\perp h_2$ and $h_1' \perp\!\!\!\perp h_2'$ in (2.9) hold true simultaneously for some T_{12}, T_{21}, and we want to prove that h_1, h_2, h_1', h_2' belong to Gaussian under (2.12). Note that a Hilbert space is a special case of Banach space. Then we use the Proposition 7. We assume that T_{12} is invertible without losing generality. Next we show that the eigenvalue of $T_{12}T_{21}$ is less than 1, which means that $I - T_{12}T_{21}$ is invertible (see Theorem 3.5.5 in HSING and EUBANK (2015)). To achieve this, we multiply (2.12) by T_{21} from the left-hand side, then we obtain

$$T_{21}\mathscr{K}_{11}T_{21}^{*} = T_{21}T_{12}\{T_{21}\mathscr{K}_{11}T_{21}^{*} + \mathscr{K}_{22}\} \, ,$$

which means that the eigenvalue of $T_{21}T_{12}$ is less than 1. Noting that $T_{21}T_{12}$ and $T_{12}T_{21}$ share the eigenvalues:

$$T_{21}T_{12}u = \lambda u \Longrightarrow T_{12}T_{21}T_{12}u = \lambda T_{12}u \Longrightarrow T_{12}T_{21}v = \lambda v$$

for $\lambda \neq 0$, $u \in \mathscr{H}_2$, and $v := T_{12}u \in \mathscr{H}_1$, we have proved that the eigenvalue of $T_{12}T_{21}$ is less than 1. Then, as we did in (2.2.1), we correspond

$$(h_1.h_2, T_{21}, I, I - T_{21}T_{12}, -T_{12}) = (\xi_1, \xi_2, A_1, A_2, B_1, B_2)) \ ,$$

where A_1, A_2, B_1, B_2 are invertible.

□

2.3.3 Causal Inference in Multivariate Scenarios

In the context of multivariate cases, we introduce two lemmas following Shimizu et al. (2011):
1. Lemma 9 identifies the exogenous function.
2. Lemma 10 establishes the causal order among residuals.
By analyzing residuals, we can determine the causal order of random functions. This is achieved after identifying an exogenous function, which, under the assumption of no latent confounders, corresponds to an independent external influence. The independence of these residuals is assessed through a series of pairwise regressions.

Lemma 9. *For multivariate case, a random function f_j is exogenous if and only if f_j is independent of its residuals $h_i^{(j)} = f_i - T_{ij}f_j$ for all $i \neq j$.*

Proof. For the sufficiency, if $f_j \perp\!\!\!\perp h_i^{(j)}$, assume f_j is not exogenous, then $f_j = \sum_{k\in P_j} T_{jk}f_k + h_j = \sum_{k\in P_j} T_{jk}\sum_{l\neq j} T_{kl}h_l + h_j$, where P_j means parents of f_j. Then $h_i^{(j)} = (I - T_{ij}T_{ji})f_i - T_{ij}\sum_{k\in P_j, k\neq i} T_{jk}f_k - T_{ij}h_j = (I - T_{ij}T_{ji})\sum_{q\neq j} T_{iq}h_q - T_{ij}\sum_{k\in P_j, k\neq i} T_{jk}\sum_{l\neq j} T_{kl}h_l - T_{ij}h_j$. The two formulas are composed of linear combinations of external influences other than h_j, from Prop. 7, all the functions are non-Gaussian, then $h_i^{(j)} \not\perp\!\!\!\perp f_j$, then it contradicts. Therefore, f_j should be exogenous; For the necessity, if f_j is exogenous, $f_j = h_j$, $f_i = T_{ij}f_j + h_i$ with $h_i \perp\!\!\!\perp f_j, h_i = \sum_{k\neq j} T_{ik}h_k$, we know the residual error $h_i^{(j)} = h_i$. Then, we know $f_j \perp\!\!\!\perp h_i^{(j)}$ from the independence of noise functions. So far, the lemma has been proven. □

Lemma 10. *Let $k_{r^{(j)}}(i)$ is the causal order of $r_i^{(j)}$, $k(i)$ denotes a causal order of f_i. Then, the same ordering of the residuals $r_i = h_i^{(1)} = f_i - T_{i1}f_1, i = 1\ldots, p-1$ is a causal ordering for the original observed functions as well: $k_{r^{(j)}}(l) < k_{r^{(j)}}(m) \Longleftrightarrow k(l) < k(m)$.*

Proof. When we determine the exogenous function f_1, we need to estimate the $p-1$ residuals of f_1: $r_i = h_i^{(1)} = f_i - T_{i1}f_1 = \sum_{j\neq 1} T_{ij}f_j + T_{i1}f_1 - T_{i1}f_1 = \sum_{j\neq 1} T_{ij}f_j, i = 1, \ldots, p-1$, which is $r_i = \sum_{j\neq 1} T_{ij}\sum_{k\neq j} T_{jk}h_k$. For the residual of $r_2 = f_2 - T_{21}f_1 = h_2$ (second function) is $r_i^{(j)} =$

$r_i - T'_{ij} r_j = \sum_{j \neq 1} T_{ij} \sum_{k \neq j,2} T_{jk} h_k + T'_{i2} h_2 - T'_{i2} h_2 = \sum_{j \neq 1} T_{ij} \sum_{k \neq j,2} T_{jk} h_k$ for all $i \neq j$. From the independence assumption of noise functions, we know $r_2 \perp\!\!\!\perp r_i^{(2)}$. Then we know the causal relationships of residuals $r_i, i = 1, \ldots, p-1$ are the same as $f_i, i = 1, \ldots, p-1$ with the T'_{ij} because what we need to do is to test the independence between r_i and its residual $r_i^{(j)}$. □

Extending the notion, we can determine the order among any number of random functions such as $f_i = \sum_{j=1}^{i-1} T_{i,j} f_j + h_i$ with non-Gaussian h_i and bounded linear operators $T_{i,j}; H_j \to H_i$ for p random functions $f_1 \in H_1, \ldots, f_p \in H_p$.

2.4 The Procedure

Consider one model from (2.9):

$$f_2 = T_{21} f_1 + h_2 \,. \tag{2.16}$$

Then let's notice the statement as follows:

Proposition 11 (HSING and EUBANK (2015)). *Let $T : \mathscr{H}_1 \to \mathscr{H}_2$ be a compact[4] bounded linear operator, $\{\lambda_j\}$ be the eigenvalues, and $\{e_{1,j}\}$ and $\{e_{2,j}\}$ be the sequences with orthonormal eigenvectors of T^*T and TT^*, respectively. Then*

$$Tf = \sum_{i=1}^{\infty} \lambda_i \langle f, e_{1,i} \rangle_{\mathscr{H}_1} e_{2,i}$$

with $f \in \mathscr{H}_1$.

Following the notation in Proposition 11, we write the three terms $T_{21} f_1 = \sum_{i=1}^{\infty} \lambda_i f_{1,i} e_{1,i}$, $f_2 = \sum_{i=1}^{\infty} f_{2,i} e_{2,i}$, $h_2 = \sum_{i=1}^{\infty} h_{2,i} e_{2,i}$. Then, (2.16) becomes:

Theorem 12. *Suppose that $T_{21} : \mathscr{H}_1 \to \mathscr{H}_2$ is compact. If we regard the bases of $\mathscr{H}_1$ and $\mathscr{H}_2$ as $\{e_{1,i}\}$ and $\{e_{2,i}\}$, respectively, then*

$$f_{2,i} = \lambda_i f_{1,i} + h_{2,i} \tag{2.17}$$

for $i = 1, 2, \ldots$, where $\lambda_1 \geq \lambda_2 \geq \cdots$.

To ensure convergence of the eigenvalue sequence $\{\lambda_i\}$, we suppose that the operator T_{21} is compact. Without compactness, the $\{\lambda_i\}$ would not be convergent. Practically, we approximate the infinite-dimensional random functions $f_1 \in \mathscr{H}_1$, $f_2, h_2 \in \mathscr{H}_2$ by finite length M random vectors. We select the bases $\{e_{1,i}\}_{i=1}^M$ and $\{e_{2,i}\}_{i=1}^M$ to minimize the approximation error.

[4] We define a bounded linear operator $T : \mathscr{H}_1 \to \mathscr{H}_2$ to be compact if, for any bounded infinite sequence $\{f_n\}$ in $\mathscr{H}_1$, the sequence $\{Tf_n\}$ has a convergent subsequence in $\mathscr{H}_2$.

FPCA offers a more effective fit than PCA for raw data dimensionality reduction, particularly with time-series data like fMRI and EEG, where dimensions vary with sampling frequency (e.g., 100Hz vs. 1Hz). As frequency increases, dimensions approach infinity. FPCA overcomes this by approximating infinite dimensions through orthogonal bases, preserving maximal original data information and capturing latent details beyond traditional sampling. Figures 2.2 and 2.3 demonstrate FPCA's necessity for functional data.

The other merit of using the FPCA (functional principal component analysis) approach is its efficiency. We assume the following procedure: first, we approximate the W time points sampled from functions by the L coefficients of the basis functions (B-spline). Then, we transform it by the M coefficients of the basis functions defined above. The time complexity is as follows. $M < L \ll W$ and the time complexity $C(M)$ of the proposed procedure is much less than $C(W)$. For example, Shimizu et al. (2011) evaluated the complexity of their method as $C(W) = O(n(Wp)^3q^2 + Wp)^4q^3)$, where q ($\ll n$) is the maximal rank found by the low-rank decomposition used in the kernel-based independence measure, although the proposed procedure requires additional $O(nL^2 + L^3)$ complexity for the covariance matrix $O(nL^2)$ and eigenvalue decomposition $O(L^3)$.

This paper primarily examines the summary causal relationships among random functions, focusing less on specific time points or partial windows in temporal data. There are three graphical representations of causal structures in temporal data, namely, the *full-time causal graph*, the *window causal graph*, and the *summary causal graph* (Gong et al., 2023). The *full-time causal graph*, illustrated on the left in Fig. 2.4, depicts a complete dynamic system, representing all vertices including components $f_1, \ldots, f_p$ at each time point t, connected through lag-specific directed links such as $f_i^{t-k} \to f_j^t$. However, due to the challenges of capturing a single observation for each series at every time point, constructing a full-time causal graph can be complex. To address this, the *window causal graph* concept is introduced, which operates under the assumption of a time-homogeneous causal structure. This graph, shown in the middle of Fig. 2.4, works within a time window corresponding to the maximum lag in the full-time graph. On the other hand, the *summary causal graph*, displayed on the right in Fig. 2.4, abstracts each time series component into a single node, illustrating inter-series causal relationships without specifying particular time lags. The complexity of this summary graph depends on the choice of multivariate dependence measure, such as mutual information or HSIC. The algorithmic complexity for generating this graph is similar to that of DirectLiNGAM. Fig. 2.4 visually compares these different types of causal graphs for multivariate time series.

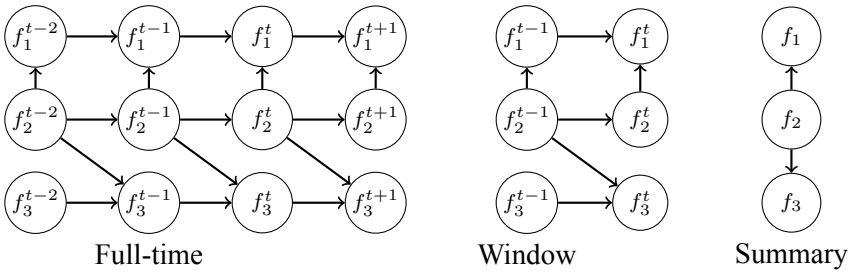

Figure 2.4: ***Illustration of Different Kinds of Multivariate Time Series Causal Graphs****. Left: Full-time; Middle: Window; Right: Summary (this paper).*

2.4.1 Algorithm

To show how to implement this method, we provide algorithm pseudocode and empirical experiments to demonstrate the efficiency. The algorithm presented in this study shares similarities with the greedy search method of DirectLiNGAM. However, it diverges in two key aspects: first, we leverage Functional Principal Component Analysis (FPCA) for data preprocessing, and second, our independence test considers multivariate relationships rather than univariate ones. This makes Func-LiNGAM straightforward to implement. For the purpose of this paper, we focus on providing a basic implementation without delving into enhancing search methods or other optimizations, as they are not the primary focus of our research. The whole algorithm is as Algorithm 1.

Note that the W means the sampled time points from one random function. As the intrinsically infinite-dimensional property of functional data, we need to approximate W with efficient finite representation (FPCA with principal component number M ($M \ll W$)). The number M can be decided by the explained variance ratio (95% or 99%). To be simple, here we let all the M of random functions be the same.

2.5 Experiment

To validate our method, we conducted comprehensive experiments using simulated data, as shown in Table 2.1. We observed an improvement in performance as the sample size increased across multiple functions. Notably, precision decreased monotonically and Structural Hamming Distance (SHD) increased monotonically as the number of functions (p) grew. Our data generation process, following the settings in Qiao et al. (2019), involved $n \times p$ random functions, defined as:

$$X_{ij}(t) = \phi(t)^T \delta_{ij} \tag{2.18}$$

Algorithm 1 Func-LiNGAM (Can be regarded as vector-based DirectLiNGAM but with FPCA preprocessing.)

Input: Each function has W time points, then construct Wp-dimensional random vector f (W: Full-time points) for p functions, a set of its variable subscripts U and a $Wp \times n$ data matrix as F, initialize an ordered list of functions $K = \emptyset$ and $m := 1$;

Output: Adjacent Matrix $\hat{T} \in \mathbb{R}^{p \times p}$

Use FPCA for finite approximating each random vector to make their dimensions from Wp to Mp, where M is the number of principal components.

repeat

(a) Perform least squares regressions of the approximating random vector $\hat{f}_i \in \mathbb{R}^M$ on $\hat{f}_j \in \mathbb{R}^M$ for all $i \in U \backslash K (i \neq j)$ and compute the residual vectors $\mathbf{r}^{(j)}$ and the residual data matrix $\mathbf{R}^{(j)}$ from the data matrix F for all $j \in U \backslash K$. Find a variable $\hat{f}_m$ that is most independent of its residuals:

$$\hat{f}_m = \arg \min_{j \in U \backslash K} MI \left(\hat{f}_j ; U \backslash K \right),$$

where MI is the independence measure such as mutual information or other measures.

(b) Append m to the end of K.

(c) Let $\hat{\mathbf{f}} := \mathbf{r}^{(m)}, \hat{F} := \mathbf{R}^{(m)}$.

until $p - 1$ subscripts are appended to K

Append the remaining variable to the end of K.

Construct a strictly lower triangular matrix $\hat{T}$ by following the order in K, and estimate the connection strengths $\hat{T}_{ij}$ by using least squares regression in this paper.

where i represents the i_{th} sample ($i = 1, \ldots, n$), and j denotes the j_{th} random vector. The vector $\delta_{ij} \in \mathbb{R}^5$ can be an arbitrary non-Gaussian random vector. Here we generated these by first creating random vectors $q_{ij} \sim \mathcal{N}(0, I_5)$, then we square each element of the vector to get δ_{ij}. The five-dimensional Fourier basis $\phi(t)$ was also used. We modeled the causal relationships in δ_i as follows:

$$\delta_{i0} = \epsilon_0, \quad \delta_{i1} = B_{1,0}\delta_{i0} + \epsilon_1, \quad \ldots, \quad \delta_{ip} = B_{p,p-1}\delta_{i(p-1)} + \epsilon_p \tag{2.19}$$

where $u_l \sim \mathcal{N}(0, I_5)$, then we square each element of the vector to get ϵ_l. To be simple, we set $B_{l,l-1} = I_5, l = 1, \ldots, p$. The sample size is $n = \{100, 200, 300, 700\}$, $p = \{5, 10, 20, 30, 50, 70\}$, and the observed values, $g_{ij}(t_k)$, follow

$$g_{ij}(t_k) = X_{ij}(t_k) + e_{ijk},$$

where e_{ijk} is derived from the square of the random variable q_{ijk}, where $q_{ijk} \sim \mathcal{N}(0, 0.25)$. Specifically, $e_{ijk} = q_{ijk}^2$. Due to the squaring of a normally distributed variable with a variance of 0.25, the resulting distribution of e_{ijk} can be described as a Gamma distribution with a shape parameter of $\frac{1}{2}$ and a scale parameter of 0.5, applicable for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Every random function is sampled at $W = 1000$ equidistant time points, $0 = t_1, \ldots, t_{1000} = 1$.

We employ B-spline bases as a fitting technique for each random function instead of the Fourier basis to represent the actual data accurately. B-spline bases offer more flexibility and can capture the complex shapes and patterns present in the data. After fitting the random functions with B-spline bases, we calculate each random function's estimated principal component scores. These scores are derived from the basis coefficients, with the number of calculated principal component scores limited to the first M components ($M \leq W$). The choice of M allows us to control the dimensionality of the data representation, providing a balance between capturing the most important variability in the data and minimizing computational complexity. By calculating these estimated principal component scores, we obtain a concise representation of the data that encapsulates its essential characteristics while reducing its dimensionality. This approach allows for efficient analysis and interpretation of the random functions within the context of our methodology. We set $M = 5$ (99% explained variance ratio) for the B-spline. Cross-validation can also obtain the optimal M. However, we set the parameters to ensure they maintain as much information as possible. We evaluate the Func-LiNGAM with Precision, Recall ratio, F1-score, and SHD (Structural Hamming Distance in Tsamardinos et al. (2006)) in 50 trials as Table 2.1. The smaller the SHD, the better the performance. To clarify, our objective is to demonstrate an implementation example rather than to propose a superior algorithm through comparison.

Data size	Metrics	Various number of functions (mean ± standard deviation)					
		$p = 5$	$p = 10$	$p = 20$	$p = 30$	$p = 50$	$p = 70$
$n = 100$	Precision	0.76 ± 0.14	0.64 ± 0.10	0.57 ± 0.09	0.40 ± 0.06	0.30 ± 0.04	0.25 ± 0.03
	Recall	0.99 ± 0.04	0.95 ± 0.0	0.90 ± 0.07	0.75 ± 0.07	0.65 ± 0.04	0.59 ± 0.04
	F1	0.85 ± 0.10	0.76 ± 0.08	0.70 ± 0.09	0.52 ± 0.07	0.41 ± 0.05	0.35 ± 0.03
	SHD	1.40 ± 0.95	5.03 ± 1.91	13.47 ± 4.17	33.47 ± 6.56	74.73 ± 9.86	119.47 ± 10.70
$n = 200$	Precision	0.83 ± 0.14	0.76 ± 0.29	0.72 ± 0.07	0.70 ± 0.06	0.54 ± 0.05	0.46 ± 0.07
	Recall	1.00 ± 0.00	0.80 ± 0.24	0.99 ± 0.01	0.97 ± 0.03	0.88 ± 0.03	0.81 ± 0.07
	F1	0.90 ± 0.08	0.78 ± 0.27	0.83 ± 0.05	0.81 ± 0.05	0.67 ± 0.05	0.59 ± 0.07
	SHD	0.97 ± 0.91	3.63 ± 4.58	7.70 ± 2.35	12.53 ± 3.36	37.03 ± 6.60	66.20 ± 12.79
$n = 300$	Precision	0.85 ± 0.13	0.79 ± 0.28	0.75 ± 0.07	0.74 ± 0.05	0.70 ± 0.05	0.60 ± 0.04
	Recall	1.00 ± 0.00	0.84 ± 0.23	1.00 ± 0.00	0.99 ± 0.01	0.99 ± 0.01	0.93 ± 0.03
	F1	0.92 ± 0.08	0.81 ± 0.26	0.86 ± 0.05	0.85 ± 0.03	0.82 ± 0.03	0.73 ± 0.04
	SHD	0.80 ± 0.75	3.17 ± 4.43	6.57 ± 2.50	10.27 ± 2.41	21.27 ± 4.36	42.90 ± 6.25
$n = 700$	Precision	0.92 ± 0.10	0.81 ± 0.08	0.80 ± 0.07	0.78 ± 0.05	0.74 ± 0.03	0.70 ± 0.02
	Recall	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.95 ± 0.05
	F1	0.96 ± 0.06	0.88 ± 0.05	0.88 ± 0.04	0.87 ± 0.03	0.85 ± 0.02	0.83 ± 0.05
	SHD	0.40 ± 0.55	2.50 ± 1.20	4.96 ± 2.06	8.80 ± 2.34	17.40 ± 2.97	32.70 ± 4.37

Table 2.1: *Evaluation of Func-LiNGAM with various number p of functions. The causal graph is as $f_1 \to f_2 \to \cdots \to f_p$ (50 trials).*

2.6 Actual Data

This section demonstrates the application of the proposed approach to analyzing brain connectomes for functional magnetic resonance imaging (fMRI) data. The fMRI data (Richardson et al., 2018) is preprocessed by downsampling it to a resolution of 4mm, with a repetition time (TR) of 2 seconds. This data consists of 155 subjects ($n = 155$), 168 time points ($W = 168$), and 17 parcels ($p = 17$). During the study, 155 participants took part in the fMRI scans. Among them, 122 participants were children, 33 were adults. The participants were instructed to watch a short animated movie that aimed to evoke various mental states and physical sensations about the characters depicted in the movie. Our objective is to investigate the causal relationships between various brain regions when individuals watch the short film, regardless of age. To check the Gaussianity of the observed functions, we performed the Shapiro–Wilk normality test (Shapiro and Wilk, 1965) on $p = 17$ parcels at each $W = 168$ time point. The null hypothesis (i.e., the observations are marginally Gaussian) was rejected for many combinations of scalp position and time point, and therefore, the non-Gaussianity of the proposed model is deemed appropriate. Next, we estimate the adjacency matrix between the parcels with the number of principal components $M = 5$. The adjacency matrix reveals the presence of connections between specific parcel pairs. To visualize the brain connectivity and causal relationships, we present a 2D graph using the `Nilearn` Python package and a 3D graph using the BrainNet Viewer (Xia et al., 2013) (Fig. 2.5).

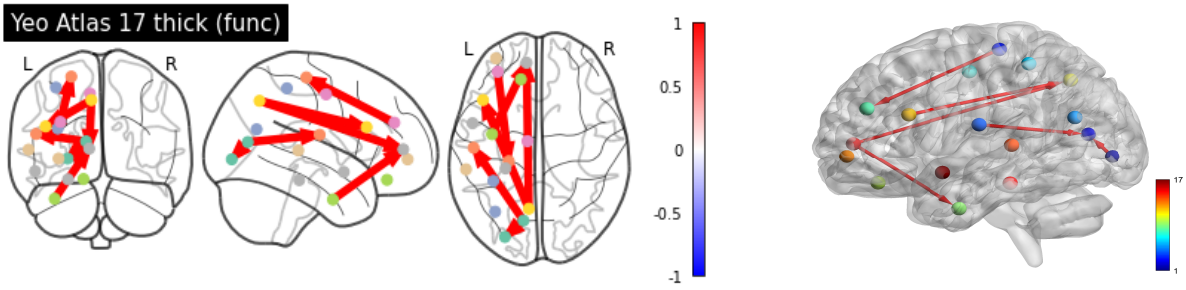

Figure 2.5: *Brain Connectivity Graphs (Left: 2D , Right: 3D).*

2.7 Conclusion

We have introduced a novel framework called Func-LiNGAM, which aims to identify causal relationships among random functions. For the theoretical foundation of Func-LiNGAM, we have proven the identifiability of both non-Gaussian random vectors (Theorem 5) and non-Gaussian processes (Theorem 8). Additionally, we have proposed a method to approximate random functions using random vectors based on Functional Principal Component Analysis (FPCA). Empirically, we demonstrate that the proposed procedure of Func-LiNGAM achieves accurate and efficient identification of causal orders among non-Gaussian random functions. Furthermore, we have preliminarily applied Func-LiNGAM to analyze brain connectivity using fMRI data. Our framework combines theoretical advancements with practical applications, showcasing its effectiveness in identifying causal relationships among random functions and its potential for various domains, such as brain connectivity.

Chapter 3

Dropout Drops Double Descent

3.1 Introduction

Recent investigations have shown that over-parameterized models, including linear regression and neural networks (Belkin et al., 2019, 2020; Hastie et al., 2019; Cun et al., 1991; Nakkiran et al., 2021a; Opper and Kinzel, 1996; Advani et al., 2020), demonstrate significant generalization capabilities, even when the labels are influenced by pure noise. This unique characteristic has attracted considerable academic attention, posing significant challenges to traditional generalization theory. A key framework, "Double Descent," helps explain this behavior (Belkin et al., 2019). In the under-parameterized realm, as we increase the number of model parameters or sample sizes, the test error initially shows a reduction, as illustrated by the peak curve in Figure 3.1. Intriguingly, as we transition into the over-parameterized domain, instead of increasing, the test error continues to decrease, revealing an unexpected secondary descent phase.

This peak phenomenon was first observed as early as three decades ago (Cun et al., 1991; Opper and Kinzel, 1996), and its re-emergence in recent years (Belkin et al., 2019; Advani et al., 2020) underlines the significant role it plays in research within the over-parameterized regime.

A primary objective of machine learning algorithms is to provide accurate out-of-sample predictions—a quality known as generalization. Traditional generalization theory presents a 'U-shaped' risk curve derived from the bias-variance trade-off (Hastie et al., 2009), which suggests the optimal model selection occurs prior to the interpolation point (when $n = p$). This trade-off suggests that a small hypothesis class lacks the expressive power necessary to include the truth function. Conversely, a larger class may introduce spurious overfitting patterns. However, in contrast to this traditional view, the double-descent

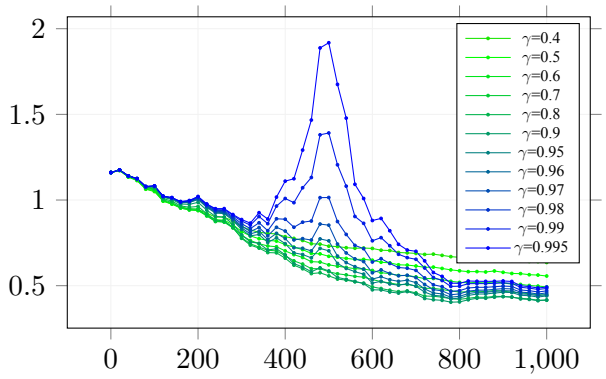

Figure 3.1: ***Test Risk of Sample-Wise Double Descent with Dropout.*** *γ denotes the probability of dropout as R. The number in the legend is the present probability. $p = 500$ and the sample size of the x-axis. The sample distribution $x \sim \mathcal{N}(0, I_p)$, $\mathrm{y} = x^T\beta^* + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.25)$, $\beta^* \sim \mathcal{U}(0, 1)$ and $||\beta^*||_2 = 1$.*

behavior, marked by a "\/\"-shaped trend with increasing model size, implies that we can discover a superior model with zero train and test error without succumbing to overfitting.

The reason behind the relatively recent surge in attention towards the double descent phenomenon is somewhat elusive, but the widespread adoption of regularization methods, such as ridge regularization (Hastie et al., 2019; Nakkiran et al., 2021b) and early stopping (Heckel and Yilmaz, 2021), designed to nullify double descent, might provide some explanation. In this study, we focus on one of the most popular regularization methods—dropout.

Dropout is a well-established regularization technique for training deep neural networks. It aims to prevent 'co-adaptation' among neurons by randomly excluding them during training (Hinton et al., 2012). Dropout's effectiveness extends across a wide range of machine learning tasks, from classification (Srivastava et al., 2014) to regression (Toshev and Szegedy, 2014). Notably, dropout was a vital component in the design of AlexNet (Krizhevsky et al., 2012), significantly outperforming its competitors in the 2012 ImageNet challenge. Due to dropout's proven efficiency in avoiding overfitting (Srivastava et al., 2014) and its broad application scope, we propose that it may significantly mitigate the double descent phenomenon. This leads us to the following question:

Under what conditions and how does dropout mitigate the double descent phenomenon?

We recognize that the double-descent phenomenon exists under both sample-wise and model-wise conditions. This paper considers its occurrence in both linear and nonlinear models to improve test performance without unexpected non-monotonic responses. The elimination of double descent has indeed become a hot research topic. For instance, ridge regularization can alleviate double descent (Nakkiran et al., 2021b),

as early stopping (Heckel and Yilmaz, 2021).

We explore a well-specified linear regression model utilizing dropout with $R_{ij} \sim \mathcal{B}er(\gamma), R \in \{0,1\}^{n\times p}, \gamma > 0, X \in \mathbb{R}^{n\times p}, y \in \mathbb{R}^n, \beta \in \mathbb{R}^p$, aiming to minimize the empirical risk:

$$L = ||y - (R * X)\beta||_2^2 ,$$

where $*$ denotes an element-wise product, serving to drop parameters during the training phase randomly. Dropout aids in preventing overfitting and offers a means to efficiently combine a wide range of different neural network architectures (Srivastava et al., 2014).

Our Contributions. Our study tackles the aforementioned question using theoretical and empirical methodologies. Theoretically, we explore the simplest linear regression with dropout regularization, which echoes the influence observed in general ridge regression (Ishwaran and Rao, 2014). When considering the test error—which includes both the bias and variance of a well-formulated linear regression model that employs dropout for isotropic Gaussian features[1]—we adopt a non-asymptotic perspective. Although we couldn't secure an exact solution to substantiate the monotonic decline of the test error, we devised an alternative approach. Through the application of Taylor series expansion, we obtained an approximate solution, providing persuasive evidence supporting the continuous decrease of the test error. On the empirical front, our numerical experiments demonstrate that the dropout technique can effectively mitigate the double descent phenomenon in both linear and nonlinear models. In more specific terms, we demonstrate:

- **Eliminating the Sample-Wise Double Descent.** We empirically validate the monotonicity of the test error as the sample size increases (see Figure 3.1) and theoretically prove the monotonicity of the second-order Neumann series test error. We plan to detail the exact solution in future work.
- **Eliminating the Model-Wise Double Descent.** We empirically demonstrate the monotonicity of the test error as the model size increases.
- **Multi-layer CNN.** We provide empirical evidence showing that dropout can alleviate the double descent in multi-layer CNNs.

3.1.1 Related works

Dropout. The purpose of dropout, as proposed in Srivastava et al. (2014), is to alleviate overfitting, and numerous variants of this technique have been further examined in Ba and Frey (2013); Wang and Manning (2013); Kingma et al. (2015); Khan et al. (2019); Li et al. (2016); Gal et al. (2017); Saito et al. (2018).

[1] Normal distribution with an identity covariance matrix.

As for the theory behind dropout, Wager et al. (2013) demonstrates that it functions as an adaptive regularization. Gal and Ghahramani (2016) postulates that dropout operates akin to a Bayesian approximation algorithm—specifically a *Gaussian Process*, incorporating an element of uncertainty into the functioning of black-box neural networks. Additionally, several studies have addressed the Rademacher complexity of dropout (Gao and Zhou, 2016), and its implicit and explicit regularization (Wei et al., 2020; Helmbold and Long, 2015).

Generalized Ridge Regression. The dropout estimator resembles a generalized ridge estimator, represented as $\hat{\beta} = (X^\top X + \lambda\Sigma_w)^{-1}X^\top y$, with Σ_w being the weighted matrix and $\lambda > 0$. Generalized ridge regression was first introduced in Hoerl and Kennard (2000), with numerous developments discussed in Casella (1980); Hemmerle (1975); Hua and Gunst (1983); Ishwaran and Rao (2014); Maruyama and Strawderman (2005); Mori and Suzuki (2018); Strawderman (1978). Nevertheless, these estimators are typically contemplated when $n > p$. Hence, their impact in high-dimensional and over-parameterized regimes is scarcely known. Wu and Xu (2020) recently provided an asymptotic view of the weighted ℓ_2 regularization in linear regression.

Dropping Double Descent. Several studies have aimed to counteract the double descent phenomenon. Heckel and Yilmaz (2021) illustrates that early stopping can attenuate double descent. Nakkiran et al. (2021b) argues that optimal ridge regularization has a similar effect in the non-asymptotic view, a finding that aligns with our study. Hastie et al. (2019) further sheds light on ridge regularization, illustrating a trend towards the same test error as the tail of double descent in model size.

3.2 Background

We consider linear regression in which p (≥ 1) covariates $x \in \mathbb{R}^p$ and response $\mathrm{y} \in \mathbb{R}$ are related by

$$\mathrm{y} = x^\top \beta_0 + \epsilon\ ,\ \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.1}$$

with unknown $\beta_0 \in \mathbb{R}^p$ and $\sigma^2 > 0$, where the occurrences of ϵ is independent from those of x, and we estimate β_0 from $n(\geq 1)$ i.i.d. training data $(x_1, \mathrm{y}_1), \ldots, (x_n, \mathrm{y}_n) \in \mathbb{R}^p \times \mathbb{R}$.

In particular, we assume that the covariates are generated by

$$x \sim \mathcal{N}(0,\ I_p)\ . \tag{3.2}$$

Thus, the covariates and response have the joint distribution $\mathcal{D}$ defined by (3.1) and (3.2), and we express $z^n := \{(x_i, \mathrm{y}_i)\}_{i=1}^n \sim \mathcal{D}^n$ for the training data. For each $\beta \in \mathbb{R}^p$, we define

$$R(\beta) := \mathop{\mathbb{E}}_{(x,\mathrm{y})\sim\mathcal{D}}[(x^\top\beta - \mathrm{y})^2], \tag{3.3}$$

where $\mathop{\mathbb{E}}_{(x,\mathrm{y})\sim\mathcal{D}}[\cdot]$ is the expectation w.r.t. the distribution $\mathcal{D}$.

Suppose we estimate β from the training data z^n by $\hat{\beta}_n : (\mathbb{R}^p \times \mathbb{R})^n \to \mathbb{R}^p$. Then, we define

$$\bar{R}(\hat{\beta}_n) := \mathop{\mathbb{E}}_{z^n\sim\mathcal{D}^n} R(\hat{\beta}_n(z^n)) = \mathop{\mathbb{E}}_{z^n\sim\mathcal{D}^n} \mathop{\mathbb{E}}_{(x,\mathrm{y})\sim\mathcal{D}} [(x^\top \hat{\beta}_n(z^n) - \mathrm{y})^2] \tag{3.4}$$

where $\mathbb{E}_{z^n\sim\mathcal{D}^n}[\cdot]$ is the expectation w.r.t. the distribution D^n. Note that (3.4) averages (3.3) over the training data as well while both evaluate the expected squared loss of the estimation.

In this paper, we consider the situation of dropout: given the training data $z^n = \{(x_i, \mathrm{y}_i)\}_{i=1}^n$, for $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n\times p}$ and $y = [\mathrm{y}_1, \ldots, \mathrm{y}_n]^\top \in \mathbb{R}^n$, we estimate β by the $\hat{\beta}(z^n)$ that minimizes the training error $\mathop{\mathbb{E}}_{R\sim\mathcal{B}er(\gamma)}[L]$ for

$$L = \|y - (r * X)\beta\|_2^2 ,$$

where $*$ denotes the element-wise product, each element of $R \in \mathbb{R}^{n\times p}$ takes one and zero with probabilities γ and $1-\gamma$, respectively, and we write $r \sim Ber(\gamma)$ for the distribution. Then, the quantity $\mathop{\mathbb{E}}_{R\sim\mathcal{B}er(\gamma)}[L]$ can be expressed by

$$\begin{aligned}
&\mathop{\mathbb{E}}_{r\sim\mathcal{B}er(\gamma)} \|y - (r * X)\beta\|_2^2 = \mathop{\mathbb{E}}_{R\sim\mathcal{B}er(\gamma)} \|y - M\beta\|_2^2 \\
&= y^\top y - 2\beta^\top \mathbb{E}(M^\top) y + \beta^\top \mathbb{E}(M^\top M)\beta \\
&= y^\top y - 2\gamma\beta^\top X^\top y + \beta^\top \mathbb{E}(M^\top M)\beta \\
&= \|y - \gamma X\beta\|_2^2 - \gamma^2\beta^\top X^\top X\beta + \beta^\top \mathbb{E}(M^\top M)\beta \\
&= \|y - \gamma X\beta\|_2^2 + \beta^\top(\mathbb{E}(M^\top M) - \gamma^2 X^\top X)\beta \\
&= \|y - \gamma X\beta\|_2^2 + (1-\gamma)\gamma\|\Gamma\beta\|_2^2
\end{aligned} \tag{3.5}$$

where $M := r * X$, $\Gamma = \mathrm{diag}(X^\top X)^{1/2}$, the final equation follows from the fact that the element-wise expectation $\mathbb{E}(M^\top M)$ is

$$\mathbb{E}\left[\sum_k m_{ik} m_{jk}\right] = \begin{cases} \gamma^2 \sum_k x_{ik} x_{jk}, & i \neq j \\ \gamma \sum_k x_{ik}^2, & i = j \end{cases}$$

for the (i,j)-th element of $M^\top M$ (the off-diagonal elements of $\mathbb{E}(M^\top M)$ and $\gamma^2 X^\top X$ are canceled out).

We can consider this as a Tikhonov regularization method. Let $\beta' = \gamma\beta$ as in Srivastava et al. (2014). Then, (3.5) becomes

$$\|y - X\beta'\|^2 + \frac{1-\gamma}{\gamma}\|\Gamma\beta'\|^2 , \tag{3.6}$$

which is minimized when β' is equal to

$$\hat{\beta}_{n,\gamma} = \left(X^\top X + \frac{1-\gamma}{\gamma} \Gamma^\top \Gamma \right)^{-1} X^\top y \,. \tag{3.7}$$

3.3 Drop Double-Descent in Linear Regression

In this section, we show the monotonicity of the solution in the sample size n with dropout in linear regression, and its proof follows in Appendix 3.6.1. Hereafter, we denote $\hat{\beta}$ by $\hat{\beta}_{n,\gamma}$ when we require n and γ to be explicit.

Before proving the claim, we notice that the test error is of the form

$$R(\hat{\beta}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\{ x^\top (\hat{\beta} - \beta_0) + \epsilon \}^2 \right] = \|\hat{\beta} - \beta_0\|_2^2 + \sigma^2 \,,$$

which is due to

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,I_d),\epsilon\sim\mathcal{N}(0,\sigma^2)} [\{(\hat{\beta} - \beta_0)^\top x + \epsilon\}^2] = \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,I_p)} [(\{(\hat{\beta} - \beta_0)^\top x\})^\top \{(\hat{\beta} - \beta_0)^\top x\}] + \sigma^2.$$

For the dropout estimator Eq. (3.7), the expected test error is

$$\begin{aligned}
\bar{R}(\hat{\beta}_{n,\gamma}) &= \mathbb{E}_X \mathbb{E}_y [R(\hat{\beta}_{n,\gamma})] = \mathbb{E}_X \mathbb{E}_y [\|\hat{\beta}_{n,\gamma} - \beta_0\|_2^2] + \sigma^2 \\
&= \mathbb{E}_X \mathbb{E}_y [\|(X^\top X + \Lambda)^{-1} X^\top y - \beta_0\|_2^2] + \sigma^2 \\
&= \mathbb{E}_X [\|(X^\top X + \Lambda)^{-1} X^\top (X\beta_0 + \epsilon) - \beta_0\|_2^2] + \sigma^2 \\
&= \mathbb{E}_X [\|((X^\top X + \Lambda)^{-1} X^\top X - I_p)\beta_0\|_2^2] + \sigma^2 \mathbb{E}_X [\|(X^\top X + \Lambda)^{-1} X^\top\|_F^2] + \sigma^2
\end{aligned}$$

where $\Lambda = \frac{1-\gamma}{\gamma} \mathrm{diag}(X^\top X)$. By neglecting the constant terms, the quantity $\bar{R}(\hat{\beta}_{n,\gamma})$ becomes

$$\beta_0^\top \mathbb{E}_X \left[\left(I + A^\top \right)^{-1} (I + A)^{-1} \right] \beta_0 + \sigma^2 \mathbb{E}_X \left[\left\| \left(X^\top X + \Lambda \right)^{-1} X^\top \right\|_F^2 \right], \tag{3.8}$$

where $A = \Lambda^{-1} X^\top X$.

We evaluate the expected test error (3.8) by taking Taylor's expansion of the matrix

$$\left(I + A^\top \right)^{-1} (I + A)^{-1} \,.$$

Then, we claim[2].

[2]We say $f(n) = O(g(n))$ if there exist $b > 0$ and $n_0 \geq 1$ such that $|f(n)| \leq b|g(n)|$ for $n \geq n_0$.

Theorem 13. *Let $\alpha = C < \frac{1}{(1+\sqrt{\frac{p}{n}})^2}$, the expected test error (3.8) is*

$$f(\alpha) = \left\{1 - 2\alpha + 3\alpha^2\frac{p}{n}\right\}\|\beta_0\|^2 + \sigma^2\alpha^2(\alpha+1)\frac{p}{n} + O(\frac{1}{n^2})$$

with $\alpha = \frac{\gamma}{1-\gamma}$.

Note the convergence of this Neumann series. We consider the condition that the eigenvalue of $A = \Lambda^{-1}X^\top X$ should be smaller than 1.

To prove the lemma, we notice some critical points.

1. Let $Q := \text{diag}(X^\top X)$, $P := Q^{-\frac{1}{2}}X^\top XQ^{-\frac{1}{2}}$, $\Lambda := \frac{1-\gamma}{\gamma}Q$, and $M := \Lambda^{-\frac{1}{2}}X^\top X\Lambda^{-\frac{1}{2}}$. Then, M and $A = \Lambda^{-1}X^\top X$ share share the same characteristic polynomial

$$\begin{aligned}\mathcal{P}_M(\lambda) &= \det(\Lambda^{-\frac{1}{2}}X^\top X\Lambda^{-\frac{1}{2}} - \lambda I) = \det(\Lambda^{-1/2})\det(X^\top X - \Lambda^{\frac{1}{2}}\lambda\Lambda^{\frac{1}{2}})\det(\Lambda^{-\frac{1}{2}}) \\ &= \det(\Lambda^{-1})\det(X^\top X - \lambda\Lambda) = \det(\Lambda^{-1}X^\top X - \lambda I) = \mathcal{P}_A(\lambda)\end{aligned},$$

   so do the eigenvalues.

2. Let λ_{max} and λ_{min} be the largest and smallest eigenvalues of M. Then, $\lambda_{\text{max}} \to (1+\sqrt{\frac{p}{n}})^2$ and $\lambda_{\text{min}} \to (1-\sqrt{\frac{p}{n}})^2$ as $n, p \to \infty$ with $\frac{p}{n} \to d \in (0, \infty)$ if $\mathbb{E}[x^4] < \infty$ (Theorem 1.1 in Jiang (2004)).

Hence, the maximum eigenvalues of matrices M and A are shown to approach $(1+\sqrt{\frac{p}{n}})^2$ asymptotically. Moreover, our empirical investigations corroborate that the largest eigenvalue of the sample correlation matrix M aligns closely with the theoretical prediction of $(1+\sqrt{\frac{p}{n}})^2$, as illustrated in Fig. 3.2. As delineated in Lemma 2, the Taylor series expansion converges when the parameter $\gamma/(1-\gamma)$ is multiplied to make the largest eigenvalue of M less than 1. The proof of Theorem 13 is in Appendix 3.6.1.

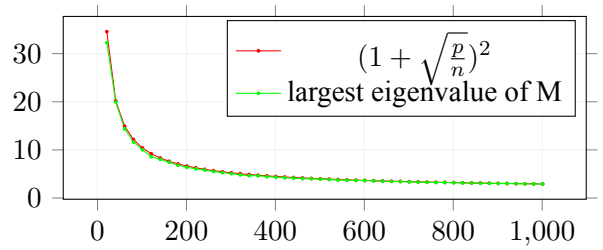

Figure 3.2: ***The Largest eigenvalue of Sample Correlation Matrix*** *($Q \in \mathbb{R}^{n\times p}$). X-axis denotes the number of sample n, Y-axis denotes the magnitude of largest eigenvalue and $n \in \mathbb{N}$, $p = 500$*

3.4 Experiments

This section provides empirical evidence that dropout with the optimal rate can effectively eliminate the double descent phenomenon in a broader range of scenarios compared to what is formally proven in Theorem 13.

3.4.1 Monotonicity in Sample Size

Elimination Double Descent in Linear Regression. (Synthetic Data)
In this part, we evaluate test error using dropout with pseudo optimal probability 0.8 (from Figure 3.1) in linear regression, the sample distribution $x \sim \mathcal{N}(0, I_p)$, $y = x^\top \beta^* + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.25)$, $\beta^* \sim \mathcal{U}(0, 1)$ and $\|\beta^*\|_2 = 1$. Moreover, the monotonic curves in Figure 3.3 show that the test error always remains monotonicity within the optimal dropout rate when the sample size increases for various dimensions p.

Random ReLU Initialization. (Fashion-MNIST)
We consider the random nonlinear features stemming from the random feature framework of Rahimi and Recht (2007). We apply random features to Fashion-MNIST (Xiao et al., 2017), an image classification dataset with 10 classes. In the preprocessing step, the input images vector $x \in \mathbb{R}^d$ are normalized and flattened to $[-1, 1]^d$ for the $d = 784$. To make the correct estimation of mean square loss, the class labels are dealt with the one-hot encoding to $y \in \{\vec{e}_1, \ldots, \vec{e}_{10}\} \subset \mathbb{R}^{10}$. According to the given number of random features D, and the number of sample data n, we are going to acquire the random classifier by performing linear regression on the nonlinear embedding: $\tilde{X} := \mathrm{ReLU}(XW^\top)$ where $X \in \mathbb{R}^{n \times d}$ and $W \in \mathbb{R}^{D \times d}$ is a matrix with every entry sampled i.i.d from $\mathcal{N}(0, 1/\sqrt{d})$, and with the nonlinear activation function ReLU applied pointwise. This is equivalent to a 2-layer fully connected neural network with a frozen (randomly initialized) first layer, trained with dropout. Figure 3.4 shows the monotonic test error.

3.4.2 Monotocity in Model Size

Like above setting, the sample distribution $x \sim \mathcal{N}(0, I_p)$, $y = x^\top \beta^* + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.25)$, $\beta^* \sim \mathcal{U}(0, 1)$ and $\|\beta^*\|_2 = 1$. The experiment result is the monotonic curves in Figure 3.5 show that the test error remains monotonicity with optimal dropout as the model size increases. For the multiple descents in Figure 3.5, the readers can find more details in Chen et al. (2021). Because we can think of the dropout

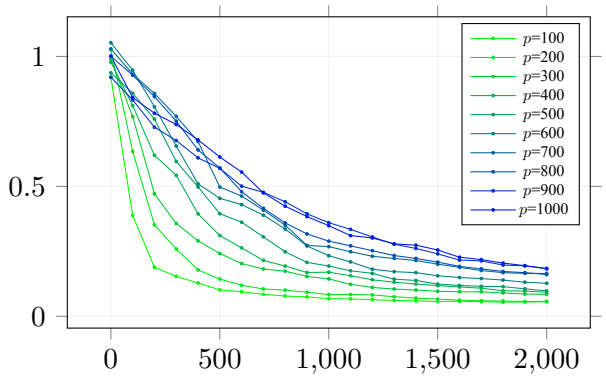

Figure 3.3: ***Test Risk with Number of Sample in linear regression with Dropout probability 0.8.*** *The test error curves decrease with the optimal dropout rate. The X-axis in this figure is the dimension of the parameter (0.8 is a pseudo-optimal value). The Y-axis is test risk.*

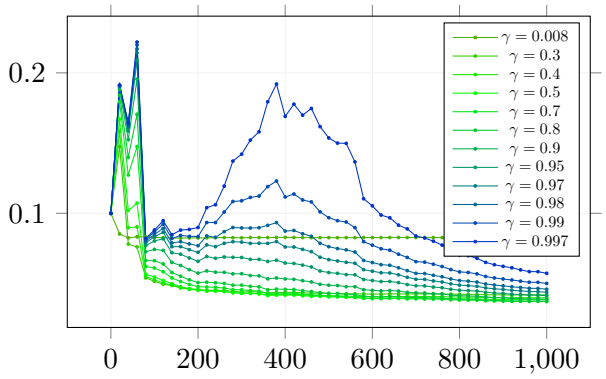

Figure 3.4: ***Test Risk with Number of Sample in Nonlinear Model with Dropout using Fashion-Mnist.*** *The test error curves are decreasing with optimal dropout. X-axis: sample size; Y-axis: Test risk.*

estimator as a generalized ridge estimator:

$$\begin{aligned}\mathop{\mathbb{E}}_{r\sim\mathcal{B}er(\gamma)}\|y-(r*X)\beta\|_2^2 &= \|y-\gamma X\beta\|_2^2+(1-\gamma)\gamma\|\Gamma\beta\|_2^2\\ &= \|y-\gamma X\Gamma^{-1}\Gamma\beta\|_2^2+(1-\gamma)\gamma\|\Gamma\beta\|_2^2\\ &= \|y-\gamma X'\beta'\|_2^2+(1-\gamma)\gamma\|\beta'\|_2^2\end{aligned} \tag{3.9}$$

We can think the covariates is also nonisotropic as $x' \sim \mathcal{N}(0, \Gamma^{-2})$. Then, we can observe the multiple descents as Nakkiran et al. (2021b). where $M := r * X$, $\Gamma = \text{diag}(X^\top X)^{1/2}$. Similar to Nakkiran et al. (2021b), we see the triple descent with the first peak around $p = 300$ dimension due to the 300-dimensional large eigenspace and the second peak at $n = p$. That is, the covariance has one "large" eigenspace and one "small" eigenspace.

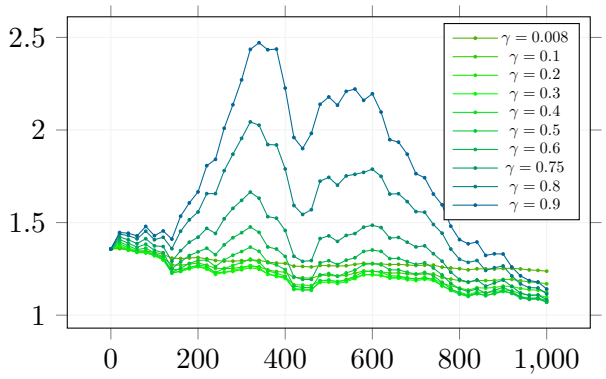

Figure 3.5: ***Test Risk with of model size in Linear Regression with Dropout.*** *The test error curves decrease with the optimal dropout rate. X-axis: the dimension of the parameter; Y-axis: Test risk.*

3.4.3 Multi-layer CNN

We use the same setups as in Nakkiran et al. (2021a). Here, we give the brief details of the model. For the full details, please check Appendix 3.7.1.

Standard CNNs: We consider a simple family of 5-layer CNNs, with 4 convolutional layers of widths $[k, 2k, 4k, 8k]$ for varying k, and a fully-connected layer. For context, the CNN with width $k = 64$, can reach over 90% test accuracy on CIFAR-10 with data augmentation. We train with cross-entropy loss and the following optimizer: Adam with 0.0001 learning rate for 10K epochs; SGD with $0.1/\sqrt{\lfloor T/512 \rfloor + 1}$ for 500K gradient steps.

Label Noise. In our experiments, label noise (Arpit et al., 2017) of probability prefers to train on samples with the correct label with probability 0%, 20%, and a uniformly random incorrect label otherwise (label noise is sampled only once and not per epoch).

Dropout layer. We add the dropout layer before the full-connected linear layer with the present rate γ (Srivastava et al., 2014). Figure 3.6 shows the test error results. The training loss is in Figure 3.7.

3.5 Discussion

Our proof considers only the non-exact solution for the expected test error. Therefore, we cannot definitively assert that the test risk decreases monotonically. However, based on our experimental results and this non-exact proof, we propose the following conjecture:

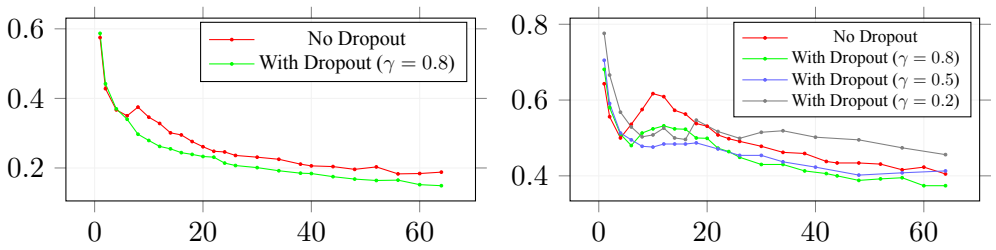

Figure 3.6: ***Test Risk with Number of width parameter in 5 layer-CNN with Dropout.*** *The x-axis is CNN width parameter (left: 0% label noise with Adam; right: 20% label noise with SGD). We can see dropout drops double descent.(γ: present rate)*

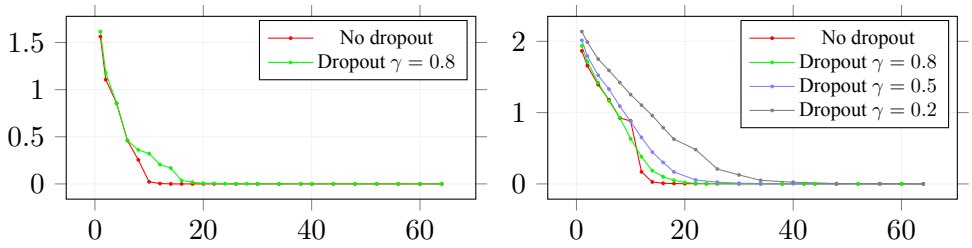

Figure 3.7: ***Train Loss with width parameter in 5 layer-CNN with Dropout (left: Adam, right: SGD). X-axis is CNN width parameter***

Conjecture 14. *For any $n, p \geq 1$, $\sigma^2 > 0$, and β_0, the expected test risk is monotonic in sample as*

$$\bar{R}(\hat{\beta}_{n+1}) \leq \bar{R}(\hat{\beta}_n). \tag{3.10}$$

In future research, we aim to prove that the exact solution with dropout can mitigate double descent.

Note the optimal hyperparameter remains in the fixed dimension p with a changeable sample size n. This is because the original data y from the model $y = X\beta + \epsilon$ will change, thus affecting the common test error. Additionally, Wainwright (2019) contains a statement about the sample covariance matrix $\text{diag}(X^\top X)$, which converges to the identity matrix for all $\delta > 0$ and $||x_i||_2 \leq \sqrt{d}$ (Corollary 6.20 in Wainwright (2019)):

$$P[\|\frac{\text{diag}(X^\top X)}{n} - I_p\|_2 \geq \delta] \leq 2p \cdot exp\left(-\frac{n\delta^2}{2d(1+\delta)}\right) \tag{3.11}$$

for the $\mathbb{E}(\text{diag}(X^\top X/n)) = I_p$, and by coupling the previous conclusions, it seems that the dropout estimator tends to the ridge estimator (LeJeune et al., 2020) and has the same asymptotic risk as the ridge estimator in Hastie et al. (2019).

3.6 Appendix

3.6.1 Proof of Theorem 13

The First term of (3.8)

Let $\Lambda := \frac{1-\gamma}{\gamma}\,\mathrm{diag}\left(X^\top X\right)$, $A := \Lambda^{-1}X^\top X$, and $\alpha = \frac{1-\gamma}{\gamma}$. We evaluate $\mathbb{E}[\left(I + A^\top\right)^{-1}(I+A)^{-1}]$. Note $\left(I + A^\top\right)^{-1}(I+A)^{-1} = I - A - A^\top + A^2 + \left(A^\top\right)^2 + A^\top A + \cdots$. For $A = (a_{i,j})$, we have $a_{i,j} = \frac{\gamma}{1-\gamma}\cdot\frac{\sum_k x_{k,i}x_{k,j}}{\sum_k x_{k,i}^2}$, and $\mathbb{E}[A] = \frac{\gamma}{1-\gamma}\cdot I$, which is due to (3.2). For $A^2 = (b_{i,j})$, we have

$$b_{i,j} = \left(\frac{\gamma}{1-\gamma}\right)^2 \cdot \sum_h \frac{\sum_k x_{k,i}x_{k,h}}{\sum_k x_{k,i}^2}\frac{\sum_k x_{k,h}x_{k,j}}{\sum_k x_{k,h}^2}$$

and $\mathbb{E}\left[A^2\right] = \left(\frac{\gamma}{1-\gamma}\right)^2\frac{p}{n}\cdot I$. Apparently, we have $\mathbb{E}\left[A^\top\right] = \frac{\gamma}{1-\gamma}\cdot I$ and $\mathbb{E}\left[\left(A^\top\right)^2\right] = \left(\frac{\gamma}{1-\gamma}\right)^2\frac{p}{n}\cdot I$. Finally, we evaluate $\mathbb{E}\left[A^\top A\right]$. For $A^\top A = (c_{i,j})$, we have

$$c_{i,j} = \left(\frac{\gamma}{1-\gamma}\right)^2 \cdot \sum_h \frac{\sum_k x_{k,i}x_{k,h}}{\sum_k x_{k,h}^2}\frac{\sum_k x_{k,h}x_{k,j}}{\sum_k x_{k,h}^2}$$

so that $E\left[c_{i,j}\right] = 0$ for $i \neq j$.

$$\begin{aligned}
\mathbb{E}\left[c_{i,i}\right] &= \left(\frac{\gamma}{1-\gamma}\right)^2 \sum_h \mathbb{E}\left(\frac{\sum_k x_{k,i}x_{k,h}}{\sum_k x_{k,h}^2}\right)^2 \\
&= \left(\frac{\gamma}{1-\gamma}\right)^2 \sum_h \mathbb{E}\left(\sum_k x_{k,i}\frac{x_{k,h}}{\sum_k x_{k,h}^2}\right)^2 \quad (x_i \perp\!\!\!\perp x_h) \\
&= \left(\frac{\gamma}{1-\gamma}\right)^2 \sum_h \mathbb{E}\left(\frac{\sum_k x_{k,h}^2 + 2\sum_{i\neq j} x_{i,h}x_{j,h}}{(\sum_k x_{k,h}^2)^2}\right) \quad \left(\mathbb{E}\left(2\sum_{i\neq j} x_{i,h}x_{j,h}\right) = 0\right) \\
&= \left(\frac{\gamma}{1-\gamma}\right)^2 \sum_h \mathbb{E}\left[\frac{1}{\sum_k x_{k,h}^2}\right],
\end{aligned}$$

where we have used $\mathbb{E}\left(\sum_r u_r\alpha_r\right)^2 = \mathbb{E}\sum_r u_r^2\alpha_r^2 = \sum_r \alpha_r^2$, when $u_r \sim N(0,1)$, $r = 1, 2, \cdots$, are independent. Then, from the inverse density function of chi-square distribution, we have $\mathbb{E}\left[A^\top A\right] = \left(\frac{\gamma}{1-\gamma}\right)^2\cdot\frac{p}{n-2}\cdot I$. Then, the first term of (3.8) is

$$\left\{1 - 2\left(\frac{\gamma}{1-\gamma}\right)\frac{p}{n} + \left(\frac{\gamma}{1-\gamma}\right)^2\left(\frac{2p}{n} + \frac{p}{n-2}\right)\right\}\|\beta_0\|^2 .$$

The Second term of (3.8)

Since

$$\left\|\left(X^\top X + \Lambda\right)^{-1}X^\top\right\|_F^2 = \mathrm{trace}\left\{\left(X^\top X + \Lambda\right)^{-1}X^\top\right\}^\top\left\{\left(X^\top X + \Lambda\right)^{-1}X^\top\right\}$$

the diagonal entries of $X\Lambda^{-1}\left\{I - A^\top - A + A^2 + \left(A^\top\right)^2 + A^\top A\right\}\Lambda^{-1}X^\top$ are

$$(X\Lambda^{-1}\Lambda^{-1}X^\top)\ldots m'_r = \left(\frac{\gamma}{1-\gamma}\right)^2 \sum_i \frac{x_{r,i}^2}{(\sum_k x_{k,i}^2)^2}$$

$$(X\Lambda^{-1}A\Lambda^{-1}X^\top)\ldots a'_r = \left(\frac{\gamma}{1-\gamma}\right)^3 \sum_i \sum_j \frac{x_{r,i}x_{r,j}a_{i,j}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2}$$

$$(X\Lambda^{-1}A^2\Lambda^{-1}X^\top)\ldots b'_r = \left(\frac{\gamma}{1-\gamma}\right)^4 \sum_i \sum_j \frac{x_{r,i}x_{r,j}b_{i,j}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2}$$

$$(X\Lambda^{-1}A^\top A\Lambda^{-1}X^\top)\ldots c'_r = \left(\frac{\gamma}{1-\gamma}\right)^4 \sum_i \sum_j \frac{x_{r,i}x_{r,j}c_{i,j}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2}$$

for $r = 1, \ldots, n$. First, we derive

$$\sum_r m'_r = \left(\frac{\gamma}{1-\gamma}\right)^2 \mathbb{E}\sum_i \frac{\sum_r x_{r,i}^2}{(\sum_k x_{k,i}^2)^2} = \left(\frac{\gamma}{1-\gamma}\right)^2 \frac{p}{n-2}$$

$$\begin{aligned}\sum_r a'_r &= \left(\frac{\gamma}{1-\gamma}\right)^3 \sum_r \sum_i \left\{\sum_{j\neq i} \frac{x_{r,i}x_{r,j}\frac{\sum_k x_{k,i}x_{k,j}}{\sum_k x_{k,i}^2}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2} + \frac{x_{r,i}^2}{\left(\sum_k x_{k,i}^2\right)^2}\right\} \\ &= \left(\frac{\gamma}{1-\gamma}\right)^3 \sum_i \left\{\sum_{j\neq i} \frac{(\sum_k x_{k,i}x_{k,j})^2}{\left(\sum_k x_{k,i}^2\right)^2 \sum_k x_{k,j}^2} + \frac{1}{\sum_k x_{k,i}^2}\right\} \\ &= \left(\frac{\gamma}{1-\gamma}\right)^3 \sum_i \frac{1}{\sum_k x_{k,i}^2}\left(\sum_{j\neq i} \hat{\rho}_{i,j}^2 + 1\right)\end{aligned}$$

Please note that the distribution of $\hat{\rho}i, j$ is independent of $x1, i, \ldots, x_{n,i}$ (as demonstrated in the derivation).Hence, the expectation of $\sum_r a'_r$ is $\left(\frac{\gamma}{1-\gamma}\right)^3 \left(\frac{p-1}{n} + 1\right) \sum_i \frac{1}{\sum_k x_{k,i}^2}$, when $x_{1,i}, \ldots x_{n,i}$ are given. Thus, we obtain

$$E\left[\sum_r a'_r\right] = \left(\frac{\gamma}{1-\gamma}\right)^3 \cdot \frac{p}{n-2} \cdot \left(\frac{p-1}{n} + 1\right)$$

On the other hand.

$$\sum_r b'_r = \left(\frac{\gamma}{1-\gamma}\right)^4 \sum_r \sum_i \sum_j \frac{x_{r,i}x_{r,j}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2} \sum_h \frac{\sum_k x_{k,i}x_{k,h}}{\sum_k x_{k,i}^2} \frac{\sum_k x_{k,h}x_{k,j}}{\sum_k x_{k,h}^2}$$

Let

$$\beta_{i,j,h} := \sum_r \frac{x_{r,i}x_{r,j}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2} \frac{\sum_k x_{k,i}x_{k,h}}{\sum_k x_{k,i}^2} \frac{\sum_k x_{k,h}x_{k,j}}{\sum_k x_{k,h}^2}$$

Then, the $\sum_h \beta_{i,j,h}$ with $i = j$ is $\frac{1}{\sum_k x_{k,i}^2} \sum_h \left(\frac{\sum_k x_{k,h}x_{k,j}}{\sqrt{\sum_k x_{k,h}^2 \sum_k x_{k,i}^2}}\right)^2$ and its expectation is $\frac{1}{n-2}\left(\frac{p-1}{n} + 1\right)$.

When $j \neq i = h$, it's $\frac{1}{\sum_k x_{k,i}^2}\left(\frac{\sum_k x_{k,i}x_{k,j}}{\sqrt{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2}}\right)^2$, its expectation is $\frac{1}{n(n-2)}$. Since the $\beta_{i,j,h}$ with i, j, h

different is

$$\sum_r \frac{x_{r,i}x_{r,j}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2} \frac{\sum_k x_{k,i}x_{k,h}}{\sum_k x_{k,i}^2} \frac{\sum_k x_{k,h}x_{k,j}}{\sum_k x_{k,h}^2}$$

its expectation is $\frac{1}{n(n-2)}$ If we take expectation w.r.t. $\{x_{k,h}\}$, then the value becomes

$$\sum_r \frac{x_{r,i}x_{r,j}}{\left(\sum_k x_{k,i}^2\right)^2 \sum_k x_{k,j}^2} \cdot \frac{1}{n} \sum_k x_{k,i}x_{k,j} \ ,$$

where the fact $E\left[\frac{Z_1}{Z_1+\cdots+Z_m}\right] = \frac{1}{m}$ for i.i.d. $Z_1, \ldots, Z_m$ has been used. Thus, the expectation is $\frac{1}{n(n-2)}$ as well. Hence, $E\left[\sum_h \beta_{i,j,h}\right]$ with $i \neq j$ is $\frac{p}{n(n-2)}$. Therefore,

$$E\left[\sum_r b_r'\right] = \left(\frac{\gamma}{1-\gamma}\right)^4 \frac{1}{n-2}\left(\frac{2p-1}{n}+1\right) \ .$$

Finally, we obtain $E\left[\sum_r c_r'\right]$. Let

$$\gamma_{i,j,h} := \sum_r \frac{x_{r,i}x_{r,j}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2} \frac{\sum_k x_{k,i}x_{k,h}}{\sum_k x_{k,h}^2} \frac{\sum_k x_{k,h}x_{k,j}}{\sum_k x_{k,h}^2}.$$

If $i = j$, we have $\sum_h \gamma_{i,j,h} := \sum_h \frac{1}{\sum_k x_{k,h}^2} \cdot \left(\frac{\sum_k x_{k,i}x_{k,h}}{\sqrt{\sum_k x_{k,i}^2}\sqrt{\sum_k x_{k,h}^2}}\right)^2$ and its expectation is $\frac{d}{n(n-2)}$. If $i \neq j, h = i$

$$\gamma_{i,j,h} := \sum_r \frac{x_{r,i}x_{r,j}}{\sum_k x_{k,i}^2 \sum_k x_{k,j}^2} \frac{\sum_k x_{k,i}x_{k,j}}{\sum_k x_{k,i}^2} = \frac{1}{\sum_k x_{k,i}^2}\left(\frac{\sum_k x_{k,i}x_{k,j}}{\sqrt{\sum_k x_{k,i}^2}\sqrt{\sum_k x_{k,j}^2}}\right)^2$$

and its expectation is $\frac{1}{n(n-2)}$ If i, j, h are different, if we fix $\{x_{k,j}\}$ and $\{x_{k,h}\}$, then the expectation of $\gamma_{i,j,h} := \frac{1}{\sum_r x_{r,i}^2}\hat{\rho}_{j,h}\hat{\rho}_{i,j}\hat{\rho}_{i,h}$ is zero. Thus, we have

$$E\left[\sum_r c_r'\right] = \left(\frac{\gamma}{1-\gamma}\right)^4 \left(\frac{p^2}{n(n-2)} + \frac{2p(p-1)}{n(n-2)}\right) = \left(\frac{\gamma}{1-\gamma}\right)^4 \frac{3p^2-2p}{n(n-2)}$$

with $\alpha = \frac{\gamma}{1-\gamma}$. Next, the test error is calculated by summing these terms, resulting in

$$\left\{1 - 2\alpha + \alpha^2 \frac{p}{n}\left(3 + \frac{2}{n-2}\right)\right\} \|\beta_*\|^2 + \alpha^2 \frac{p}{n-2}$$
$$+ \alpha^3 \frac{p}{n-2}\left(\frac{p-1}{n}+1\right) + \alpha^4 \frac{4p^2-2p-1+n}{n(n-2)}$$

3.7 Experiment Details

3.7.1 Models

Standard CNNs. We consider a simple family of 5-layer CNNs, with four Conv-Batch Norm-ReLU-MaxPool layers and a fully-connected output layer. We scale the four convolutional layer widths as

$[k, 2k, 4k, 8k]$. The MaxPool is $[1, 2, 2, 8]$. For all the convolution layers, the kernel size = 3, stride = 1, and padding = 1. This architecture is based on the "backbone" architecture from Page (2018). Fork= 64, this CNN has 1558026 parameters and can reach$>$ 90% test accuracy on CIFAR-10 (Krizhevsky, 2009) with data augmentation. The scaling of model size with k is shown in "Figure 13" of Nakkiran et al. (2021a).

Chapter 4

Conclusion and Future work

4.1 Conclusion

In this comprehensive research, we have delved into the multifaceted world of linear operators within machine learning, contributing significantly from both causal discovery and linear regression analysis perspectives.

Our journey began with the introduction of Functional Linear Non-Gaussian Acyclic Model (Func-LiNGAM), a pioneering development in causal discovery. Func-LiNGAM extends the conventional LiNGAM framework to encompass infinite-dimensional spaces, including vectors and functions. This expansion unlocks new horizons for uncovering causal relationships within complex datasets, such as fMRI and EEG. The research's theoretical underpinnings, including guarantees of identifying causal relationships in infinite-dimensional Hilbert spaces, provide a solid foundation for its practical applications. Additionally, incorporating functional principal component analysis addresses the sparsity challenge in these datasets. Our experimental results, including the analysis of brain connectivity patterns from real fMRI data, underscore the efficacy of Func-LiNGAM in unveiling causal connections among multivariate functions.

Simultaneously, our exploration ventured into linear regression, particularly addressing the enigmatic double descent phenomenon. By introducing dropout layers alongside fully connected linear layers, we have illuminated a novel approach to mitigating fluctuations in prediction error rates as sample size or model complexity increases. While we did not provide rigorous mathematical proof, empirical evidence revealed a consistent relationship between dropout rate and optimal test error, thus offering insights into the strategic use of dropout regularization in linear regression. Our pioneering investigation into the

connection between dropout and the double descent phenomenon enriches our understanding of machine learning model performance.

In conclusion, this comprehensive research underscores the remarkable versatility of linear operators in machine learning. By advancing the frontiers of causal discovery through Func-LiNGAM and shedding light on perplexing phenomena within linear regression, we contribute theoretically and practically to the field. These studies not only deepen our understanding of linear regression but also provide pragmatic methodologies for handling complex, high-dimensional datasets, showcasing the immense potential of linear operators in advancing machine learning research.

4.2 Future work

4.2.1 Double Descent

In Wainwright (2019), there is statement about the sample covariance matrix $\text{diag}(X^TX)$ that it converges to identity matrix for all $\delta > 0$ and $||x_i||_2 \leq \sqrt{d}$ (Corollary 6.20 in Wainwright (2019)):

$$P[||\frac{\text{diag}(X^TX)}{n} - I_p||_2 \geq \delta] \leq 2p \cdot exp\left(-\frac{n\delta^2}{2d(1+\delta)}\right) \tag{4.1}$$

for the $\mathbb{E}(\text{diag}(X^TX/n)) = I_p$ and by coupling the previous conclusions, we say that the dropout estimator is actually equal to the ridge estimator for all n, the number of samples.

In addition, we should draw attention to this transform, which if we write $\text{diag}(X^TX) = \Lambda$, then the loss function of original model using dropout is:

$$\begin{aligned} L &= ||X\beta - y||_2^2 + \frac{1-\gamma}{\gamma}||\Lambda^{1/2}\beta||_2^2 \\ &= ||X\Lambda^{-1/2}\Lambda^{1/2}\beta - y||_2^2 + \frac{1-\gamma}{\gamma}||\Lambda^{1/2}\beta||_2^2 \\ &= ||X'\beta' - y||_2^2 + \frac{1-\gamma}{\gamma}||\beta'||_2^2 \end{aligned} \tag{4.2}$$

which the data has changed its distribution into $x' \sim \mathcal{N}(0, \Sigma)$, with the general nonisotropic covariates, this part may be finished in the future work.

4.2.2 Functional Data

Proposing a novel dimension reduction approach for approximating the infinite-dimensional functional data with kernel dimension reduction, originally designed for supervised problems, to unsupervised dimensionality reduction. This paper uses kernel-based independence measures to derive low-dimensional representations that maximally capture information from functional data and minimize the redundancy among the chosen features. We demonstrate that whenever the coefficients of functional data exhibit a linear or nonlinear relationship, our method achieves better results for FPCA. Moreover, our method outperforms FPCA even when the functional data is more complex.

List of Publications

• Tianle Yang, Joe Suzuki (2022) The Functional LiNGAM. The 11th International Conference on Probabilistic Graphical Models. https://proceedings.mlr.press/v186/yang22a.html

• Tian-Le Yang, Joe Suzuki (2023) Functional Linear Non-Gaussian Acyclic Model for Causal Discovery. Behaviormetrika. (PGM Extended, Accept but needs final editing)

• Tian-Le Yang, Joe Suzuki (2023) Dropout Drops Double Descent. Japanese Journal of Statistics and Data Science. (Accept) https://arxiv.org/abs/2305.16179

Acknowledgements

I am profoundly grateful for the guidance and support I received throughout the journey of this doctoral research. First and foremost, I extend my deepest gratitude to my advisor, Joe Suzuki, whose expertise and insightful guidance have been invaluable. I also wish to thank the members of my research team for their collaboration and contributions to this work. Special thanks are due to JST for their financial support. I am deeply appreciative of my family for their endless encouragement and understanding. Lastly, I acknowledge the support from my friends and colleagues, who have been a constant source of motivation and inspiration.

References

Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2020.08.022. URL https://www.sciencedirect.com/science/article/pii/S0893608020303117.

Devansh Arpit, Stanisław Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arpit17a.html.

Lei Jimmy Ba and Brendan Frey. Adaptive dropout for training deep neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3084–3092, Red Hook, NY, USA, 2013. Curran Associates Inc.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1903070116.

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. doi: 10.1137/20M1336072. URL https://doi.org/10.1137/20M1336072.

George Casella. Minimax Ridge Regression Estimation. *The Annals of Statistics*, 8(5):1036 – 1056, 1980. doi: 10.1214/aos/1176345141. URL https://doi.org/10.1214/aos/1176345141.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman

Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8898–8912. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/4ae67a7dd7e491f8fb6f9ea0cf25dfdb-Paper.pdf.

Yann Le Cun, Ido Kanter, and Sara A. Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Phys. Rev. Lett.*, 66:2396–2399, May 1991. doi: 10.1103/PhysRevLett.66.2396. URL https://link.aps.org/doi/10.1103/PhysRevLett.66.2396.

G. Darmois. Analyse generale des liaisons stochastiques. *Rev. Inst. Intern. Stat*, 21:2–8, 1953.

I. Ebert-Uphoff and Y. Deng. Causal Discovery for Climate Research Using Graphical Models. *Journal of Climate*, 25(17):5648–5665, 2012.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/84ddfb34126fc3a48ee38d7044e87276-Paper.pdf.

Wei Gao and Zhi-Hua Zhou. Dropout rademacher complexity of deep neural networks. *Science China Information Sciences*, 59(7):072104, 2016. doi: 10.1007/s11432-015-5470-z. URL https://doi.org/10.1007/s11432-015-5470-z.

S. Ghurye and I. Olkin. A characterization of the multivariate normal distribution. *Ann. Math. Statist.*, 33(2):533–541, June 1962.

Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, Jingping Bi, Lun Du, and Jin Wang. Causal discovery from temporal data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5803–5804, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599552. URL https://doi.org/10.1145/3580305.3599552.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer New York, NY, 2 edition, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation, 2019. URL https://arxiv.org/abs/1903.08560.

Reinhard Heckel and Fatih Furkan Yilmaz. Early stopping in deep networks: Double descent and how to eliminate it. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tlV90jvZbw.

David P. Helmbold and Philip M. Long. On the inductive bias of dropout. *Journal of Machine Learning Research*, 16(105):3403–3454, 2015. URL http://jmlr.org/papers/v16/helmbold15a.html.

William J. Hemmerle. An explicit solution for generalized ridge regression. *Technometrics*, 17(3):309–314, 1975. ISSN 00401706. URL http://www.jstor.org/stable/1268066.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012. URL https://arxiv.org/abs/1207.0580.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000. ISSN 00401706. URL http://www.jstor.org/stable/1271436.

T. HSING and R. EUBANK. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester, 2015.

Tsushung A. Hua and Richard F. Gunst. Generalized ridge regression: a note on negative ridge parameters. *Communications in Statistics - Theory and Methods*, 12(1):37–45, 1983. doi: 10.1080/03610928308828440. URL https://doi.org/10.1080/03610928308828440.

Hemant Ishwaran and J. Sunil Rao. Geometry and properties of generalized ridge regression in high dimensions. 2014.

Tiefeng Jiang. The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, 66(1):35–48, 2004. ISSN 09727671. URL http://www.jstor.org/stable/25053330.

Salman H. Khan, Munawar Hayat, and Fatih Porikli. Regularization of deep neural networks with spectral dropout. *Neural Networks*, 110:82–90, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2018.09.009. URL https://www.sciencedirect.com/science/article/pii/S0893608018302715.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/bc7316929fe1545bf0b98d114ee3ecb8-Paper.pdf.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. The implicit regularization of ordinary least squares ensembles. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3525–3535. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/lejeune20b.html.

B. Li. Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics*, 46(1):79–103, 2018. doi: https://doi.org/10.1002/cjs.11329. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11329.

Zhe Li, Boqing Gong, and Tianbao Yang. Improved dropout for shallow and deep learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2531–2539, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Peter J. Lucas, Linda C. Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and healthcare. *Artificial Intelligence in Medicine*, 30:201–214, 04 2004. doi: 10.1016/j.artmed.2003.11.001.

Yuzo Maruyama and William E. Strawderman. A new class of generalized Bayes minimax ridge regression estimators. *The Annals of Statistics*, 33(4):1753 – 1770, 2005. doi: 10.1214/009053605000000327. URL https://doi.org/10.1214/009053605000000327.

Yuichi Mori and Taiji Suzuki. Generalized ridge estimator and model selection criteria in multivariate linear regression. *Journal of Multivariate Analysis*, 165:243–261, 2018. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2017.12.006. URL https://www.sciencedirect.com/science/article/pii/S0047259X17307819.

M. V. Myronyuk. On the Skitovich-Darmois theorem and Heyde theorem in a Banach space. *Ukr Math J*, 60:1437–1447, 2008.

Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent, 2019. URL https://arxiv.org/abs/1912.07242.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, dec 2021a. doi: 10.1088/1742-5468/ac3a74. URL https://dx.doi.org/10.1088/1742-5468/ac3a74.

Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=7R7fAoUygoa.

Manfred Opper and Wolfgang Kinzel. *Statistical Mechanics of Generalization*, pages 151–209. Springer New York, New York, NY, 1996. ISBN 978-1-4612-0723-8. doi: 10.1007/978-1-4612-0723-8_5. URL https://doi.org/10.1007/978-1-4612-0723-8_5.

David Page. How to train your resnet 4: Architecture, 2018. URL https://myrtle.ai/how-to-train-your-resnet-4-architecture/.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge University Press, 2000.

X. Qiao, S. Guo, and G. M. James. Functional Graphical Models. *J. Amer. Statist. Assoc.*, 114(525): 211–222, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2005. ISBN 9780387400808. URL http://www.worldcat.org/isbn/9780387400808.

Hilary Richardson, Grace Lisandrelli, Alexa Riobueno-Naylor, and Rebecca Saxe. Development of the social brain from age three to twelve years. *Nature Communications*, 9(1):1027, 2018.

Saptarshi Roy, Raymond K. W. Wong, and Yang Ni. Directed cyclic graph for causal discovery from multivariate functional data, 2023.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005. doi: 10.1126/science.1105809. URL https://www.science.org/doi/abs/10.1126/science.1105809.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJIoJWZCZ.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. ISSN 00063444. URL http://www.jstor.org/stable/2333709.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011.

V. P. Skitivic. On a property of the normal distribution. *Dokl. Akad. Nauk SSSR (N.S.)*, (89):217–219, 1953.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2000.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

William E. Strawderman. Minimax adaptive generalized ridge regression estimators. *Journal of the American Statistical Association*, 73(363):623–627, 1978. ISSN 01621459. URL http://www.jstor.org/stable/2286612.

Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014. doi: 10.1109/CVPR.2014.214.

I. Tsamardinos, L.E. Brown, and C.F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Mach Learn*, 65:31–78, 2006.

R.S. Tsay and M. Pourahmadi. Modelling structured correlation matrices. *Biometrika*, 104(1):237–242, 2017.

Jan van Neerven. Stochastic Evolution Equations, 3 2020. URL https://ocw.tudelft.nl/courses/stochastic-evolution-equations/subjects/lecture-4-gaussian-random-variables/.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization, 2013. URL https://arxiv.org/abs/1307.1493.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Sida Wang and Christopher Manning. Fast dropout training. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 118–126, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/wang13a.html.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10181–10192. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wei20d.html.

Denny Wu and Ji Xu. On the optimal weighted \ell_2 regularization in overparameterized linear regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10112–10123. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/72e6d3238361fe70f22fb0ac624a7072-Paper.pdf.

M. Xia, J. Wang, and Y. He. BrainNet Viewer: A network visualization tool for human brain connectomics. *PLoS ONE*, 8(7):e68910, 2013.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Tianle Yang and Joe Suzuki. The functional lingam. In Antonio Salmerön and Rafael Rumï, editors, *The 11th International Conference on Probabilistic Graphical Models*, volume 186, pages 25–36. PMLR, 05–07 Oct 2022.

Fangting Zhou, Kejun He, Kunbo Wang, Yanxun Xu, and Yang Ni. Functional bayesian networks for discovering causality from multivariate functional data, 2022.