A Survey on High-Dimensional Regression

Caroline Wang* Zhi Yu Yuan†

June 1, 2026

Abstract

Modern applications of statistical theory and methods often require models that are both interpretable and flexible enough to capture complex data-generating processes. Semiparametric models provide such a framework by combining finite-dimensional parameters of interest with infinite-dimensional or otherwise unspecified nuisance components. In this article, we introduce the motivation behind semiparametric modeling and explain how it differs from fully parametric and fully nonparametric approaches. We then present examples of semiparametric models, including restricted moment models and nonparametric models in which the parameter of interest is a finite-dimensional functional of an infinite-dimensional distribution. To develop the geometric tools needed for semiparametric inference, we introduce Hilbert spaces of mean-zero, square-integrable random functions, together with the notions of inner products, orthogonality, and projections. Finally, we connect this geometry to influence functions, showing how the first-order behavior and asymptotic variance of semiparametric estimators can be studied through elements of a Hilbert space. This perspective provides the foundation for understanding efficiency, nuisance parameters, and further applications of semiparametric theory, including problems involving missing data.

Contents

*Department of Mathematics and Statistics, caroline.wang2@mail.mcgill.ca
†Department of Mathematics and Statistics, zhiyu.yuan@mail.mcgill.ca

1 Introduction

1.1 Motivation

For most statistical problems, probability models are used to describe the data-generating process. Let $Z_1, \ldots, Z_n$ denote the observed data, where each random vector Z_i contains the variables measured on the i-th individual in a sample of n individuals drawn from a population of interest. Throughout this discussion, we assume that $Z_1, \ldots, Z_n$ are independent and identically distributed, with common density p_0. A statistical model is then defined as a collection of densities

$$\mathcal{P} = \{p_\theta : \theta \in \Theta\},$$

where each density in $\mathcal{P}$ is considered a plausible candidate for the true data-generating density p_0.

In classical parametric statistics, the model is indexed by a finite-dimensional parameter θ. That is, the parameter space Θ is a subset of some Euclidean space $\mathbb{R}^p$, where $p < \infty$. Such a model can be written as

$$\mathcal{P} = \{p(z;\theta) : \theta \in \Omega \subset \mathbb{R}^p\}.$$

The parameter θ is used to characterize the possible densities in the model, and statistical inference is typically concerned with estimating θ, or some function of θ, from the observed data. For example, in normal linear regression, the model is described by a finite number of parameters, such as the regression coefficients β and the error variance σ^2. The main goal is often to estimate β, which describes the relationship between the response variable and the covariates.

However, finite-dimensional parametric models can be restrictive. They require the statistician to specify a relatively small class of possible distributions in advance. If the chosen model is misspecified, then the resulting estimators and conclusions may be biased or misleading. To obtain more flexible models, it is natural to consider larger model classes in which some components of the distribution are left unspecified. This leads us to models indexed by infinite-dimensional parameters.

In such settings, the full parameter θ may no longer belong to a finite-dimensional Euclidean space. Instead, θ may contain an infinite-dimensional component, such as an unknown density, regression function, baseline hazard function, or error distribution. Importantly, we may still be interested in estimating only a finite-dimensional quantity. We therefore write the parameter as

$$\theta = (\beta, \eta),$$

where $\beta \in \mathbb{R}^q$ is the finite-dimensional parameter of interest, and η is an infinite-dimensional nuisance parameter. The nuisance parameter η is not the main object of inference, but it is necessary to describe the full statistical model.

A model of this form is called a semiparametric model. More generally, a semiparametric model is a statistical model that contains both a finite-dimensional parametric component and an infinite-dimensional nonparametric component:

$$\mathcal{P} = \{p(z;\beta,\eta) : \beta \in \mathcal{B} \subset \mathbb{R}^q,\ \eta \in \mathcal{H}\},$$

where $\mathcal{H}$ is an infinite-dimensional space. In some cases, the parameter of interest may also be written as a functional of the full parameter,

$$\beta = \beta(\theta).$$

The motivation for studying semiparametric models is that they provide a compromise between fully parametric and fully nonparametric approaches. The finite-dimensional component β allows us to focus on an interpretable parameter of scientific interest, while the infinite-dimensional component η gives the model enough flexibility to avoid imposing unnecessary assumptions on the data-generating process. By allowing part of the parameter space to be infinite-dimensional, semiparametric models reduce model restrictions and can lead to inference procedures that are more robust and broadly applicable.

1.2 Infinite-dimensional Spaces

The parameter spaces considered in semiparametric theory are typically subsets of linear spaces. More precisely, we write $\Omega \subset \mathcal{S}$, where $\mathcal{S}$ is a linear vector space. A space $\mathcal{S}$ is said to be linear if, for any two elements $\theta_1, \theta_2 \in \mathcal{S}$ and any scalars $a, b \in \mathbb{R}$, the linear combination $a\theta_1 + b\theta_2$ also belongs to $\mathcal{S}$.

A linear space is finite-dimensional if it can be spanned by a finite number of elements. That is, $\mathcal{S}$ is finite-dimensional if there exist elements $s_1, \ldots, s_m \in \mathcal{S}$ such that every element $s \in \mathcal{S}$ can be written as

$$s = \sum_{j=1}^{m} a_j s_j$$

for some constants $a_1, \ldots, a_m \in \mathbb{R}$. The dimension of a finite-dimensional linear space is the minimum number of linearly independent elements required to span the entire space. Conversely, a linear space that cannot be spanned by any finite set of elements is called infinite-dimensional.

An example of an infinite-dimensional linear space is the space of continuous real-valued functions on the real line. Let $\mathcal{S} = C(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous}\}$. This is a linear space because if $f, g \in C(\mathbb{R})$ and $a, b \in \mathbb{R}$, then

$$af + bg$$

is also continuous, and therefore $af + bg \in C(\mathbb{R})$.

To see that $C(\mathbb{R})$ is infinite-dimensional, it is enough to show that it contains finite-dimensional subspaces of arbitrarily large dimension. For each integer $m \geq 0$, consider the space of polynomials of degree at most m,

$$\mathcal{S}_m = \left\{ f(x) = \sum_{j=0}^{m} a_j x^j : a_0, \ldots, a_m \in \mathbb{R} \right\}.$$

Since every polynomial is continuous, we have $\mathcal{S}_m \subset C(\mathbb{R})$. The space $\mathcal{S}_m$ is spanned by the functions $1, x, x^2, \ldots, x^m$. Moreover, these functions are linearly independent. Indeed, suppose that for some constants $a_0, \ldots, a_m$,

$$\sum_{j=0}^{m} a_j x^j = 0 \quad \text{for all } x \in \mathbb{R}.$$

Then the polynomial on the left-hand side is identically zero, which implies that all of its coefficients must be zero:

$$a_0 = a_1 = \cdots = a_m = 0.$$

Therefore, $1, x, x^2, \ldots, x^m$ form a basis for $\mathcal{S}_m$, and so

$$\dim(\mathcal{S}_m) = m + 1.$$

Since m can be chosen arbitrarily large, the space $C(\mathbb{R})$ contains subspaces of arbitrarily large finite dimension. Hence, $C(\mathbb{R})$ cannot itself be spanned by any finite number of functions. Therefore,$C(\mathbb{R})$ is an infinite-dimensional linear space.

This example illustrates why infinite-dimensional parameter spaces naturally arise in semiparametric models. When a component of the data-generating process is an unknown function, such as an unknown regression function, density, or baseline hazard function, it cannot generally be described by finitely many parameters. Semiparametric models therefore allow us to retain a finite-dimensional parameter of interest, such as β, while allowing the nuisance component η to belong to an infinite-dimensional function space. This added flexibility reduces the need for restrictive parametric assumptions and makes the resulting model more robust to misspecification.

1.3 Example of Semiparametric Models

Restricted Moment Models

A common statistical problem is to model the relationship between a response variable Y, which may be vector-valued, and a vector of covariates X. Throughout, we use the convention that an unindexed random vector Z represents a single observation. In this setting, we write $Z = (Y, X)$, and consider a family of probability distributions for Z satisfying the conditional moment restriction

$$E(Y \mid X) = \mu(X, \beta),$$

where $\mu(X, \beta)$ is a known function of the covariates X and of an unknown finite-dimensional parameter $\beta \in \mathbb{R}^q$. The function μ may be linear or nonlinear in β. For example, in a linear regression model, one may take

$$\mu(X, \beta) = \beta^\intercal X,$$

whereas in a log-linear model, one may take $\mu(X, \beta) = \exp(\beta^\intercal X)$.

The restriction $E(Y \mid X) = \mu(X, \beta)$ does not specify the full conditional distribution of Y given X. It only specifies its conditional mean. This distinction is important because two conditional distributions may have the same mean while differing in variance, skewness, tail behavior, or overall shape. Thus, the model imposes a restriction on only one feature of the conditional distribution, namely its first conditional moment, while leaving many other features unspecified.

Equivalently, the model may be written as

$$Y = \mu(X, \beta) + \epsilon,$$

where $\epsilon = Y - \mu(X, \beta)$. The conditional moment restriction then becomes

$$E(\epsilon \mid X) = 0.$$

This means that, after accounting for the systematic component $\mu(X, \beta)$, the remaining error term has conditional mean zero. In other words, the model assumes that $\mu(X, \beta)$ correctly describes the conditional average of Y given X, but it does not require a full specification of the conditional distribution of the error term.

Models defined by such conditional moment restrictions are called *restricted moment models*. They are semiparametric because the parameter of interest β is finite-dimensional, while the remaining components of the data-generating distribution are left unspecified and are therefore represented by infinite-dimensional nuisance parameters.

To make this explicit, suppose for simplicity that Y is a one-dimensional continuous random variable. Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be independent and identically distributed observations from the joint distribution of (Y, X). For a single observation, the joint density may be written as

$$p_{Y,X}\{y, x; \beta, \eta(\cdot)\},$$

where $\eta(\cdot)$ denotes the infinite-dimensional nuisance component describing the parts of the distribution that are not determined by β.

Since

$$\epsilon = Y - \mu(X, \beta),$$

we can express the joint density of (Y, X) in terms of the joint density of (ϵ, X). For fixed x and β, the transformation from Y to ϵ is simply a translation:

$$\epsilon = y - \mu(x, \beta).$$

Since this transformation has Jacobian equal to one, we obtain

$$p_{Y,X}(y, x) = p_{\epsilon,X}\{y - \mu(x, \beta), x\}.$$

The joint density of (ϵ, X) can be decomposed as

$$p_{\epsilon,X}(\epsilon, x) = p_{\epsilon|X}(\epsilon \mid x) p_X(x).$$

The conditional density $p_{\epsilon|X}$ must satisfy

$$p_{\epsilon|X}(\epsilon \mid x) \geq 0 \quad \text{for all } \epsilon, x,$$

$$\int p_{\epsilon|X}(\epsilon \mid x)\, d\epsilon = 1 \quad \text{for all } x,$$

$$\int \epsilon p_{\epsilon|X}(\epsilon \mid x)\, d\epsilon = 0 \quad \text{for all } x,$$

and

$$\int p_X(x) d\nu_X(x) = 1.$$

The first two conditions are the usual requirements for a valid conditional density, while the third condition encodes the model assumption

$$E(\epsilon \mid X = x) = 0.$$

Similarly, the marginal density of X must satisfy

$$p_X(x) \geq 0$$

and

$$\int p_X(x)\, d\nu_X(x) = 1,$$

where ν_X denotes an appropriate dominating measure for X.

The central point is that the conditional mean restriction does not determine the full conditional distribution of ϵ given X. It only requires this distribution to have mean zero for every value of X.

Therefore, many different conditional densities $p_{\epsilon|X}$ are compatible with the same finite-dimensional parameter β.

To see why this leads to an infinite-dimensional nuisance parameter, consider the following construction. Begin with an arbitrary positive function

$$h^{(0)}(\epsilon, x) > 0,$$

subject to suitable regularity and integrability conditions. Since $h^{(0)}$ is not necessarily a density, normalize it by defining

$$h^{(1)}(\epsilon, x) = \frac{h^{(0)}(\epsilon, x)}{\int h^{(0)}(u, x)\, du}.$$

Then

$$\int h^{(1)}(\epsilon, x)\, d\epsilon = 1 \quad \text{for all } x.$$

Let ϵ^* be a random variable whose conditional density given $X = x$ is $h^{(1)}(\epsilon, x)$. Its conditional mean is

$$m(x) = \int \epsilon h^{(1)}(\epsilon, x)\, d\epsilon.$$

This random variable does not necessarily satisfy the required mean-zero condition. To construct an error variable that does, define $\epsilon = \epsilon^* - m(X)$. Then, for every x,

$$E(\epsilon \mid X = x) = E(\epsilon^* \mid X = x) - m(x) = 0.$$

Since the transformation from ϵ^* to ϵ is a translation, its Jacobian is equal to one. Therefore, the conditional density of ϵ given $X = x$ is

$$\eta_1(\epsilon, x) = h^{(1)}\{\epsilon + m(x), x\},$$

where

$$m(x) = \int u h^{(1)}(u, x)\, du.$$

By construction, η_1 satisfies

$$\eta_1(\epsilon, x) > 0,$$

$$\int \eta_1(\epsilon, x)\, d\epsilon = 1 \quad \text{for all } x,$$

and

$$\int \epsilon \eta_1(\epsilon, x)\, d\epsilon = 0 \quad \text{for all } x.$$

Because the initial function $h^{(0)}(\epsilon, x)$ was chosen almost arbitrarily from a class of positive functions, the resulting class of conditional densities $\eta_1(\epsilon, x)$ is infinite-dimensional. In other words, the conditional mean restriction still allows infinitely many possible shapes for the conditional error distribution.

Similarly, the marginal density of X may be written as

$$p_X(x) = \eta_2(x),$$

where

$$\eta_2(x) > 0$$

and

$$\int \eta_2(x)\, d\nu_X(x) = 1.$$

As long as the support of X is sufficiently rich, the collection of possible marginal densities η_2 is also infinite-dimensional.

Therefore, the restricted moment model can be parametrized by

$$\{\beta, \eta_1, \eta_2\},$$

where $\beta \in \mathbb{R}^q$ is the finite-dimensional parameter of interest, $\eta_1(\epsilon, x) = p_{\epsilon|X}(\epsilon \mid x)$ is the conditional density of the error given the covariates, and $\eta_2(x) = p_X(x)$ is the marginal density of the covariates. The joint density of (Y, X) can then be written as

$$p_{Y,X}\{y, x; \beta, \eta_1, \eta_2\} = \eta_1\{y - \mu(x, \beta), x\}\eta_2(x).$$

This shows that restricted moment models are semiparametric. The model contains a finite-dimensional parameter of interest, β, which determines the conditional mean structure. At the same time, it contains infinite-dimensional nuisance parameters, η_1 and η_2, which describe the conditional distribution of the error and the marginal distribution of the covariates. Thus, restricted moment models preserve the interpretability of a parametric regression model while allowing greater flexibility in the underlying data-generating distribution.

Nonparametric Model

In the previous example, the probability model was indexed by an infinite-dimensional parameter that could be decomposed as $\theta = (\beta, \eta)$, where $\beta \in \mathbb{R}^q$ was the finite-dimensional parameter of interest, and η was an infinite-dimensional nuisance parameter. This type of decomposition is useful when the parameter of interest appears explicitly as one component of the full parameter.

We now consider a different type of semiparametric problem, where the object of interest is not necessarily obtained by separating the full parameter into a finite-dimensional component and a nuisance component. Instead, suppose that we observe a single random variable Z, and that we want to estimate certain moments of its distribution. In this setting, we impose no parametric restrictions on the distribution of Z, except that the moments of interest must exist.

Let the density of Z be denoted by $\theta(z)$. Here, $\theta(\cdot)$ represents the full distribution of Z. We allow $\theta(z)$ to be any valid density satisfying

$$\theta(z) \geq 0$$

and

$$\int \theta(z)\, d\nu_Z(z) = 1,$$

together with any additional integrability conditions required for the moments of interest to be well-defined. For example, if we are interested in the mean of Z, then we require

$$\int |z|\theta(z)\, d\nu_Z(z) < \infty.$$

If we are interested in the variance, then we require the corresponding second moment to be finite.

Because $\theta(\cdot)$ is allowed to vary over a broad class of densities, the parameter space is infinite-dimensional whenever the support of Z is infinite. Indeed, specifying an unrestricted density generally requires specifying an entire function, not merely a finite vector of real numbers.

Suppose now that the parameter of interest is a functional of the density $\theta(\cdot)$, denoted by $\beta(\theta)$. For instance, if the parameter of interest is the mean of Z, then

$$\beta(\theta) = E_\theta(Z) = \int z\theta(z)\, d\nu_Z(z).$$

If the parameter of interest is the variance of Z, then

$$\beta(\theta) = \int \{z - E_\theta(Z)\}^2 \theta(z)\, d\nu_Z(z).$$

In this formulation, β is not treated as a separate component of the full parameter θ. Rather, β is a finite-dimensional summary of the infinite-dimensional object $\theta(\cdot)$. Therefore, it is often more natural to work directly with the functional $\beta(\theta)$ instead of trying to partition the parameter space into a parameter of interest and a nuisance parameter.

This gives another way of viewing semiparametric models. The full model is nonparametric because the distribution $\theta(\cdot)$ is unrestricted and infinite-dimensional. However, the parameter of interest $\beta(\theta)$ is finite-dimensional. Thus, the model is semiparametric in the sense that we conduct inference on a finite-dimensional feature of an otherwise infinite-dimensional distribution.

1.4 Semiparametric Estimators

In a semiparametric model, a semiparametric estimator of β, denoted by $\widehat{\beta}_n$, is an estimator that, loosely speaking, is both consistent and asymptotically normal. Consistency means that, under the probability law indexed by the true parameters (β, η), the estimator converges in probability to the true value of β:

$$\widehat{\beta}_n - \beta \xrightarrow{P_{\beta,\eta}} 0.$$

Equivalently,

$$\widehat{\beta}_n \xrightarrow{P_{\beta,\eta}} \beta.$$

Asymptotic normality means that, after scaling by $\sqrt{n}$, the estimation error converges in distribution to a normal random vector:

$$\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{\mathcal{D}\{\beta,\eta\}} N\left(0, \Sigma(\beta, \eta)\right),$$

where $\Sigma(\beta, \eta)$ is a $q \times q$ covariance matrix that may depend on both the finite-dimensional parameter of interest β and the infinite-dimensional nuisance parameter η.

Several important questions arise when studying semiparametric models. In particular, we are interested in the following problems:

1. How can we construct semiparametric estimators of the finite-dimensional parameter of interest β? More fundamentally, under what conditions do such estimators exist?

2. Among all possible semiparametric estimators of β, how can we determine which estimator is optimal, or most efficient?

Both questions require an understanding of the geometry underlying semiparametric estimation. In particular, the behavior of a semiparametric estimator is closely connected to its *influence function*, which describes the first-order effect of small perturbations in the underlying distribution on the estimator.

The study of influence functions naturally leads to geometric ideas. To compare estimators, we need to understand notions such as projection, orthogonality, and minimum distance in appropriate function spaces. These concepts are most conveniently developed in the framework of Hilbert spaces. In this setting, efficient estimation can be understood geometrically: an efficient estimator is one whose influence function has the smallest possible variance among all valid influence functions.

Therefore, before discussing semiparametric efficiency in detail, we introduce the geometry of Hilbert spaces. We will focus on the notions of inner products, orthogonality, projections, and minimum distance, and then explain how these ideas are used to characterize efficient estimators in semiparametric models.

2 Hilbert Space for Random Vectors

2.1 Space of Mean-Zero q-dimensional Radom Functions

Let Z denote the random vector corresponding to a single observation. As usual, we assume that Z is defined on an underlying probability space $(\mathcal{F}, \mathcal{A}, P)$, where $\mathcal{F}$ is the sample space, $\mathcal{A}$ is the associated σ-algebra, and P is the probability measure. For the moment, we do not consider a statistical model consisting of a family of probability measures. Instead, we fix P and assume that it is the true probability measure generating the observation Z.

We now consider the space of q-dimensional random functions of Z. More precisely, let h be a measurable function of the form

$$h : \mathcal{F} \to \mathbb{R}^q,$$

or, equivalently, a random vector $h(Z)$. We restrict attention to functions satisfying

$$E\{h(Z)\} = 0$$

and

$$E\{h^\intercal(Z)h(Z)\} < \infty.$$

The first condition means that $h(Z)$ has mean zero. This centering condition is important because influence functions are typically required to have expectation zero. The second condition means that $h(Z)$ has finite second moment. Since $h^\intercal(Z)h(Z) = \|h(Z)\|^2$, this condition ensures that $h(Z)$ has finite variance.

These two assumptions are essential for asymptotic theory. In particular, finite second moments allow us to apply central limit theorem arguments to averages of such functions. Without this condition, the variance may be infinite, and the usual Gaussian approximations used to construct confidence intervals and hypothesis tests may fail.

Since the elements of this space are random functions, we will often write h instead of $h(Z)$ when there is no risk of confusion. Thus, when we refer to an element h, we implicitly mean the random vector obtained by evaluating the function at the observation Z.

The collection of all such functions forms a linear space. Indeed, suppose h_1 and h_2 both satisfy

$$E\{h_1(Z)\} = 0, \qquad E\{h_2(Z)\} = 0,$$

and

$$E\{h_1^\intercal(Z)h_1(Z)\} < \infty, \qquad E\{h_2^\intercal(Z)h_2(Z)\} < \infty.$$

Then, for any constants $a, b \in \mathbb{R}$, the linear combination $ah_1 + bh_2$ also has mean zero, since

$$E\{ah_1(Z) + bh_2(Z)\} = aE\{h_1(Z)\} + bE\{h_2(Z)\} = 0.$$

Moreover, $ah_1 + bh_2$ also has finite second moment. Therefore, it belongs to the same space.

Thus, the space of mean-zero, square-integrable q-dimensional random functions of Z is a linear space. This space will serve as the basic geometric setting in which we define inner products, orthogonality, projections, and ultimately the influence functions used to study semiparametric efficiency.

2.2 Hilbert Space

A Hilbert space, denoted by $\mathcal{H}$, is a complete normed linear vector space whose norm is induced by an inner product. In addition to the linear structure, a Hilbert space allows us to define geometric notions such as distance, angles, projections, and orthogonality. These notions are introduced through the inner product.

Definition 1. *Let $\mathcal{H}$ be a real linear vector space. An inner product on $\mathcal{H}$ is a function*

$$\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$$

that assigns to each pair of elements $h_1, h_2 \in \mathcal{H}$ a scalar $\langle h_1, h_2 \rangle$, and satisfies the following properties:

1. ***Symmetry:***
$$\langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle.$$

2. ***Additivity:***
$$\langle h_1 + h_2, h_3 \rangle = \langle h_1, h_3 \rangle + \langle h_2, h_3 \rangle.$$

3. ***Homogeneity:***
$$\langle \lambda h_1, h_2 \rangle = \lambda \langle h_1, h_2 \rangle,$$
for any scalar $\lambda \in \mathbb{R}$.

4. ***Positive definiteness:***
$$\langle h, h \rangle \geq 0,$$
with equality if and only if $h = 0$.

The first three properties express linearity of the inner product in one of its arguments, while the fourth property ensures that the inner product can be used to define a meaningful notion of length. Indeed, once an inner product is defined, the norm of an element $h \in \mathcal{H}$ is defined by

$$\|h\| = \langle h, h \rangle^{1/2}.$$

The distance between two elements $h_1, h_2 \in \mathcal{H}$ is then given by

$$\|h_1 - h_2\|.$$

In some settings, the function $\langle \cdot, \cdot \rangle$ may satisfy the algebraic properties above, but $\langle h, h \rangle = 0$ may only imply that $h = 0$ almost everywhere, rather than pointwise. This occurs frequently in spaces of random variables, where two functions that differ only on an event of probability zero are statistically indistinguishable. In such cases, we define the Hilbert space in terms of equivalence classes of functions, where two functions are identified if they are equal almost surely.

For the linear space of q-dimensional measurable random functions with mean zero and finite second moment, we define the inner product between two elements h_1 and h_2 by

$$\langle h_1, h_2 \rangle = E\{h_1^\intercal(Z) h_2(Z)\}.$$

We refer to this as the *covariance inner product*. Since the functions have mean zero, this inner product corresponds to the covariance-type quantity between $h_1(Z)$ and $h_2(Z)$. In particular,

$$\langle h, h \rangle = E\{h^\intercal(Z) h(Z)\} = E\{\|h(Z)\|^2\}.$$

Thus, the norm induced by this inner product is

$$\|h\| = \left[E\{h^\intercal(Z) h(Z)\}\right]^{1/2}.$$

This inner product satisfies symmetry, additivity, and homogeneity by the linearity of expectation and the usual properties of matrix multiplication. The positive definiteness condition holds up to almost sure equality. More precisely,$\langle h, h \rangle = 0$ implies $E\{\|h(Z)\|^2\} = 0$, which implies that

$$h(Z) = 0 \quad \text{almost surely.}$$

Therefore, we identify two elements h_1 and h_2 whenever

$$h_1(Z) = h_2(Z) \quad \text{almost surely,}$$

or equivalently,

$$P\{h_1(Z) \neq h_2(Z)\} = 0.$$

We will generally not emphasize these measure-theoretic subtleties and will simply treat functions that are equal almost surely as the same element of the space.

Finally, the inner product allows us to define orthogonality. Two elements $h_1, h_2 \in \mathcal{H}$ are said to be orthogonal if

$$\langle h_1, h_2 \rangle = 0.$$

In the present setting, this means

$$E\{h_1^\intercal(Z) h_2(Z)\} = 0.$$

This notion of orthogonality will be central in semiparametric theory, because efficient influence functions can be characterized using projections onto orthogonal subspaces of a Hilbert space.

2.3 Projection Theorem

Let $\mathcal{U} \subset \mathcal{H}$. We say that $\mathcal{U}$ is a linear subspace of $\mathcal{H}$ if, whenever $u_1, u_2 \in \mathcal{U}$, we also have

$$a u_1 + b u_2 \in \mathcal{U}$$

for all scalars $a, b \in \mathbb{R}$. In particular, every linear subspace must contain the origin. Indeed, by choosing $a = b = 0$, we obtain $0u_1 + 0u_2 = 0 \in \mathcal{U}$.

One of the most important geometric results in Hilbert spaces is the projection theorem. This theorem formalizes the idea that, given an element $h \in \mathcal{H}$ and a closed linear subspace $\mathcal{U}$, there is a unique element of $\mathcal{U}$ that is closest to h.

Theorem 1 (Projection Theorem)**.** *Let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{U} \subset \mathcal{H}$ be a closed linear subspace. Then, for every $h \in \mathcal{H}$, there exists a unique element $u_0 \in \mathcal{U}$ such that*

$$\|h - u_0\| \leq \|h - u\| \quad \textit{for all } u \in \mathcal{U}.$$

Moreover, the residual $h - u_0$ is orthogonal to $\mathcal{U}$, meaning that

$$\langle h - u_0, u \rangle = 0 \quad \textit{for all } u \in \mathcal{U}.$$

The element u_0 is called the projection of h onto $\mathcal{U}$, and we denote it by $\Pi(h \mid \mathcal{U})$. Furthermore, u_0 is the only element $u \in \mathcal{U}$ such that

$$h - u$$

is orthogonal to $\mathcal{U}$.

The projection theorem is the Hilbert space analogue of the familiar idea of projecting a vector onto a subspace in finite-dimensional Euclidean space. In $\mathbb{R}^n$, the closest point in a subspace is obtained by dropping a perpendicular line from the vector to the subspace. The same intuition applies here: the projection u_0 is the element of $\mathcal{U}$ closest to h, and the difference

$$h - u_0$$

is perpendicular, or orthogonal, to every element of $\mathcal{U}$.

The assumption that $\mathcal{U}$ is closed is essential. Without closedness, a closest element may fail to exist, even if there are elements of $\mathcal{U}$ that get arbitrarily close to h. The completeness of the Hilbert space is also important because it guarantees that limiting arguments used to construct the projection remain inside the space.

A direct consequence of orthogonality is the Pythagorean theorem, which extends naturally from Euclidean geometry to Hilbert spaces.

Theorem 2 (Pythagorean Theorem)**.** *If $h_1, h_2 \in \mathcal{H}$ are orthogonal, that is,*

$$\langle h_1, h_2 \rangle = 0,$$

then

$$\|h_1 + h_2\|^2 = \|h_1\|^2 + \|h_2\|^2.$$

Indeed, this follows directly from the definition of the norm induced by the inner product:

$$\|h_1 + h_2\|^2 = \langle h_1 + h_2, h_1 + h_2 \rangle.$$

Expanding the right-hand side gives

$$\|h_1 + h_2\|^2 = \langle h_1, h_1 \rangle + 2\langle h_1, h_2 \rangle + \langle h_2, h_2 \rangle.$$

Since h_1 and h_2 are orthogonal, $\langle h_1, h_2 \rangle = 0$, and therefore

$$\|h_1 + h_2\|^2 = \|h_1\|^2 + \|h_2\|^2.$$

These results will be central in semiparametric theory. In particular, projections in Hilbert spaces will allow us to decompose influence functions into components that lie in a nuisance tangent space and components that are orthogonal to it. This geometric decomposition is the key to understanding efficient influence functions and efficient estimators.

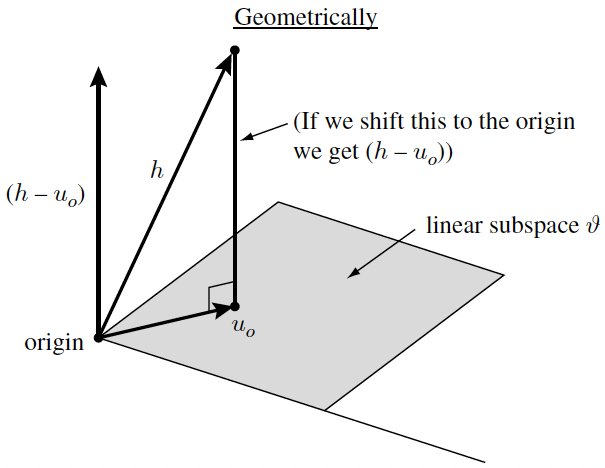

Fig. 2.1. Projection onto a linear subspace

3 The Geometry of Influence Functions

We now connect the previous geometric discussion of Hilbert spaces to the study of influence functions. Recall that an estimator $\widehat{\beta}_n$ of a finite-dimensional parameter $\beta_0 \in \mathbb{R}^q$ is said to be asymptotically linear if it admits a representation of the form

$$\widehat{\beta}_n - \beta_0 = \frac{1}{n}\sum_{i=1}^{n} \varphi(Z_i) + o_p(n^{-1/2}),$$

where $\varphi(Z)$ is a mean-zero, square-integrable q-dimensional random function:

$$E\{\varphi(Z)\} = 0, \qquad E\{\varphi^\intercal(Z)\varphi(Z)\} < \infty.$$

The function φ is called the *influence function* of the estimator.

Multiplying the asymptotically linear representation by $\sqrt{n}$, we obtain

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varphi(Z_i) + o_p(1).$$

Thus, the first-order asymptotic behavior of $\widehat{\beta}_n$ is entirely determined by its influence function. In particular, by the multivariate central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varphi(Z_i) \xrightarrow{\mathcal{D}} N\left(0, E\{\varphi(Z)\varphi^\intercal(Z)\}\right).$$

Therefore, by Slutsky's theorem,

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N\left(0, E\{\varphi(Z)\varphi^\intercal(Z)\}\right).$$

This shows that, asymptotically, an asymptotically linear estimator can be studied through its influence function. The influence function determines both the limiting distribution and the asymptotic variance of the estimator. This observation is fundamental in semiparametric theory because it allows us to compare estimators by comparing their influence functions.

We now show that the influence function of an asymptotically linear estimator is unique, up to almost sure equality.

Theorem 3. *An asymptotically linear estimator has a unique influence function, up to almost sure equality.*

Proof. Suppose, for contradiction, that the estimator $\widehat{\beta}_n$ admits two influence functions, $\varphi(Z)$ and $\varphi^*(Z)$. Then both functions are mean-zero and square-integrable, and we have

$$\widehat{\beta}_n - \beta_0 = \frac{1}{n}\sum_{i=1}^{n} \varphi(Z_i) + o_p(n^{-1/2})$$

and

$$\widehat{\beta}_n - \beta_0 = \frac{1}{n}\sum_{i=1}^{n} \varphi^*(Z_i) + o_p(n^{-1/2}).$$

Subtracting the two representations gives

$$\frac{1}{n}\sum_{i=1}^{n}\{\varphi(Z_i) - \varphi^*(Z_i)\} = o_p(n^{-1/2}).$$

Multiplying both sides by $\sqrt{n}$, we obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi(Z_i) - \varphi^*(Z_i)\} = o_p(1).$$

Define

$$\Delta(Z) = \varphi(Z) - \varphi^*(Z).$$

Since both φ and φ^* have mean zero, we have

$$E\{\Delta(Z)\} = 0.$$

Also, since both functions are square-integrable, Δ is square-integrable. By the multivariate central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Delta(Z_i) \xrightarrow{\mathcal{D}} N\left(0, E\{\Delta(Z)\Delta^\intercal(Z)\}\right).$$

However, from the previous equation, this same sequence converges in probability to zero. Therefore, its limiting distribution must be degenerate at zero. This is possible only if

$$E\{\Delta(Z)\Delta^\intercal(Z)\} = 0.$$

Equivalently,

$$E\left[\{\varphi(Z) - \varphi^*(Z)\}\{\varphi(Z) - \varphi^*(Z)\}^\intercal\right] = 0.$$

In particular,

$$E\left[\|\varphi(Z) - \varphi^*(Z)\|^2\right] = 0.$$

Hence,

$$\varphi(Z) = \varphi^*(Z) \quad \text{almost surely.}$$

Therefore, the influence function is unique up to almost sure equality. □

The representation of estimators through their influence functions naturally leads to a geometric interpretation. Since influence functions are mean-zero, square-integrable random functions, they belong to the Hilbert space introduced earlier. Consequently, comparing estimators can be viewed as comparing elements of this Hilbert space. In particular, notions such as distance, orthogonality, projection, and minimum variance will allow us to characterize efficient influence functions and, consequently, efficient estimators.

4 Final Remarks

Semiparametric models provide a powerful framework for statistical inference because they combine the interpretability of finite-dimensional parameters with the flexibility of infinite-dimensional nuisance components. We have seen that this structure makes semiparametric models especially useful when the main scientific question concerns a low-dimensional parameter, while other parts of the data-generating process are too complex or unrealistic to specify parametrically. The Hilbert space formulation further shows that semiparametric inference has a natural geometric interpretation: influence functions can be viewed as elements of a function space, and efficiency can be studied through projections, orthogonality, and variance minimization.

One important direction in which semiparametric methods become especially valuable is the study of missing data. In many real-world datasets, some variables are only partially observed due to nonresponse, dropout, loss to follow-up, or incomplete measurement. A fully parametric approach would require strong assumptions about both the outcome model and the missingness mechanism, which may be difficult to justify in practice. Semiparametric models offer a more flexible alternative: they allow the parameter of interest, such as a population mean, treatment effect, or regression coefficient, to remain finite-dimensional while treating the distribution of the observed data and the missingness process as nuisance components.

This makes missing data problems a natural continuation of the ideas developed in this paper. The concepts of influence functions, nuisance parameters, and Hilbert space projections provide the foundation for constructing estimators that remain consistent and efficient even when part of the data is missing. In particular, semiparametric theory helps clarify how much information is lost due to missingness, what assumptions are needed to recover the parameter of interest, and how to construct estimators that use the observed data as efficiently as possible. Thus, missing data provides an important and practical setting in which the geometric tools of semiparametric inference can be further explored.

5 References

[Tsi06] Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.