

Article

Geometric Residual Projection in Linear Regression: Rank-Aware Operators and a Geometric Multicollinearity Index

Mais Alkhateeb ^{1,2,*}  and Samir Brahim Belhaouari ¹ ¹ College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar; sbelhaouari@hbku.edu.qa² College of Education and Arts, Lusail University, Doha 9717, Qatar

* Correspondence: malakhateeb@lu.edu.qa

Abstract

Residuals play a central role in linear regression, yet their geometry is often hidden by inverse- and pseudoinverse-based formulas. We develop a rank-aware framework for residual projection that makes the underlying orthogonality explicit. When the design matrix has codimension one, the unexplained component of the response lies along a single unit normal to the predictor space, and the residual projector reduces to the rank-one operator $nn^\top$, avoiding matrix inversion. For general designs, the residual lies in a higher-dimensional orthogonal complement spanned by an orthonormal basis N , and the residual projector factorizes as $NN^\top$. Using cross-products, wedge products, and Gram determinants, we provide basis-independent characterizations of the residual subspace. We further introduce the Geometric Multicollinearity Index (GMI), a scale-invariant diagnostic derived from the polar sine that quantifies the collapse of predictor-space volume under multicollinearity. Synthetic perturbation studies and an illustrative real-data experiment show that the proposed projectors reproduce ordinary least squares residuals, that GMI responds predictably to controlled collinearity, and that the projector viewpoint clarifies the distinction between regression residuals and PCA reconstruction residuals in both full-rank and rank-deficient settings.

Keywords: linear regression; orthogonal projection; geometric residuals; multicollinearity; polar sine; Gram determinant; geometric multicollinearity index (GMI); null-space

MSC: 62J05; 62H20; 15A03



Academic Editors: Fajie Wang, Ji Lin and Yingbin Chai

Received: 29 November 2025

Revised: 26 January 2026

Accepted: 29 January 2026

Published: 17 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Linear regression remains one of the most widely used modeling tools in statistics, data science, and machine learning [1]. In its familiar form

$$Y = X\beta + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0, \quad X \in \mathbb{R}^{n \times p}, \quad Y \in \mathbb{R}^n, \quad \beta \in \mathbb{R}^p, \quad (1)$$

the ordinary least squares (OLS) estimator chooses coefficients that minimize the squared discrepancy between observations and predictions. When the design matrix X has the full column rank, the fitted values and the associated orthogonal projector onto the column space $C(X)$ take the classical form

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y, \quad \hat{Y} = X\hat{\beta}, \quad P = X(X^\top X)^{-1} X^\top, \quad (2)$$

with P symmetric and idempotent [2]. More generally, the orthogonal projector onto $C(X)$ can be written as $P = XX^+$ in terms of the Moore–Penrose pseudoinverse X^+ [2].

The fitted component and residual are

$$\hat{Y} = PY, \quad r = (I - P)Y, \quad (3)$$

so that $r \in C(X)^\perp$ and

$$Y = PY + r,$$

is the orthogonal decomposition of the response into a part explained by $C(X)$ and a part orthogonal to it. This algebraic formulation is standard, but it can obscure the underlying geometry that governs the behavior of the residuals, and inversion-based formulas can be fragile in high-dimensional or nearly singular regimes [2].

A classical identity from numerical linear algebra makes the geometry explicit:

$$I - P = NN^\top, \quad (4)$$

where the columns of N form an orthonormal basis of $C(X)^\perp$ [2]. When $\rho = n - 1$, the orthogonal complement is one-dimensional and the projector reduces to the rank-one form

$$I - P = nn^\top, \quad (5)$$

where n is the unit vector (ambiguous sign) perpendicular to $C(X)$. In this codimension one setting, the residual is the projection of Y into a single direction, and the operator (5) completely avoids matrix inversion. For higher codimension, the residual lies in a subspace of dimension $k = n - \rho$, and the factorization (4) makes this structure explicit.

The construction of normal directions motivates generalizations of the classical cross-product to $\mathbb{R}^n$. Belhaouari et al. formalize a determinant-based vector product that yields a vector orthogonal to $n - 1$ linearly independent vectors in $\mathbb{R}^n$ [3]. In regression problems with $\dim C(X)^\perp = 1$, a single normal vector uniquely defines the rank-one projector (5). For higher codimensions, these constructions naturally combine with null-space and QR-based methods to form an orthonormal residual basis N .

Multicollinearity is a persistent challenge in regression modeling, as highly dependent predictors inflate variances, destabilize coefficient estimates, and impede interpretation [4,5]. Classical diagnostics such as the variance inflation factor (VIF) quantify this effect but offer limited geometric insight. We therefore introduce the Geometric Multicollinearity Index (GMI), a scale-invariant, volume-based measure derived from the polar sine and the Gram determinant. For a full column-rank matrix $X = [x_1, \dots, x_p]$ with $p \leq n$, the polar sine normalizes $\det(X^\top X)$ by the product of the column norms, taking values in $[0, 1]$: it equals 1 for orthogonal predictors and approaches 0 as multicollinearity increases.

In rank-deficient or over-parameterized regimes, the Gram determinant vanishes and the GMI saturates at 1, indicating extreme geometric degeneracy. As a normalized volume measure of the parallelotope spanned by the columns of X , the GMI provides a compact, scale-invariant geometric diagnostic across dimensions. In this regime, $\text{GMI} = 1$ should be interpreted as a diagnostic flag for complete geometric collapse, rather than as a measure of severity beyond rank deficiency.

Although both regression and principal component analysis (PCA) rely on orthogonal projections, their objectives differ fundamentally [6]. PCA projects X onto directions of maximum variance, yielding residuals $(I - P_{\text{PCA}})X$ that remain in $C(X)$. By contrast, regression projects the response, decomposing Y into a fitted component PY and a residual $r = (I - P)Y \in C(X)^\perp$ [2]. This distinction between feature-space and response-space residuals is essential for interpreting linear-model diagnostics.

Scope of the present paper is the identities $I - P = NN^\top$ and $I - P = nn^\top$ are standard results from orthogonal projection theory. The contribution of this paper lies in a rank-aware geometric interpretation of these identities via multivector and cross-product constructions, together with the introduction of the volume-based Geometric Multicollinearity Index (GMI) as a complementary geometric diagnostic.

This work is theoretical in nature. We develop a rank-aware residual projector, characterize the associated residual subspaces, and formalize GMI as an intrinsic measure of multicollinearity. Algorithmic, large-scale, regularized, and nonlinear extensions are beyond the scope of this paper and are left for future work.

The main contributions of this theoretical paper are as follows:

1. Rank-aware geometric formulation is the first one where we develop a unified, rank-aware description of the residual projector based on the identity $I - P = NN^\top$, which reduces to the rank-one form $nn^\top$ when $\dim C(X)^\perp = 1$ and generalizes to arbitrary rank.
2. Multivector perspective on residual spaces such that we reformulate $C(X)^\perp$ and the associated projector using multivectors, wedge products, and cross-products, yielding a basis-independent geometric characterization valid in both codimension-one and higher-codimension settings.
3. In Geometric Multicollinearity Index (GMI) We introduce a scale-invariant, volume-based diagnostic derived from the Gram determinant and polar sine, which quantifies geometric degeneracy in both full-rank and rank-deficient regimes and complements classical VIF-type measures.
4. Also, We clarify the conceptual distinction between regression residuals and PCA reconstruction residuals, emphasizing the impact of projecting the response versus the predictors on residual geometry and interpretation.
5. Finally, we provide simple numerical examples illustrating rank-one and multivector residual projections and the behavior of GMI under controlled perturbations.

The organization of the paper is provided as follows Section 2 introduces the notation and basic geometric concepts. Section 3 develops the main theoretical results on residual projectors and the geometric structure of $C(X)^\perp$. Section 4 presents numerical illustrations, including the behavior of GMI under controlled multicollinearity. Sections 5 and 6 discuss implications and conclude.

2. Notation and Preliminaries

We briefly fix notation and recall the algebraic and geometric concepts that underpin the proposed framework. We follow standard linear algebra conventions as in Strang [7], Lay et al. [8], and Golub and Van Loan [2].

2.1. Basic Notation and Subspaces

The transpose of a matrix or vector is denoted by the superscript $\top$. Lowercase letters (e.g., x , n) denote column vectors, and uppercase letters (e.g., X , N) denote matrices. For a matrix $X \in \mathbb{R}^{n \times p}$, we write $C(X)$ for its column space and denote its rank by ρ . Throughout, r denotes the residual vector while ρ denotes the rank of X , to avoid notational ambiguity. All notation is defined once and used consistently throughout the manuscript; in particular, N denotes an orthonormal basis of the orthogonal complement $C(X)^\perp$.

We reserve r exclusively for the residual vector in (8).

The orthogonal complement of a subspace $V \subseteq \mathbb{R}^n$ is denoted by $V^\perp$, and the null space of a matrix A is

$$\mathcal{N}(A) = \{z \in \mathbb{R}^m : Az = 0\}.$$

A summary of all symbols and subspace conventions is provided in Abbreviations.

Given $n \in \mathbb{R}^n$, the matrix $nn^\top \in \mathbb{R}^{n \times n}$ denotes its outer product. The identity matrix in $\mathbb{R}^{n \times n}$ is denoted by I .

2.2. Projection and Residual Operators

Let $X \in \mathbb{R}^{n \times p}$ have rank $\rho \leq \min\{n, p\}$; when X has full column rank ($\rho = p$), the orthogonal projector admits the classical closed form $P = X(X^\top X)^{-1}X^\top$.

The orthogonal projector onto the column space $\mathcal{C}(X)$ is the unique symmetric idempotent matrix P that satisfies

$$\mathcal{C}(P) = \mathcal{C}(X), \quad P^\top = P, \quad P^2 = P.$$

When X has full column rank ρ , this projector admits the well-known closed form

$$P = X(X^\top X)^{-1}X^\top, \quad (6)$$

which is symmetric and idempotent [2]. In the general case, including rank-deficient designs, P can be written as

$$P = XX^+, \quad (7)$$

where X^+ denotes the Moore–Penrose pseudoinverse of X [2]. Throughout, X^+ denotes the Moore–Penrose pseudoinverse of X , computed via an SVD-based routine.

Both (6) and (7) implement the same geometric operation: orthogonal projection onto $\mathcal{C}(X)$.

When $X^\top X \succ 0$, the same orthogonal projector can equivalently be constructed by Cholesky factorization of $X^\top X$, producing identical residuals while requiring positive definiteness of the Gram matrix.

For any response vector $Y \in \mathbb{R}^n$, the fitted component is

$$\hat{Y} = PY,$$

and the residual is

$$r = (I - P)Y, \quad (8)$$

which obeys $X^\top r = 0$ and hence lies in the orthogonal complement $\mathcal{C}(X)^\perp$ [2].

Let $\rho = \text{rank}(X)$ and set $k = n - \rho = \dim \mathcal{C}(X)^\perp$. Choose any matrix $N \in \mathbb{R}^{n \times k}$ whose columns form an orthonormal basis of $\mathcal{C}(X)^\perp$:

$$N^\top N = I_k, \quad X^\top N = 0. \quad (9)$$

Then the residual projector admits the geometric factorization

$$I - P = NN^\top, \quad (10)$$

and the null space of $X^\top$ is

$$\mathcal{N}(X^\top) = \{z \in \mathbb{R}^n : X^\top z = 0\} = \mathcal{C}(X)^\perp. \quad (11)$$

This characterization is standard in numerical linear algebra and regression theory [2] and forms the algebraic backbone of our geometric residual formulation. Equation (10) makes explicit that the residual r lies in a k -dimensional orthogonal complement spanned by the columns of N .

2.3. Cross-Product and Cross-Wedge-QR Method

To construct explicit orthogonal directions, especially in the case of codimension-one where $\dim \mathcal{C}(X)^\perp = 1$, we employ the cross-product in $\mathbb{R}^n$ [3]. Given $n - 1$ linearly independent vectors

$$x_1, \dots, x_{n-1} \in \mathbb{R}^n,$$

the determinant-based construction of Belhaouari et al. produces

$$\text{cross}(x_1, \dots, x_{n-1}) \in \mathbb{R}^n,$$

which is orthogonal to each x_i :

$$x_i^\top \text{cross}(x_1, \dots, x_{n-1}) = 0, \quad i = 1, \dots, n-1.$$

This extends earlier n -dimensional vector products and related constructions in eigenanalysis and geometry [9,10].

In the regression setting, when $\rho = n - 1$ the predictor space $\mathcal{C}(X)$ has codimension one and $\mathcal{C}(X)^\perp$ is spanned by a single unit vector n . Choosing any basis of $\mathcal{C}(X)$ and applying the cross-product yields a nonzero vector orthogonal to $\mathcal{C}(X)$; normalizing it gives a unit normal n spanning $\mathcal{C}(X)^\perp$. In this case, the residual projector reduces to the rank-one form

$$I - P = nn^\top, \quad (12)$$

and the residual is simply $r = (n^\top Y) n$.

In this codimension-one setting, the rank-one projector $I - P = nn^\top$ coincides with the classical construction of the cross-product, and we refer to this case as the residual projector of the cross-product (rank one).

A key geometric property of the cross-product is that its norm equals the $(n - 1)$ -dimensional volume of the parallelotope spanned by its arguments [3]:

$$\|\text{cross}(v_1, v_2, \dots, v_{n-1})\| = \text{vol}_{n-1}(v_1, \dots, v_{n-1}). \quad (13)$$

Thus, the cross-product simultaneously provides an orthogonal direction and encodes a volume. When $\dim \mathcal{C}(X)^\perp > 1$, a single cross-product yields at most one null direction; in that case additional null directions must be obtained by other means (e.g., null-space methods and QR), as discussed later.

2.4. Wedge Products, Gram Matrices, and Polar Sine

To describe volumes and linear independence in a coordinate-free manner, we rely on standard constructions from linear algebra and multilinear geometry [2,7,8]. Given vectors $u_1, \dots, u_k \in \mathbb{R}^n$, their wedge product $u_1 \wedge \dots \wedge u_k$ encodes the oriented k -dimensional parallelotope spanned by $\{u_i\}$. Geometrically, $u_1 \wedge \dots \wedge u_k = 0$ if and only if the vectors are linearly dependent, and its norm $\|u_1 \wedge \dots \wedge u_k\|$ equals the volume k of the associated parallelotope. This interpretation underlies many geometric treatments of subspaces, volumes, and orthogonality in $\mathbb{R}^n$ [2,7].

For computational purposes, such volumes can be expressed using Gram matrices. Given vectors $v_1, \dots, v_m \in \mathbb{R}^n$, the associated Gram matrix is

$$G = \begin{bmatrix} v_1^\top v_1 & \cdots & v_1^\top v_m \\ \vdots & \ddots & \vdots \\ v_m^\top v_1 & \cdots & v_m^\top v_m \end{bmatrix} = [v_1, \dots, v_m]^\top [v_1, \dots, v_m]. \quad (14)$$

The m -dimensional volume of the parallelotope spanned by $\{v_i\}$ is given by

$$\text{vol}_m(v_1, \dots, v_m) = \sqrt{\det(G)}, \quad (15)$$

a classical identity that plays a central role in covariance analysis and principal component analysis [2,6].

We quantify the angular separation of p predictor vectors using the polar sine. Let $X = [x_1, \dots, x_p] \in \mathbb{R}^{n \times p}$ with Gram matrix $G = X^\top X$. Following geometric constructions based on determinants and cross-products [3,9,10], the polar sine is defined as

$$\text{psin}(x_1, \dots, x_p) = \frac{\sqrt{\det(X^\top X)}}{\|x_1\| \cdots \|x_p\|}. \quad (16)$$

When $p \leq n$ and the columns of X are linearly independent, $\det(X^\top X) > 0$ and $\text{psin}(x_1, \dots, x_p) \in (0, 1]$; it equals 1 if and only if the vectors are mutually orthogonal. If the columns of X are linearly dependent or $p > n$, then $X^\top X$ is singular and $\det(X^\top X) = 0$, and we adopt the convention $\text{psin}(x_1, \dots, x_p) = 0$. Thus, the polar sine takes values in $[0, 1]$, is scale-invariant, and decreases toward zero as the set of columns becomes geometrically degenerate.

Within the present framework, the polar sine provides a normalized geometric measure of predictor-space collapse under multicollinearity and forms the basis of the Geometric Multicollinearity Index (GMI) introduced in the following section.

2.5. Projection Strategies: QR and Cross-Product Formulations

Two projection strategies play a central role in the remainder of the paper.

First, QR decomposition is used when an entire orthonormal basis of $\mathcal{C}(X)$ or $\mathcal{C}(X)^\perp$ is required. If $X = QR$ is a (thin) QR factorization with orthonormal columns in Q , then

$$P = QQ^\top,$$

is the orthogonal projector onto $\mathcal{C}(X)$, and orthonormal complements can be obtained by extending Q to an orthogonal basis of $\mathbb{R}^n$ [2]. QR-based projectors are numerically stable and well suited for high-dimensional or nearly singular problems.

Second, when the residual space is one-dimensional (for example, when $\rho = n - 1$), the cross-product provides a compact alternative. It produces a normal vector n directly from which the residual projector is derived:

$$I - P = nn^\top.$$

This follows without forming or inverting $X^\top X$ [3]. In higher-codimension settings ($\dim \mathcal{C}(X)^\perp > 1$), cross-products can still be used to generate individual null directions, but a full residual basis is obtained more naturally from $\mathcal{N}(X^\top)$ and orthonormalised via QR.

When null directions are generated sequentially via cross-products and accumulated into a multivector basis that is subsequently orthonormalised using QR, we refer to the resulting construction as the Recursive Cross-Wedge-QR method.

These geometric primitives—orthonormal bases, cross-products, wedge products, Gram matrices, and polar sine—constitute the toolkit on which the residual projection framework developed in the following sections is built.

Here s denotes the sketch size (number of random probe vectors) used in the Gaussian range-finding procedure.

Sketch–QR Baseline (Reproducibility)

The Sketch–QR baseline follows a standard randomized range-finding procedure. Given $X \in \mathbb{R}^{n \times p}$ and a sketch size s , we draw a Gaussian sketching matrix $\Omega \in \mathbb{R}^{p \times s}$ with independent and identically distributed $\mathcal{N}(0, 1)$ entries and form $Y = X\Omega \in \mathbb{R}^{n \times s}$. We then compute a thin QR factorization $Y = \tilde{Q}R$, where $\tilde{Q} \in \mathbb{R}^{n \times s}$ has orthonormal columns. The approximate projector is $\tilde{P} = \tilde{Q}\tilde{Q}^\top$, and the approximate residual is $\tilde{r} = (I - \tilde{P})y$. Unless stated otherwise, we use a single sketch without power iterations or oversampling [11,12].

3. Core Theory

3.1. Rank-One Residual Projection

Assume that $X \in \mathbb{R}^{n \times p}$ has rank $n - 1$. Then $\dim \mathcal{C}(X) = n - 1$ and $\dim \mathcal{C}(X)^\perp = 1$. There exists a (unique up to sign) unit vector $n \in \mathbb{R}^n$ such that

$$X^\top n = 0, \quad \|n\| = 1.$$

Proposition 1 (Rank-One Residual Projection). *Let $X \in \mathbb{R}^{n \times p}$ have $\rho = n - 1$, and let P denote the orthogonal projector onto $\mathcal{C}(X)$ (for example, $P = XX^+$ in terms of the Moore–Penrose pseudoinverse X^+). Then the orthogonal residual projector is*

$$I - P = nn^\top.$$

Proof. Because $\mathcal{C}(X)^\perp$ is one-dimensional, any residual $r \in \mathcal{C}(X)^\perp$ can be written as $r = \alpha n$ for some scalar α . Since $n \in \mathcal{C}(X)^\perp$ and P is the orthogonal projector onto $\mathcal{C}(X)$, we have $Pn = 0$ and therefore $(I - P)n = n$. For any $Y \in \mathbb{R}^n$, we may write $Y = u + v$ with $u \in \mathcal{C}(X)$ and $v \in \mathcal{C}(X)^\perp = \text{span}\{n\}$. Then

$$(I - P)Y = (I - P)(u + v) = v,$$

while

$$nn^\top Y = nn^\top(u + v) = nn^\top u + nn^\top v = 0 + v = v,$$

because $n^\top u = 0$ and v is a scalar multiple of n . Thus, $(I - P)Y = nn^\top Y$ for all Y , and therefore $I - P = nn^\top$, which is symmetric and idempotent. $\square$

3.2. Generalization to n Dimensions

The rank-one projector $I - P = nn^\top$ extends naturally to the general setting in which the predictor matrix $X \in \mathbb{R}^{n \times p}$ has $\rho < n$. In this case, the orthogonal complement $\mathcal{C}(X)^\perp$ has dimension $k = n - \rho$, and the residual is no longer confined to a single direction but lies in a k -dimensional subspace.

Theorem 1 (Multivector Residual Projection). *Let X have rank $\rho < n$ and set $k := n - \rho$. Let $N \in \mathbb{R}^{n \times k}$ have orthonormal columns spanning $\mathcal{C}(X)^\perp$ (so $N^\top N = I_k$). Then, the residual projector is $I - P = NN^\top$ and $r = (I - P)y = NN^\top y$.*

Proof. Since $N^\top N = I_{n-\rho}$, the matrix $NN^\top$ is symmetric and satisfies

$$(NN^\top)^2 = N(N^\top N)N^\top = NI_{n-\rho}N^\top = NN^\top,$$

so $NN^\top$ is an orthogonal projector. Its range equals $\mathcal{C}(N) = \mathcal{C}(X)^\perp$.

On the other hand, P is the orthogonal projector onto $\mathcal{C}(X)$, so $I - P$ is the orthogonal projector onto $\mathcal{C}(X)^\perp$. For any $Y \in \mathbb{R}^n$, write $Y = u + v$ with $u \in \mathcal{C}(X)$ and $v \in \mathcal{C}(X)^\perp$. Then

$$(I - P)Y = v.$$

Since $v \in \mathcal{C}(N)$ and the columns of N are orthonormal, we also have

$$NN^\top Y = NN^\top(u + v) = NN^\top u + NN^\top v = 0 + v = v,$$

because $N^\top u = 0$ and $NN^\top$ acts as the identity on $\mathcal{C}(N)$. Thus, $(I - P)Y = NN^\top Y$ for all Y , and hence $I - P = NN^\top$. $\square$

Figure 1 illustrates the situation in a simple three-dimensional example. The predictor space occupies a lower-dimensional subspace (blue), while the residual lies in its orthogonal complement (orange), spanned by multiple orthonormal directions when $\dim \mathcal{C}(X)^\perp > 1$.

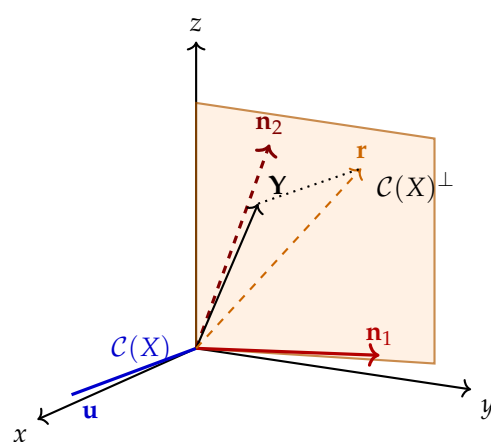


Figure 1. Geometric interpretation of the residual projection in a three-dimensional example with $\dim \mathcal{C}(X) = 1$ and $\dim \mathcal{C}(X)^\perp = 2$. The design matrix X has a one-dimensional column space $\mathcal{C}(X)$ (blue line), while the residual space $\mathcal{C}(X)^\perp$ is a two-dimensional plane (orange), spanned by an orthonormal basis $\mathbf{n}_1, \mathbf{n}_2$. For a response vector $\mathbf{Y}$, the residual $\mathbf{r} = (I - P)\mathbf{Y} = NN^\top \mathbf{Y}$ is its orthogonal projection onto $\mathcal{C}(X)^\perp$.

3.2.1. Geometric Multicollinearity Index (GMI)

The multivector formulation of the residual projector reflects the angular configuration of the columns of X . When the predictors are nearly linearly dependent, the volume of the parallelotope spanned by these columns collapses, signaling instability. To capture this behavior geometrically, we introduce the Geometric Multicollinearity Index (GMI), a scale-invariant measure based on the polar sine.

Definition 1 (Polar Sine and GMI). Let $X = [x_1, \dots, x_p] \in \mathbb{R}^{n \times p}$ have nonzero columns $x_j \neq 0$, and let $\text{psin}(x_1, \dots, x_p)$ be the polar sine defined in Equation (16).

We adopt the shorthand

$$\text{psin}(X) := \text{psin}(x_1, \dots, x_p), \quad (17)$$

and define the Geometric Multicollinearity Index by

$$\text{GMI}(X) = 1 - \text{psin}(X). \quad (18)$$

Lemma 1 (Basic properties). Let $X = [x_1, \dots, x_p] \in \mathbb{R}^{n \times p}$ have nonzero columns, and let $G = X^\top X$ denote its Gram matrix. Then:

- (i) $\text{psin}(X) \in [0, 1]$ and hence $\text{GMI}(X) \in [0, 1]$;
- (ii) $\text{psin}(X) = 1$ (and $\text{GMI}(X) = 0$) if and only if the columns of X are mutually orthogonal;
- (iii) $\text{psin}(X) = 0$ (and $\text{GMI}(X) = 1$) whenever the columns are linearly dependent or $p > n$.

Proof. By Hadamard's inequality,

$$\det(G) \leq \prod_{j=1}^p \|x_j\|^2,$$

with equality if and only if the columns of X are mutually orthogonal. Substituting this bound into the polar-sine definition Equation (16) yields

$$0 \leq \text{psin}(X) = \frac{\sqrt{\det(G)}}{\|x_1\| \cdots \|x_p\|} \leq 1,$$

which proves (i). Moreover, $\text{psin}(X) = 1$ holds precisely when equality in Hadamard's inequality holds, i.e., when the columns are pairwise orthogonal, giving (ii).

If the columns of X are linearly dependent or $p > n$, then G is singular and $\det(G) = 0$, so by definition $\text{psin}(X) = 0$ in that case, and (iii) follows immediately. The corresponding statements for $\text{GMI}(X)$ are then direct consequences of the definition $\text{GMI}(X) = 1 - \text{psin}(X)$. $\square$

Lemma 2 (Scale invariance). Let $A = \text{diag}(a_1, \dots, a_p)$ with each $a_j \neq 0$. Then $\text{psin}(XA) = \text{psin}(X)$ and hence $\text{GMI}(XA) = \text{GMI}(X)$.

Proof. We have $(XA)^\top(XA) = A^\top GA$ and $\det(A^\top GA) = \det(A)^2 \det(G)$, while $\|a_j x_j\| = |a_j| \|x_j\|$ for each j . The scale factors cancel in the quotient (17), so $\text{psin}(XA) = \text{psin}(X)$, and (18) yields the claim for GMI . $\square$

Remark 1. For orthonormal predictors, $\text{psin}(X) = 1$ and $\text{GMI}(X) = 0$. For nearly collinear predictors, $\det(G) \approx 0$, so $\text{psin}(X) \approx 0$ and $\text{GMI}(X) \approx 1$. Intermediate angular configurations (e.g., vectors separated by $\approx 45^\circ$) yield values such as $\text{psin} \approx 0.71$ and $\text{GMI} \approx 0.29$, as illustrated in Figure 2. In this sense, GMI is a normalized version of determinant-based multicollinearity measures (such as the generalized variance $\det(G)$ or $\det$ of the correlation matrix), rescaled to lie in $[0, 1]$ and made explicitly scale-invariant via the polar-sine normalization.

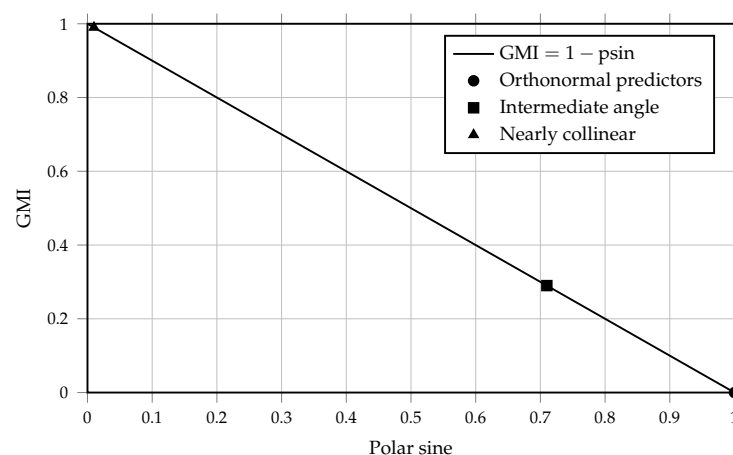


Figure 2. Geometric Multicollinearity Index (GMI) as a function of the polar sine. The index is zero for orthonormal predictors, increases steadily as angular separation decreases, and approaches one under near-collinearity.

GMI provides a direct geometric diagnostic that complements classical VIF-type measures. Whereas VIF assesses how much the variance of each individual regression coefficient is inflated by linear dependence among the predictors, GMI treats the entire predictor set as a single geometric object and summarizes its angular degeneracy in a single number in $[0, 1]$. Because it depends only on the angular configuration of the predictors and is invariant under independent rescaling of the columns, it can be compared meaningfully across differently scaled datasets.

3.2.2. Uniqueness of the Residual Projection

The multivector formulation $I - P = NN^\top$ expresses the residual projector in terms of any orthonormal basis of the orthogonal complement $\mathcal{C}(X)^\perp$. The next result confirms that this projector is independent of the specific basis chosen.

Remark 2 (Independence of the residual projector from the basis). *Let $X \in \mathbb{R}^{n \times p}$ have rank $r < n$, and let $N, \tilde{N} \in \mathbb{R}^{n \times (n-r)}$ have orthonormal columns spanning $\mathcal{C}(X)^\perp$. Then there exists an orthogonal $Q \in \mathbb{R}^{(n-r) \times (n-r)}$ such that $\tilde{N} = NQ$, and hence*

$$\tilde{N}\tilde{N}^\top = NQQ^\top N^\top = NN^\top.$$

In particular, the representation $I - P = NN^\top$ from Theorem 1 is independent of the orthonormal basis chosen of $\mathcal{C}(X)^\perp$: it is simply the (unique) orthogonal projector onto $\mathcal{C}(X)^\perp$.

3.3. Generalized Projection in Rank-Deficient Spaces via Null-Spaces and Wedge Products

When the design matrix X is rank-deficient or when $n - \rho > 1$, the orthogonal complement $\mathcal{C}(X)^\perp$ has dimension greater than one, and a single normal vector—as in the rank-one case—is no longer sufficient to describe the residual direction. In such settings, a full orthonormal basis of $\mathcal{C}(X)^\perp$ is required. This subsection explains how standard null-space computations, combined with QR orthonormalization, yield such a basis, and how wedge products provide a coordinate-free way to reason about independence and volume.

Definition 2 (Residual Projection in the Rank-Deficient Case). *Let $X \in \mathbb{R}^{n \times p}$ have rank $r < n$, and set $k = n - \rho = \dim \mathcal{C}(X)^\perp$. If $N \in \mathbb{R}^{n \times k}$ has orthonormal columns spanning $\mathcal{C}(X)^\perp$, then the residual projector is*

$$I - P = NN^\top.$$

Proposition 2 (Limitation of the cross-product). *The cross-product of Belhaouari [3] produces a single vector orthogonal to $n - 1$ independent vectors in $\mathbb{R}^n$. This suffices when $\dim \mathcal{C}(X)^\perp = 1$, but for $n - r > 1$ it yields only one null direction. Additional directions are needed to span the entire residual space.*

Proof. If $n - \rho = 1$, then $\mathcal{C}(X)^\perp$ is one-dimensional, and a single cross-product applied to a basis of $\mathcal{C}(X)$ recovers the unique (up to sign) normal vector n , giving $I - P = nn^\top$ as in the rank-one case. When $n - \rho > 1$, the orthogonal complement has dimension $k > 1$, so multiple linearly independent null directions are required. A single cross-product produces at most one such direction and therefore cannot span a k -dimensional space. Additional vectors must be constructed, for instance, via null-space computations and subsequent orthonormalization. $\square$

Theorem 2 (Construction of a Residual Basis via null-space and QR). *Let $X \in \mathbb{R}^{n \times p}$ have rank r and set $k = n - \rho$, so that $\mathcal{C}(X)^\perp = \mathcal{N}(X^\top)$ has dimension k . Then there exists an orthonormal basis $N \in \mathbb{R}^{n \times k}$ of $\mathcal{C}(X)^\perp$ such that*

$$I - P = NN^\top,$$

where P is the orthogonal projector onto $\mathcal{C}(X)$. Moreover, such an N can be obtained by the null-space + QR procedure in Algorithm 1.

Algorithm 1 Residual Projector via Null Space and QR (Explicit SVD/RRQR)

Require: Design matrix $X \in \mathbb{R}^{n \times p}$; numerical tolerance $\tau > 0$.

Ensure: Orthonormal basis N of $\mathcal{N}(X^\top)$ and residual projector $(I - P) = NN^\top$.

- 1: **Estimate numerical rank:** compute ρ using either SVD or RRQR (below); set $k \leftarrow n - \rho$.
- 2: **Construct null-space basis:** compute $B \in \mathbb{R}^{n \times k}$ spanning $\mathcal{N}(X^\top)$ using exactly one of the following:

(SVD option: robust)

Compute a full SVD $X = U\Sigma V^\top$ with singular values $\sigma_1 \geq \sigma_2 \geq \dots$.

Set $\rho := \#\{i : \sigma_i \geq \tau \sigma_1\}$ and $k := n - \rho$.

Set $B \leftarrow U(:, \rho + 1 : n) \triangleright \mathcal{N}(X^\top)$ spanned by left singular vectors with near-zero σ_i

(RRQR option: efficient)

Compute a column-pivoted QR factorization $X^\top \Pi = QR$.

Set $\rho := \max\{i : |R_{ii}| \geq \tau |R_{11}|\}$ and $k := n - \rho$.

Partition $R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$.

Solve $R_{11}Z = R_{12}$ for Z (do not form R_{11}^{-1}).

Form $B_0 = \begin{bmatrix} -Z \\ I_k \end{bmatrix}$ and set $B \leftarrow \Pi B_0$.

Implementation note: Row reduction is sufficient for symbolic or toy examples; for floating-point computations, use SVD (robust) or RRQR (fast).

- 3: (Optional) Verify $\text{rank}(B) = k$.
 - 4: Compute a thin QR factorization $B = Q_B R_B$ with orthonormal columns in Q_B .
 - 5: Set $N \leftarrow Q_B$.
 - 6: **return** N and the residual projector $(I - P) = NN^\top$.
-

Proof. By construction, the columns of B span $\mathcal{N}(X^\top)$, which is equal to $\mathcal{C}(X)^\perp$. If $b_1 \wedge \dots \wedge b_k \neq 0$ then the b_i 's are linearly independent and form a basis of this k -dimensional subspace. The thin QR factorization $B = QR$ with orthonormal columns in Q then produces $N := Q$ with $N^\top N = I_k$ and $\mathcal{C}(N) = \mathcal{C}(B) = \mathcal{C}(X)^\perp$. The matrix $NN^\top$ is therefore the orthogonal projector onto $\mathcal{C}(X)^\perp$, so $I - P = NN^\top$ by the uniqueness of the orthogonal projector onto a fixed subspace. $\square$

Remark 3. In practice, the null-space basis in Step 1 is typically obtained from a numerically stable factorization such as SVD or a rank-revealing QR. The wedge product need not be computed explicitly; it serves primarily as a geometric way to interpret the independence check. When $\dim \mathcal{C}(X)^\perp = 1$, the construction reduces to the case of rank one, where a normal vector can be obtained either from the null space of $X^\top$ or by applying a cross-product to a basis of $\mathcal{C}(X)$. The numerical tolerance used for the determination of rank (e.g., $\rho = \tau = 10^{-12}$ relative to the largest singular value) follows standard practice and does not affect the conceptual geometric results, which hold exactly at the algebraic level.

Remark 4 (Implementation note: computing $\mathcal{N}(X^\top)$ in floating-point arithmetic). The instruction “compute a basis of $\mathcal{N}(X^\top)$ ” requires (i) a numerical rank decision and (ii) an explicit extraction rule. In this work we use either:

(a) SVD. Compute $X^\top = U\Sigma V^\top$, choose the numerical rank $\rho = \#\{i : \sigma_i \geq \tau\sigma_1\}$, and set $B = V(:, r+1 : n)$.

(b) RRQR. Compute a column-pivoted QR factorization $X^\top \Pi = QR$, choose rank ρ from the diagonal of R (e.g., $|R_{ii}| \geq \tau|R_{11}|$), partition $R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$, and set $B = \Pi \begin{bmatrix} -R_{11}^{-1}R_{12} \\ I_{n-\rho} \end{bmatrix}$.

SVD is typically more robust under severe ill conditioning, while RRQR is often faster when pivoted QR is sufficient. Row reduction is used only for symbolic/toy demonstrations.

Figure 3 illustrates the residual subspace and two independent null directions before orthonormalization.

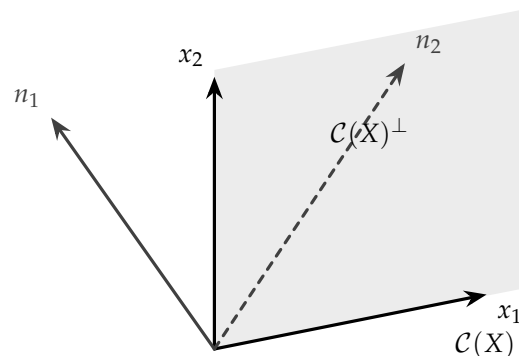


Figure 3. Two-dimensional schematic of the residual subspace generated by the cross-wedge-QR construction. The vectors x_1 and x_2 span the predictor space $\mathcal{C}(X)$, while n_1 and n_2 indicate two independent directions in the orthogonal complement $\mathcal{C}(X)^\perp$ prior to QR orthonormalization.

3.4. Projection Method Variants and Complexity

Definition 3 (Variants of Residual Projection). Let $X \in \mathbb{R}^{n \times p}$ be a real-valued design matrix of rank $\rho < n$. The residual projection operator $I - P$ onto the orthogonal complement $\mathcal{C}(X)^\perp$ can be constructed by several exact procedures, each suited to different regimes of dimension, conditioning, and interpretability.

The following variants highlight that null-space, wedge-based, and determinant-based approaches provide geometrically transparent alternatives to inversion-based formulas in linear regression.

Proposition 3 (Wedge-QR Null-Space Variant). Let $X \in \mathbb{R}^{n \times p}$ have rank ρ , and let $k = n - \rho$. Using the null-space + QR construction of Theorem 2, we obtain an orthonormal residual basis $N \in \mathbb{R}^{n \times k}$ spanning $\mathcal{C}(X)^\perp$ and hence the projector $I - P = NN^\top$. If, in addition, we compute a thin QR factorization $X = QR$, then the Gram matrix satisfies

$$G = X^\top X = R^\top R, \quad \det(G) = \det(R)^2,$$

so the squared volume of the parallelotope spanned by the columns of X is given directly by $\det(R)^2$. The general complexity of forming N and R is $\mathcal{O}(np^2 + nk)$.

Proof. The existence of an orthonormal basis N of $\mathcal{C}(X)^\perp$ and the identity $I - P = NN^\top$ follow from Theorem 2. The factorization $X = QR$ implies $G = X^\top X = R^\top R$, so $\det(G) = \det(R)^2$, which yields the stated volume formula. The complexity estimate is inherited from the null space + QR construction. $\square$

When geometric interpretations involving volume are necessary—such as evaluating multicollinearity through the Geometric Multicollinearity Index (GMI)—factorizations of the Gram matrix are particularly convenient.

Proposition 4 (Cholesky-Based Volume and Projection). *Let $G = X^\top X$ and assume that X has full-column-rank so that $G \succ 0$. Compute the Cholesky factorization $G = LL^\top$, which yields*

$$\det(G) = \det(L)^2 = \prod_i L_{ii}^2.$$

Let N be any orthonormal basis of $\mathcal{C}(X)^\perp$ constructed as in Theorem 2. Then $I - P = NN^\top$ is the orthogonal residual projector. The dominant cost is $\mathcal{O}(p^3 + nk)$: $\mathcal{O}(p^3)$ for the Cholesky factorization and $\mathcal{O}(nk)$ for forming and applying N .

Proof. The positive definiteness of G ensures the existence of a unique Cholesky factor L with positive diagonal entries, and the determinant formula follows from $\det(G) = \det(L)^2 = \prod_i L_{ii}^2$. The construction of N and the identity $I - P = NN^\top$ is provided by Theorem 2; any such N produces an orthogonal projector onto $\mathcal{C}(X)^\perp$. $\square$

In the special case where the residual subspace is one-dimensional, the projector admits an especially simple representation.

Proposition 5 (Cross-product (Rank-One Case)). *If $\dim(\mathcal{C}(X)^\perp) = 1$, then*

$$I - P = nn^\top, \quad (19)$$

where $n \in \mathbb{R}^n$ satisfies $X^\top n = 0$ and $\|n\| = 1$. The one-time cost of building n is $\mathcal{O}(n^2)$, and each subsequent projection $(I - P)Y$ costs $\mathcal{O}(n)$.

Proof. When $\mathcal{C}(X)^\perp$ is one-dimensional, any nonzero vector in this space is a scalar multiple of a unit normal n . The orthogonal projector onto the line spanned by n is precisely $nn^\top$. Symmetry and idempotence follow from $\|n\| = 1$, and orthogonality is ensured by $X^\top n = 0$. $\square$

Example 1. Consider $n = 4$, $p = 3$ and the residual codimension $k = n - p = 1$. In this codimension-one setting, the classical OLS residual projector based on $(X^\top X)^{-1}$ (assuming that X has full column rank) involves a matrix inversion together with several matrix–vector multiplications, which require on the order of a few hundred floating-point operations in this small example. By contrast, computing a unit normal vector n via the geometric cross-product costs $\mathcal{O}(n^2)$ operations, and each application of the rank-one projector $nn^\top$ to a new response Y costs only $\mathcal{O}(n)$ operations. Once n is constructed, the projection of additional responses is therefore substantially cheaper than recomputing residuals through $(X^\top X)^{-1}$, illustrating the efficiency of geometric formulation in symbolic or resource-limited regimes.

Unless stated otherwise, we use a single Gaussian sketch without power iterations or oversampling; orthonormalization is performed via a standard thin QR factorization, see Algorithm 2.

Algorithm 2 Sketch–QR Baseline (Gaussian Range Finder)

Require: Design matrix $X \in \mathbb{R}^{n \times p}$; sketch size $s \ll n$.

Ensure: Approximate orthonormal basis $\tilde{Q}$ and projector $\tilde{P}$.

- 1: Draw a Gaussian sketching matrix $\Omega \in \mathbb{R}^{p \times s}$ with i.i.d. $\mathcal{N}(0, 1)$ entries.
 - 2: Form the sketched matrix $Y = X\Omega \in \mathbb{R}^{n \times s}$.
 - 3: Compute a thin QR factorization $Y = \tilde{Q}R$.
 - 4: Define the approximate projector $\tilde{P} = \tilde{Q}\tilde{Q}^\top$.
 - 5: Compute the approximate residual $\tilde{r} = (I - \tilde{P})y$.
-

Table 1 compares four exact ways to build the residual projector $(I - P)$ for a given design matrix X . Classical OLS and the Cholesky-based variant both work through $(X^\top X)^{-1}$ and treat the residual subspace only implicitly, with a cubic dependence on the number of predictors p and the restriction that $X^\top X$ be invertible. In contrast, the Wedge-QR and rank-one cross-product methods construct $(I - P)$ from an explicit basis of $\mathcal{C}(X)^\perp$; the former handles general rank-deficient cases, while the latter is tailored to the codimension-one setting and achieves the lowest per-response cost when many responses must be projected.

Table 1. Exact residual projection variants for $X \in \mathbb{R}^{n \times p}$ with rank ρ and residual dimension $k = n - \rho$. The “one-off” cost refers to constructing the projector, while the “per-response” cost refers to applying the residual mapping $(I - P)Y$ to a new response vector Y . Methods involving $(X^\top X)^{-1}$ assume that X has full column rank so that $X^\top X$ is invertible.

Method	Residual $(I - P)Y$	One-Off Cost	Per-Response Cost
Classical OLS (via $(X^\top X)^{-1}$)	$(I - X(X^\top X)^{-1}X^\top)Y$	$\mathcal{O}(np^2 + p^3)$	$\mathcal{O}(np)$
QR / null-space basis	$NN^\top Y$	$\mathcal{O}(np^2 + nk)$	$\mathcal{O}(nk)$
Cholesky-based (when $X^\top X \succ 0$)	$(I - X(X^\top X)^{-1}X^\top)Y$	$\mathcal{O}(p^3)$ (form $X^\top X$ and $LL^\top$)	$\mathcal{O}(np)$
Rank-one cross-product ($\dim \mathcal{C}(X)^\perp = 1$)	$nn^\top Y$	$\mathcal{O}(n^2)$ (construct n)	$\mathcal{O}(n)$

Notes: Classical OLS and the Cholesky-based variant treat the residual subspace implicitly. The QR/null space basis method also applies to rank-deficient X and explicitly constructs a basis N of $\mathcal{C}(X)^\perp$. The rank-one cross-product method applies when $\dim \mathcal{C}(X)^\perp = 1$ and is particularly efficient once the normal n is known.

In summary, Wedge-QR and Cholesky-based constructions are well suited for when one needs both an exact residual projector and access to volume-based quantities such as $\det(X^\top X)$ or the GMI. The formulation of a rank-one cross-product is preferable whenever $\dim \mathcal{C}(X)^\perp = 1$, as it produces a simple geometric projector $nn^\top$ with linear complexity per response and, when $p \ll n$, is usually the cheapest option to repeatedly project many responses. For general rank-deficient designs, a QR/null-space construction offers a numerically safer alternative to explicitly forming and inverting $X^\top X$. Together, the four variants in Table 1 provide a practical menu of exact geometric projectors that avoid unnecessary matrix inversion and can be chosen according to the rank of X , the conditioning of $X^\top X$ and the available computational budget.

We emphasize that the exact residual projector $I - P$ can be constructed via several algebraically equivalent routes (normal-equations OLS when applicable, rank-one cross products, null-space/QR, wedge-based constructions and Cholesky-based variants), which differ only in numerical assumptions and computational cost; see Table 1.

4. Illustrative Numerical Examples

When the design matrix X is rank deficient, the normal equations $X^\top X\beta = X^\top Y$ do not admit a unique solution. In all rank-deficient experiments, we therefore compute the minimum-norm least-squares solution using the Moore–Penrose pseudoinverse, i.e., $\hat{\beta} = X^+Y$, and refer to this baseline as OLS (Moore–Penrose pseudoinverse). The full implementation and experimental scripts are publicly available [13].

Predictive performance is assessed using K -fold cross-validation with $K = 5$, reporting CV MSE and CV R^2 . Unless stated otherwise, all randomized procedures use a fixed random seed (seed = 42) to ensure reproducibility.

Coefficient stability is evaluated via $B = 100$ perturbation trials with the design matrix X held fixed. In each trial, the response is perturbed as

$$\tilde{y}^{(b)} = y + \eta^{(b)}, \quad \eta^{(b)} \sim \mathcal{N}(0, \sigma_y^2 I), \quad \sigma_y = 0.01 \text{ std}(y), \quad b = 1, \dots, B.$$

The model is refit for each trial to obtain coefficients $\hat{\beta}^{(b)}$. We let SD_j denote the sample standard deviation of the j th coefficient $\{\hat{\beta}_j^{(b)}\}_{b=1}^B$ across trials. We summarize coefficient stability by

$$\text{medianSD}(\hat{\beta}) := \text{median}_{j=1,\dots,p}(SD_j),$$

i.e., the median across coefficients of their across-trial standard deviations.

To validate the geometric projector, we verify agreement between the classical OLS residual $r_{\text{OLS}} = (I - P)y$ and the geometric residual $r_{\text{geo}} = NN^T y$ by monitoring $\|r_{\text{OLS}} - r_{\text{geo}}\|_2$ up to numerical tolerance.

Sketch-QR [11] yields an approximate basis $\tilde{Q}$ and projector $\tilde{P} = \tilde{Q}\tilde{Q}^T$, producing the approximate residual $\tilde{r} = (I - \tilde{P})y$. Agreement with the exact residual is summarized using cosine similarity $\cos(r_{\text{OLS}}, \tilde{r})$ and the residual discrepancy $\|\tilde{r} - r_{\text{OLS}}\|_2$ (and, when reported, a principal-angle distortion between $\mathcal{C}(X)$ and $\mathcal{C}(\tilde{Q})$).

4.1. Rank-One Residual Projection in $\mathbb{R}^3$

This example explicitly shows how the normality of the cross-product reproduces the ordinary least-squares (OLS) residual in the rank-one case, in line with Proposition 1.

We consider a rank-one residual setting where the orthogonal complement of the column space is one-dimensional. Let us

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 \\ 2 \\ \frac{11}{2} \end{bmatrix},$$

with $X \in \mathbb{R}^{3 \times 2}$ of full-column-rank (so $X^T X$ is invertible). The ordinary least-squares projection matrix is

$$P = X(X^T X)^{-1}X^T, \quad r_{\text{OLS}} = (I - P)Y.$$

A direct calculation yields

$$r_{\text{OLS}} = \begin{bmatrix} \frac{3}{14} \\ -\frac{9}{28} \\ \frac{3}{28} \end{bmatrix}, \quad r_{\text{OLS}} \in \mathcal{C}(X)^\perp.$$

To recover the same residual in purely geometric form, we construct a unit normal vector to the column space of X . Let us

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}, \quad n_0 = x_1 \times x_2 = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}, \quad n = \frac{n_0}{\|n_0\|} = \frac{1}{\sqrt{14}} \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}.$$

Then $X^T n = 0$ and $\|n\| = 1$, so Proposition 1 applies, and

$$I - P = nn^T.$$

The residual can therefore be written as the rank-one projection

$$r = (I - P)Y = nn^T Y = (n^T Y)n = \begin{bmatrix} \frac{3}{14} \\ -\frac{9}{28} \\ \frac{3}{28} \end{bmatrix},$$

which coincides with r_{OLS} and lies on the line spanned by n .

From a computational point of view, this example highlights the attraction of the geometric formulation in low dimension: the cross-product and the multiplication of a scalar vector are sufficient to obtain the residual, avoiding explicit matrix inversion. Once $\mathbf{n}$ is known, applying the projector $(\mathbf{I} - \mathbf{P}) = \mathbf{n}\mathbf{n}^\top$ to a new response vector reduces to one dot product and one scaled copy of $\mathbf{n}$, i.e., $\mathcal{O}(n)$ work per residual.

4.2. Multivector Residual Projection in $\mathbb{R}^4$

This example illustrates the multivector residual projection via an orthonormal basis of $\mathcal{C}(\mathbf{X})^\perp$ as in Theorem 2 and links it to the Geometric Multicollinearity Index.

We now consider a multivector setting in which the residual subspace has a dimension greater than one. Let us

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 5 \end{bmatrix}, \quad \mathbf{X} \in \mathbb{R}^{4 \times 2}, \quad \rho = 2.$$

The ordinary least-squares projector is again

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top, \quad \mathbf{r}_{\text{OLS}} = (\mathbf{I} - \mathbf{P})\mathbf{Y},$$

and a straightforward computation gives

$$\mathbf{r}_{\text{OLS}} = \begin{bmatrix} \frac{1}{5} \\ -\frac{1}{10} \\ -\frac{2}{5} \\ \frac{3}{10} \end{bmatrix}.$$

Since $\rho = 2$, its orthogonal complement has dimension $\dim \mathcal{C}(\mathbf{X})^\perp = 2$. A basis for the null space of $\mathbf{X}^\top$ is

$$\mathbf{n}_1 = \begin{bmatrix} 1 \\ -2 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{n}_2 = \begin{bmatrix} 2 \\ -3 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{X}^\top \mathbf{n}_i = 0, \quad i = 1, 2.$$

Applying Gram–Schmidt yields an orthonormal basis

$$\mathbf{u}_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -2 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{30}} \begin{bmatrix} 2 \\ -1 \\ -4 \\ 3 \end{bmatrix}, \quad \mathbf{N} = [\mathbf{u}_1 \ \mathbf{u}_2] \in \mathbb{R}^{4 \times 2},$$

with $\mathbf{N}^\top \mathbf{N} = \mathbf{I}_2$ and $\mathbf{X}^\top \mathbf{N} = 0$. The general residual projection theorem then gives

$$\mathbf{I} - \mathbf{P} = \mathbf{N}\mathbf{N}^\top, \quad \mathbf{r}_{\text{geom}} = \mathbf{N}\mathbf{N}^\top \mathbf{Y}.$$

Evaluating $\mathbf{r}_{\text{geom}}$ for this $\mathbf{X}$ and $\mathbf{Y}$ reproduces the residual OLS:

$$\mathbf{r}_{\text{geom}} = \mathbf{r}_{\text{OLS}} = \begin{bmatrix} \frac{1}{5} \\ -\frac{1}{10} \\ -\frac{2}{5} \\ \frac{3}{10} \end{bmatrix},$$

making explicit that the residual lives in the two-dimensional subspace spanned by $\mathbf{u}_1$ and $\mathbf{u}_2$ and realizing the null space + QR construction in Theorem 2, in a concrete low-dimensional setting.

From a complexity point of view, building $\mathbf{N}$ on a null-space basis followed by QR (or Gram–Schmidt) is a one-off cost of order $\mathcal{O}(np^2 + nk)$ for $\mathbf{X} \in \mathbb{R}^{n \times p}$ with residual dimension $k = n - \rho$, while applying the projector $(\mathbf{I} - \mathbf{P}) = \mathbf{N}\mathbf{N}^\top$ to a new response vector requires only $\mathcal{O}(nk)$ operations.

4.2.1. GMI for the $\mathbb{R}^4$ Example

For this design matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2]$, the Gram matrix is

$$\mathbf{G} = \mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}, \quad \det(\mathbf{G}) = 20.$$

The column norms are $\|\mathbf{x}_1\| = 2$ and $\|\mathbf{x}_2\| = \sqrt{30}$, so the polar sine and the Geometric Multicollinearity Index are

$$\text{psin}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sqrt{\det(\mathbf{X}^\top \mathbf{X})}}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|} = \frac{\sqrt{20}}{2\sqrt{30}} = \frac{\sqrt{6}}{6}, \quad \text{GMI} = 1 - \frac{\sqrt{6}}{6} \approx 0.59.$$

In this small example, the GMI value $\text{GMI} \approx 0.59$ indicates a moderate degree of collinearity between the two regressors; the residual subspace $\mathcal{C}(\mathbf{X})^\perp$ has dimension two (as $n - r = 2$), and, compared to the orthonormal case ($\text{GMI} = 0$), the fitted coefficients are more sensitive to perturbations of $\mathbf{Y}$.

This example isolates the effect of controlled collinearity on GMI in a minimal symbolic setting.

Consider the family of matrices

$$\mathbf{X}_\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \varepsilon & \varepsilon \end{bmatrix}, \quad \varepsilon \in [0, 1]. \quad (20)$$

For every $\varepsilon \in [0, 1]$ the two columns are linearly independent, so $\mathbf{X}_\varepsilon^\top \mathbf{X}_\varepsilon$ is positive definite, and the polar-sine is well defined. A direct computation gives

$$\text{psin}(\mathbf{X}_\varepsilon) = \frac{\sqrt{1 + 2\varepsilon^2}}{1 + \varepsilon^2}, \quad (21)$$

and hence

$$\text{GMI}(\varepsilon) = 1 - \frac{\sqrt{1 + 2\varepsilon^2}}{1 + \varepsilon^2}. \quad (22)$$

As ε increases, the columns become more collinear, psin decreases, and GMI increases, reflecting the growing multicollinearity in a purely geometric way. In families where columns eventually become exactly collinear, the Gram determinant vanishes, the polar

sine drops to zero, and GMI reaches 1, in agreement with the rank-deficient discussion in Section 3.

4.2.2. Large-Scale Multivector Residual Projection

To complement the low-dimensional symbolic example above, we consider a larger synthetic design in which the residual subspace has a dimension greater than one and the multivector formulation $I - P = NN^T$ is essential.

We generate a rank-deficient design matrix $X \in \mathbb{R}^{n \times p}$ with $n = 1000$, $p = 50$, and $\rho = 45$, so that $\dim \mathcal{C}(X)^\perp = n - \rho = 955$. All predictors are standardized, and the response is constructed so that the exact residual r_{exact} has a prescribed norm, ensuring a controlled comparison between methods.

Table 2 reports residual norms, cosine similarity to the exact residual, and run times for the variants of the Article 1 projector only. The exact constructions (null space/QR, rank-one cross-product when applicable, and the Moore–Penrose pseudoinverse baseline for OLS) reproduce the same residual up to numerical precision, confirming the multivector identity $I - P = NN^T$ in a high-dimensional setting.

Table 2. Large-scale synthetic multivector residual projection (core methods). The design matrix has $n = 1000$, $p = 50$, $\rho = 45$, and residual codimension $\dim \mathcal{C}(X)^\perp = 955$. All predictors are standardized. Exact projector variants reproduce the same residual up to numerical precision, while Sketch–QR produces a controlled approximation whose agreement is summarized by cosine similarity. Entries marked as FAIL indicate that the algebraic assumptions of the method (e.g., positive definiteness of $X^T X$ for Cholesky-based constructions) are violated, rather than an algorithmic failure.

#	Method	$\ r\ _2$	$\cos(r, r_{\text{exact}})$	Runtime (s)	Status
1	OLS (Moore–Penrose pseudoinverse)	0.5092	1.000	0.0002	OK
2	Cross-product (rank one)	NaN	NaN	NaN	N/A
3	Multivector QR (Exact)	0.5092	1.000	0.0000	OK
4	Recursive Cross–Wedge–QR	0.5092	1.000	0.0008	OK
5	Wedge–QR (+GMI)	0.5092	1.000	0.0000	OK
6	Cholesky–QR	NaN	NaN	0.0016	FAIL
7	Sketch–QR (Gaussian, $s = 150$)	0.5092	1.000	0.0111	OK

Dataset diagnostics: $\text{cond}(X) \approx 6.28 \times 10^{13}$, $\kappa(X^T X) \approx 3.95 \times 10^{15}$, $\text{GMI}(X) = 1.000$.

Sketch–QR yields an approximate residual, whose agreement with the exact residual is summarized by cosine similarity. Entries marked NaN or N/A indicate methods that are not applicable in this regime (e.g., rank-one cross-product when $\dim \mathcal{C}(X)^\perp \neq 1$), while FAIL indicates that the required algebraic conditions (e.g., $X^T X \succ 0$) are violated.

This experiment demonstrates that the geometric residual formulation extends naturally from symbolic low-dimensional examples to realistic large- n , rank-deficient designs, without relying on regularization or nonlinear modeling.

Together with the symbolic R^4 example, this large-scale synthetic experiment confirms that the multivector residual projector NN^T is exact in dimensions, while Sketch–QR provides a principled approximation when a reduced basis is desired, setting the stage for the real-data illustrations in Section 4.3.

4.3. Illustrative Real Data Use of GMI

Real-data experiments use two regression benchmarks: Boston Housing and Auto MPG. Both are standardized via centering and scaling before forming X , and they provide realistic multicollinearity patterns suitable for illustrating GMI as a purely geometric, response-independent diagnostic.

This example illustrates how GMI can serve as a purely geometric multicollinearity diagnostic on real data and how it can be interpreted alongside coefficient stability.

We use the Boston Housing dataset because it is a widely used regression benchmark with well-known predictor correlations, making it a convenient setting for illustrating multicollinearity diagnostics and residual geometry on realistic data.

To illustrate the geometric diagnostics on real data, we use the classical Boston Housing dataset, with median house price as the response and a set of standardized socio-economic and structural predictors. The protocol used in this subsection is as follows:

- (a) Standardize all predictors to zero mean and unit variance, and form the design matrix X .
- (b) For selected subsets of predictors, form the corresponding design matrix X , compute the Gram matrix $G = X^T X$, and evaluate the polar sine and GMI using the definitions in (18).
- (c) Report predictive metrics using five-fold cross-validation (CV) so that MSE and R^2 are out-of-sample summaries rather than in-sample fit statistics.
- (d) Fit ordinary least-squares models on these subsets and monitor the variation of the estimated coefficients under small perturbations of Y (e.g., additive Gaussian noise or resampling).

The purpose of this subsection is to demonstrate (i) that the geometric residual projector $(I - P) = NN^T$ reproduces the classical OLS residuals on a realistic dataset and (ii) that GMI provides a compact, response-independent summary of predictor-space degeneracy.

4.3.1. Boston Housing and Auto MPG: Increasing Model Size

Real-data experiments use two regression benchmarks: Boston Housing and Auto MPG. Both are standardized through centering and scaling before forming X and provide realistic multicollinearity patterns suitable to illustrate GMI as a purely geometric and response-independent diagnostic.

Tables 3 and 4 report results for increasing the size of the model ($p = 1$, $p = 4$, $p \geq 6$, and in full where applicable). For each subset, we report out-of-sample predictive metrics using five-fold cross-validation (CV MSE and CV R^2), geometric multicollinearity (GMI), and classical diagnostics (max VIF and $\kappa(X^T X)$), along with a summary of coefficient stability.

We restrict algorithmic comparisons to the projector variants defined in Table 1 (OLS/Cholesky when applicable, null space/QR, rank-one cross-product when $\text{codim} = 1$, and Sketch-QR).

Table 3. Boston Housing illustration (standardized predictors). Predictive metrics are reported using 5-fold cross-validation (CV MSE and CV R^2). Geometric diagnostics (GMI) are response-independent and depend only on the design matrix X . We additionally report max VIF and $\kappa(X^T X)$, and summarize coefficient stability by $\text{medianSD}(\hat{\beta})$ under resampling/perturbation, to connect GMI with classical multicollinearity measures.

Model	p	CV MSE	CV R^2	GMI	max VIF	$\kappa(X^T X)$	$\text{medianSD}(\hat{\beta})$
$p = 1$ (Area/RM $\rightarrow$ Price/MEDV)	1	44.02	0.467	0.000	1.00	1.00	0.342
$p = 4$ (Area/RM, TAX, PTRATIO, LSTAT)	4	28.02	0.655	0.434	2.09	7.85	0.257
$p \geq 6$ (Area/RM, LSTAT, PTRATIO, TAX, NOX, INDUS)	6	28.20	0.653	0.790	3.25	17.22	0.266
Full model (all predictors)	13	23.67	0.712	0.988	9.01	96.47	0.235

Tables 3 and 4 show that increasing the size of the model improves predictive performance (lower CV MSE, higher CV R^2) while typically increasing multicollinearity. In Boston, moving from $p = 1$ to the full model reduces CV MSE from 44.02 to 23.67 and increases CV R^2 from 0.467 to 0.712, while GMI increases from 0 to 0.988 and $\kappa(X^T X)$ increases from 1 to 96.47. In Auto MPG the collinearity is even stronger: the full model

attains CV MSE 11.31 and CV R^2 0.810, but the corresponding diagnostics are $\text{GMI} \approx 0.980$, $\max \text{VIF} \approx 21.84$, and $\kappa(X^\top X) \approx 136.84$.

Table 4. Auto MPG illustration (standardized predictors), using the same preprocessing and evaluation protocol as in Section 4.3. Predictive metrics are reported using 5-fold cross-validation (CV MSE and CV R^2). Geometric diagnostics (GMI) are response-independent and depend only on the design matrix X . We additionally report $\max \text{VIF}$ and $\kappa(X^\top X)$, and summarize coefficient stability by $\text{medianSD}(\hat{\beta})$ under resampling/perturbation.

Model	p	CV MSE	CV R^2	GMI	$\max \text{VIF}$	$\kappa(X^\top X)$	$\text{medianSD}(\hat{\beta})$
$p = 1$ model (Auto)	1	18.91	0.682	0.000	1.00	1.00	0.118
$p = 3$ engine-core (Auto)	3	18.22	0.693	0.843	10.31	45.06	0.367
$p \geq 6$ model (Auto)	6	12.05	0.797	0.973	19.64	117.03	0.325
Full model (Auto)	7	11.31	0.810	0.980	21.84	136.84	0.355

These results illustrate the intended role of GMI: it is computed purely from X (independent of Y) and tracks familiar multicollinearity measures (VIF and condition number), while the coefficient stability summary provides a concrete link between predictor space degeneracy and sensitivity of fitted parameters.

4.3.2. Implementation Note (Stable GMI Computation)

When the Gram matrix $G = X^\top X$ is poorly conditioned, the evaluation of direct determinants is numerically fragile. We therefore compute the polar sine and GMI in the log-domain with a small jitter $\varepsilon > 0$:

$$\log \text{psin}_\varepsilon(G) = \frac{1}{2} \log \det(G + \varepsilon I) - \frac{1}{2} \sum_{i=1}^p \log(G_{ii} + \varepsilon), \quad \text{GMI}_\varepsilon(G) = 1 - \exp(\log \text{psin}_\varepsilon(G)).$$

In implementation, $\log \det(G + \varepsilon I)$ is evaluated using a numerically stable log-determinant routine (equivalently, a Cholesky factorization when $G + \varepsilon I \succ 0$). Unless stated otherwise, we use $\varepsilon = 10^{-12}$.

Figure 4 makes the equivalence of the exact geometric projector and classical OLS residuals explicit: the discrepancy curve for the exact QR-basis construction is at a numerical noise level in both the $p = 1$ and $p = 4$ models. In contrast, Sketch-QR constructs an approximate projector $\tilde{P}$, so its residual $\tilde{r}$ differs slightly from r_{OLS} ; the magnitude of this discrepancy decreases as the size of the sketch s increases. Overall, this visualization complements Tables 3 and 4 by separating the exact residual-geometry identities (Article 1 methods) from the approximate Sketch-QR construction while keeping the presentation compact and reproducible.

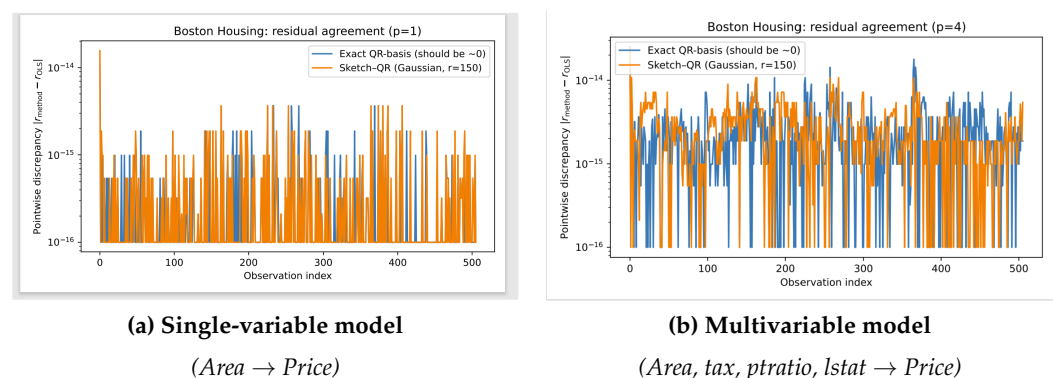


Figure 4. Residual agreement diagnostics on Boston Housing (standardized). For each model, we plot the pointwise absolute residual discrepancy $|r_{\text{method}} - r_{\text{OLS}}|$ across observations. Exact projector

variants (null space/QR and, when applicable, Cholesky/OLS) agree with OLS up to numerical precision, while Sketch-QR (Gaussian sketch, $s = 150$) yields a small but nonzero approximation error. For readability, we visualize discrepancies rather than overlaying multiple residual traces.

5. Discussion

This paper presents a rank-aware geometric interpretation of linear regression residuals via orthogonal projectors. In the case of codimension one, the residual is fully determined by a unit normal $n \in \mathcal{C}(X)^\perp$, resulting in $r = (I - P)Y = nn^\top Y$; in general rank settings, the residual lies in a higher-dimensional orthogonal complement spanned by an orthonormal basis N , with $(I - P) = NN^\top$. This formulation makes the residual subspace explicit and decouples basis construction from repeated projection.

An immediate consequence is that all exact computational routes—normal-equation OLS, Cholesky-based solvers, and null space or QR constructions—must yield identical residuals up to numerical tolerance, as they implement the same projector onto $\mathcal{C}(X)^\perp$. We confirm this numerically through residual agreement metrics and discrepancy visualizations, showing that differences arise only from conditioning and the computational regime. In contrast, Sketch-QR yields an approximate projector and residual, whose deviation from the exact solution is quantified using cosine similarity and norm-based discrepancies.

We further link residual geometry to predictor-space conditioning through the Geometric Multicollinearity Index (GMI), a scale-invariant, normalized volume measure derived from the polar sine. Across the Boston Housing and Auto MPG benchmarks, higher GMI values consistently align with larger max VIF and $\kappa(X^\top X)$, while coefficient stability summaries illustrate increased parameter sensitivity under geometric degeneracy.

Although both regression and PCA rely on orthogonal projections, their objectives differ fundamentally: PCA projects predictors to minimize reconstruction error, with residuals remaining in $\mathcal{C}(X)$, whereas regression projects the response, yielding residuals in $\mathcal{C}(X)^\perp$.

Finally, this work focuses on unregularized least squares and exact orthogonal projection. With regularization, the operators are no longer orthogonal and the residual subspace becomes non-Euclidean or data-adaptive, motivating future work on regularized and sketch-aware residual geometry while preserving the geometric diagnostics enabled by GMI.

6. Conclusions

This paper presented a rank-aware geometric interpretation of linear regression residuals. In the case of codimension-one, the residual projector reduces to the rank-one form $(I - P) = nn^\top$, where n spans $\mathcal{C}(X)^\perp$. For general rank, the residual lies in a $k = n - \rho$ dimensional orthogonal complement, with a projector $(I - P) = NN^\top$ for an orthonormal basis N of $\mathcal{C}(X)^\perp$. Although algebraically equivalent to the classical OLS formulation, this representation makes the residual subspace explicit and decouples basis construction from repeated projection.

Building on this viewpoint, we introduced the Geometric Multicollinearity Index (GMI), a scale-invariant diagnostic derived from the polar sine and Gram determinant that quantifies predictor-space degeneracy. Synthetic experiments confirm predictable behavior under controlled perturbations, while real-data benchmarks on Boston Housing and Auto MPG show that increasing model size typically improves predictive performance while increasing multicollinearity; GMI captures this trade-off and aligns with max VIF and $\kappa(X^\top X)$. Across all experiments, exact projector constructions yield identical residuals up to numerical tolerance, whereas Sketch-QR produces a controlled approximation whose deviation is quantified via cosine similarity and residual discrepancy norms.

In general, these results elevate $\mathcal{C}(X)^\perp$ to a central geometric object in regression analysis and position GMI as a concise indicator of departure from orthogonal designs. Extensions to sketch-based, large-scale, streaming, and nonlinear settings are deferred to companion work.

Author Contributions: Conceptualization, M.A. and S.B.B.; methodology, M.A.; formal analysis, M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.A. and S.B.B.; supervision, S.B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The Boston Housing dataset is available from Kaggle [14] and was originally introduced by Harrison and Rubinfeld [15]. The Auto MPG dataset is available from the UCI Machine Learning Repository [16]. No new data were created. The code used to reproduce the experiments and tables in this manuscript is publicly available as a GitHub repository [13].

Acknowledgments: The authors gratefully acknowledge the academic support of Hamad Bin Khalifa University and Lusail University. They also thank Qatar National Library (QNL) for access to scholarly resources and research databases essential to this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OLS	Ordinary Least Squares
GMI	Geometric Multicollinearity Index
PCA	Principal Component Analysis
VIF	Variance Inflation Factor
CV	Cross-Validation (K-fold)
QR	QR (Orthogonal–Triangular) Decomposition
RRQR	Rank-Revealing QR Decomposition
SVD	Singular Value Decomposition
FLOPs	Floating-Point Operations
GRP	Geometric Residual Projection

References

1. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009. [CrossRef]
2. Golub, G.H.; Van Loan, C.F. *Matrix Computations*, 4th ed.; Johns Hopkins University Press: Baltimore, MD, USA, 2013.
3. Belhaouari, S.; Kahalan, Y.C.; Aksikas, I.; Hamdi, A.; Belhaouari, I.; Haoudi, E.M.N.; Bensmail, H. Generalizing the cross-product to N Dimensions: A Novel Approach for Multidimensional Analysis and Applications. *Mathematics* **2025**, *13*, 514. [CrossRef]
4. Chan, J.Y.-L.; Leow, S.M.H.; Bea, K.T.; Cheng, W.K.; Phoong, S.W.; Hong, Z.-W.; Chen, Y.-L. Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics* **2022**, *10*, 1283. [CrossRef]
5. Kim, J.H. Multicollinearity and misleading statistical results. *Korean J. Anesthesiol.* **2019**, *72*, 558–569. [CrossRef] [PubMed]
6. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150202. [CrossRef] [PubMed]
7. Strang, G. *Introduction to Linear Algebra*; Wellesley–Cambridge Press: Wellesley, MA, USA, 1993.
8. Lay, D.C.; Lay, S.R.; McDonald, J.J. *Linear Algebra and Its Applications*, 5th ed.; Pearson: Boston, MA, USA, 2016.
9. Mortari, D. n -Dimensional cross-product and its application to matrix eigenanalysis. *J. Guid. Control Dyn.* **1996**, *20*, 509–515. [CrossRef]
10. Walsh, B. The scarcity of cross-products on Euclidean spaces. *Am. Math. Mon.* **1967**, *74*, 188–194. [CrossRef]
11. Mahoney, M.W. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.* **2011**, *3*, 123–224. [CrossRef]

12. Halko, N.; Martinsson, P.-G.; Tropp, J.A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **2011**, *53*, 217–288. [[CrossRef](#)]
13. Alkhateeb, M.H.; Belhaouari, S.B. Geometric Residual Projection (GRP): Reproducible Code and Experiments. *GitHub Repository*, 2025. Available online: <https://github.com/MaisAlkhateeb/Geometrics-Residual-Projection> (accessed on 16 January 2026).
14. Kaggle. Boston Housing Dataset. Available online: <https://www.kaggle.com/datasets/altavish/boston-housing-dataset> (accessed on 21 November 2025).
15. Harrison, D.; Rubinfeld, D.L. Hedonic prices and the demand for clean air. *J. Environ. Econ. Manag.* **1978**, *5*, 81–102. [[CrossRef](#)]
16. Quinlan, R. *Auto MPG [Dataset]*; UCI Machine Learning Repository: Irvine, CA, USA, 1993. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.