Supplemental Materials for "Structured Nonlinear Principal Component Analysis for Multivariate Functional Data"

This supplementary document shows detailed proofs of the theoretical result and additional information in the main paper.

1 Theoretical Results with Technical Details

This supplementary appendix provides the rigorous mathematical proofs for the theoretical properties of the SN-FPCA framework presented in the main manuscript. We begin by establishing a unified notation system and restating the foundational assumptions.

1.1 Introduction and Notation

Throughout this appendix, let $(\mathcal{T}, \mathcal{B}(\mathcal{T}), \mu)$ be a compact measure space, where $\mathcal{T}$ represents the functional domain (typically a closed interval $[0, 1]$). We denote $\mathcal{H} = L_2(\mathcal{T}, \mu)$ as the Hilbert space of square-integrable functions with the standard inner product $\langle f, g \rangle = \int_{\mathcal{T}} f(t)g(t)d\mu(t)$ and the induced norm $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle}$.

For the multivariate functional setting, we consider the product space $\mathcal{H}^p = \mathcal{H} \times \cdots \times \mathcal{H}$. For any $\mathbf{X} = [X^{(1)}, \ldots, X^{(p)}]^\top \in \mathcal{H}^p$, the aggregate norm is defined as $\|\mathbf{X}\|^2_{\mathcal{H}^p} = \sum_{m=1}^p \|X^{(m)}\|^2_{\mathcal{H}}$.

The functional neural network is modeled as a class of nonlinear operators $\mathcal{F}$. A specific network configuration is denoted by Θ, where Θ is the parameter space. We define:

- L: The number of hidden layers (depth) of the network.
- S: The maximum number of active weights/parameters (network sparsity).
- $\sigma(\cdot)$: The ReLU activation function, $\sigma(x) = \max(0, x)$.
- $\mathbf{\Phi}(t) = [\phi_1(t), \ldots, \phi_M(t)]^\top$: The pre-specified B-spline basis functions spanning a finite-dimensional subspace $\mathcal{V}_M \subset \mathcal{H}$.

For the penalty analysis, $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the standard ℓ_1 and ℓ_2 norms for vectors in Euclidean space $\mathbb{R}^k$. The expectation $\mathbb{E}[\cdot]$ is taken with respect to the true underlying data-generating distribution $\mathbb{P}$.

1.2 Restatement of Assumptions

To ensure the self-containment of the proofs, we restate the primary assumptions from the main text.

Assumption 1. *The underlying true functional signals* $\mathbf{X}_i(t) = [X_i^{(1)}(t), \ldots, X_i^{(p)}(t)]^\top$ *are uniformly bounded and each* $X_i^{(m)}$ *belongs to a Hölder space of smoothness* $\alpha > 0$*. The generative mappings* $g^* : \mathbb{R}^d \to \mathcal{H}$ *and* $h^{*(m)} : \mathbb{R}^d \to \mathcal{H}$ *are themselves of smoothness* $\beta > 0$ *(i.e.,* β*-times differentiable with the highest derivative being Hölder continuous of the fractional part when* β *is non-integer). Moreover, the data-generating process is driven by a compact latent domain* $\mathcal{Z} \subset \mathbb{R}^d$*. For the product space* $\mathcal{H}^p = \mathcal{H} \times \cdots \times \mathcal{H}$*, the norm is* $\|\mathbf{X}\|_{\mathcal{H}^p}^2 = \sum_{m=1}^p \|X^{(m)}\|_{\mathcal{H}}^2$*. Here, Hölder continuity of* g^* *and* $h^{*(m)}$ *is understood with respect to the* L_2 *norm on* $\mathcal{H}$*. The functional domain is* $\mathcal{T} \subset \mathbb{R}^D$*,* $\mathcal{T} = [a, b]^D$ *(a rectangular domain in* $\mathbb{R}^D$*) for some integer* $D \geq 1$*.*

Assumption 2. *For the purpose of analyzing the penalty mechanism, consider the idealized estimation-layer setting where the working models* $\mathcal{S}$ *and* $\mathcal{P}^{(m)}$ *admit an exact representation in the dictionaries* $\mathbf{\Psi}^{(sh)}$ *and* $\mathbf{\Psi}^{(pr,m)}$*. That is, there exist coefficient vectors* $\boldsymbol{\alpha}^{*(sh)}$ *and* $\boldsymbol{\alpha}^{*(pr,m)}$ *such that*

$$\mathcal{S}(t) = [\mathbf{\Psi}^{(sh)}(t)]^\top \boldsymbol{\alpha}^{*(sh)}, \qquad \mathcal{P}^{(m)}(t) = [\mathbf{\Psi}^{(pr,m)}(t)]^\top \boldsymbol{\alpha}^{*(pr,m)}.$$

Under this representation, we assume that the true decomposition is structurally parsimonious: the active sets of the shared coefficients and the private coefficients are disjoint, i.e.,

$$supp(\boldsymbol{\alpha}^{*(sh)}) \cap supp(\boldsymbol{\alpha}^{*(pr,m)}) = \varnothing, \quad \forall m.$$

This condition formalizes the requirement that any functional variation common across all variates is captured by the shared component alone, and any variation specific to a single variate is captured by that variate's private component.

1.3 Auxiliary Lemmas

To establish the overarching approximation bound in Theorem 1, we decouple the functional neural network's error into two distinct mathematical components: the functional basis truncation error and the latent network approximation error. We introduce two auxiliary lemmas to bound each respectively.

Lemma 1 (Multivariate Basis Truncation Error)**.** *Under Assumption 1, let the functional domain be a compact subset* $\mathcal{T} \subset \mathbb{R}^D$ *with domain dimension* $D \geq 1$*. Let* $\mathcal{V}_M$ *be the space spanned by* M *tensor-product B-spline basis functions of order* $q > \alpha$*. Let* $Proj_{\mathcal{V}_M}\mathbf{X}$ *denote the optimal* L_2 *projection of the true signal* $\mathbf{X}(t)$ *onto* $\mathcal{V}_M$*. Then, there exists a constant* $C_1 > 0$ *such that the projection error is bounded by:*

$$\mathbb{E}\left[\|\mathbf{X} - Proj_{\mathcal{V}_M}\mathbf{X}\|_{\mathcal{H}^p}^2\right] \leq C_1 M^{-2\alpha/D}. \tag{1}$$

Proof. By Assumption 1, each true functional component $X^{(m)}(t) \in C^\alpha(\mathcal{T})$ is defined on a D-dimensional domain. According to classical approximation theory for multivariate polynomial splines in DeVore and Lorentz (1993), constructing a tensor-product

B-spline space $\mathcal{V}_M$ with a total of M basis functions implies approximately $M^{1/D}$ basis functions per physical dimension. The effective knot spacing (mesh size) is $h \sim M^{-1/D}$.

For any function f in the Hölder space $C^\alpha(\mathcal{T})$ and a spline space of order $q > \alpha$, the uniform approximation error is governed by the mesh size: $\inf_{s\in\mathcal{V}_M} \|f - s\|_\infty \leq ch^\alpha \leq c' M^{-\alpha/D}$ for some constant $c' > 0$.

Since $\mathcal{T}$ is a compact measure space, the L_2 norm is bounded by the uniform norm up to a finite measure constant: $\|f\|_{L_2}^2 \leq \mu(\mathcal{T})\|f\|_\infty^2$. Therefore, the L_2 projection error for a single variate is bounded by $\mathcal{O}(M^{-2\alpha/D})$. Summing this error across all p bounded variates yields the aggregated squared error bound:

$$\sum_{m=1}^{p} \|X^{(m)} - \mathrm{Proj}_{\mathcal{V}_M} X^{(m)}\|_{\mathcal{H}}^2 \leq p \cdot \mu(\mathcal{T})(c')^2 M^{-2\alpha/D} = C_1 M^{-2\alpha/D},$$

where $C_1 = p\mu(\mathcal{T})(c')^2$. Summing over the p variates yields a conservative upper bound that holds regardless of any cross-variate correlation structure. □

Lemma 2 (Expected Coefficient Estimation Error)**.** *Let $\tilde{g}^* : \mathcal{Z} \to \mathbb{R}^{pM}$ denote the mapping from the latent variable $\mathbf{z}$ to the stacked B-spline coefficient vector of the full multivariate functional signal $\mathbf{X}(\mathbf{z})$. Under Assumption 1, both g^* and $h^{*(m)}$ are β-Hölder continuous, hence their sum $X^{(m)}$ is also β-Hölder continuous, and $\tilde{g}^*$ inherits the same smoothness. Suppose the network depth and active parameter count scale with sample size N and output dimension pM as $L \asymp \log N$ and $S \asymp pM \cdot N^{\frac{d}{2\beta+d}}$. Then there exists a sparse ReLU network architecture $\hat{f}_{\Theta^*}$ such that the expected squared error satisfies*

$$\mathbb{E}_{\mathbf{z},\, data}\big[\|\tilde{g}^*(\mathbf{z}) - \hat{f}_{\Theta^*}(\mathbf{z})\|_2^2\big] \leq C_2\, pM\, N^{-\frac{2\beta}{2\beta+d}}, \tag{2}$$

where β is the Hölder smoothness of the mapping g^*, and $C_2 > 0$ is a constant independent of M and N.

Proof. Under Assumption 1, the stacked coefficient map $\tilde{g}^* \in \mathbb{R}^{pM}$ is β-Hölder. We decompose the expected squared risk into an approximation error (inherent to the network architecture) and an estimation error (arising from finite-sample training):

$$\mathbb{E}_{\mathbf{z},\,\mathrm{data}}\big[\|\tilde{g}^* - \hat{f}_{\Theta^*}\|_2^2\big] \leq 2 \inf_{f_\Theta\in\mathcal{F}} \underbrace{\mathbb{E}_{\mathbf{z}}\big[\|\tilde{g}^*(\mathbf{z}) - f_\Theta(\mathbf{z})\|_2^2\big]}_{\text{Approximation error}} + 2 \underbrace{\mathbb{E}_{\mathbf{z},\,\mathrm{data}}\big[\|f_\Theta - \hat{f}_{\Theta^*}(\mathbf{z})\|_2^2\big]}_{\text{Estimation error}}. \tag{3}$$

Approximation error. By the deep ReLU network approximation theorem (Yarotsky, 2017), approximating a β-Hölder mapping with a pM-dimensional output space requires a network sparsity (active weights) scaling linearly with the output dimension pM. Specifically, there exists a target network $f_\Theta \in \mathcal{F}$ such that the squared approximation error is bounded by $\mathcal{O}\left(pM \cdot (S/pM)^{-2\beta/d}\right)$.

Estimation error. Standard results from nonparametric regression with sparse ReLU networks Schmidt-Hieber (2020) establish that for a network with sparsity S and depth L, the expected estimation error satisfies $\mathbb{E}\big[\|f_\Theta - \hat{f}_{\Theta^*}\|_2^2\big] \leq C_e \frac{SL\log N}{N}$.

Balancing the two terms. Substituting $L \asymp \log N$ and $S \asymp pM \cdot N^{\gamma}$, we balance the approximation error $\mathcal{O}(pM \cdot N^{-2\beta\gamma/d})$ and the estimation error $\mathcal{O}(pM \cdot N^{\gamma-1} \log N)$. Equating exponents yields $\gamma = d/(d+2\beta)$. With this choice, both terms attain the rate $pM \cdot N^{-2\beta/(2\beta+d)}$ up to a logarithmic factor, completing the proof. □

1.4 Proof of Theorem 1

For the reader's convenience, we first restate the theorem from the main text before presenting its formal proof.

Theorem 1. *Suppose Assumption 1 holds. Let the SN-FPCA architecture project the input onto M B-spline basis functions of order $q > \alpha$, and map the projection coefficients through a deep ReLU network with depth L and active parameters S. Assume the bottleneck dimension satisfies $K \geq d$. By judiciously balancing the spatial truncation error and the latent estimation error, we select the optimal basis dimension as $M \asymp N^{\gamma}$, where $\gamma = \frac{2\beta D}{(D+2\alpha)(2\beta+d)}$. Consequently, if the network scales as $L \asymp \log N$ and $S \asymp pM \cdot N^{\frac{d}{2\beta+d}}$, there exist network parameters Θ^* such that the expected reconstruction error satisfies:*

$$\mathbb{E}\Big[\|\mathbf{X} - \hat{\mathbf{X}}_{\Theta^*}\|_{\mathcal{H}^p}^2\Big] \leq C \cdot N^{-\frac{4\alpha\beta}{(D+2\alpha)(2\beta+d)}},$$

where $C > 0$ is a constant depending on the smoothness parameters, domain size, and variate count p, but strictly independent of N.

Proof. We bound the total expected functional reconstruction error $\mathbb{E}[\|\mathbf{X}(\cdot) - \hat{\mathbf{X}}_{\Theta^*}(\cdot)\|_{\mathcal{H}^p}^2]$.

Let $\mathrm{Proj}_{\mathcal{V}_M}\mathbf{X}$ denote the optimal L_2 spline projection of the true signal onto the space spanned by the M B-spline basis functions, and let $\hat{\mathbf{X}}_{\Theta^*}$ be the functional output synthesized by the network. By adding and subtracting the optimal projection $\mathrm{Proj}_{\mathcal{V}_M}\mathbf{X}$, we apply the inequality $(a+b)^2 \leq 2a^2 + 2b^2$:

$$\|\mathbf{X} - \hat{\mathbf{X}}_{\Theta^*}\|_{\mathcal{H}^p}^2 \leq 2\underbrace{\|\mathbf{X} - \mathrm{Proj}_{\mathcal{V}_M}\mathbf{X}\|_{\mathcal{H}^p}^2}_{T_1} + 2\underbrace{\|\mathrm{Proj}_{\mathcal{V}_M}\mathbf{X} - \hat{\mathbf{X}}_{\Theta^*}\|_{\mathcal{H}^p}^2}_{T_2}. \tag{4}$$

Taking the expectation on both sides, we bound each term separately.

Bounding T_1 (Basis Truncation Error): By Lemma 1, the expected error between the true Hölder continuous function and its optimal spline projection on a D-dimensional domain is bounded by:

$$\mathbb{E}[T_1] \leq C_1 M^{-2\alpha/D}. \tag{5}$$

Bounding T_2 (Network Coefficient Error): Both $\mathrm{Proj}_{\mathcal{V}_M}\mathbf{X}$ and $\hat{\mathbf{X}}_{\Theta^*}$ belong to the same finite-dimensional subspace $\mathcal{V}_M \subset \mathcal{H}$ spanned by the M B-spline basis functions $\{\phi_k\}_{k=1}^M$. For B-splines of fixed order on a compact domain, the basis $\{\phi_k\}$ forms a Riesz basis (Schumaker, 2007). Consequently, there exists a universal constant

$C_\Phi > 0$ (the upper Riesz bound) such that for any function $f = \sum_{k=1}^{M} c_k \phi_k \in \mathcal{V}_M$,

$$\|f\|_{\mathcal{H}} \leq C_\Phi \, \|\mathbf{c}\|_2,$$

where $\mathbf{c} = (c_1, \ldots, c_M)^\top \in \mathbb{R}^M$ is the coefficient vector and $\|\cdot\|_2$ denotes the Euclidean norm.

Applying this inequality to the difference $\text{Proj}_{\mathcal{V}_M}\mathbf{X} - \hat{\mathbf{X}}_{\Theta^*}$ across all p variates yields:

$$\|\text{Proj}_{\mathcal{V}_M}\mathbf{X} - \hat{\mathbf{X}}_{\Theta^*}\|_{\mathcal{H}^p}^2 \leq C_\Phi^2 \, \|\mathbf{c}(\text{Proj}_{\mathcal{V}_M}\mathbf{X}) - \mathbf{c}(\hat{\mathbf{X}}_{\Theta^*})\|_2^2,$$

where $\mathbf{c}(\cdot)$ extracts the stacked coefficient vector across all p variates.

By Lemma 2, the expected squared coefficient error is bounded as $\mathbb{E}\big[\|\tilde{g}^*(\mathbf{z}) - \hat{f}_{\Theta^*}(\mathbf{z})\|_2^2\big] \leq C_2 \, pM \, N^{-\frac{2\beta}{2\beta+d}}$. Since the coefficient vector of $\text{Proj}_{\mathcal{V}_M}\mathbf{X}$ is exactly $\tilde{g}^*(\mathbf{z})$ and that of $\hat{\mathbf{X}}_{\Theta^*}$ is $\hat{f}_{\Theta^*}(\mathbf{z})$ by construction, we obtain:

$$\mathbb{E}[T_2] \leq C_\Phi^2 \, C_2 \, pM \, N^{-\frac{2\beta}{2\beta+d}}.$$

By absorbing the variate count p and the Riesz constant C_Φ^2 into a new constant $C_2' \equiv C_\Phi^2 C_2 p$, we elegantly isolate the basis dimension M:

$$\mathbb{E}[T_2] \leq C_2' \, M \, N^{-\frac{2\beta}{2\beta+d}}.$$

Synthesis: Synthesis and Optimal Rate: Combining the bounds for T_1 and T_2, we obtain the intermediate expected risk bound:

$$\mathbb{E}\left[\|\mathbf{X}(\cdot) - \hat{\mathbf{X}}_{\Theta^*}(\cdot)\|_{\mathcal{H}^p}^2\right] \leq 2C_1 M^{-2\alpha/D} + 2C_2' M N^{-\frac{2\beta}{2\beta+d}}. \tag{6}$$

To achieve the optimal asymptotic convergence rate, we must balance the spatial basis truncation error (which decays with M) and the network estimation error (which grows linearly with M). Equating the asymptotic rates of the two terms yields the optimal scaling for the basis dimension M:

$$M^{-2\alpha/D} \asymp M N^{-\frac{2\beta}{2\beta+d}} \implies M^{1+\frac{2\alpha}{D}} \asymp N^{\frac{2\beta}{2\beta+d}}.$$

Solving for the optimal basis dimension M^* gives:

$$M^* \asymp N^{\frac{2\beta D}{(D+2\alpha)(2\beta+d)}}.$$

Substituting this optimal scaling M^* back into the intermediate risk bound, both terms converge at the identical optimal rate:

$$(M^*)^{-2\alpha/D} \asymp \left(N^{\frac{2\beta D}{(D+2\alpha)(2\beta+d)}}\right)^{-\frac{2\alpha}{D}} = N^{-\frac{4\alpha\beta}{(D+2\alpha)(2\beta+d)}}.$$

Consequently, by absorbing the constants, we obtain the final joint optimal approximation-estimation bound:

$$\mathbb{E}\left[\|\mathbf{X}(\cdot) - \hat{\mathbf{X}}_{\Theta^*}(\cdot)\|_{\mathcal{H}^p}^2\right] \leq C \cdot N^{-\frac{4\alpha\beta}{(D+2\alpha)(2\beta+d)}},$$

where $C > 0$ is a universal constant independent of N. This completes the proof. □

1.5 Proof of Proposition 1

Proposition 1 (Structural Decomposition via Subgradient Bottleneck)**.** *Consider an idealized orthogonal dictionary design where the empirical observation associated with a specific basis element for the m-th variate is denoted by $z_m = z^* + \epsilon_m$. Here, z^* represents the true underlying cross-variate shared consensus signal, and ϵ_m denotes the variate-specific perturbation. The unconstrained coordinate-wise minimization problem is formalized as follows:*

$$\min_{s,\{c_m\}_{m=1}^p} J(s, \mathbf{c}) = \frac{1}{2}\sum_{m=1}^{p}(z_m - s - c_m)^2 + \lambda|s| + \lambda\sum_{m=1}^{p}|c_m|, \tag{7}$$

where $\lambda > 0$ is the regularizing penalty parameter. Assuming the shared signal is of sufficient magnitude ($|z^*| \gg \lambda$), then for any system with $p \geq 2$ variates, the asymmetric penalty structure creates a strict subgradient bottleneck that mathematically prevents the shared consensus signal z^* from being redundantly represented across multiple private branches $\hat{c}_m$ simultaneously.

Proof. Since the unconstrained objective function $J(s, \mathbf{c})$ defined in (7) is continuously subdifferentiable and strictly convex, a candidate solution $(\hat{s}, \hat{\mathbf{c}})$ constitutes the unique global minimum if and only if it satisfies the Karush-Kuhn-Tucker (KKT) subgradient optimality conditions. Let $u \in \partial|\hat{s}|$ and $v_m \in \partial|\hat{c}_m|$ denote the respective scalar elements residing within the standard L_1 subdifferentials, which are strictly bounded such that $u \in [-1, 1]$ and $v_m \in [-1, 1]$.

Setting the partial subgradients of $J(s, \mathbf{c})$ with respect to s and each c_m to zero yields the following system of structural equations:

$$\sum_{m=1}^{p}(z_m - \hat{s} - \hat{c}_m) = \lambda u, \tag{8}$$

$$z_m - \hat{s} - \hat{c}_m = \lambda v_m, \quad \forall m \in \{1, \ldots, p\}. \tag{9}$$

Summing the variate-specific condition (9) over all $m = 1, \ldots, p$, we obtain:

$$\sum_{m=1}^{p}(z_m - \hat{s} - \hat{c}_m) = \lambda\sum_{m=1}^{p} v_m. \tag{10}$$

By directly equating the left-hand side of (10) with the shared structural condition (8) and dividing by the positive scalar λ, we derive the fundamental subgradient

balance equation:

$$u = \sum_{m=1}^{p} v_m. \tag{11}$$

We now establish the structural exclusion property via proof by contradiction. Suppose the network fails to achieve structural separation and attempts to redundantly capture the global shared consensus signal z^* entirely through the private branches. This corresponds to a degenerate scenario where $\hat{s} = 0$, and all p private components are actively driven to manifest this shared variation. Under a sufficiently strong signal z^*, these p active private components must share the same sign as the underlying signal to minimize the reconstruction loss. This boundary activation implies that their corresponding subgradients reach the extremity of the subdifferential, i.e., $v_m = \mathrm{sgn}(z^*)$ for all $m = 1, \ldots, p$.

Consequently, the magnitude of the right-hand side of the balance equation (11) rigorously evaluates to:

$$\left| \sum_{m=1}^{p} v_m \right| = |p \cdot \mathrm{sgn}(z^*)| = p. \tag{12}$$

However, by definition, the left-hand side of (11) is the subgradient of the single shared branch, which is strictly constrained by the mathematical bound:

$$|u| \leq 1. \tag{13}$$

Combining these two results yields the structural inequality $p \leq 1$, which directly contradicts the multivariate premise that $p \geq 2$. This contradiction mathematically proves that the asymmetric L_1 penalty operates as an impassable subgradient bottleneck. It ensures that the global consensus structure cannot be redundantly absorbed by the private branches, compelling the optimizer to route cross-variate consensus exclusively into the shared component $\hat{s}$. □

2 Simulation I: Selection of the Latent Dimension K

Figure S1 displays the out-of-sample reconstruction and decomposition performance across a grid of latent dimensions $K \in \{2, 4, 6, \ldots, 16\}$ for all three simulation scenarios. The left column shows the Total RMSE of the five compared methods (MFPCA, KPCA, Unconstrained-AE, FAE and SN-FPCA), and the right column shows the Shared and Private RMSE of SN-FPCA.

Across all scenarios, the Total RMSE of SN-FPCA, Unconstrained-AE, and MFPCA drops sharply from $K = 2$ to $K = 4$ and plateaus thereafter, indicating that the intrinsic dimensionality of the data has been adequately captured. KPCA exhibits higher and more slowly decreasing RMSE, due to the well-known instability of the pre-image mapping when projecting back from the kernel-induced feature space to the functional domain. Notably, SN-FPCA achieves the lowest Total RMSE at every K, confirming that the functional basis projection and the structured decomposition jointly contribute to more accurate out-of-sample reconstruction.

The right column further shows that the Shared and Private RMSE of SN-FPCA follow the same saturation pattern: both component-wise errors reach their minima

at $K = 4$ and remain essentially flat for larger K. This simultaneous saturation of reconstruction and separation accuracy at $K = 4$ is expected, because the true data-generating process involves exactly four independent latent scores (two shared and two private) in all three scenarios.

These results provide the empirical basis for reporting all quantitative benchmarks in the main text at the parsimonious choice $K = 4$.

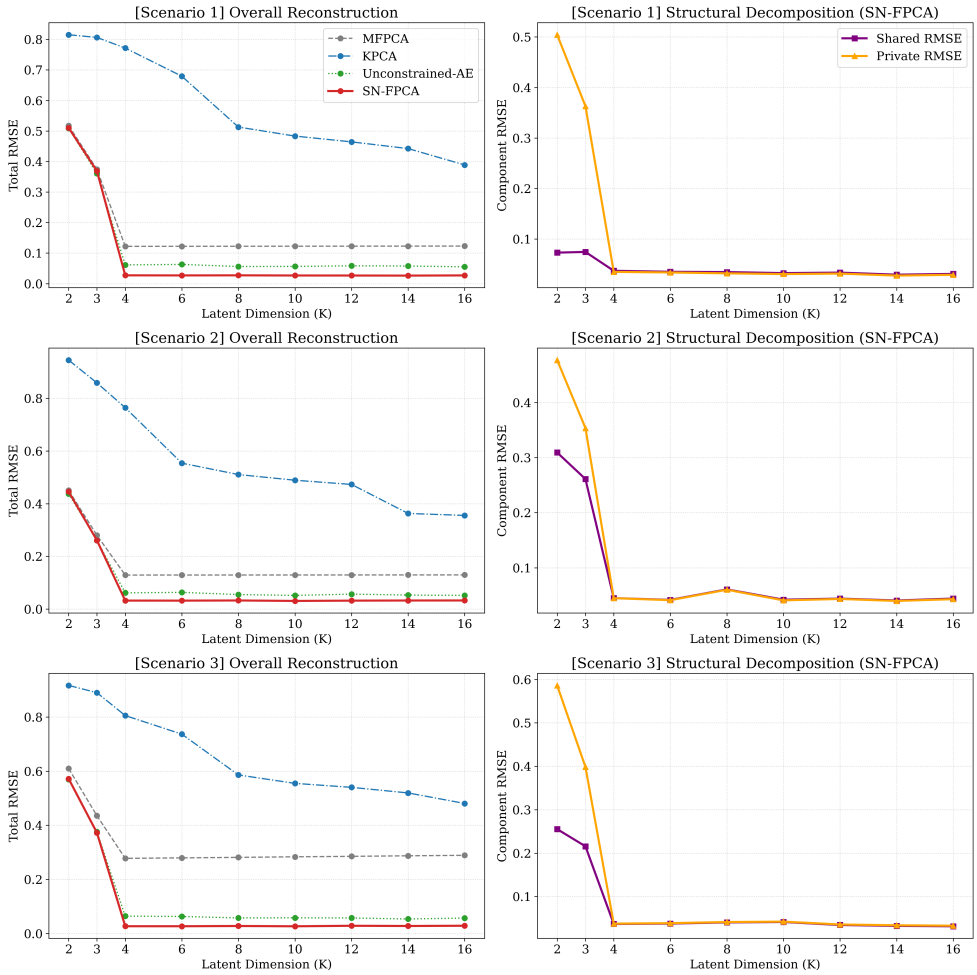

Fig. S1 Test set reconstruction and decomposition performance across varying latent dimensions (K) for all three simulation scenarios.

3 Implementation Details and Reproducibility

To ensure the reproducibility of our empirical results and to provide full transparency regarding the computational framework, this section details the network architectures, hyperparameter configurations, and the implementation specifics of our SN-FPCA model utilized in Section 4 and Section 5.

Simulations (I & II): For the synthetic bivariate functional datasets (20×20 grid), the SN-FPCA data projection utilizes $M = 10$ B-spline bases per spatial dimension, yielding an input coefficient vector of dimension $2 \times 10 \times 10 = 200$. Both the Encoder and Decoder are implemented as lightweight 2-layer Multi-Layer Perceptrons (MLPs) with a hidden dimension of 128, utilizing standard ReLU activations.

Real-World Application (GTSRB): To accommodate the complex, high-frequency environmental perturbations in the multivariate traffic sign images ($3 \times 32 \times 32$ pixels), we employ a deeper FNN architecture. The observations are projected onto a denser 20×20 B-spline basis, resulting in an input coefficient dimension of $3 \times 20 \times 20 = 1200$. The Encoder consists of a 3-layer MLP: `Linear(1200, 1024)` $\rightarrow$ `Linear(1024, 512)` $\rightarrow$ `Linear(512, K)`. The Decoder symmetrically employs a 3-layer MLP: `Linear(K, 256)` $\rightarrow$ `Linear(256, 512)` $\rightarrow$ `Linear(512,` N_w`)`, where N_w is the total number of structural loadings ($N_{sh} + 3 \times N_{pr}$). To stabilize the optimization over the complex real-world manifold, Batch Normalization (BatchNorm1d) and Leaky ReLU (negative slope = 0.2) activations are applied after each hidden layer.

A comprehensive summary of the training hyperparameters across all experiments is provided in Table S1.

Hyperparameter	**Simulation I**	**Simulation II**	**GTSRB Application**
Basis Number per Axis (M)	10	10	20
Shared Capacity (N_{sh})	12	12	32
Private Capacity (N_{pr})	6	6	16
Sparsity Penalty (λ)	10^{-3}	10^{-6}	10^{-2}
Learning Rate (Adam)	0.002	0.002	0.002
Batch Size	250	250	64
Maximum Epochs	200	200	200

Table S1 Summary of hyperparameter configurations for the SN-FPCA framework across different experiments.

References

DeVore, R.A., Lorentz, G.G.: Constructive approximation. In: Grundlehren der Mathematischen Wissenschaften (1993)

Schumaker, L.: Spline Functions: Basic Theory, (2007)

Schmidt-Hieber, J.: Nonparametric regression using deep neural networks with relu activation function. The Annals of Statistics **48**(4), 1875–1897 (2020)

Yarotsky, D.: Error bounds for approximations with deep relu networks. Neural Networks **94**, 103–114 (2017)