Proposer of the vote of thanks to Whiteley et al. and contribution to the Discussion of ‘Statistical exploration of the Manifold Hypothesis’

Tengyao Wang*

December 2025

Whiteley et al. revisit one of the most influential ideas in modern statistical learning: the manifold hypothesis. While the hypothesis is routinely invoked to justify the success of high-dimensional data analysis, it is rarely given a probabilistic explanation. The central contribution of this paper is to show how low-dimensional geometric structure can emerge naturally from statistical dependence among random coordinate functions, rather than being imposed *a priori*.

The proposed Latent Metric Model (LMM) provides a mathematically coherent framework linking latent structure, kernel geometry, and observed data geometry. In this sense, the paper offers not merely a reformulation of the manifold hypothesis, but a principled explanation for why manifold-like data structures may arise in practice.

1 The latent metric model

The LMM assumes that the observed data are generated by evaluating a collection of centred random functions on a latent space $\mathcal{Z}$. Each observation corresponds to a latent location Z_i, with the observed vector formed by evaluating p random coordinate functions at that location, possibly with added noise. Dependence across coordinates is captured through the mean correlation kernel

$$f(z, z') = \frac{1}{p} \sum_{i=1}^{p} \mathbb{E}\{X_j(z) X_j(z')\},$$

which induces a reproducing kernel Hilbert space (RKHS) feature map $\phi : \mathcal{Z} \to \mathcal{H}$. This construction naturally gives rise to three related geometric objects: the latent space $\mathcal{Z}$, the feature manifold $\mathcal{M} = \{\phi(z) : z \in \mathcal{Z}\}$ and the observed data manifold $\mathcal{Y}$.

*Department of Statistics, London School of Economics and Political Science. Email: t.wang59@lse.ac.uk

A key strength of the model lies in the transparency with which it links these three spaces. Geometry is not assumed but induced by the correlation structure encoded in the kernel.

2 Links between latent space, feature manifold and data manifold

The connection between $\mathcal{Z}$ and $\mathcal{M}$ is established through classical RKHS theory. Under mild regularity conditions, continuity and positive definiteness of the kernel imply that the feature map is continuous and injective, yielding a homeomorphism between $\mathcal{Z}$ and $\mathcal{M}$.

More strikingly, when the kernel is isotropic and stationary, the feature map is locally isometric. Local distances in the feature space are proportional to those in the latent space, so that the latent geometry is faithfully preserved at small scales. In addition, smoothness of the kernel implies approximate low rank structure of the associated covariance operator, providing a link between smoothness assumptions and effective dimension reduction. These results show that correlations encode not only topology but also local geometry of the latent space.

The second geometric link arises from random projection arguments. Each observed vector can be viewed as a random linear projection of the feature representation $\phi(Z_i)$ into $\mathbb{R}^p$. The authors show that this projection preserves distances in expectation, with concentration improving as dimension p increases. This phenomenon is closely related to the Johnson–Lindenstrauss lemma. As the ambient dimension grows, the observed data manifold becomes approximately isometric to the feature manifold, and hence to the latent space itself. As a result, the essential geometric structure survives the projection from the infinite-dimensional feature space to the finite-dimensional observation space.

3 Implications for practice

Beyond its theoretical elegance, the LMM sheds light on why common data-analytic pipelines often perform well. In practice, nonlinear dimension reduction methods such as t-SNE or UMAP are frequently preceded by a linear reduction step, typically PCA.

Within the LMM framework, this workflow is well motivated. Smooth correlation kernels induce approximately low-rank structure, justifying an initial linear projection to a moderate-dimensional space. The low intrinsic dimensionality of the resulting manifold then motivates the use of nonlinear embedding methods. The model thus provides a coherent theoretical rationale for procedures that are widely used but rarely justified formally.

4 Discussion and open questions

The paper raises several natural questions for further discussion. From an interpretative perspective, one may ask whether the LMM should be viewed as a mechanistic explanation for manifold structure, or primarily as a flexible mathematical representation.

From a modelling standpoint, the assumption of isotropic and stationary kernels plays a central role in obtaining approximate isometry. In realistic data settings, these assumptions may be violated. It would be interesting to understand how sensitive the geometric conclusions are to such departures, and what weaker conditions might still yield meaningful geometric guarantees.

Finally, there are important statistical questions. How can the induced kernel geometry be estimated from finite samples? Can aspects of the model be tested empirically? Addressing these questions would further strengthen the practical relevance of the framework.

5 Concluding remarks

Whiteley et al. present a technically sophisticated and conceptually satisfying contribution that bridges statistical dependence, kernel methods, and geometric learning. By showing how manifold structure can emerge naturally from correlation among random features, the paper provides a fresh perspective on a foundational assumption in modern data analysis. It is a thought-provoking piece that is likely to stimulate further theoretical and methodological developments.