Referee Report on the PhD Thesis

“Estimating local intrinsic dimension via density estimation”

Candidate: Piotr Tempczyk
Institution: University of Warsaw
Report by Paweł Dłotko

1. Bibliographic and contextual information

The dissertation addresses the problem of *local intrinsic dimension* (LID) estimation in high-dimensional data, motivated by the manifold hypothesis and by applications where the ambient dimension may be in the thousands (e.g. image data). The thesis develops a likelihood-based methodology for LID estimation using deep generative density models and complements it with a unifying theoretical framework based on diffusion (Wiener process) viewpoints and a careful benchmarking methodology.

2. Summary of aims and main results

The overarching goal of the thesis is to provide a principled and scalable method for estimating local intrinsic dimension, understood as the effective number of degrees of freedom near a given point and at a chosen scale. The thesis:

- formalizes an approach to LID estimation based on the scaling of Gaussian-smoothed densities with respect to the noise scale,
- proposes and analyzes the estimator **LIDL** (Local Intrinsic Dimension using approximate Likelihood), implemented with modern density estimators (normalizing-flow type models),
- develops a diffusion/heat-equation perspective that clarifies relationships between a broad class of contemporary neural intrinsic-dimension estimators, and
- proposes a benchmarking framework that bridges classical “toy” manifolds and realistic datasets through controlled transformations that preserve or modify intrinsic dimension in a known way.

3. Scientific contributions and originality

In my assessment, the thesis makes several original and valuable contributions to the theory and practice of intrinsic dimension estimation in machine learning:

3.1. LIDL: scale-by-noise likelihood slope as a local dimension estimator

The central methodological idea is to estimate dimension through how the (log-)density of a point changes under controlled Gaussian perturbations. In essence, if the data concentrate on (or near) a d-dimensional structure embedded in $\mathbb{R}^D$, then the Gaussian-smoothed density exhibits a characteristic scaling in the small-noise limit. LIDL turns this scaling into a practical estimator by fitting a density model and regressing the log-density against the log-noise scale over a selected window of scales, exploiting that $\log \rho_\delta(x)$ scales like $-(D-d)\log\delta$ for sufficiently small δ.

3.2. A unifying diffusion (Wiener-process) viewpoint

A substantial theoretical part of the thesis reframes Gaussian perturbation as diffusion in time and uses the heat equation to connect temporal derivatives of smoothed densities to spatial differential operators. This provides a clean conceptual language that unifies several families of neural LID estimators (likelihood-slope methods, Jacobian-spectrum methods, and score/geometry methods), and clarifies which components are "local" versus "holistic" in nature.

3.3. Benchmarking methodology beyond simplistic synthetic manifolds

The thesis argues convincingly that existing benchmarking practices for intrinsic dimension are often insufficient: they either use overly idealized manifolds or employ real data where ground-truth LID is unknown. The proposed solution is a family of controlled transformations (and a synthetic-to-image generation procedure) that preserve realism while allowing known changes in intrinsic dimension. This is both practically useful and scientifically important, because it makes failure modes visible and comparable across methods.

4. Proven results and their significance

The thesis contains a number of formally stated and proved results (theorems, propositions, and supporting lemmas). I summarize below the most relevant ones and explain their role.

4.1. Core scaling theorem underpinning LIDL

A central theoretical statement (Theorem 2.2.7, "core estimate") establishes that for data supported on a smooth d-dimensional submanifold embedded in $\mathbb{R}^D$, the Gaussian-smoothed density $\rho_\delta(x)$ satisfies the asymptotic relation

$$\log \rho_\delta(x) = (d - D)\log \delta + O(1) \quad \text{as } \delta \to 0.$$

This result rigorously justifies the slope-based estimation principle used by LIDL and clarifies the nature of the bias terms and regularity assumptions. The proof is supported by a sequence of lemmas that formalize local tangent/normal structure and control the contribution from shrinking neighborhoods.

4.2. Extensions to more general geometric settings

The thesis goes beyond the simplest embedded-manifold model. In particular, it introduces conditions for so-called "good immersions" and proves (Proposition 2.4.4) that the core asymptotic behavior extends to such settings. This is relevant in practice because real datasets may exhibit self-intersections, multi-chart representations, or unions of components. The result also clarifies how the effective dimension behaves at points with multiple preimages (e.g. it is governed by the smallest relevant dimension in certain cases).

4.3. Controlled examples and finite-sample/scale effects

The thesis includes results that sharpen intuition about when slope-based estimation works well and when it becomes delicate. For example, Proposition 2.5.1 analyzes anisotropic Gaussian data and identifies regimes in which the estimated dimension depends on spectral gaps relative to the noise scale. Proposition 2.5.2 studies discrete sampling on a line and derives bounds that show how the admissible noise range depends on sample size and accuracy. These examples, while stylized, provide concrete insight into practical hyperparameter choices.

4.4. Heat-equation identities and estimator unification

In Chapter 3, several results connect the diffusion viewpoint to operational estimators. Lemma 3.2.1 and its corollary provide explicit Laplacian expressions for diffused densities in split coordinates. Propositions 3.3.1–3.3.3 establish equivalences between power-law characterizations and relate the slope quantity to ratios involving $\Delta\rho_t/\rho_t$, yielding identities of the form

$$\beta_t(x) = d - D + t \cdot \frac{\Delta_x(\psi * \varphi_t^d)(x)}{(\psi * \varphi_t^d)(x)},$$

which clarifies bias terms and links slope-based and derivative-based viewpoints. Additional results (e.g. Lemma 3.4.2) quantify how distant mixture components vanish exponentially fast in the small-time limit, supporting locality claims.

Assessment: The theoretical part is nontrivial and technically competent. Importantly, the results are not “theory for theory’s sake”; they directly motivate algorithm design choices (noise-scale regression, scale windows, and diagnostics) and explain empirical phenomena observed in the experiments.

5. Experimental evaluation and benchmarking

5.1. Synthetic benchmarks and scalability

The thesis evaluates LIDL on synthetic datasets where ground-truth intrinsic dimension is known, including high-dimensional settings where classical estimators often deteriorate. The experiments emphasize scalability with ambient dimension and demonstrate that likelihood-based methods can remain effective in regimes beyond the practical reach of many traditional neighborhood-based estimators.

5.2. Real datasets and “advanced” high-dimensional data

The thesis includes experiments on widely used real image datasets, including MNIST, FashionMNIST, and CelebA, using density estimators to compute approximate likelihoods. The reported analyses (including sorting by estimated LID and class-conditional comparisons) support the interpretation that LID correlates with perceptual or structural complexity and interacts meaningfully with downstream tasks (e.g. reconstruction behavior in generative models and classification difficulty).

5.3. Controlled real-data benchmarks via transformations

A particularly valuable part of the work is the transformation-based benchmarking framework: starting from a real dataset, one applies carefully designed transformations that preserve or modify intrinsic dimension in a known manner (e.g. auxiliary dimension injection, ambient-space extension by upscaling, monotone embeddings). This yields realistic test cases with controlled “dimension changes” and exposes estimator fragilities that would be invisible on naive benchmarks.

5.4. Realistic synthetic-to-image benchmark (“Arrows”)

To combine realism with known ground truth, the thesis introduces an image-generation procedure that produces structured 2D images from a low-dimensional parameterization (with dimension scaling with the number of objects). This benchmark is a reasonable compromise between pure toy manifolds and large-scale real datasets with unknown intrinsic dimension, and it provides an additional stress test for the proposed method and competitors.

Assessment: The experimental program is extensive and thoughtfully designed, with emphasis on stress-testing assumptions rather than only reporting favorable cases.

6. Candidate's publication record in relation to the thesis

The candidate has an active and coherent publication record that aligns closely with the dissertation topic and supports the maturity of the presented research program.

- The paper *"LIDL: Local Intrinsic Dimension Estimation Using Approximate Likelihood"* (ICML 2022) introduces the LIDL approach that constitutes the central methodological contribution of the dissertation. The thesis goes substantially beyond the conference publication by providing deeper theoretical analysis, broader geometric scope, and a significantly expanded empirical evaluation.
- The paper *"A Wiener Process Perspective on Local Intrinsic Dimension Estimation Methods"* (AAAI 2025) is tightly connected to the diffusion-based framework developed in Chapter 3 and supports the thesis's unification of intrinsic dimension estimators.
- Additional publications in related areas of machine learning (e.g. semi-supervised segmentation and Bayesian neural networks) demonstrate breadth in deep learning methodology and large-scale experimentation, which is reflected in the technical execution of the dissertation.
- The candidate has also contributed to benchmark/competition-style work involving real-world datasets (e.g. transfer learning with heterogeneous EEG datasets), which resonates with the thesis's attention to realistic evaluation protocols and benchmarking pitfalls.

7. Strengths ad weaknesses

The dissertation exhibits a number of clear strengths. First, it is built on a solid theoretical foundation: the core asymptotic scaling result is proved under transparent assumptions and is complemented by nontrivial extensions and illuminating examples. Second, the proposed estimator LIDL is methodologically relevant: it is conceptually straightforward, practically implementable with modern density estimators, and explicitly multi-scale through the noise parameter. Third, the diffusion/heat-equation viewpoint provides a unifying perspective that clarifies connections between several contemporary intrinsic-dimension estimators and places them within a common analytic framework. Fourth, the benchmarking component is particularly valuable: the transformation-based evaluation protocol is well motivated and helps uncover failure modes that standard synthetic benchmarks often fail to reveal. Finally, the overall presentation and scope indicate a mature research program, reinforced by peer-reviewed publications closely aligned with the thesis topic.

Despite the high quality of the work, several limitations (many of which are also acknowledged by the author) merit discussion. A primary practical bottleneck is the choice of the noise scale: although the theory motivates the $\delta \to 0$ regime, real data typically admit only a narrow operating window in which density estimates remain reliable and the manifold-based approximation is not dominated by discreteness or ambient effects. Developing a more systematic, possibly adaptive, scale-selection strategy would further strengthen the applicability of the method. A second limitation is the dependence on the quality of the underlying density model; as with other likelihood-based approaches, performance can be sensitive to model expressiveness and training quality, which is an inherent constraint rather than a weakness of the thesis itself, but it should be kept in mind when positioning the approach.

8. Formal aspects and presentation

The dissertation is well structured, with a clear separation between theoretical development, methodological design, and empirical evaluation. The notation is mostly consistent, and the chapters build a coherent narrative from geometric foundations to practical algorithms and benchmarks.

9. Conclusion and recommendation

In conclusion, the thesis presents a substantial, original, and well-executed contribution to intrinsic dimension estimation, combining rigorous analysis, algorithmic development, and thoughtful benchmarking. The results are relevant to contemporary machine learning and data analysis, and the candidate's publication record corroborates the significance and maturity of the research.

Recommendation: I recommend the dissertation for the public defense and awarding the doctoral degree.

Signed by /
Podpisano przez:

Paweł Dłotko

Date / Data:
2026-02-01 10:38