

# Novel high intrinsic dimensionality estimators

A. Rozza · G. Lombardi · C. Ceruti · E. Casiraghi ·  
P. Campadelli

Received: 29 October 2011 / Accepted: 1 May 2012 / Published online: 26 May 2012  
© The Author(s) 2012

**Abstract** Recently, a great deal of research work has been devoted to the development of algorithms to estimate the intrinsic dimensionality (id) of a given dataset, that is the minimum number of parameters needed to represent the data without information loss. id estimation is important for the following reasons: the capacity and the generalization capability of discriminant methods depend on it; id is a necessary information for any dimensionality reduction technique; in neural network design the number of hidden units in the encoding middle layer should be chosen according to the id of data; the id value is strongly related to the model order in a time series, that is crucial to obtain reliable time series predictions.

Although many estimation techniques have been proposed in the literature, most of them fail on noisy data, or compute underestimated values when the id is sufficiently high. In this paper, after reviewing some of the most important id estimators related to our work, we provide a theoretical motivation of the bias that causes the underestimation effect, and we present two id estimators based on the statistical properties of manifold neighborhoods, which have been developed in order to reduce this effect. We exhaustively evaluate the proposed techniques on synthetic and real datasets, by employing an objective evaluation measure to compare their performance with those achieved by state of the art algorithms; the results show that the proposed methods are promising, and produce reliable estimates also in the difficult case of datasets drawn from non-linearly embedded manifolds, characterized by high id.

**Keywords** Intrinsic dimensionality estimation · Dimensionality reduction · Manifold learning

## 1 Introduction

At the present, pattern recognition techniques applied to solve real life classification problems (such as face recognition, Abate et al. 2007, protein subcellular localization, Rozza et

---

Editors: Dimitrios Gunopoulos, Donato Malerba, and Michalis Vazirgiannis.

A. Rozza (✉) · G. Lombardi · C. Ceruti · E. Casiraghi · P. Campadelli  
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39-41,  
20135 Milan, Italy  
e-mail: [rozza@dsi.unimi.it](mailto:rozza@dsi.unimi.it)