Bayesian Composite Likelihood Inference for the Intrinsic Dimension

Brutti Pierpaolo [1] & De Santis Fulvio [1]

[1] *Department of Statistics, "Sapienza" University of Rome, Piazzale Aldo Moro 5, 00185 Rome, pierpaolo.brutti@uniroma1.it*

Résumé. Dans ce travail nous proposons une méthode bayésienne pour l'inférence sur la dimension intrinsèque d'un nuage de points échantillonnés à partir d'une structure de dimension faible, plongée dans un espace de grande dimension. L'ingrédient essentiel de notre "recette" bayésienne est une vraisemblance marginale composite construite sous l'hypothèse d'indépendance, comme suggérée par MacKay et Ghahramani (2005), afin d'améliorer une proposition antérieure qui utilise des approximations locales basées sur le processus de Poisson (Levina et Bickel, 2005). Pour obtenir une distribution "a posteriori" avec un comportement et une courbure asymptotiques approximativement correctes, nous calibrons cette pseudovraisemblance comme dans Pauli et al. (2001) et ensuite, à partir d'exemples réels et simulés, nous comparons une méthode MCMC standard avec une variante de la méthode bayésienne de référence décrit dans Ventura et al. (2013).

Mots-clés. Dimension intrinsèque, Vraisemblance non–calculable, Vraisemblance composite marginale, Inférence de Bayes.

Abstract. In this work we propose a new Bayesian method for inference on the intrinsic dimension of point–cloud data sampled from a low–dimensional structure embedded in a high–dimensional ambient space. The basic ingredient of our Bayesian recipe is a composite marginal likelihood built under independence assumptions, that was suggested by MacKay and Ghahramani (2005) to improve on an earlier proposal based on local Poisson process approximations (Levina and Bickel, 2005). In order to get a posterior with approximately correct asymptotic behavior and curvature, we calibrate this pseudolikelihood as in Pauli et al. (2011). In simulated and real examples we compare a standard MCMC method against a variation of the default Bayesian technique described in Ventura et al. (2013).

Keywords. Intrinsic dimension, Intractable likelihood, Composite marginal likelihood, Bayesian inference.

1 Introduction

The need for analyzing massive high–dimensional datasets is widespread nowadays and has spawned a flurry of research papers that tackle the problem from different statistical perspectives ranging from the more theoretical to the more applied corners of the discipline. In spite of this abundance, though, there is always a single crucial assumption, essentially shared by all these techniques, that make them work; that is, the data are not genuinely high–dimensional but, in a way or another, can be squeezed back to a lower dimension, their intrinsic dimension, without losing any relevant portion of information. In computer vision and image processing, for example, the intrinsic dimension of a sequence of n pictures taken at a typical resolution of 720×480, say, may represent the (small) number of degrees of freedom needed to capture the dynamic features hidden in these $D = 345,600$ dimensional signals, such as different exposure levels or roto-translations of single elements. From these basic facts, it becomes almost self-evident how important practically is to get a reliable estimate of this fundamental descriptor of a dataset.

In this work we propose a Bayesian method for inference on the intrinsic dimension having as basic ingredient a composite marginal likelihood of independence suggested by MacKay and Ghahramani (2005) to improve on an earlier proposal detailed in Levina and Bickel (2005). In order to get a posterior with approximately correct asymptotic behavior and curvature, we calibrate this pseudolikelihood as in Pauli et al. (2011), and then compare the performance of a standard MCMC method against a variation of the default Bayesian framework described in Ventura et al. (2013).

2 Background

In this section we will briefly review some of the likelihood–based approaches to estimate the intrinsic dimension already available in the literature and more strictly related to our developments. As in Levina and Bickel (2005), let $\mathbb{X}_n = \left\{ \mathbf{X}_i \in \mathbb{R}^D \right\}_{i=1}^n$ be a set of i.i.d. observations that represent a *sufficiently smooth* embedding of a lower–dimensional sample $\mathbb{Y}_n = \left\{ \mathbf{Y}_i \in \mathbb{R}^d \right\}_{i=1}^n$ drawn from an *unknown* smooth density $f(\cdot) = f_Y(\cdot)$ supported on $\mathbb{R}^d$ with $d << D$. Then, if $R_k(\mathbf{x})$ denotes the Euclidean distance from a fixed point $\mathbf{x}$ to its k-th nearest–neighborhood (NN) in the sample $\mathbb{X}_n$, from the smoothness of the embedding we can see that the proportion k/n of points that fall into a ball of radius $R_k(\mathbf{x})$ around $\mathbf{x}$ satisfies the following

$$\frac{k}{n} \approx f(\mathbf{x})\, V(d) \left[R_k(\mathbf{x}) \right]^d,$$

where $V(d)$ is the volume of the unit sphere in $\mathbb{R}^d$. With this basic fact at hand, we can proceed by considering the (inhomogeneous) point process $N(R, \mathbf{x})$ which counts the number of $\mathbb{X}_n$–samples falling into a small D–dimensional sphere $\mathcal{B}_R(\mathbf{x})$ of radius R centered around $\mathbf{x}$. This is a binomial process that under appropriate conditions can be approximated by a suitable Poisson process. Hence, finally, if we assume $f(\mathbf{x}) \approx \text{const}$ inside a small enough sphere $\mathcal{B}_R(\mathbf{x})$, the rate λ of $N(R, \mathbf{x})$ can be written as

$$\lambda(R, \mathbf{x}) = f(\mathbf{x})\, V(d)\, d\, R^{d-1},$$

and the associated *local* log–likelihood takes the form (Snyder and Miller, 1991)

$$L\big(d(\mathbf{x}), \theta(\mathbf{x})\big) = \int_0^R \log \lambda(r, \mathbf{x}) \mathrm{d}N(r, \mathbf{x}) - \int_0^R \log \lambda(r, \mathbf{x}) \mathrm{d}r,$$

where $\theta(\mathbf{x}) = \log f(\mathbf{x})$. This is an *exponential family* whose MLEs solve the following likelihood equations

$$\begin{cases} \frac{\partial L}{\partial \theta} = N(R, \mathbf{x}) - \mathrm{e}^{\theta}\, V(d)\, R^d = 0 \\ \frac{\partial L}{\partial d} = \left(\frac{1}{d} + \frac{V'(d)}{V(d)}\right) N(R, \mathbf{x}) + \int_0^R \log r \,\mathrm{d}N(r, \mathbf{x}) - \mathrm{e}^{\theta}\, V(d)\, R^d \left(\log R + \frac{V'(d)}{V(d)}\right) = 0 \end{cases},$$

and are equal to

$$\begin{cases} \widehat{d}(\mathbf{x}) = \left[\frac{1}{N(R,\mathbf{x})} \sum_{j=1}^{N(R,\mathbf{x})} \log \frac{R}{R_j(\mathbf{x})}\right]^{-1} \\ \widehat{f}(\mathbf{x}) = \frac{N(R,\mathbf{x})}{V(\widehat{d}(\mathbf{x}))R^{\widehat{d}(\mathbf{x})}} \end{cases}.$$

The problem now is to combine the local estimates $\widehat{d}(\mathbf{x}_i)$ obtained in the neighborhood of each of the n datapoints into a single estimator of the overall intrinsic dimension d. The original proposal by Levina and Bickel was simply to *average* these n estimators, but this solution showed a surprisingly strong bias at small radius R that MacKay and Ghahramani (2005) fixed just by considering a *composite likelihood* built under the working assumption of independence, which gives the following new set of estimates

$$\begin{cases} \widetilde{d} = \left[\frac{1}{\sum_i N(R,\mathbf{x}_i)} \sum_{i=1}^{n} \sum_{j=1}^{N(R,\mathbf{x}_i)} \log \frac{R}{R_j(\mathbf{x}_i)}\right]^{-1}, \\ \widetilde{f}(\mathbf{x}_i) = \frac{N(R,\mathbf{x}_i)}{V(\widetilde{d})R^{\widetilde{d}}}, \quad \forall\, i \in \{1, \ldots, n\}. \end{cases}$$

These early developments have been followed by a number of variations and extensions (e.g. Haro et al., 2006, 2007, 2008) but, to the best of our knowledge, no Bayesian procedure is available at this time to *directly* estimate the intrinsic dimension of data, although such an estimate may actually arise as a *byproduct* of approaches with broader modeling scopes. For example, since a compact d–dimensional Riemannian manifold can always be covered by a finite number of d–dimensional balls, Chen et al. (2010) adopt a Bayesian nonparametric framework to fit a particular mixture of Gaussians to the data. In this model, each cluster in the mixture may have a different dimensionality – with the overall intrinsic dimension estimated as the average intrinsic dimension of the clusters – and the algorithm seeks to minimize both the number of clusters and their intrinsic dimension by adjusting the relevant posterior log–probabilities.

3 Our proposal

When the full likelihood function is too difficult to handle analytically because of complex dependencies, but, as in the present case, it is possible to compute the likelihood function for some subsets of the data, it may be useful and effective to resort to a class of approximate likelihoods called *composite likelihoods* (Varin et al., 2011). In general a composite likelihood $CL(\boldsymbol{\psi})$ is defined as $CL(\boldsymbol{\psi}) = \prod_i f(y \in A_i; \boldsymbol{\psi})^{w_i}$ where $\{w_i\}_i$ are positive weights and $\{A_i\}_i$ are all measurable events in the sample space. $CL(\boldsymbol{\psi})$ can essentially be interpreted as a proper likelihood but for a *misspecified* model. For this reason, a composite likelihood do not satisfy the so called information identity, and this typically implies that it is wrongly too concentrated. As a result, before we can crunch a composite likelihood $CL(\boldsymbol{\psi})$ into some sort Bayesian machinery, we necessarily need to adjust it tweaking the weights $\{w_i\}_i$ in order to get (approximately) the right asymptotic behavior. To this end, here we calibrate MacKay and Ghahramani's pseudolikelihood as suggested in Pauli et al. (2013) – see also Pace et al. (2011) – to obtain a final composite likelihood defined as follow

$$CL(\boldsymbol{\psi}) = CL(d, \boldsymbol{\theta}) = \prod_{i=1}^{n} L\big(d, \theta(\mathbf{x}_i)\big)^{1/\bar{\lambda}}$$

where $\bar{\lambda} = (1/p) \sum_{j=1}^{p} \lambda_j(\widehat{\boldsymbol{\psi}})$, $p = \mathtt{dim}(\boldsymbol{\psi})$, $\widehat{\boldsymbol{\psi}}$ is the maximum composite likelihood estimator, $\{\lambda_j(\boldsymbol{\psi})\}_j$ are the eigenvalues of $I(\boldsymbol{\psi})^{-1}\, J(\boldsymbol{\psi})$, $I(\boldsymbol{\psi}) = \mathbb{E}\big(-\nabla \mathbf{u}(\boldsymbol{\psi}; \mathbb{X}_n)\big)$ is the expected (Fisher) information matrix, $J(\boldsymbol{\psi}) = \mathbb{V}ar\big(\mathbf{u}(\boldsymbol{\psi}; \mathbb{X}_n)\big)$ the variability matrix and $\mathbf{u}(\boldsymbol{\psi}) = \nabla \log CL(d, \boldsymbol{\theta})$ denotes the score function associated to $CL(\boldsymbol{\psi})$. Then, once we choose a suitable prior density $\pi(d, \boldsymbol{\theta})$, we may formally get a genuine posterior as

$$\pi_{CL}(d, \boldsymbol{\theta} \,|\, \mathbb{x}_n) \propto \pi(d, \boldsymbol{\theta})\, CL(d, \boldsymbol{\theta}).$$

In practice this approach would require not only the use of possibly complex MCMC schemes to explore the posterior density, but also the specification of a prior distribution over the whole high dimensional nuisance parameter $\boldsymbol{\theta} = \log \boldsymbol{f} = \big[\log f(x_1), \ldots, \log f(x_n)\big]^{\mathtt{T}}$. As a workaround to these familiar Bayesian quirks, we also consider the default Bayesian analysis of Ventura et al. (2009, 2013).

Bibliography

[1] Chen, H., Silva, J., Dunson, D., Carin, L. (2010), Hierarchical Bayesian Embeddings for Analysis and Synthesis of Dynamic Data.
Available online: http://people.ee.duke.edu/∼lcarin/Haojun_JMLR12.pdf
[2] Haro, G., Randall, G., Sapiro G. (2006), Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. *In Advances in NIPS 19, Vancouver, Canada.*
[3] Haro, G., Randall, G., Sapiro G. (2007), Regularized mixed dimensionality and density learning in computer vision. In *Proceedings of 1st Workshop on Component Analysis Methods for C.C.M. and E.P. in C.V., in conjunction with CVPR.*

[4] Haro, G., Randall, G., Sapiro G. (2008), Translated Poisson Mixture Model for Stratification Learning. *Int. J. Comput. Vis.*, 80, 358–374.

[5] Levina, E., Bickel, P.J. (2005), Maximum likelihood estimation of intrinsic dimension. In *Advances in NIPS 17, Vancouver, Canada.*

[6] MacKay D. J.C., Ghahramani Z.(2005). Comments on "Maximum Likelihood Estimation of Intrinsic Dimension" by E. Levina and P. Bickel.
Available online: http://www.inference.phy.cam.ac.uk/mackay/dimension/

[7] Pace, L., Salvan, A., Sartori, N. (2011), Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21, 129-148.

[8] Pauli, F., Racugno, W., Ventura, L. (2011), Bayesian composite marginal likelihoods, *Statistica Sinica*, 21, 149-164.

[9] Snyder, D.L., Miller, M.I. (1991), *Random Point Processes in Time and Space.* Springer–Verlag.

[10] Varin, C., Reid, N., Firth, D. (2011), An overview of composite likelihood methods, *Statistica Sinica*, 21, 5-42.

[11] Ventura, L., Cabras, S., Racugno, W. (2009), Prior distributions from pseudo-likelihoods in the presence of nuisance parameters, *Journal of the American Statistical Association*, 104, 768–774.

[12] Ventura, L., Sartori, N., Racugno, W. (2013), Objective Bayesian higher–order asymptotics in models with nuisance parameters. *Computational Statistics & Data Analysis*, 60, 90-96.