GEOMETRIC COMPLEXITY FOR STATISTICAL MANIFOLDS

BRUNO MERA, PAULO MATEUS, AND ALEXANDRA M. CARVALHO

Abstract. Model complexity plays an essential role in its selection by choosing a model that fits the data and is also succinct. Two-part codes and the minimum description length have successfully delivered procedures to single out the best models, avoiding overfitting. In this talk, we pursue this approach and complement it by performing further assumptions in the parameter space. Concretely, we assume that the parameter space is a smooth manifold, and by using tools of Riemannian geometry, we derive a sharper expression than the standard one given by the stochastic complexity, where the scalar curvature of the Fisher information metric plays a dominant role. Furthermore, we present a sharper approximation to the capacity for exponential families and apply our results to derive optimal dimensional reduction in the context of principal component analysis. The main results were published in [13].

Introduction

Two-part codes are an essential tool in model selection. Not only do they optimize the likelihood of the data given the model, but they also take into account model complexity. There has been a line of research where one considers, in the most abstract setting, families of distributions satisfying minimal requirements and derives an expression for model complexity, such as the stochastic complexity, among others [16, 18]. These formulas are sharp to the extent of the absence of assumptions in the assignment of a probability distribution to each point in the parameter space. Moreover, it is a rather usual assumption that this parameter space has the topology of an open subset in $\mathbb{R}^n$.

In this talk, we show that by making additional assumptions on the parameter space and endowing it with natural information geometric structures, we can arrive at sharper results by applying techniques from Riemannian geometry. In practice, the parameters of the distributions are usually taken to live on a smooth manifold, and the distribution is assumed to vary smoothly with the parameters. However, usually one takes the simplification that this manifold is a trivial open subset of the Euclidean space. In this work, we will drop this assumption, hence allowing for non-trivial topologies. Moreover, Information Theory endows the manifold with a positive (semi-)definite covariant 2-tensor, namely a Riemannian metric – the Fisher information [2, 1]. Since we are given a Riemannian structure, we have a natural notion of a uniform distribution over the manifold of parameters, which corresponds to what is known in the literature as Jeffreys' prior [12, 7].

The second author has been partially supported by Instituto de Telecomunicações, namely by the Fundação para a Ciência e a Tecnologia (FCT) through national funds, by FEDER, COMPETE 2020, and by Regional Operational Program of Lisbon, under UIDB/50008/2020.

The talk at the 8IMM 2022 has been given by the second author.

In the literature, when the parameter space is just a bounded open set in $\mathbb{R}^n$, one can find the (normalized) maximum likelihood code, defined by

$$p^*(x^N) = \frac{p(x^N|\hat{\theta})}{\int_{y^N \in \mathcal{X}^N} p(y^N|\hat{\theta}) dy^N}. \tag{1}$$

The associated length was firstly given by Rissanen [16], computed through Laplace's formula, and has the form

$$\begin{aligned} L^*(x^N) &= -\log(p^*(x^N)) \\ &= -\log p(x^N|\hat{\theta}) + \frac{n}{2}\log\left(\frac{N}{2\pi}\right) + \log \int \sqrt{|I(\theta)|} d\theta + o(1), \end{aligned} \tag{2}$$

where the expansion is stated in terms of the size of the dataset N. Note that throughout the text, we will consider log to be the natural logarithm which is more convenient in the context of the present geometrical approach. While in Rissanen's original work, he considered x^N beyond i.i.d. processes, in the present work, we will only focus on this case. Observe that Eq. (2) does not account for the possible dependence of the $o(1)$ term in the dimension of the parameter space. Indeed, in this work, using techniques from Riemannian Geometry, we find the sharper formula to the codelength

$$L^*(x^N) \simeq_{\text{a.e.}} -\log p(x^N|\hat{\theta}) + \frac{n}{2}\log\left(\frac{N}{2\pi}\right) + \log \text{vol}_g(M) + \underbrace{\frac{1}{6N} R(\hat{\theta}) + \text{O}\left(\frac{1}{N^2}\right)}_{o(1) \text{ as a function of } N}, \tag{3}$$

where three classical geometric invariants can be easily identified, namely: (i) the dimension of the manifold n, (ii) the Riemannian volume $\text{vol}_g(M)$; and (iii) the Ricci scalar curvature $R(\hat{\theta})$ evaluated at the maximum likelihood estimate $\hat{\theta}$. The above equation holds almost everywhere in the sense that the left-hand side and the right-hand side converge with probability 1 to the same random variable. While in Eq.(2) the term $\log \int \sqrt{|I(\theta)|} d\theta$ is precisely the logarithm of the Riemannian volume, we choose to write it explicitly to highlight its geometric nature. Note that the scalar curvature might be very large as a function of the dimensionality of the data involved. We will illustrate this effect by showing how the latter quantity explicitly depends on n for the case of Gaussian models. Indeed, nowadays, data is becoming very high dimensional, in particular, the number of variables (related to n) is becoming comparable to the number of samples N. Hence, terms that were before negligible might have a very strong dependence on n and one needs sharper formulae that take into account this dependence. Furthermore, we remark that the methods from Riemannian geometry used in the current work to write down the sharper formulas can be systematically used to derive higher orders approximations. In particular, near $\hat{\theta}$, one can use the expansion in normal coordinates of Jeffreys' prior, i.e., the square root of the determinant of the metric, together with a higher order Taylor expansion of $-\log p(x|\theta)$ and use standard Gaussian integration methods to do so. This is reminiscent of the methods used in asymptotic expansions of heat kernels in geometry used for instance, in modern proofs of instances of the Atiyah-Singer index theorem [5]. Although geometric approaches [6, 19, 15] and very sharp formulas [21], some of them based on average properties [8, 9], were already derived in the past, they did not attain the sharpness of $\text{O}(1/N^2)$ achieved in this work

To derive the codelength given by Eq. (3), motivated by the results in [3], we follow a Bayesian approach considering Jeffreys' prior and we adapt Laplace's method to manifolds, using canonical Riemann normal coordinates to our advantage. The complexity of density estimation and their relation to two-part codes has been made explicit by Barron and Cover [4]. With this observation, Eq. (3) can be interpreted as a codelength of a two-part code, where the stochastic complexity [16] is refined taking into account the geometry of the statistical model, and therefore we call such refinement the *Geometric Complexity*.

Along the same lines of the above results, one can obtain a sharper expansion of the average case minimax redundancy $\widetilde{R}_N(M)$ in the context of a statistical manifold M. In particular, we will consider exponential families parametrized by a bounded open subset of Euclidean space with the non-trivial geometry provided by the Fisher metric. We stress that the geometry induced by the Fisher metric is in general non-trivial due to the following observation: for dimensions greater than two, if one takes a Riemannian metric different from the standard Euclidean one, the associated Riemann tensor will be non-vanishing. Nevertheless, the hypotheses for the capacity theorem by Haussler [10] are still satisfied and hence the average case minimax redundancy equals the capacity for the case at hand.

We show that the capacity $C_N(M)$ is given by

$$\widetilde{R}_N(M) = C_N(M) = \log \mathrm{vol}_g(M) + \frac{n}{2}\log\left(\frac{N}{2\pi e}\right) + \frac{1}{6N\mathrm{vol}_g(M)}S(g) + \mathrm{O}\left(\frac{1}{N^2}\right),$$

where $S(g) = \int_M R(\theta)dV_g$ is the Einstein-Hilbert action functional [11, 20, 14] evaluated at the Fisher information metric g. This provides yet another hint on modern views relating information, complexity and gravity [17].

We apply our results to a very well-established method for dimensional reduction, namely, Principal Component Analysis (PCA). In particular, our results yield a natural criterion for the choice of the optimal dimension by adapting the two-part code given in Eq. (3) to zero mean Gaussian families with varying covariance. The underlying parameter space is the manifold $\mathscr{P}_m$ of positive definite matrices, with reduced dimension $m \times m$ which we want to optimize, equipped with the Fisher metric. We considered a bounded subset $M(s)$ of $\mathscr{P}_m$, controlled by an integer s that is the smallest integer such that $I_d \leq \Sigma \leq 2^{2s} I_d$, where $\Sigma = XX^T/N$ is the empirical covariance matrix and I_d is the $d \times d$ identity matrix. We also assume that each component of the data is written as an integer multiple of the precision for each variable, and therefore the volume depends on the precision and not in a particular system of units. For this particular case, the formula becomes

$$L^*(x^N) \simeq_{\text{a.e.}} -\log p(x^N|\hat{Q}) + \frac{m(m+1)}{4}\log\left(\frac{N}{2\pi}\right) + \log \mathrm{vol}_g(M(s)) - \frac{(m+2)m(m-1)}{24N},$$

where

$$\begin{aligned}\log \mathrm{vol}_g\left(M(s)\right) = &-\frac{3}{2}m\log(2) - \log(m!) + m\log(2) + \frac{m(m+1)}{4}\log(\pi) \\ &- \log\left(\frac{\pi^{\frac{1}{4}} A^{\frac{3}{2}} G\left(\frac{m}{2} + \frac{3+(-1)^m}{4}\right)}{2^{\frac{1}{24}} e^{\frac{1}{8}}}\right) - \log\left(G\left(\left\lfloor \frac{m}{2} \right\rfloor + 1\right)\right) + \log I(s),\end{aligned}$$

A is the Glaisher constant, G is the Barnes G-function, and

$$I(s) = s^m \left(\log(2)\right)^m 8^{\frac{m(m-1)}{4}} \times \int_{[0,1]^m} \prod_{1\leq i<j\leq m} \sinh\left(s\log(2)|u_i - u_j|\right) \prod_{i=1}^{m} du_i,$$

whose asymptotic behavior with s is studied in the published paper [13]. Notice that the fourth term of Eq. (3) does not appear in our expression, since it is exactly zero for Gaussian models. Remarkably, the curvature term is negative due to the hyperbolic nature of the geometry of Gaussian statistical models, which brings a negative correction to $L^*(x^N)$. This correction is expected to be particularly relevant for high dimensional data.

References

[1] S. Amari. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 2012.

[2] S. Amari and H. Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[3] V. Balasubramanian. MDL, Bayesian inference, and the geometry of the space of probability distributions. In *Advances in minimum description length: Theory and applications*, pages 81–98. MIT Press, 2005.

[4] Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

[5] Nicole Berline, Ezra Getzler, and Michele Vergne. *Heat kernels and Dirac operators*. Springer Science & Business Media, 2003.

[6] Nikolai Nikolaevich Cencov. *Statistical decision rules and optimal inference*. Number 53. American Mathematical Soc., 2000.

[7] B. Clarke and A. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 08 1994.

[8] Bertrand Clarke. Information optimality and Bayesian modelling. *Journal of Econometrics*, 138(2):405–429, 2007.

[9] Bertrand Clarke. Comment on Article by Sancetta. *Bayesian Analysis*, 7(1):37–44, 2012.

[10] D. Haussler. A general minimax result for relative entropy. *IEEE Transactions on Information Theory*, 43(4):1276–1280, 1997.

[11] David Hilbert. *The Foundations of Physics (First Communication)*, pages 1925–1938. Springer Netherlands, Dordrecht, 2007.

[12] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[13] Bruno Mera, Paulo Mateus, and Alexandra M. Carvalho. Model complexity in statistical manifolds: The role of curvature. *IEEE Trans. Inf. Theory*, 68(9):5619–5636, 2022.

[14] Charles W Misner, Kip S Thorne, and John Archibald Wheeler. *Gravitation*. Macmillan, 1973.

[15] Kohei Miyamoto, Andrew R Barron, and Jun'ichi Takeuchi. Improved MDL estimators using local exponential family bundles applied to mixture families. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1442–1446. IEEE, 2019.

[16] J. Rissanen. Fisher information and stochastic complexity. *IEEE transactions on information theory*, 42(1):40–47, 1996.

[17] Leonard Susskind. Three lectures on complexity and black holes. arXiv preprint arXiv:1810.11563, 2018.

[18] A. Suzuki and K. Yamanishi. Exact calculation of normalized maximum likelihood code length using Fourier analysis. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1211–1215, 2018.

[19] Jun'ichi Takeuchi and Andrew R Barron. Asymptotically minimax regret for models with hidden variables. In *2014 IEEE International Symposium on Information Theory*, pages 3037–3041. IEEE, 2014.

[20] Steven Weinberg. *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. Wiley, New York, NY, 1972.

[21] Jingge Zhu. Semi-supervised learning: the case when unlabeled data is equally useful. In *Conference on Uncertainty in Artificial Intelligence*, pages 709–718. PMLR, 2020.

Bruno Mera; Tohoku University
Email address: xx@xx.uu

Paulo Mateus; Instituto de Telecomunicações, Instituto Superior Técnico, University of Lisboa...
Email address: paulo.mateus@tecnico.ulisboa.pt

Alexandra M. Carvalho; Instituto de Telecomunicações, IST, University of Lisbon
Email address: alexandra.carvalho@tecnico.ulisboa.pt