A MAP ESTIMATION ALGORITHM FOR BAYESIAN POLYNOMIAL REGRESSION ON RIEMANNIAN MANIFOLDS

Prasanna Muralidharan[1] *Jacob Hinkle*[2] *P. Thomas Fletcher*[1]

[1] School of Computing & SCI Institute, University of Utah,
[2] Computational Science Center, National Renewable Energy Lab

ABSTRACT

In this paper, we present a Bayesian formulation of polynomial regression on a Riemannian manifold. Previous methods for fitting a curve to manifold-valued data have been formulated as geometric, least-squares estimation problems. We show that least-squares estimation on manifolds, much like the familiar Euclidean case, suffers from overfitting when using higher-order polynomials. Our Bayesian model mitigates this overfitting by placing a prior on the polynomial coefficients that shrinks their magnitude, analogous to Bayesian Euclidean regression with a Gaussian prior on the coefficients. We develop an algorithm for computing maximum a posteriori estimates of polynomial coefficients and the noise variance. Experiments on synthetically generated sphere data and a real shape regression problem demonstrate the advantages of our approach.

1. INTRODUCTION

Several manifold regression models [1, 2, 3, 4, 5] have been recently introduced, where the goal is to estimate a smooth curve on a manifold that best explains the relationship between a real-valued independent parameter (such as time) and a manifold-valued response variable. A primary application for such models is in studies of the temporal changes to biological or anatomical shape caused by growth, disease, or other processes. Manifold representations of shape, or *shape spaces* [6], have been effective at modeling the nonlinear variability inherent to shape. Manifold data also arise in applications involving directional data (spheres), and other geometric objects (rotations, affine transforms, diffeomorphisms, symmetric positive-definite tensors, structured covariance matrices [7], etc.).

Jupp and Kent [4] originally introduced a spherical regression model formulated as data unrolling, which was later extended to shape spaces by Kume et al. [8]. Miller [9] proposed a piece-wise geodesic growth model on the space of diffeomorphic transformations. Nonparametric formulations of the regression problem on manifolds have also been developed [1, 10]. More recently, parametric regression models have been proposed, including geodesic regression [2, 5] and polynomial regression [3], formulated as least-squares minimization problems on manifolds. Such models provide easy to interpret parameters relating to actual physical changes in the dependent variable. As is typical in regression problems, there is a trade-off in model flexibility (e.g., polynomials are more flexible than geodesics) versus generalizability, that is, avoiding overfitting.

We first show that least-squares fitting of higher-order polynomials can suffer from overfitting to the data even on a Riemannian manifold. This is not surprising, given that adding more high-order terms to the polynomial increases the flexibility of the curve, just as in the Euclidean case. We instead propose a Bayesian formulation of the Riemannian polynomial regression problem, where we introduce a Bayesian prior to control the magnitude of polynomial coefficients. Unlike in $\mathbb{R}^n$, the resulting posterior distribution does not have a closed-form solution for manifold-valued data. We present a maximum a posteriori (MAP) estimation procedure and demonstrate it's ability to prevent the overfitting that occurs with least-squares estimation.

2. BACKGROUND

Recall that a Riemannian manifold (M, g) is a differentiable manifold M equipped with a metric g, which provides a smoothly varying inner product on the tangent spaces of M. For a thorough treatment of concepts in Riemannian geometry, we refer the reader to [11]. Here, we briefly review what it means to be a polynomial on a Riemannian manifold. Hinkle, et., al [3] define a kth-order polynomial on a Riemannian manifold M as a curve γ satisfying

$$(\nabla_{\dot{\gamma}(t)})^k \dot{\gamma}(t) = 0 \quad \text{s.t.} \quad (\nabla_{\dot{\gamma}(0)})^i \gamma(0) = c_i \tag{1}$$

for all $t \in [0, 1]$ and fixed scalars $c_i, i = 0, 1, \ldots, k$. Equation (1) can be rewritten as a system of $k + 1$ coupled equa-

This work was funded by NSF CAREER Grant 1054057 and NIH grant 1R01EB022876-01.

tions,

$$\begin{aligned} \dot{\gamma}(t) &= \beta_1(t), \\ \nabla_{\dot{\gamma}(t)}\beta_i(t) &= \beta_{i+1}(t), \quad i = 1, \ldots, k-1, \\ \nabla_{\dot{\gamma}(t)}\beta_k(t) &= 0. \end{aligned} \tag{2}$$

As with polynomials in $\mathbb{R}^n$, specifying initial conditions $\beta_0(0), \beta_1(0), \ldots, \beta_k(0)$, namely, the initial position in M and initial values of all orders of derivatives (in $T_{\beta_0(0)}M$) up to order k determines a unique k^{th}-order polynomial. The space of all possible $\beta_0, \ldots, \beta_k$ is also a smooth manifold, which we denote $\mathcal{P}_M^k$. For instance, the space of "order zero" polynomials is the manifold M itself. When $k = 1$, polynomial curves are just geodesics, and $\mathcal{P}_M^k = TM$, the tangent bundle. Just as TM can be constructed as a disjoint union of all tangent spaces of M, $\mathcal{P}_M^k \equiv \amalg_{p\in M}(T_pM)^k$, i.e., the smooth manifold of polynomial parameters is built as the disjoint union of all products of k copies of tangent spaces. We denote a specific set of initial conditions as $\beta \in \mathcal{P}_M^k$ and the resulting polynomial curve as γ_β.

Given data $(x_i, y_i), i = 1, \ldots, n$, [3] define the least-squares estimate as the polynomial curve that minimizes the sum-of-squared geodesic distances to the data, i.e.,

$$\hat{\beta} = \arg \min_{\beta \in \mathcal{P}_M^k} \sum_{i=1}^{n} d_M(\gamma_\beta(x_i), y_i)^2. \tag{3}$$

where d_M is distance metric defined on the manifold M and $\hat{\beta} \in \mathcal{P}_M^k$ corresponds to the polynomial curve that minimizes the sum-of-squared geodesic distances to the data. See that, for $k = 1$, the above optimization problem reduces to estimating the best-fit least-squares geodesic given manifold data, addressed by [12] and [5]. [12] shows that the least-squares geodesic is also the maximum likelihood estimate under certain distributional assumptions.

The main drawback of least-squares estimation of higher order polynomials for manifold data is that it suffers from data-overfitting, much like in the Euclidean case. This is expected as a complex/flexible model has many free parameters and thus can always fit the data better. This can be seen on the sphere in Figure 1. Here, we generate data on the sphere about a cubic (in red), and estimate a 5th-order least-squares regression solution (in blue) as per (3). See that the blue curve overfits the data on the sphere just as in the Euclidean example on the left. The overfitting manifests as a blow up in the norms of coefficients as we go up in polynomial order, as seen in Table 1 for the least-squares estimate.

3. THE MODEL

In contrast to the least-squares approach, where polynomial parameters β are deterministic, we instead treat β as a random variable and propose a Bayesian model with appropriate sparsity constraints to mitigate overfitting. Our first principle is that the model should automatically choose the appropriate polynomial coefficients to include in a data-driven fashion. The second principle is Occam's razor, that the model should be no more complex than is needed to explain the data. Therefore, analogous to sparse Euclidean regression [13, 14], we formulate an automatic relevance determination (ARD) prior on higher order polynomial coefficients that results in an automatic selection of the most relevant polynomial coefficients, driving the others to zero.

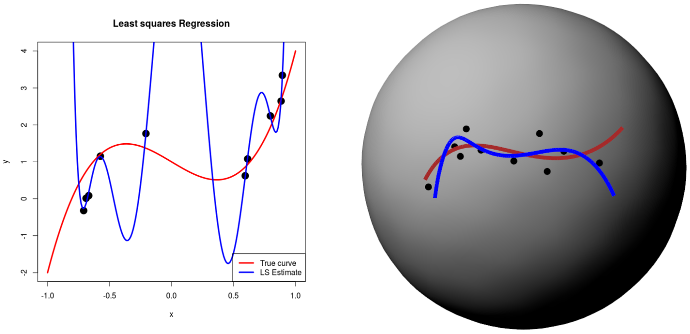

Fig. 1. Ground truth cubic (red), synthetically generated data (black). Least-squares polynomial regression estimate - Left: $8th$ order polynomial (blue) on Euclidean space; Right: 5th order polynomial (blue) on the sphere

Coeff. norms	$\|\beta_1\|$	$\|\beta_2\|$	$\|\beta_3\|$	$\|\beta_4\|$	$\|\beta_5\|$
Ground truth	1.56	6.08	12.37	0.0	0.0
Least-squares	2.89	23.1	132.3	473.0	818.6

Table 1. Coefficient norms on the sphere: Ground-truth Vs Least-squares estimates. $\beta_i \in T_{\beta_0}M, i = 1, \ldots, k$ are tangent vectors at the initial position β_0 that represent higher order derivatives. For instance β_1 represents velocity, β_2 acceleration and so on.

Let $D = \{(x_i, y_i), i = 1, \ldots, n\}$ with $x_i \in \mathbb{R}$ being the explanatory variable and $y_i \in M$ the manifold response variable. Given polynomial parameters, β, with the associated polynomial, γ_β, we propose the following generative Bayesian regression model:

Likelihood. Given independent variable x_i and the polynomial coefficients β, we model y_i as a Riemannian normal random variable about the location $\gamma_\beta(x_i) \in M$

$$p(y_i|\beta; \tau, x_i) = \frac{1}{C(\tau, \beta, x_i)} \exp\left(-\frac{\tau}{2} d_M(\gamma_\beta(x_i), y_i)^2\right), \tag{4}$$

where τ is a dispersion parameter for this distribution, similar to the precision of a Euclidean Gaussian, and

$$C(\tau, \beta, x_i) = \int_M \exp\left(-\frac{\tau}{2} d_M(\gamma_\beta(x_i), y_i)^2\right) dy \tag{5}$$

is the associated normalization coefficient. The value $C(\tau, \beta, x_i)$

only depends on τ when M is a Riemannian homogeneous space (i.e., a manifold with transitive isometry group) [12]. We therefore write $C(\tau)$ instead of $C(\tau, \beta, x_i)$ for this setting. Note that the coefficient $C(\tau)$ is finite for compact manifolds, but in general, the integral is not computable in closed-form. For symmetric spaces, Zhang and Fletcher [15] outline an approximation to $C(\tau)$, which we use in this paper. Due to the independence of the normalizing coefficient on β, maximizing $\log p(D|\beta;\tau)$, the log of the joint data likelihood, is equivalent to the least-squares minimization formulation in (3), as described by [3]. For manifolds that are not homogeneous spaces, this property doesn't hold.

Prior on β. We place a uniform prior on the initial position β_0, but penalize the higher order derivatives $\beta_j \in T_{\beta_0}M, j = 1, \ldots, k$, by placing a Gaussian prior on them.

$$p(\beta|\omega) = \left(\prod_{j=1}^{k} \frac{\omega_j}{2\pi}\right)^{\frac{d}{2}} \exp\left(-\sum_{j=1}^{k} \frac{\omega_j}{2} \|\beta_j\|^2\right)$$

where d is the dimension of the manifold M and ω_j is the precision of a d-dimensional isotropic Gaussian prior on $T_{\beta_0}M$ for the coefficient β_j. The variable ω_j moderates the strength of the coefficient β_j in the model. Note that this distribution is proper for compact manifolds, but improper for non-compact ones (e.g., this is a uniform Jeffrey's prior on the intercept when $M = \mathbb{R}^n$).

Posterior for β. Given data D, and the prior on β, the posterior distribution for β is

$$p(\beta|D;\tau,\omega) = \frac{p(D|\beta,\tau)p(\beta|\omega)}{\int p(D|\beta,\tau)p(\beta|\omega)d\beta} = \frac{1}{C_{\text{post}}} \exp\left(-U(\beta)\right), \tag{6}$$

where

$$U(\beta) = \frac{\tau}{2}\sum_{i=1}^{n} d(\gamma_\beta(x_i), y_i)^2 + \sum_{j=1}^{k} \frac{\omega_j}{2}\|\beta_j\|^2, \tag{7}$$

$$C_{\text{post}} = \int_{\mathcal{P}_M^k} \exp(-U(\beta))\, d\beta. \tag{8}$$

Note that the above distribution is surely integrable for compact homogeneous spaces.

4. INFERENCE

Lets consider Bayesian regression with Euclidean data, i.e., the dependent response variable $y_i \in \mathbb{R}^d$. As in our setup, lets assume a polynomial regression model, with Gaussian errors about the curve. In addition to a Gaussian data likelihood and Gaussian priors on higher order coefficients $\beta_j, j \geq 1$, lets also place a Gaussian prior on position β_0 instead of the uniform prior as in our setup. The resulting posterior in this case is also Gaussian due to conjugacy. The marginal likelihood $p(D|\omega,\tau)$ is also Gaussian in this case, and therefore has an explicit functional form. Maximizing this marginal likelihood results in closed form expressions for parameters τ and ω. However, for data lying on a general manifold, there isn't a natural way to define conjugate priors. Also, the posterior distribution (6) does not have an explicit functional form. Although the posterior normalizing constant C_{post} is finite for compact homogeneous spaces, we can't compute it explicitly, and therefore exact inference of hyperparameters is not possible. In such a setting, it is natural to devise strategies to sample from the posterior distribution and use the samples to infer model parameters. This is a challenging task on Riemannian manifolds. Some methods have been developed to sample from distributions on manifolds by [16]. But extending those methods to sampling on $\mathcal{P}_M^k$ is beyond the scope of this paper.

Instead, we compute the MAP, i.e., the mode of the posterior, as a point estimate for the polynomial regression coefficients. Note that maximizing the log posterior is equivalent to minimizing the objective function U given in (7). Analogous to [3], we minimize U using a gradient descent scheme. Computing gradients involves first forward integrating the polynomial, and then backward integrating adjoint variables, λ_j, to arrive at gradients for the initial conditions at time $t = 0$. Solving the variational problem for the adjoint variables leads to the following system of differential equations (see [3] for a derivation):

$$\begin{aligned} \nabla_{\dot\gamma_\beta(t)}\lambda_0(t) &= -\sum_{i=1}^{k} R(\beta_i(t), \lambda_i(t))\beta_1(t), \\ \nabla_{\dot\gamma_\beta(t)}\lambda_i(t) &= \lambda_{i-1}(t), \end{aligned}$$

where R is the Riemannian curvature tensor on M. Now, the gradients of U w.r.t. the β_j can be written as

$$\begin{aligned} \nabla_{\beta_0}U &= -\lambda_0(0), \\ \nabla_{\beta_j}U &= -\lambda_j(0) - \omega_j\beta_j. \end{aligned} \tag{9}$$

We propose a max-max algorithm to simultaneously infer the noise precision τ, the prior parameters ω_j, and the MAP estimate for the polynomial coefficients β. Here, we alternatively fix τ and ω_j, estimating the MAP for β, and then fix β to estimate τ and ω_j until convergence. This procedure is summarized in Algorithm 1.

Algorithm 1 Max-max algorithm for MAP of β, τ and ω

```
Initialize $(\tau, \omega)$. Output: $\hat\beta, \hat\tau, \hat\omega$
while Until $\tau, \omega$ and $\hat\beta$ converge do
    Fix $\tau, \omega$, update $\beta$: $\hat\beta = \arg_\beta \min U(\beta)$
    Fix $\beta$, update $\tau, \omega$:
    $\hat\tau = \arg_\tau \min \left(n \log C(\tau) + \frac{\tau}{2}\sum_{i=1}^{n} d(\gamma_{\hat\beta}(x_i), y_i)^2\right)$
    $\hat\omega_j = \frac{d}{\|\hat\beta_j\|^2}$
end while
```

5. RESULTS

We validate the proposed model on two example manifolds, namely the sphere S^2 and Kendall shape space. We run our experiments on synthetic data for the sphere and on rat calivarium data, first studied by Bookstein [17]. For basics about sphere and Kendall shape space geometry, see [2, 18].

5.1. Comparing MAP and MLE on the sphere

Here, we generated data from the likelihood model given by (4) from a cubic polynomial (in red) (see Figure 2). We computed a 5th-order MLE and a 5th-order MAP estimate. Also, from Tables 1- 2, note that the coefficient norms of the MLE blow up as we go up in order, whereas the MAP is much more stable at higher orders. For the MLE, we set the noise parameter $\tau = 500$, whereas for the MAP, we set $\tau = \hat{\tau} = 446.3$. $\hat{\tau}$ is the estimated τ parameter from our max max algorithm. Notice that in terms of standard deviation, $\sigma = \frac{1}{\sqrt{\tau}} = 0.0447$, and $\hat{\sigma} = \frac{1}{\sqrt{\hat{\tau}}} = 0.0473$ are very similar.

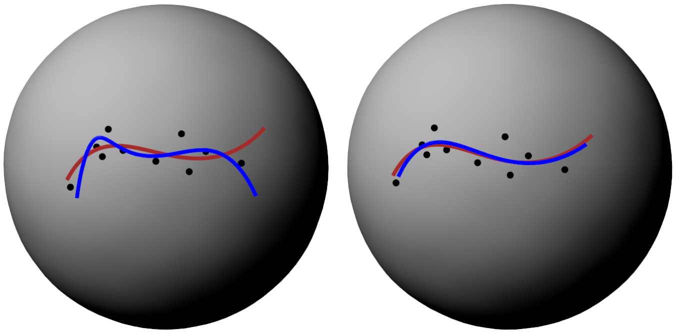

Fig. 2. Regression using least squares (left, in blue) and maximum a posteriori (right, in blue) estimation. Ground truth curve (in red), with synthetic data shown in black.

Coeff. norms	$\|\beta_1\|$	$\|\beta_2\|$	$\|\beta_3\|$	$\|\beta_4\|$	$\|\beta_5\|$
Ground truth	1.56	6.08	12.37	0.0	0.0
MAP	1.52	6.83	13.13	1.87	0.46

Table 2. Ground truth vs MAP: Coefficient norms

5.2. Comparing MAP and MLE on rat calivarium data

In this experiment, we model the rat calivarium data on $2D$-Kendall shape space with eight particle positions. We choose a single rat with eight time-points (shown in Figure 3) with 8 shapes going from blue to purple). We estimate 4^{th} order polynomial regression curves - MLE (in red) and MAP (in blue). Figure 4 shows zoomed views of four of the particle positions. Just as in the sphere case, see that the ML estimate overfits the data (top-left zoom-in) whereas the MAP is more resilient to higher order inflections. We also estimated $\tau \approx 7124, \omega_1 \approx 230, \omega_2 \approx 3130$. ω_3, ω_4 values were high enough to force the 3rd and 4th order terms close to zero. Therefore the MAP is essentially quadratic.

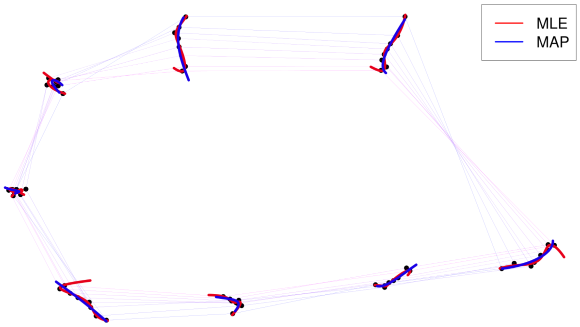

Fig. 3. Bookstein's rat calivarium - MLE(red) vs MAP (blue): Eight landmarks and their 4^{th} order trajectories

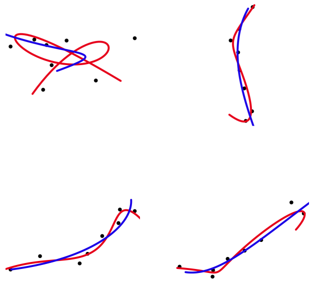

Fig. 4. Detailed zoomed views for four of the landmarks in Figure 3

6. CONCLUSION

We presented a Bayesian formulation of polynomial regression for data lying on a Riemannian manifold. Our Bayesian model mitigates overfitting to data by including a prior on polynomial coefficients that shrinks their magnitude. We develop a gradient descent algorithm for computing the MAP estimate for the regression problem, along with estimates of the noise precision and prior parameters. Future work will go beyond point estimates and investigate Monte Carlo sampling techniques to sample from the full posterior distribution.

7. REFERENCES

[1] B. Davis, P. T. Fletcher, E. Bullitt, and S. Joshi, "Population shape regression from random design data," in *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[2] P. T. Fletcher, "Geodesic regression on Riemannian manifolds," in *MICCAI Workshop on Mathematical Foundations of Computational Anatomy*, 2011, pp. 75–86.

[3] J. Hinkle, P. Muralidharan, P. T. Fletcher, and S. C. Joshi, "Polynomial regression on Riemannian manifolds," in *ECCV (3)*, 2012, pp. 1–14.

[4] P. E. Jupp and J. T. Kent, "Fitting smooth paths to spherical data," *Applied Statistics*, vol. 36, no. 1, pp. 34–46, 1987.

[5] M. Niethammer, Y. Huang, and F.-X. Viallard, "Geodesic regression for image time-series," in *Medical Image Computing and Computer Assisted Intervention*, 2011.

[6] L. Younes, "Spaces and manifolds of shapes in computer vision: An overview," *Image and Vision Computing*, vol. 30, no. 6, pp. 389–397, 2012.

[7] Salem Said, Hatem Hajri, Lionel Bombrun, and Baba C Vemuri, "Gaussian distributions on riemannian symmetric spaces: statistical learning with structured covariance matrices," *arXiv preprint arXiv:1607.06929*, 2016.

[8] A. Kume, I. L. Dryden, and H. Le, "Shape-space smoothing splines for planar landmark data," *Biometrika*, vol. 94, no. 3, pp. 513–528, 2007.

[9] M. Miller, "Computational anatomy: shape, growth, and atrophy comparison via diffeomorphisms," *NeuroImage*, vol. 23, pp. S19–S33, 2004.

[10] Monami Banerjee, Rudrasis Chakraborty, Edward Ofori, Michael S Okun, David E Viallancourt, and Baba C Vemuri, "A nonlinear regression technique for manifold valued data with applications to medical image analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4424–4432.

[11] M. do Carmo, *Riemannian geometry*, Birkhäuser, 1992.

[12] P. T Fletcher, "Geodesic regression and the theory of least squares on riemannian manifolds," *International Journal of Computer Vision*, vol. 105, no. 2, pp. 171–185, 2013.

[13] David JC MacKay, *Bayesian methods for adaptive models*, Ph.D. thesis, California Institute of Technology, 1992.

[14] Michael E Tipping, "Sparse bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.

[15] Miaomiao Zhang and P. Thomas Fletcher, "Probabilistic principal geodesic analysis," in *Advances in Neural Information Processing Systems*, 2013, pp. 1178–1186.

[16] Mark Girolami and Ben Calderhead, "Riemann manifold langevin and hamiltonian monte carlo methods," *J. of the Royal Statistical Society, Series B (Methodological*, vol. 73, pp. 123–214, 2011.

[17] F. L. Bookstein, *Morphometric Tools for Landmark Data: Geometry and Biology*, Cambridge Univ. Press, 1991.

[18] B. O'Neill, "The fundamental equations of a submersion," *Michigan Math. J.*, vol. 13, pp. 459–469, 1966.