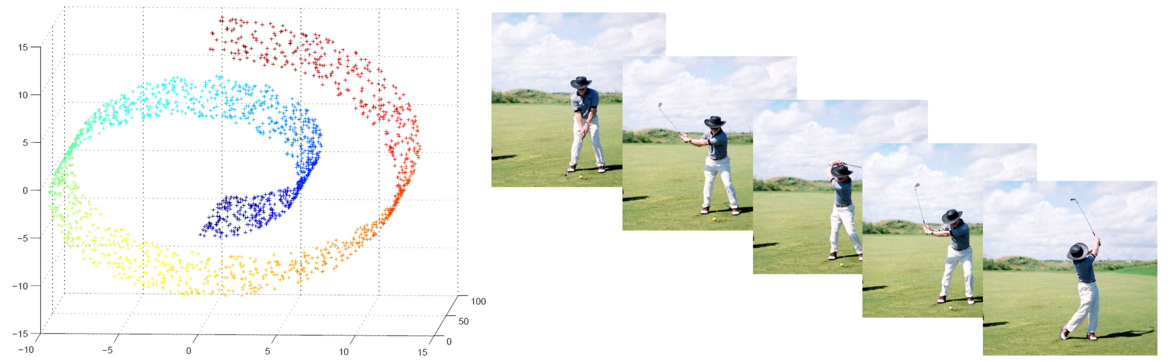
Probabilistic Curve Learning: Coulomb Repulsion and the Electrostatic Gaussian Process

Ye Wang, David Dunson

Department of Statistics, Duke University

Overview

Manifold learning is a common approach of learning low dimensional structure in multidimensional data. However, there is a clear lack of probabilistic methods that allow learning of the manifold along with the generative distribution of the observed data. The best attempt is the Gaussian process latent variable model (GP-LVM) [1], but identifiability issues lead to poor performance. We solve these issues by proposing a novel Coulomb repulsive process (Corp) for locations of points on the manifold, inspired by physical models of electrostatic interactions among particles. Combining this process with a GP prior for the mapping function yields a novel electrostatic GP (electroGP) process.



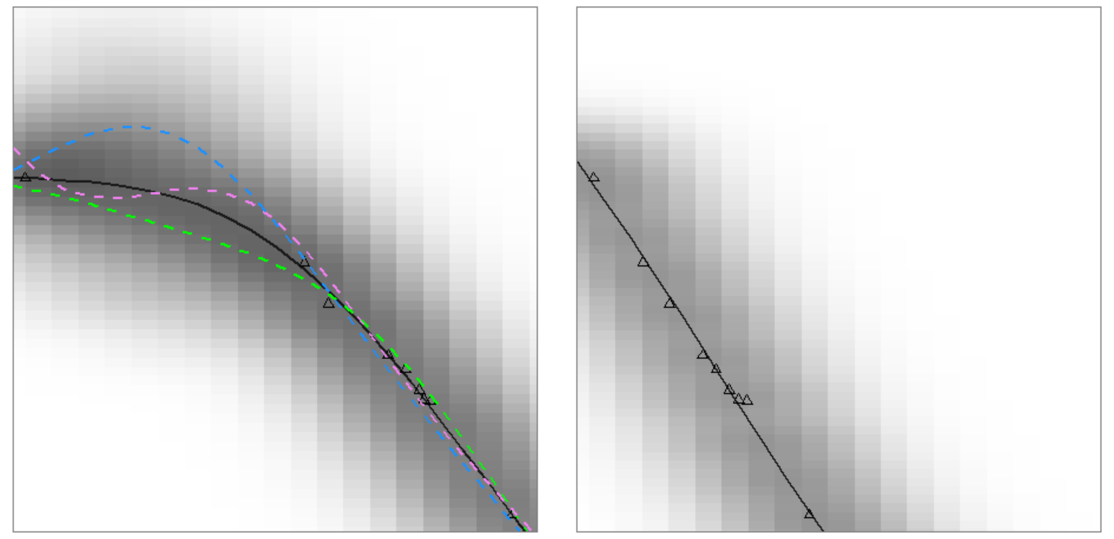
(Source: http://www.golfswingphotos.com)

Gaussian Process Latent Variable Model (GP-LVM)

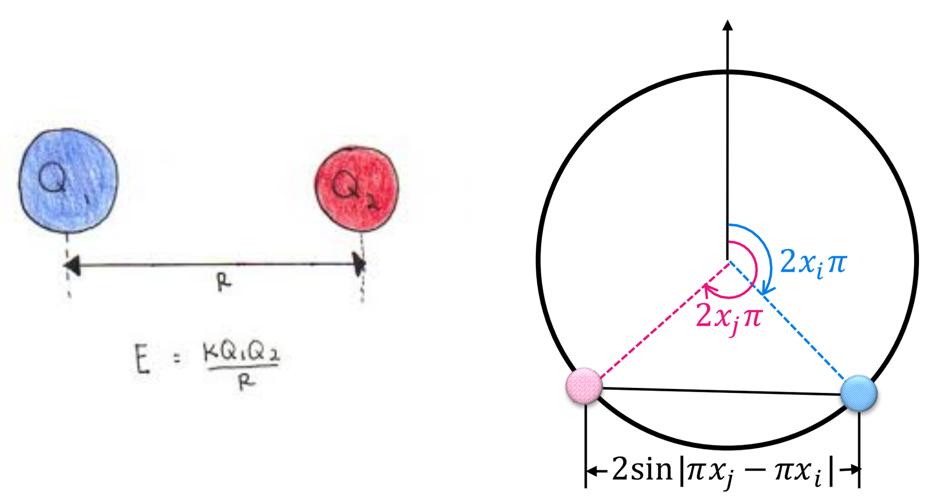
- Preliminaries
  - $\boldsymbol{Y} = \{y_{ij}\} \in R^{n\times d}$: row i represents the ith d-dimensional data vector.
  - $\boldsymbol{x}_i \in R^p$: the latent coordinates corresponding to the ith data vector.
  - $\mu_j(\cdot): R^p \mapsto R$: the continuous mapping from $\boldsymbol{x}_i$ to y_{ij}.
  - $GP(0, K^j)$: a Gaussian process with the covariance function being $K^j(\cdot,\cdot)$.
- GP-LVM and Bayesian GP-LVM [2]

$$y_{ij} = \mu_j(\boldsymbol{x}_i) + \epsilon_{ij}$$
$$\boldsymbol{x}_i \sim N(0, \boldsymbol{I})$$
$$\mu_j \sim GP(0, K^j)$$
$$\epsilon_{ij} \sim N(0, \sigma_j^2)$$

- Advantages
  - A probabilistic generative model.
  - A non-linear generalization of probabilistic PCA.
  - Bayesian GP-LVM automatically learns p.
- Identifiability Issues
  - Rotation and translation.
  - *preserving dissimilarity* vs. *preserving similarity* [3,4].



Coulomb Repulsion



Definition A univariate process is a Coulomb repulsive process (Corp) if and only if for every finite set of indices $t_1, \dots, t_k$ in the index set N_+,

$$x_{t_1} \sim Unif(0,1),$$
$$p(x_{t_i} | x_{t_1}, \dots, x_{t_{i-1}}) \propto \Pi_{j=1}^{i-1} sin^{2r}(\pi x_{t_i} - \pi x_{t_j}) 1_{x_{t_i} \in [0,1]},$$

where $r > 0$ is the repulsive parameter. The process is denoted as $x_t \sim Corp(r)$.

Lemma 1 Corp is well defined based on Kolmogorov extension theorem since

- The finite-dimensional distributions are well defined.
- The finite-dimensional distributions are exchangeable.
- For any finite set of indices $t_1, \dots, t_k, \dots, t_{k+m}$,
  $p(x_{t_1}, \dots, x_{t_k}, \dots, x_{t_{k+m}}) = p(x_{t_1}, \dots, x_{t_k}) \Pi_{j=1}^{m} p(x_{t_{k+j}} | x_{t_1}, \dots, x_{t_{k+j-1}})$

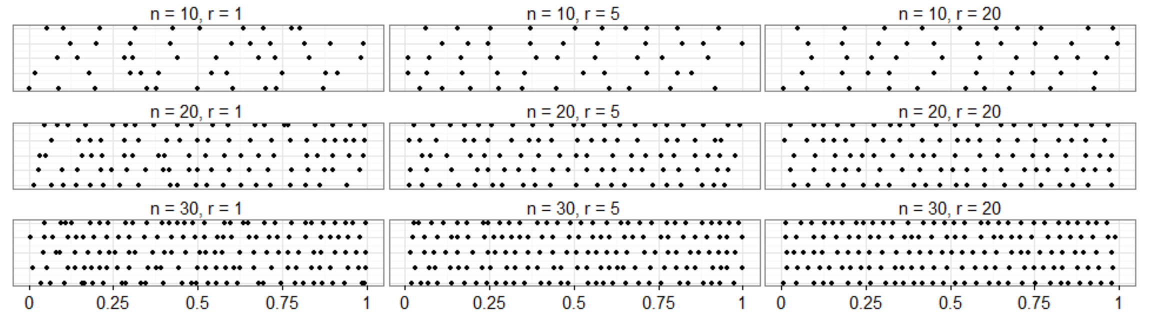
Lemma 2 For any $n \in \mathbb{N}_+$, any $1 \le i \le n$ and any $\epsilon > 0$, we have

$$p(X_n \in \mathcal{B}(X_i, \epsilon) | X_1, \dots, X_{n-1}) < \frac{2\pi^2 \epsilon^{2r+1}}{2r+1},$$

where $\mathcal{B}(X_i, \epsilon) = \{X \in [0,1] : d(X, X_i) < \epsilon\}$ and $d(x,y) = \sin|\pi x - \pi y|$.

Lemma 3 For any $n \in \mathbb{N}_+$, the joint p.d.f. of $X_1, \dots, X_n$ (due to the exchangeability, we can assume $X_1 < X_2 < \dots < X_n$ without loss of generality) is maximized when and only when

$$d(X_i, X_{i-1}) = \sin\left(\frac{1}{n+1}\right) for\ all\ 2 \le i \le n.$$



Electrostatic Gaussian Process (electroGP)

$$y_{ij} = \mu_j(x_i) + \epsilon_{ij}$$
$$x_i \sim Corp(r)$$
$$\mu_j \sim GP(0, K^j)$$
$$\epsilon_{ij} \sim N(0, \sigma_j^2)$$

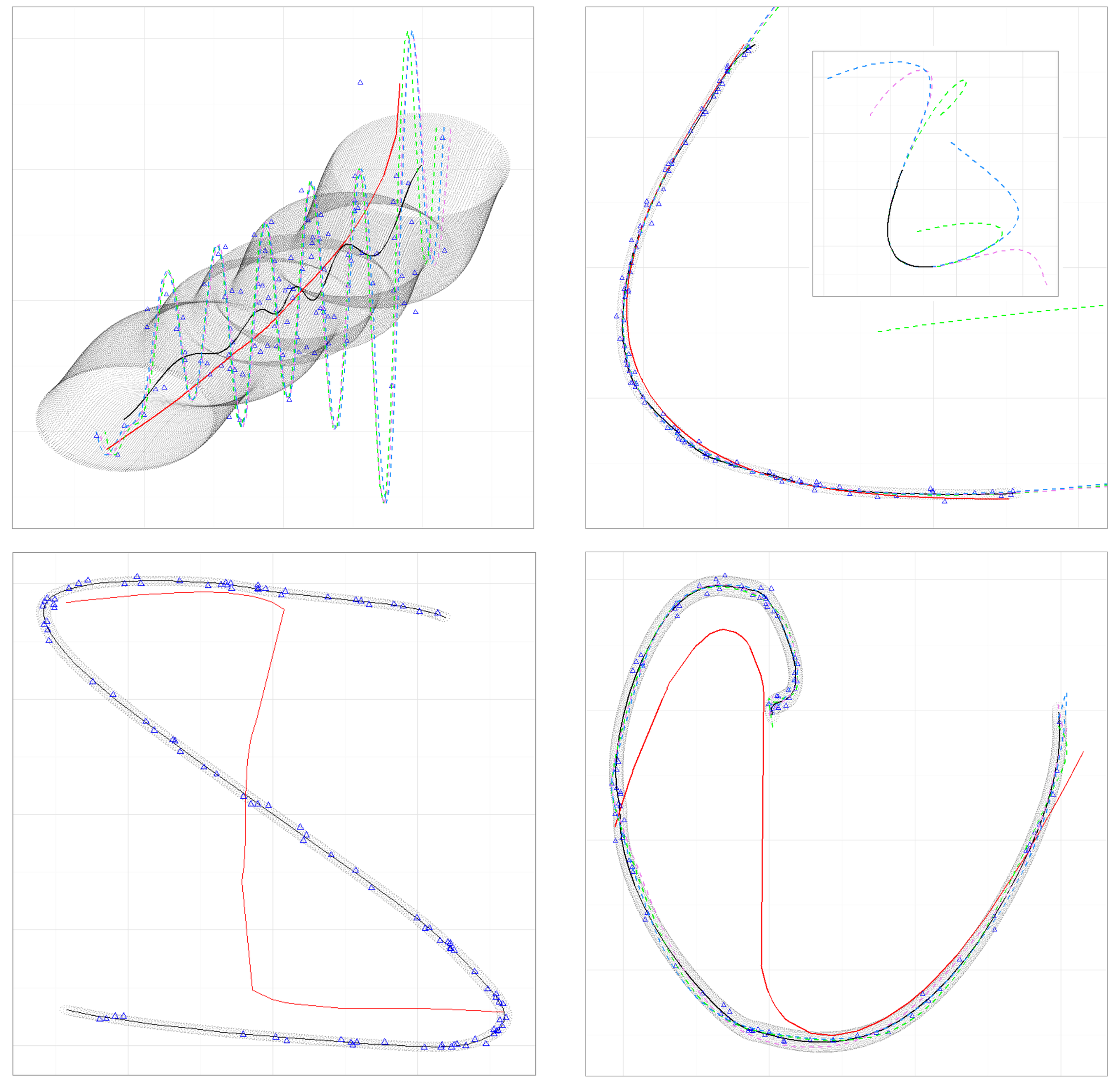
Fitting electroGP

- Preparations
  - We use the radial basis function (RBF) kernel, $K^j(x,y) = \phi_j e^{-\alpha_j (x-y)^2}$.
  - Model hyperparameters $\Theta = (\sigma_1^2, \phi_1, \alpha_1, \dots, \sigma_d^2, \phi_d, \alpha_d)$.
- From a optimization viewpoint

$$(\widehat{\boldsymbol{x}}, \widehat{\Theta}) = \underset{\boldsymbol{x},\Theta}{\operatorname{argmax}}\ \ell(\boldsymbol{y}_{1:n}) + \log[\pi(\boldsymbol{x})]$$

- ***Self-truncation***: It can be easily checked that $\log[\pi(x_i = x_j)] = -\infty$, for any i and j, meaning that these electrostatic charges cannot get past each other! We refer to this as the *self-truncation property*.
- Algorithm
  - Step 1: Learn the one dimensional coordinate $\boldsymbol{x}_0$ by your favorite distance-preserving manifold learning algorithm and rescale $\boldsymbol{x}_0$ into $(0,1)$;
  - Step 2: Solve $\Theta_0 = \underset{\Theta}{\operatorname{argmax}}\, \ell(\boldsymbol{y}_{1:n} | \boldsymbol{x}_0)$ using scaled conjugate gradient descent (SCG);
  - Step 3: Using SCG, setting $\boldsymbol{x}_0$ and Θ_0 to be the initial values, solve $\widehat{\boldsymbol{x}}$ and $\widehat{\Theta}$ w.r.t. $(\widehat{\boldsymbol{x}}, \widehat{\Theta}) = \underset{\boldsymbol{x},\Theta}{\operatorname{argmax}}\, \ell(\boldsymbol{y}_{1:n}) + \log[\pi(\boldsymbol{x})]$.

Simulation

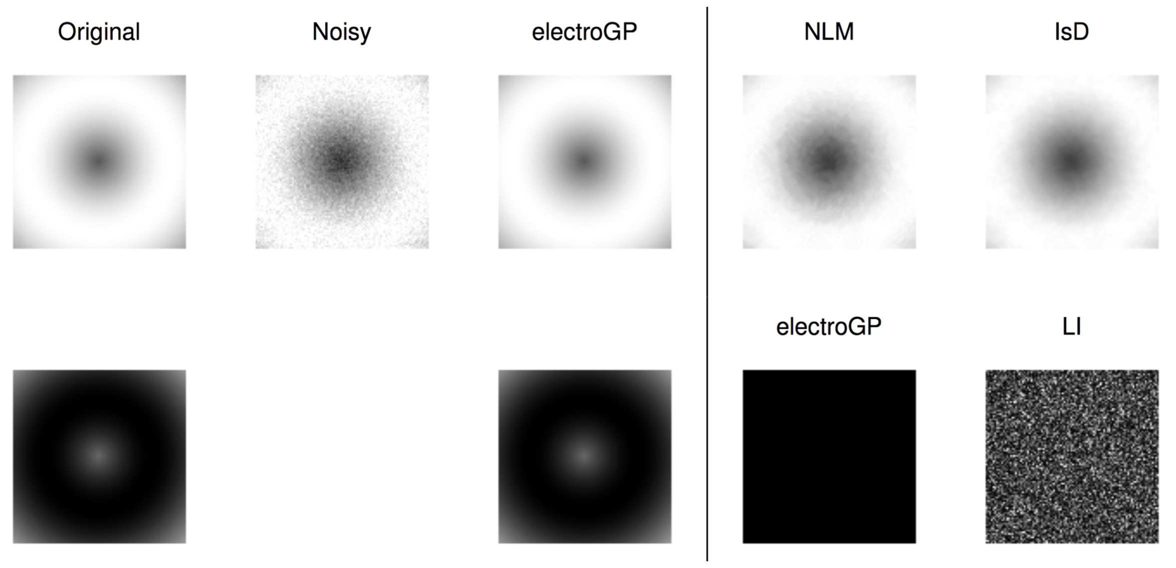


electroGP (black), principal curve [5] (red) and GP-LVM (color)

Video Denoising & Super-resolution

100 consecutive frames (of size 100×100 with gray color) were collected from a video of a shrinking shockwave. Frame 51 to 55 were assumed completely missing and the other 95 frames were observed with the original time order with strong white noises. The shockwave is homogeneous in all directions from the center; hence, the frames roughly lie on a curve. The electroGP was applied for two tasks: 1. Frame denoising; 2. Improving resolution by interpolating frames in between the existing frames.

Result

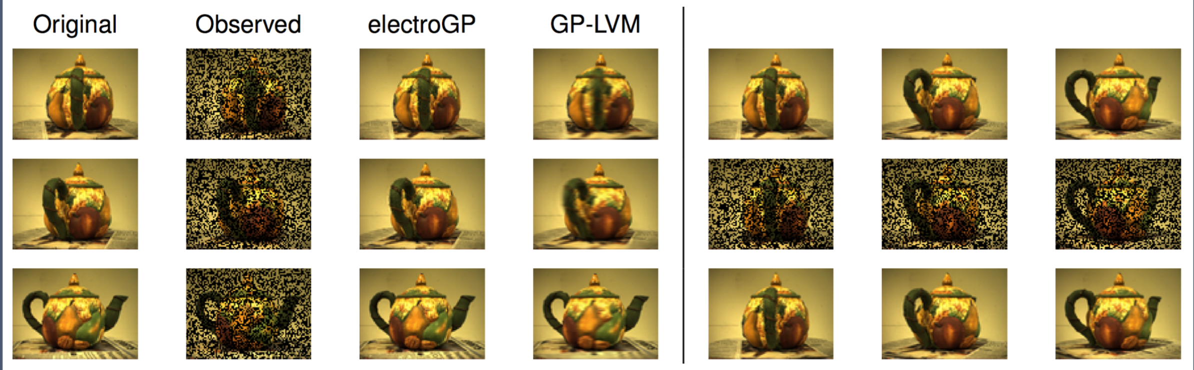


Non-local mean filter (NLM) [6] and isotropic diffusion (IsD) [7]

Video Inpainting

200 consecutive frames (of size 76×101 with RGB color) [7] were collected from a video of a teapot rotating 1800. Clearly these images roughly lie on a curve. 190 of the frames were assumed to be fully observed in the natural time order of the video, while the other 10 frames were given without any ordering information. Moreover, half of the pixels of these 10 frames were missing. The electroGP was fitted based on the other 190 frames and was used to reconstruct the broken frames and impute the reconstructed frames into the whole frame series with the correct order.

Result



Discussion

Manifold learning has dramatic importance in many applications where high-dimensional data are collected with unknown low dimensional manifold structure. While most of the methods focus on finding lower dimensional summaries or characterizing the joint distribution of the data, there is (to our knowledge) no reliable method for probabilistic learning of the manifold. This turns out to be a daunting problem due to major issues with identifiability leading to unstable and generally poor performance for current probabilistic non-linear dimensionality reduction methods. It is not obvious how to incorporate appropriate geometric constraints to ensure identifiability of the manifold without also enforcing overly-restrictive assumptions about its form.

We tackled this problem in the one-dimensional manifold (curve) case and built a novel electro- static Gaussian process model based on the general framework of GP-LVM by introducing a novel Coulomb repulsive process.

Future Work

- Extending to multi-dimensional manifold.
- A more scalabel algorithm.
- Mixutre of electroGP to account for potential multi-modality.

References

1. Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research*, 6, 1783-1816.
2. Titsias, M. K., & Lawrence, N. D. (2010). Bayesian Gaussian process latent variable model. In *International Conference on Artificial Intelligence and Statistics*, 844-851.
3. Lawrence, N. D., & Quiñonero-Candela, J. (2006, June). Local distance preservation in the GP-LVM through back constraints. In *Proceedings of the 23rd international conference on Machine learning*, 513-520.
4. Urtasun, R., Fleet, D. J., Geiger, A., Popović, J., Darrell, T. J., & Lawrence, N. D. (2008, July). Topologically-constrained latent variable models. In *Proceedings of the 25th international conference on Machine learning*, 1080-1087.
5. Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406), 502-516.
6. Buades, A., Coll, B., & Morel, J. M. (2005, June). A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2, 60-65
7. Perona, P., & Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *12*(7), 629-639.
8. Weinberger, K. Q., & Saul, L. K. (2006, July). An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*, 6 1683-1686.

Contact

Ye Wang
Department of Statistics
Duke University
Durham, NC, USA, 27705
eric.ye.wang@duke.edu

David Dunson
Department of Statistics
Duke University
Durham, NC, USA, 27705
dunson@stat.duke.edu