School of Mathematical and Statistical Sciences
Arizona State University

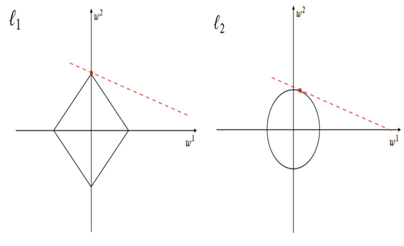
Bayesian Regularization of Latent Representation [a]

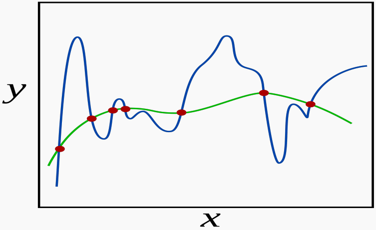
Chukwudi Paul Obite — cobite@asu.edu
Zhi Chang — zchang7@asu.edu
Keyan Wu — keyanwu@asu.edu
Shiwei Lan* — slan@asu.edu

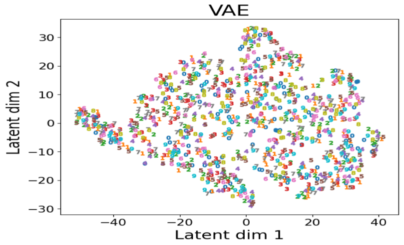
APR 24-28, 2025 ICLR @ EXPO Singapore

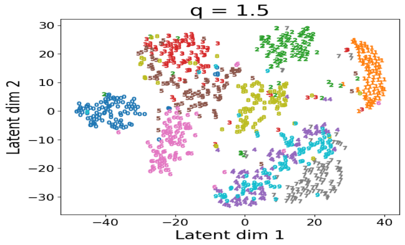
[a] https://openreview.net/pdf?id=VOoJEQlLW5

Regularization









Regularization on Latent Representations

literature review

- Principal component analysis (PCA) [7, 2, 3] has long been used as a classic technique in factor analysis.
- Probabilistic PCA [8, 6] is a generalization that has PCs as maximum likelihood estimation (MLE).
- Gaussian process latent variable model (GP-LVM) [4, 5, 10] considers the dual problem of probabilistic PCA and replaces the linear mapping with nonlinear Gaussian mapping.
- We propose a novel LVM based on *q-exponential process (QEP-LVM)* which can
  1. flexibly regularize the learned latent representations;
  2. effectively identify the intrinsic latent dimensionality;
  3. learn latent representations with better interpretability and adaptability for supervised/semi-supervised learning.

Q-Exponential Process (Q-EP)

flexible regularization on function spaces

Definition

A multivariate q-exponential distribution, denoted as $\mathrm{q-ED}_d(\boldsymbol{\mu}, \mathbf{C})$*, has the following density*

$$p(\mathbf{u}|\boldsymbol{\mu}, \mathbf{C}, q) = \frac{q}{2}(2\pi)^{-\frac{d}{2}}|\mathbf{C}|^{-\frac{1}{2}} \boxed{r^{(\frac{q}{2}-1)\frac{d}{2}}} \exp\left\{-\frac{r^{\frac{q}{2}}}{2}\right\}, \tag{1}$$

$$r(\mathbf{u}) = (\mathbf{u}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{u}-\boldsymbol{\mu})$$

- If $\mathbf{u} \sim \mathrm{q-ED}_d(0, \mathbf{C})$, then we denote $\mathbf{u}^* \sim \mathrm{q-ED}_d^*(0, \mathbf{C})$ following a *scaled* q-exponential distribution.

Definition (Q-EP)

A (centered) q-exponential process $u(x)$ *with kernel* $\mathcal{C}$, $\mathrm{q-}\mathcal{EP}(0, \mathcal{C})$*, is a collection of random variables such that any finite set,* $\mathbf{u} = (u(x_1), \cdots u(x_d))$*, follows a scaled multivariate q-exponential distribution, i.e.* $\mathbf{u} \sim \mathrm{q-ED}_d^*(0, \mathbf{C})$*.*

Q-EP Latent Variable Model

latent representation $\mathbf{X}$

- Denote data by $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_D]_{N\times D}$, latent variable by $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_Q]_{N\times Q}$, and latent function as $\mathbf{F} = [f_1(\mathbf{X}), \cdots, f_D(\mathbf{X})]_{N\times D}$.

$$\begin{aligned} \text{likelihood}: &\quad \text{vec}(\mathbf{Y})|\mathbf{F} \sim \text{q}-\text{ED}_{ND}(\text{vec}(\mathbf{F}), \mathbf{I}_D \otimes \Sigma), \\ \text{prior on latent function}: &\quad f \sim \text{q}-\mathcal{EP}(0, \mathcal{C}, \mathbf{I}_D). \end{aligned} \tag{2}$$

- Marginalizing f yields a stochastic mapping from $\mathbf{X}$ to $\mathbf{Y}$:

$$\text{marginal likelihood}: \quad \text{vec}(\mathbf{Y})|\mathbf{X} \sim \text{q}-\text{ED}_{ND}(\mathbf{0}, \mathbf{I}_D \otimes \mathbf{K}), \quad \mathbf{K} = \mathbf{C}_\mathbf{X} + \Sigma. \tag{3}$$

Theorem

Suppose $\mathbf{Y}\mathbf{Y}^\mathsf{T}/D$ has eigen-decomposition $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T}$ with $\boldsymbol{\Lambda}$ being the diagonal matrix with eigenvalues $\{\lambda_i\}_{i=1}^N$. Then the MLE for (3) is

$$\mathbf{X}^* = \mathbf{U}_Q\mathbf{L}\mathbf{V}, \ \mathbf{L} = \text{diag}(\{\sqrt{\alpha(c\lambda_i - \beta^{-1})}\}_{i=1}^Q), \ c(q) = D^{1-\frac{2}{q}}(D \wedge Q)\left[\frac{q}{2(D\wedge Q) + (q-2)N}\right]^{\frac{2}{q}},$$

where $\mathbf{U}_Q$ is an $N \times Q$ matrix with the first Q eigen-vectors in $\mathbf{U}$, $\mathbf{V}$ is an arbitrary $Q \times Q$ orthogonal matrix, and $a \wedge b := \min\{a, b\}$.

Bayesian QEP-LVM

latent representation $\mathbf{X}$

- The Bayesian QEP-LVM has a Q-EP prior on latent variable $\mathbf{X}$

$$\begin{aligned} \text{marginal likelihood :} \quad & \mathrm{vec}(\mathbf{Y})|\mathbf{X} \sim \mathrm{q-ED}(\mathbf{0}, \mathbf{I}_D \otimes \mathbf{K}), \\ \text{prior on latent variable :} \quad & \mathrm{vec}(\mathbf{X}) \sim \mathrm{q-ED}(\mathbf{0}, \mathbf{I}_{NQ}). \end{aligned}$$

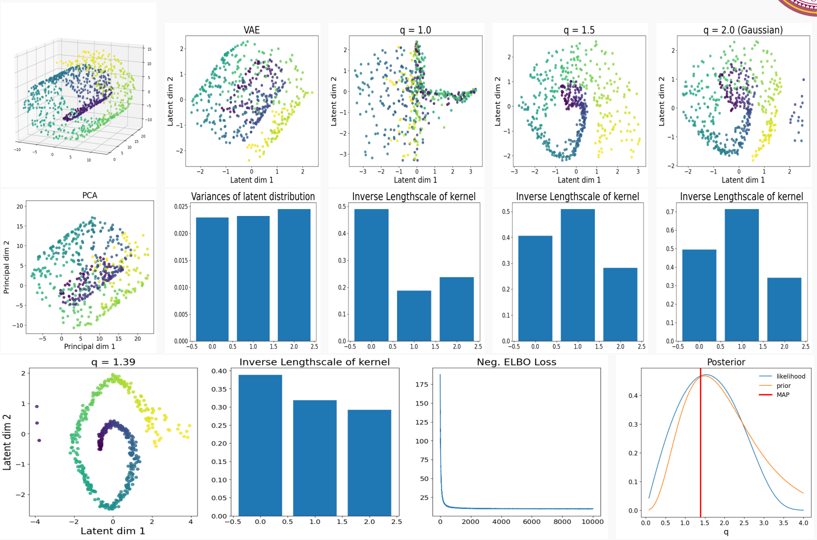
- And we use variational Bayes to approximate the posterior $\mathbf{X}|\mathbf{Y}$ with

$$\text{variational distribution for latent variable :} \quad q(\mathbf{X}) \sim \mathrm{q-ED}(\boldsymbol{\mu}, \mathrm{diag}(\{\mathbf{S}_n\})).$$

- Sparse variational approximation by inducing points [9, 1] is adopted to derive the evidence lower bound (ELBO).
- In addition to variational parameters, we jointly optimize the regularization parameter $q > 0$.
- Generative models can be built based on QEP-LVM to predict labels.
- Refer to `https://openreview.net/pdf?id=VOoJEQlLW5` for more details.

Swiss Roll

2 intrinsic dimensions



MNIST

more interpretable latent representation

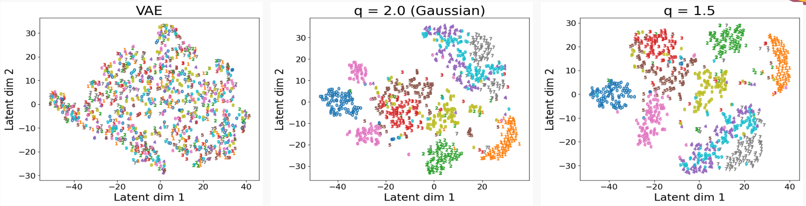


Figure: Latent representations of MNIST database by VAE (left), GP-LVM (middle), QEP-LVM with $q = 1.5$ (right). For the convenience of visualization, 10-dimensional latent spaces learned by these algorithms are projected to 2-d subspace by t-SNE respectively.

Table: Classification on handwritten digits (MNIST) based on learned latent representation: accuracy (ACC), area under ROC curve (AUC), adjusted rand index (ARI) score, normalized mutual information (NMI) score, log predictive probability (LPP) values, and running time (per class) by various Bayesian QEP-LVMs. Results in each cell are averaged over 10 experiments with different random seeds; values after $\pm$ are standard errors of these repeated experiments.

Model (q)	ACC	AUC	ARI	NMI	LPP	time/class
1.5	**0.973** $\pm$ 0.012	**0.986** $\pm$ 0.0065	**0.943** $\pm$ 0.024	**0.952** $\pm$ 0.020	**-26453.5** $\pm$ 481.9	129.91 $\pm$ 4.60
2.0 (Gaussian)	0.965 $\pm$ 0.012	0.981 $\pm$ 0.0062	0.923 $\pm$ 0.025	0.938 $\pm$ 0.019	-279414.9 $\pm$ 5184.6	127.18 $\pm$ 0.29

Conclusion

- We propose a novel Bayesian LVM based on q-exponential process (Q-EP) as a generalization of GP-LVM to regularize the learning of latent representations via a parameter $q > 0$.
- Smaller q tends to contract the latent space and leads to a more compact latent representation.
- Compared with GP-LVM, QEP-LVM learns more interpretable latent representations with effective determination of intrinsic dimensions.
- In future, we will extend this work to deep probabilistic models.

https://github.com/lanzithinking/Reg_Rep

References I

[1] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA, 09–12 May 2015. PMLR.

[2] I. T. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics. Springer-Verlag, 1986.

[3] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, April 2016.

[4] Neil Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

[5] Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(60):1783–1816, 2005.

[6] Thomas Minka. Automatic choice of dimensionality for pca. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[7] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, November 1901.

[8] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 09 1999.

[9] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[10] Michalis Titsias and Neil D. Lawrence. Bayesian gaussian process latent variable model. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Thank you !

https://math.la.asu.edu/∼slan