School of Mathematical and Statistical Sciences
Arizona State University

Bayesian Learning via Q-Exponential Process [a]

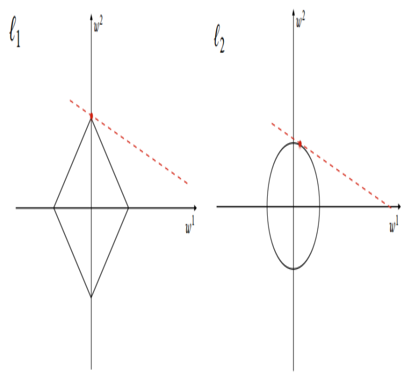
Shuyi Li, Michael O'Connor Shiwei Lan*

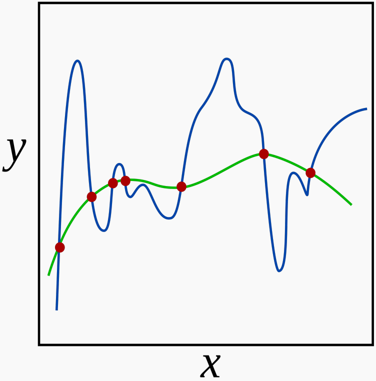
shuyili3@asu.edu mfoconn1@asu.edu slan@asu.edu

DEC 10-16, 2023 NIPS @ New Orleans

[a]Shuyi Li, Michael OConnor and Shiwei Lan*, https://arxiv.org/abs/2210.07987

Regularization





Regularization on Function Spaces

- Regularization is one of the most fundamental topics in optimization, statistics and machine learning.
- To get sparsity in estimating a parameter $u \in \mathbb{R}^d$, an ℓ_q penalty term, $\|u\|_q$, is usually added to the objective function.
- What is the probabilistic distribution corresponding to such ℓ_q penalty?
- What is the *correct* stochastic process corresponding to $\|u\|_q$ when we model functions $u \in L^q$?
- This is important for statistically modeling high-dimensional objects such as images, with penalty to preserve certain properties, e.g. edges in the image.

Regularization on Function Spaces

literature review

- Gaussian process (GP) can be viewed as L_2 regularization on function spaces, sometimes over-smooth [23, 14].
- L_1 penalty based priors include Laplace random field [22, 20, 18] and Besov process [19, 10, 15, 11].
- Student-t process (TP) [26] and elliptical process [1] with heavy tail are proposed as alternatives to GP.
- We propose the *q-exponential process (Q-EP)* based on q-exponential distribution with density proportional to $\exp\left(-\frac{1}{2}|u|^q\right)$.

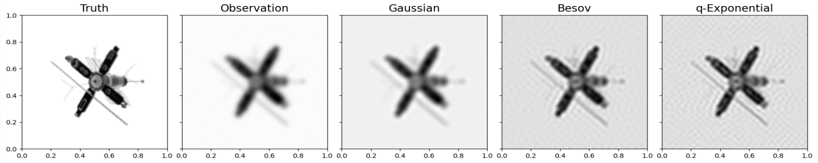


Figure: Image of satellite: true image, blurred observation, and reconstructions by GP, Besov and Q-EP models with relative errors 75.19%, 21.94% and 20.35% respectively.

Besov Process and Q-exponential Distribution

- Besov process [19, 10] is proposed to impose L_1 regularization as an “edge-preserving" prior for images:

$$u(x) = \sum_{\ell=1}^{\infty} \gamma_\ell u_\ell \phi_\ell(x), \quad u_\ell \overset{iid}{\sim} \pi_q(\cdot) \propto \exp\left(-\frac{1}{2}|\cdot|^q\right) \tag{1}$$

- How can we generalize it to a multivariate distribution and further to a stochastic process?
- By the Kolmogorov’ extension theorem [21], one should require
  1. **exchangeability** of the joint distribution, i.e. $p(\boldsymbol{\xi}_{1:J}) = p(\boldsymbol{\xi}_{\tau(1:J)})$ for any finite permutation τ;
  2. **consistency** of marginalization, i.e. $p(\boldsymbol{\xi}_1) = \int p(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) d\boldsymbol{\xi}_2$.
- Gomez [13] provided one possibility of a multivariate EP distribution, denoted as $\mathrm{EP}_d(\boldsymbol{\mu}, \mathbf{C}, q)$, with the following density:

$$p(\mathbf{u}|\boldsymbol{\mu}, \mathbf{C}, q) = \frac{q\Gamma\left(\frac{d}{2}\right)}{2\Gamma\left(\frac{d}{q}\right)} 2^{-\frac{d}{q}} \pi^{-\frac{d}{2}} |\mathbf{C}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(\mathbf{u}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{u}-\boldsymbol{\mu})\right]^{\frac{q}{2}}\right\} \tag{2}$$

Generalization of Q-exponential Distribution

marginalization consistency

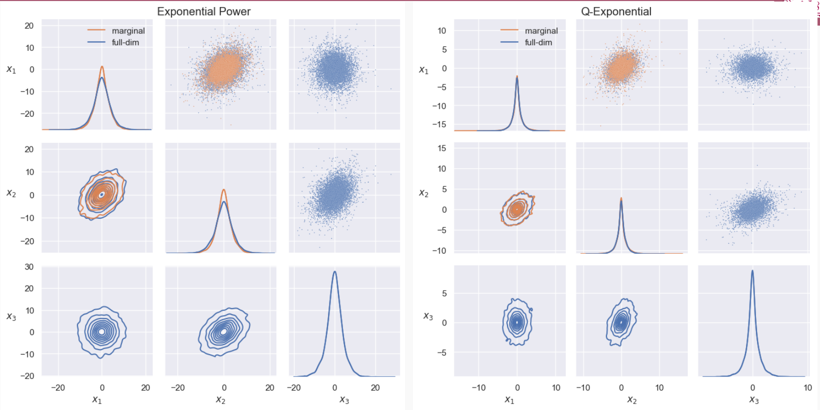


Figure: Inconsistent (Gomez's) EP distribution $\mathrm{EP}_d(\boldsymbol{\mu}, \mathbf{C}, q)$ (left) vs. consistent Q-exponential distribution $\mathrm{q-ED}_d(\boldsymbol{\mu}, \mathbf{C})$ (right). Both can be sampled using (??) with $R^q \sim \Gamma(\alpha = \frac{d}{q}, \beta = \frac{1}{2})$ and $R^q \sim \Gamma\left(\alpha = \frac{d}{2}, \beta = \frac{1}{2}\right)$ respectively. Note there is significant discrepancy between the marginalization of $\mathrm{EP}_3(\boldsymbol{\mu}, \mathbf{C}, q)$ and $\mathrm{EP}_2(\boldsymbol{\mu}, \mathbf{C}, q)$. However, the marginalization of $\mathrm{q-ED}_3(\boldsymbol{\mu}, \mathbf{C})$ coincides with $\mathrm{q-ED}_2(\boldsymbol{\mu}, \mathbf{C})$. Empirical densities are estimated based on 10000 samples (shown as dots).

Definition

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

$$p(\mathbf{u}|\boldsymbol{\mu}, \mathbf{C}, q) = \frac{q}{2}(2\pi)^{-\frac{d}{2}}|\mathbf{C}|^{-\frac{1}{2}} \boxed{r^{(\frac{q}{2}-1)\frac{d}{2}}} \exp\left\{-\frac{r^{\frac{q}{2}}}{2}\right\}, \quad r(\mathbf{u}) = (\mathbf{u}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{u}-\boldsymbol{\mu}) \tag{3}$$

- 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.* $\mathbf{u} \sim \mathrm{q-ED}_d^*(0, \mathbf{C})$*.*

Connection to Besov Process

direct control on the correlation structure through $\mathcal{C}$

- Q-EP and Besov share equivalent series representations.

Theorem (Karhunen-Loéve)

If $u(x) \sim \text{q}-\mathcal{EP}(0, \mathcal{C})$ with $\mathcal{C}$ having eigen-pairs $\{\lambda_\ell, \phi_\ell(x)\}_{\ell=1}^\infty$ such that $\mathcal{C}\phi_\ell(x) = \phi_\ell(x)\lambda_\ell$, $\|\phi_\ell\|_2 = 1$ for all $\ell \in \mathbb{N}$ and $\sum_{\ell=1}^\infty \lambda_\ell < \infty$, then we have the following series representation for $u(x)$:

$$u(x) = \sum_{\ell=1}^{\infty} u_\ell \phi_\ell(x), \quad u_\ell := \int_D u(x)\phi_\ell(x) \overset{ind}{\sim} \text{q}-\text{ED}^*(0, \lambda_\ell) \tag{4}$$

where $\text{E}[u_\ell] = 0$ and $\text{Cov}(u_\ell, u_{\ell'}) = \lambda_\ell \delta_{\ell\ell'}$ with Dirac function $\delta_{\ell\ell'} = 1$ if $\ell = \ell'$ and 0 otherwise.

- If we factor $\sqrt{\lambda_\ell}$ out of u_ℓ, we have the following expansion for Q-EP more comparable to (1) for Besov:

$$u(x) = \sum_{\ell=1}^{\infty} \sqrt{\lambda_\ell} u_\ell \phi_\ell(x), \quad u_\ell \overset{iid}{\sim} \text{q}-\text{ED}(0, 1) \propto \pi_q(\cdot) \tag{5}$$

- Let $L(\cdot; 0, \Sigma)$ be the likelihood model, and μ_0 be the prior.

$$\begin{aligned} y &= u(x) + \varepsilon, \quad \varepsilon \sim L(\cdot; 0, \Sigma) \\ u &\sim \mu_0(du) \end{aligned} \tag{6}$$

- **Conjugate case**: $\mu_0 = \text{q}-\mathcal{EP}(0, \mathcal{C})$ and $L(\cdot; 0, C) = \text{q}-\text{ED}(\mathbf{0}, \mathbf{C})$

Theorem (Posterior Prediction)

Given covariates $\mathbf{x} = \{x_i\}_{i=1}^N$ and observations $\mathbf{y} = \{y_i\}_{i=1}^N$ following $\text{q}-\text{ED}$ in the model (6) with $\text{q}-\mathcal{EP}$ prior for the same $q > 0$, we have the following posterior predictive distribution for $u(x_*)$ at (a) new point(s) x_*:

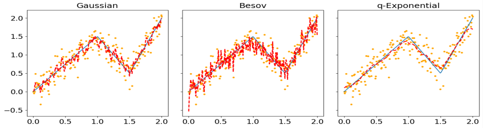
$$u(x_*)|\mathbf{y}, \mathbf{x}, x_* \sim \text{q}-\text{ED}(\boldsymbol{\mu}^*, \mathbf{C}^*), \ \boldsymbol{\mu}^* = \mathbf{C}_*^{\mathrm{T}}(\mathbf{C} + \Sigma)^{-1}\mathbf{y}, \ \mathbf{C}^* = \mathbf{C}_{**} - \mathbf{C}_*^{\mathrm{T}}(\mathbf{C} + \Sigma)^{-1}\mathbf{C}_* \tag{7}$$

where $\mathbf{C} = \mathcal{C}(\mathbf{x}, \mathbf{x})$, $\mathbf{C}_* = \mathcal{C}(\mathbf{x}, x_*)$, and $\mathbf{C}_{**} = \mathcal{C}(x_*, x_*)$.

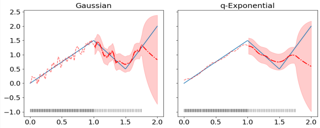
- **Non-conjugate case**: posterior sampling by dimension-independent MCMC algorithms [9, 6, 3, 4, 5] with the pushforward $\mu_0 = T^{\#}\nu_0$:

$$\mathbf{u} = T(\mathbf{z}) = \mathbf{L}\mathbf{z}\|\mathbf{z}\|^{\frac{2}{q}-1}, \quad \mathbf{z} = T^{-1}(\mathbf{u}) = \mathbf{L}^{-1}\mathbf{u}\|\mathbf{L}^{-1}\mathbf{u}\|^{\frac{q}{2}-1}, \quad \mathbf{z} \sim \nu_0 \tag{8}$$

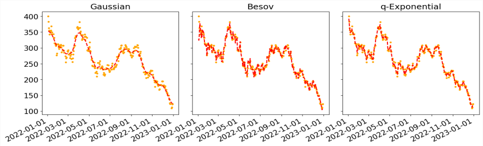
Time Series Modeling
modeling jumps or turnings



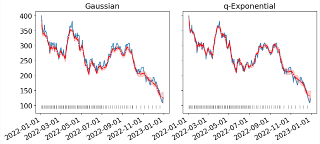
(a) Time series with sharp turnings (model fitting).



(b) Time series with turnings (prediction).



(c) Tesla stock prices in 2022 (model fitting).



(d) Tesla stock prices in 2022 (prediction).

Figure: (a)(c) MAP estimates by GP (left), Besov (middle) and Q-EP (right) models. (b)(d) Predictions by GP (left) and Q-EP (right) models. Orange dots are actual realizations (data points). Blue solid lines are true trajectories. Black ticks indicate the training data points. Red dashed lines are MAP estimates. Red dot-dashed lines are predictions with shaded region being credible bands.

Computed Tomography Imaging
preserving the edges

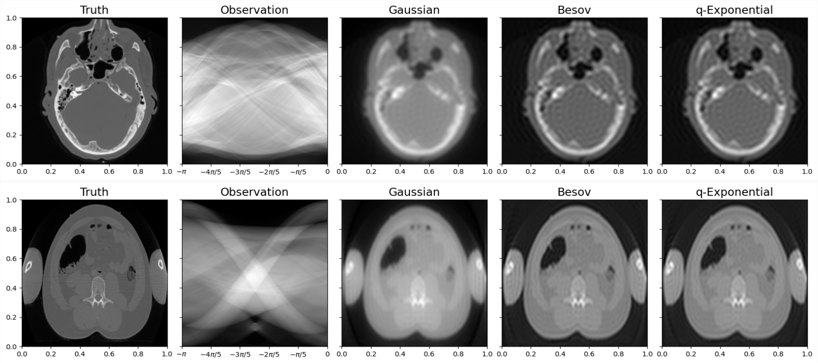


Figure: CT of human head (upper) and torso (lower): true image, observation (sinogram), and MAP estimates by GP, Besov and Q-EP models with relative errors 29.99%, 22.41% and 22.24% (for head) and 26.11%, 21.77% and 21.53% (for torso) respectively.

Computed Tomography Imaging

preserving the edges

Table: Posterior estimates of Shepp–Logan phantom by GP, Besov and Q-EP prior models: relative error, RLE $:= \|\hat{u} - u^{\dagger}\|/\|u^{\dagger}\|$, of MAP ($\hat{u} = u^{*}$) and posterior mean ($\hat{u} = \overline{u}$) respectively, log-likelihood (LL), peak signal-to-noise ratio (PSNR) [12], structured similarity index (SSIM) [28], Haar wavelet-based perceptual similarity index (HaarPSI) [24]. Numbers in the bracket are standard deviations obtained repeating the experiments for 10 times with different random seeds.

	MAP			Posterior Mean		
	GP	Besov	Q-EP	GP	Besov	Q-EP
RLE	0.6810	0.7027	**0.4087**	0.4917(6.16e-7)	0.4894(3.53e-5)	**0.4890**(4.79e-5)
LL	-1.55e+6	-1.54e+6	-1.57e+5	-5.21e+5(8.47)	-4.80e+5(196.34)	-4.56e+5(307.97)
PSNR	15.5531	15.2806	**19.9887**	18.3826(1.09e-5)	18.4226(6.27e-4)	**18.4303**(8.51e-4)
SSIM	0.4028	0.3703	**0.5967**	**0.5561**(3.92e-7)	0.5535(2.38e-4)	0.5403(5.26e-4)
HaarPSI	0.0961	0.0870	**0.3105**	0.3126(1.52e-8)	**0.3126**(3.36e-4)	0.3122(3.06e-4)

Conclusion

- In this work, we propose the q-exponential process (Q-EP) as a prior on L^q functions with a flexible parameter $q > 0$ to control the degree of regularization.
- Usually, $q = 1$ is adopted to capture abrupt changes or sharp contrast in data such as edges in the image.
- Compared with GP, Q-EP can impose sharper regularization through q.
- Compared with Besov, Q-EP enjoys the explicit formula with more control on the correlation structure as GP.
- In future, we will extend this work to spatiotemporal domain to model dynamically changing images.

github.com/lanzithinking/Q-EXP

References I

[1] Maria Bånkestad, Jens Sjölund, Jalil Taghia, and Thomas Schön. The elliptical processes: a family of fat-tailed stochastic processes. 03 2020.

[2] Johnathan M. Bardsley. Applications of a nonnegatively constrained iterative method with statistically based stopping rules to ct, pet, and spect imaging. *Electron. Trans. Numer. Anal.*, 38:34–43, 2011.

[3] A. Beskos, F. J. Pinski, J. M. Sanz-Serna, and A. M. Stuart. Hybrid Monte-Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121:2201–2230, 2011.

[4] Alexandros Beskos. A stable manifold MCMC method for high dimensions. *Statistics & Probability Letters*, 90:46–52, 2014.

[5] Alexandros Beskos, Mark Girolami, Shiwei Lan, Patrick E. Farrell, and Andrew M. Stuart. Geometric MCMC for infinite-dimensional inverse problems. *Journal of Computational Physics*, 335(Supplement C):327 – 351, 2017.

[6] Alexandros Beskos, Gareth Roberts, Andrew Stuart, and Jochen Voss. MCMC methods for diffusion bridges. *Stochastics and Dynamics*, 8(03):319–350, 2008.

[7] Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

[8] Victor Chen, Matthew M. Dunlop, Omiros Papaspiliopoulos, and Andrew M. Stuart. Dimension-robust mcmc in bayesian inverse problems. 03 2018.

[9] Simon L Cotter, Gareth O Roberts, AM Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.

[10] Masoumeh Dashti, Stephen Harris, and Andrew Stuart. Besov priors for bayesian inverse problems. *Inverse Problems and Imaging*, 6(2):183–200, may 2012.

References II

[11] Masoumeh Dashti and Andrew M. Stuart. *The Bayesian Approach to Inverse Problems*, pages 311–428. Springer International Publishing, Cham, 2017.

[12] Osama S. Faragallah, Heba El-Hoseny, Walid El-Shafai, Wael Abd El-Rahman, Hala S. El-Sayed, El-Sayed M. El-Rabaie, Fathi E. Abd El-Samie, and Gamal G. N. Geweid. A comprehensive survey analysis for present solutions of medical image fusion and future directions. *IEEE Access*, 9:11358–11371, 2021.

[13] E. Gómez, M.A. Gomez-Viilegas, and J.M. Marín. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics - Theory and Methods*, 27(3):589–600, jan 1998.

[14] A. P. Dawid J. M. Bernardo, J. O. Berger and A. F. M. Smith. Regression and classification using gaussian process priors. *Bayesian Statistics*, 6:475–501, 1998.

[15] Junxiong Jia, Jigen Peng, and Jinghuai Gao. Bayesian approach to inverse problems for functions with a variable-index besov prior. *Inverse Problems*, 32(8):085006, 2016.

[16] Mark E. Johnson. *Multivariate Statistical Simulation*, chapter 6 Elliptically Contoured Distributions, pages 106–124. Probability and Statistics. John Wiley & Sons, Ltd, 1987.

[17] Y. Kano. Consistency property of elliptic probability density functions. *Journal of Multivariate Analysis*, 51(1):139–147, 1994.

[18] Tomasz J. Kozubowski, Krzysztof Podgórski, and Igor Rychlik. Multivariate generalized laplace distribution and related random fields. *Journal of Multivariate Analysis*, 113:59–72, 2013. Special Issue on Multivariate Distribution Theory in Memory of Samuel Kotz.

[19] Matti Lassas, Eero Saksman, and Samuli Siltanen. Discretization-invariant bayesian inversion and besov space priors. *Inverse Problems and Imaging*, 3(1):87–122, 2009.

[20] Felix Lucka. Fast markov chain monte carlo sampling for sparse bayesian inference in high-dimensional inverse problems using l1-type priors. *Inverse Problems*, 28(12):125012, nov 2012.

References III

[21] Bernt Øksendal. *Stochastic Differential Equations*. Springer Berlin Heidelberg, 2003.

[22] Krzysztof Podgórski and Jörg Wegener. Estimation for stochastic models driven by laplace motion. *Communications in Statistics - Theory and Methods*, 40(18):3281–3302, sep 2011.

[23] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[24] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43, 2018.

[25] I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39:811–841, 1938.

[26] Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t Processes as Alternatives to Gaussian Processes. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 877–885, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[27] L. A. Shepp and B. F. Logan. The fourier reconstruction of a head section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, 1974.

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Thank you !

https://math.la.asu.edu/∼slan