Predicting Spatiotemporal Counts of Opioid-related Fatal Overdoses via Zero-Inflated Gaussian Processes

Kyle Heuton[1], **Shikhar Shrestha**[2], **Thomas J. Stopka**[2],
Jennifer Pustz[2], **Li-Ping Liu**[1], and **Michael C. Hughes**[1]

[1]Dept. of Computer Science, Tufts University, Medford, MA, USA
[2]Dept. of Public Health and Community Medicine, Tufts University School of Medicine, Boston, MA, USA

Abstract

Recently, *zero-inflated* Gaussian processes (GPs) have been proposed as probabilistic machine learning models for observed spatio-temporal data that contain many close-to-zero entries. In this work, we extend zero-inflated GPs to sparse count data via the zero-inflated Poisson likelihood. This change no longer admits a closed-form computation of the training objective, so we use automatic differentiation variational inference to perform approximate posterior estimation. Our motivating application is the prediction of the number of opioid-related overdose deaths that will occur in the next 3 months in each of 1620 census tracts across the state of Massachusetts, given historical decedent data and socio-economic covariates. We find zero-inflated GPs can prioritize regions in need of near-term public health interventions better than alternative models at finer spatial and temporal resolutions than most prior efforts. Surprisingly, we find that this model is successful even when using Normal likelihoods instead of the zero-inflated Poisson.

1 Introduction

Across many applications of spatiotemporal modeling, an essential feature of the observed data is the fact that many observations are nearly or exactly zero. Examples include measurements of rainfall, incidents of disease (Torabi, 2017), or counts of animals in ecological surveys (Ancelet et al., 2010). Many "zero-inflated" models over the years have sought to capture the excess of zeros seen is such data (Lambert, 1992; Agarwal et al., 2002; Liu and Blei, 2017).

Notable recent work by Hegde et al. (2018) proposed the *zero-inflated Gaussian process* (ZIGP) model to provide a flexible Bayesian nonparametric model for real values that may often be quite close to zero. By taking advantage of the Gaussian distribution and its synergy with the probit link function, that model enjoyed closed-form computation of its training objective. However, as a generative model, the ZIGP is suitable only for real-valued observations that contain regions of *close-to-zero* values, not exact zeros. For applications with integer-valued observations with exact zeros, such as count data, other likelihoods are required. In this work, we show how to elegantly combine the zero-inflated GP with a zero-inflated-Poisson likelihood. While the resulting model is non-conjugate, we demonstrate how to fit our model at scale with automatic differentiation variational inference (Kucukelbir et al., 2017).

Our motivating application is addressing the public health emergency caused by opioid-related overdose deaths. Overdose has recently become the leading cause of accidental death in the United States, ahead of vehicle crashes and gun violence (Kaafarani et al., 2017). In this paper, we focus on the state of Massachusetts, where opioid-related overdose deaths have increased six-fold over the past two decades, with more than 2,000 per year since 2016 (MA Dept of Public Health, 2022).

We develop a zero-inflated GP model that can learn from historical death records to predict near-term risk of future overdose deaths in all 1620 census tracts across the state of Massachusetts. Our analysis pursues finer spatial resolution than previous analyses of overdose events at the scale of

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

counties (Acharya et al., 2022; Sumetsky, 2017) or zip-codes (Herlands et al., 2018). Notable efforts have pursued neighborhood-level modeling of opioid incidents (Ertugrul et al., 2019), but focus on less than 60 neighborhoods. We develop performance measures tailored to assess how models could inform targeted public health interventions under limited resources, a strategy concurrently being pursued in a state-level randomized-trial (Marshall et al., 2022). We show that ZIGPs produce top-100 rankings of locations where spikes appear imminent better than alternatives, with surprisingly good performance using a Normal likelihood even though it is not specialized to count data.

2 Model

ZIGP with Normal likelihood. As proposed by Hegde et al. (2018), the zero-inflated GP is a regression model for any provided set of N inputs, indexed by n. Let $\mathbf{x}_n \in \mathbb{R}^D$ denote the (assumed known) feature vector observed at each input. The model defines three random variables at each input. First, there is a real-valued observation $\tilde{y}_n \in \mathbb{R}$. Then, there are two latent function values, which we'll call the nonsparsity-level $g_n \in \mathbb{R}$ and the signal $f_n \in \mathbb{R}$. The model assumes the following generative process defining the joint distribution $p_\theta(\mathbf{g}, \mathbf{f}, \mathbf{y})$ over N-dimensional vectors $\mathbf{g}, \mathbf{f}, \mathbf{y}$,

$$\mathbf{f}|\mathbf{g} \sim \mathcal{N}(\mathbf{0}_{1:N}, \mathbf{K}^F_{1:N,1:N} \odot \Phi(\mathbf{g})\Phi(\mathbf{g})^T), \qquad \mathbf{g} \sim \mathcal{N}(\mathbf{0}_{1:N}, \mathbf{K}^G_{1:N,1:N}),$$
$$\tilde{y}_n|f_n \sim \mathcal{N}(f_n, \sigma^2), \text{ for } n \in 1, \dots N. \tag{1}$$

Here, $\mathcal{N}$ denotes a multivariate normal distribution, $\odot$ denotes elementwise multiplication, $\Phi(\cdot)$ denotes the elementwise probit function (the CDF of the standard univariate Normal), and $\mathbf{K}^G, \mathbf{K}^F$ denote matrices of suitable size built from chosen kernel functions $k^G(\mathbf{x}_i, \mathbf{x}_j; \gamma^G), k^F(\mathbf{x}_i, \mathbf{x}_j; \gamma^F)$ that map pairs of feature vectors to scalars. Parameters, compactly denoted via vector θ, include the kernel hyperparameters γ^G, γ^F and the likelihood standard deviation $\sigma > 0$.

ZIGP with Zero-inflated Poisson likelihood. We adapt the above to a new model appropriate for generating *count* values $y_n \in \{0, 1, \dots\}$ instead of real values $\tilde{y}_n$. We can use the same generative model for f and g, but the likelihood becomes

$$p(y_n = k|f_n, g_n) = \begin{cases} \Phi(g_n)\text{PoiPMF}(0|r(f_n)) + (1 - \Phi(g_n)) & k = 0 \\ \Phi(g_n)\text{PoiPMF}(k|r(f_n)) & k \in 1, 2, \dots \end{cases} \tag{2}$$

This form matches the standard zero-inflated Poisson (Lambert, 1992), but with GP-generated values g_n and f_n controlling its parameters. With probability $1 - \Phi(g_n)$, the count is generated from a pure-zero process, so g_n values far below zero imply a zero-valued count y_n. Alternatively, when g_n is far above zero, we generate count y_n from a Poisson, with positive mean determined by a monotonic transformation $r(\cdot)$ of real-value f_n. We set r to the softplus in all experiments. The roles of f and g are intuitively illustrated in App. Fig. 2.

2.1 Automatic Differentiation Variational Inference

We are interested in estimating the posterior $p(\mathbf{f}, \mathbf{g}|\mathbf{y})$ for the function values at the N training inputs and using it to make predictions of new observations at input x_* via the posterior predictive $p(y_*|\mathbf{y})$. We'll first follow the inducing points strategy for scalable GPs (Hensman et al., 2013, 2015), introducing M random variables $\mathbf{h}^F$ representing signal function outputs and another M random variables $\mathbf{h}^G$ for nonsparsity function outputs, each with corresponding input locations z^F, z^G. The model's expanded joint distribution is now defined by factors

$$p(\mathbf{h}^G) = \mathcal{N}(0, K^G_{1:M,1:M}), \quad p(\mathbf{g}|\mathbf{h}^G) = \mathcal{N}\left(A^G\mathbf{h}^G, B^G_{1:N,1:N}\right), \tag{3}$$
$$p(\mathbf{h}^F) = \mathcal{N}(0, K^F_{1:M,1:M}), \quad p(\mathbf{f}|\mathbf{g}, \mathbf{h}^F) = \mathcal{N}\left(\Phi(\mathbf{g}) \odot A^F\mathbf{h}^F, \Phi(\mathbf{g})\Phi(\mathbf{g})^T \odot B^F_{1:N,1:N}\right). \tag{4}$$

where suitably-sized matrices A^G, B^G and A^F, B^F are defined in terms of kernel matrices K^F, K^G (see supplement). Parameters θ now include the input features z^G, z^F of the M inducing points.

To apply variational inference (Blei et al., 2017), we assume an approximate posterior q with factors $q(\mathbf{h}^G, \mathbf{h}^F, \mathbf{g}, \mathbf{f}) = q(\mathbf{h}^G)q(\mathbf{h}^F)p(\mathbf{g}|h^G)p(\mathbf{f}|\mathbf{g}, \mathbf{h}^F), \quad q(\mathbf{h}^G) = \mathcal{N}(m^G, S^G), \; q(\mathbf{h}^F) = \mathcal{N}(m^F, S^F)$, where the colored factors indicate that we reuse the same density as the generative model. Beyond θ, the additional free parameters of q are denoted compactly as $\nu = \{m^G, S^G, m^F, S^F\}$.

Our evidence lower bound optimization objective (ELBO), $\mathcal{L}(\theta, \nu) \leq \log p(\mathbf{y}; \theta)$, becomes

$$\mathcal{L}(\theta, \nu) = \mathbb{E}_{q_{\theta,\nu}(\mathbf{f},\mathbf{g})} \underbrace{\left[\sum_{n=1}^N \log p(y_n|f_n, g_n)\right]}_{\mathcal{L}_1} + \underbrace{\text{KL}(q_\nu(\mathbf{h}^G, \mathbf{h}^F)||p_\theta(\mathbf{h}^G, \mathbf{h}^F))}_{\mathcal{L}_2} \tag{5}$$

The first term is the challenging one. The second term, $\mathcal{L}_2$ is a KL divergence between Gaussian densities with known mean and covariances. Its value and its gradients with respect to θ and ν can be computed exactly using well-known formulas (Hegde et al., 2018). We omit the ELBO's final term $\mathrm{KL}(q(\mathbf{f}, \mathbf{g}|\mathbf{h})||p(\mathbf{f}, \mathbf{g}|\mathbf{h}))$ because it cancels to zero due to setting $q(\mathbf{f}, \mathbf{g}|\mathbf{h})$ equal to $p(\mathbf{f}, \mathbf{g}|\mathbf{h})$.

We'd now like to compute the first term and its gradients as a function of model parameters θ and variational parameters ν. Following the ADVI recipe (Kucukelbir et al., 2017), we can use a Monte Carlo estimator from minibatches of size B to approximate $\mathcal{L}_1$:

$$\mathcal{L}_1(\theta, \nu) \approx \frac{N}{B} \sum_{n \in \mathcal{B}} \frac{1}{S} \sum_{s=1}^{S} \log p(y_n | f_n(w_n^s, u_n^s; \theta, \nu), g_n(u_n^s; \theta, \nu)), \quad u_n^s, w_n^s \sim \mathcal{N}(0, I) \tag{6}$$

Here, the deterministic functions $g_n(\cdot), f_n(\cdot)$ produce samples of scalars g_n, f_n from per-example marginal posteriors $q(g_n)$ and $q(f_n|g_n)$. Both these q factors are Normally-distributed, and thus samples can be obtained by transforming mean-zero, unit-variance normal samples u, w via the reparameterization trick (Rezende et al., 2014) that underlies ADVI. See supplement for details. Gradients $\nabla_\theta \mathcal{L}_1, \nabla_\nu \mathcal{L}_1$ follow by applying automatic differentiation to Eq. (6).

3 Application: Preventing Fatal Overdoses involving Opioids

3.1 Data and Task Description

Data. We obtained death certificate data from the Massachusetts Registry of Vital Records and Statistics for opioid-involved overdose deaths between 2001 and 2019. These deaths were defined as unintentional, intentional, and undetermined drug poisonings containing an opioid code as an underlying cause-of-death. In Massachusetts, death certificate data are publicly available upon request and may be provided to researchers at the individual level, allowing us obtain a residential street address for each decedent. We used only decedant data, so our institution's IRB gave the project a Not Human Research Determination.

	# deaths	sparsity (%)
2000	356	94.7
2010	524	92.3
2019	1837	76.4

Table 1: **Dataset summary.** Fatal overdose deaths involving opioids in MA, for select years from 2000-2019. Sparsity is reported across the product of all 1620 census tracts and all four quarters per year.

We divided the state into the $S = 1620$ spatial tracts used in the 2020 U.S. census. Each tract by design contains typically 4000 people (range 1200-8000) (US Census Bureau, 2022). We divided time into quarters (3 month periods, beginning Jan-Mar.). We then computed the observed number of death events $y_{s,t}$ at time unit t for individuals residing in spatial tract s, using open tools (Freeman, 2022) that call the US Census Geocoding API to identify the correct census tract given a residential street address. Tab. 1 summarizes our dataset's total death events and per-tract sparsity over the years.

Prediction task. Given all historical count data $\mathbf{y}_{1:S,1:T}$ for a time period from $t \in \{1, \dots T\}$, we wish to predict the counts in the next time period $\mathbf{y}_{1:S,T+1}$. Concretely, we used each quarter of 2019 as a heldout test set, keeping all previous quarters as the training set in each case (e.g. $T = 72$ for Q1 2019). These granularity levels were deliberately chosen to formulate an *actionable* prediction task where a model could identify high-risk tracts and enable interventions to reduce future harm. Example targeted public health interventions include overdose education and naloxone distribution programs or increased access to effective medications for opioid use disorder.

Covariates. Basic covariates include a vector ℓ_s giving the numerical latitude and longitude location of geocentroid of each census tract, and the scalar start time τ_t of each period (measured in quarters since 2000). Additionally, for each census tract s at time t we have a vector $\mathbf{x}_{s,t} = [x_{st1}, x_{st2}, \dots x_{st5}]$ representing its percentile rankings (higher is more "vulnerable") across four thematic dimensions – socioeconomic, age-related demographics, minority status, and housing – as well as a 5th dimension representing a composite total vulnerability. These features are from the five published of the Social Vulnerability Index (CDC ATSDR, 2018) between 2000 and 2018.

3.2 Results and Analysis

Performance Metrics. First, we assess prediction via mean absolute error, the difference between predicted counts $\hat{\mathbf{y}}_{1:S,T+1}$ to actual counts $\mathbf{y}_{1:S,T+1}$ for test period $T + 1$. Second, we assess *Percentage of Best-Possible Reach* using the top-K tracts, $\%\mathrm{BPR}(K)$, a metric devised in conversation with public health experts to reflect the applied need to spend limited intervention resources on the K highest-priority regions identified by the model. We compute this metric as:

$$\%\mathrm{BPR}(K) = (\textstyle\sum_{k=1}^{K} y_{\mathrm{top\_tract}(k, \hat{\mathbf{y}}_{:,T+1}), T+1}) / (\sum_{k=1}^{K} y_{\mathrm{top\_tract}(k, \mathbf{y}_{:,T+1}), T+1}) \cdot 100\% \tag{7}$$

method	MAE	%BPR K=100
allzero	0.28	16.0
lastyear	0.42	26.4
GLM+Poisson	0.44	25.2
RF+Poisson	0.41	25.1
ZIGP+Normal	0.39	34.5
ZIGP+ZIPoisson	0.34	31.9

Table 2: **Comparison of methods for predicting fatal overdose counts (per census tract, per quarter) in MA in 2019.** We report mean absolute error (MAE, lower is better) and percentage of best-possible reach (%BPR, higher is better, see Eq. (7)), averaged across the 4 quarters of 2019. BPR assesses the public health intervention potential of each model's ability to identify K=100 tracts with highest predicted deaths.

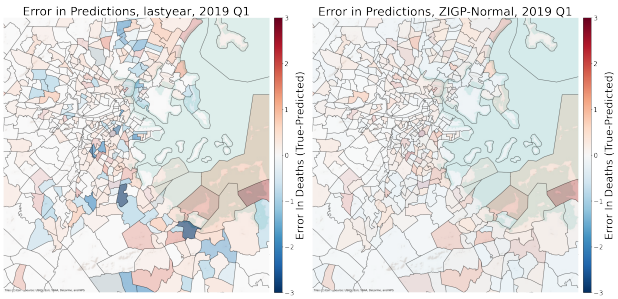

Figure 1: **Visualization of model errors at the census tract level for 2019 Q1 in metro Boston**. On the left are the residuals from predicting using the strongest baseline: last year's mortality. On the right are the residuals from the ZIGP with a Normal likelihood. Darker colors meaning larger errors. Blue hue indicates tracts where the model predicted too-high mortality; red hue indicates too-low.

where function top_tract$(k, \mathbf{a})$ produces an index corresponding to the k-th largest element in provided array $\mathbf{a}$ of size S. %BPR's numerator counts all truly at-risk individuals that could be reached by an intervention deployed for the top-K spatial tracts recommended by the model. %BPR's denominator gives the maximum possible count of all size-K sets, assuming perfect knowledge of the future data. The produced BPR value will be between 0 and 100%; larger values mean better top-K rankings. In the case of ties for the numerator, we average over all indices corresponding to the k-th best value.

Results. Tab. 2 compares several variants of Gaussian Process, including our proposed ZIGP with ZIP likelihood, with other baselines (described in App. C). Here we see the difficulty of this problem in the strength of the naive baselines. Simply guessing 0 deaths produces the model with the lowest MAE. Here the *lastyear* model represents an intuitive baseline; it assumes that the locations with the highest mortality one year ago will have the highest mortality today. More complex models such as the GLM and random forest cannot outperform this method. Here we see the success of the ZIGP model with a Normal likelihood, in that the tracts it predicts will have the highest mortality have the highest %BPR of any method.

To further gain understanding, we provide a detailed qualitative view of the model's predictions across the city of Boston in Fig. 1. The red census tracts in the east show that both models failed to anticipate a rise in deaths. In the lastyear model, we see many dark blue tracts where the model predicted too-high mortality. We see that the flexible nature of the ZIGP allowed it to avoid these errors, while still performing well at identifying high-mortality tracts by the %BPR metric.

4 Discussion

Opioid-related overdose deaths represent a compelling public health problem in need of solutions. Typical public health responses have been reactive, based on surveillance data from previous years. While helpful, risk landscapes and distributions of opioid-related overdoses can change rapidly across geographies and time, based on locally-evolving risk factors (e.g., contaminated drug source, access to treatment). Our modeling approach allows for much-needed fine granularity across an entire U.S. state. Local level prediction of risk is vital to reducing the impact of the opioid overdose crisis - a limitation that is common in other prediction models which are at the level of zip codes or cities.

Predictive approaches are of course no panacea, but we believe careful modeling that pinpoints future locations of impending overdose spikes can inform targeting of preemptive public health interventions. The model developed in this study can identify and inform policymakers and planners to channel limited resources to areas that have high risk of opioid overdose. The availability of medication for opioid use disorders such as methadone, buprenorphine, and naltrexone along with addition ancillary services can help reduce the risk of opioid overdose. Improving the availability of naloxone, an overdose rescue medication, and fentanyl testing strips can further provide additional protective measures among individuals at risk. These protective measures work best when directed at high-risk locations at the right time and with tailored responses, thereby providing the necessary tools to people who use drugs to reduce their risk of opioid-related overdose. We hope this work is a first step toward using scalable spatio-temporal machine learning to reduce harm among vulnerable populations.

Acknowledgments and Disclosure of Funding

Authors KH, TS, LL, and MCH gratefully acknowledge support from the U.S. National Science Foundation under grant NSF IIS-1908617. KH is also supported by NSF award NRT-HDR 2021874.

References

A. Acharya, A. M. Izquierdo, S. F. Gonçalves, R. A. Bates, F. S. Taxman, M. P. Slawski, H. S. Rangwala, and S. Sikdar. Exploring County-level Spatio-temporal Patterns in Opioid Overdose related Emergency Department Visits. *medRxiv: The Preprint Server for Health Sciences*, 2022.

D. K. Agarwal, A. E. Gelfand, and S. Citron-Pousty. Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9(4):341–355, 2002.

S. Ancelet, M.-P. Etienne, H. Benoît, and E. Parent. Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process. *Environmental and Ecological Statistics*, 17(3): 347–376, 2010.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. http://arxiv.org/abs/1601.00670.

CDC ATSDR. Social Vulnerability Index 2018 Database for Massachusetts., 2018.

A. M. Ertugrul, Y.-R. Lin, and T. Taskaya-Temizel. CASTNet: Community-Attentive Spatio-Temporal Networks for Opioid Overdose Forecasting. In *Machine Learning and Knowledge Discovery in Databases: European Conference (ECML PKDD)*, 2019. http://arxiv.org/abs/1905.04714.

N. Freeman. Censusgeocode: Thin Python wrapper for the US Census Geocoder, 2022. https://github.com/fitnr/censusgeocode.

P. Hegde, M. Heinonen, and S. Kaski. Variational zero-inflated Gaussian processes with sparse kernels. In *Uncertainty in Artificial Intelligence*, 2018. https://auai.org/uai2018/proceedings/papers/148.pdf.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Conference on Uncertainty in Artificial Intellegence*, pages 282–290, 2013.

J. Hensman, A. G. d. G. Matthews, and Z. Ghahramani. Scalable Variational Gaussian Process Classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.

W. Herlands, E. McFowland III, A. G. Wilson, and D. B. Neill. Gaussian Process Subset Scanning for Anomalous Pattern Detection in Non-iid Data. In *Artificial Intelligence and Statistics*, 2018. http://proceedings.mlr.press/v84/herlands18a/herlands18a.pdf.

H. M. A. Kaafarani, E. Weil, S. Wakeman, and D. Ring. The Opioid Epidemic and New Legislation in Massachusetts: Time For a Culture Change in Surgery? *Annals of Surgery*, 265(4):731–733, 2017.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017. http://jmlr.org/papers/v18/16-107.html.

D. Lambert. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*, 34(1), 1992.

L.-P. Liu and D. M. Blei. Zero-Inflated Exponential Family Embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2140–2148. PMLR, 2017. https://proceedings.mlr.press/v70/liu17a.html.

MA Dept of Public Health. Data Brief: Opioid-Related Overdose Deaths among Massachusetts Residents, 2022. https://www.mass.gov/doc/opioid-related-overdose-deaths-among-ma-residents-june-2022/download.

B. D. L. Marshall, N. Alexander-Scott, J. L. Yedinak, B. D. Hallowell, W. C. Goedel, B. Allen, R. C. Schell, Y. Li, M. S. Krieger, C. Pratty, J. Ahern, D. B. Neill, and M. Cerdá. Preventing Overdose Using Information and Data from the Environment (PROVIDENT): protocol for a randomized, population-based, community intervention trial. *Addiction (Abingdon, England)*, 117 (4):1152–1162, Apr. 2022. ISSN 1360-0443. doi: 10.1111/add.15731.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning*, pages 1278–1286, 2014. http://proceedings.mlr.press/v32/rezende14.pdf.

N. Sumetsky. Spatiotemporal modeling of opioid abuse and dependence outcomes using Bayesian hierarchical methods. Master's thesis, University of Pittsburgh Graduate School of Public Health, 2017.

M. Torabi. Zero-inflated spatio-temporal models for disease mapping. *Biometrical Journal. Biometrische Zeitschrift*, 59(3):430–444, 2017.

US Census Bureau. Glossary, 2022. https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13.

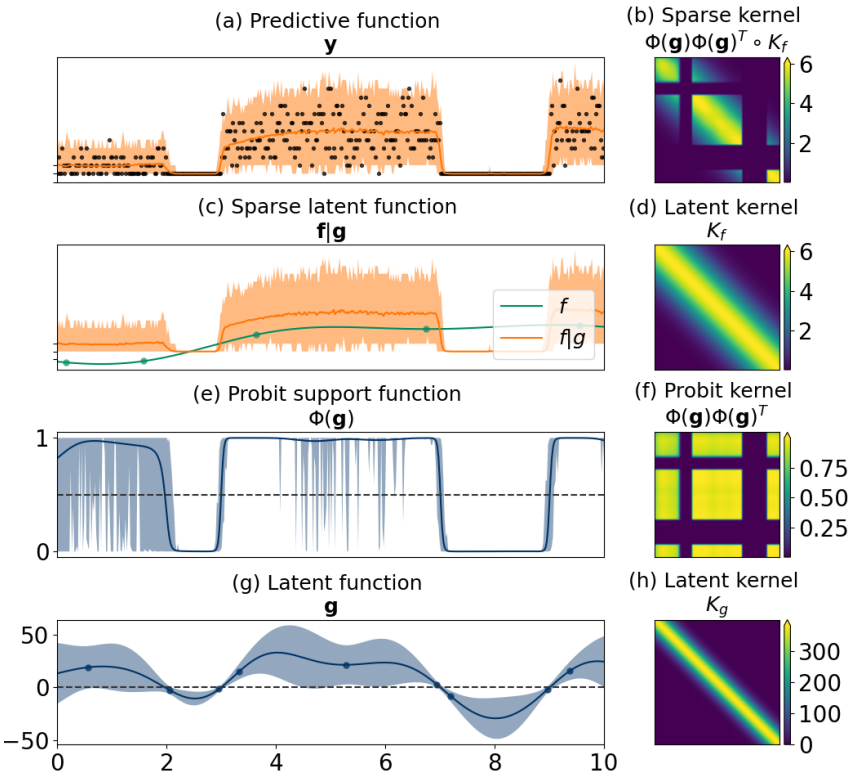

Figure 2: **Diagram of learned posteriors on y, f, and g for Zero-inflated Gaussian Process with Zero-inflated Poisson likelihood on toy dataset.** Large negative values of the nonsparsity function g lead to regions of all-zero y. Otherwise, if g is near zero or positive, the signal function f drives the mean of a Poisson density on y. This toy dataset is adapted from one in Hegde et al. (2018) originally used for real-valued data; this dataset has count observations for y (top panel).

A Expanded Model Details

Toy data illustration of the ZIGP-ZIPoisson. In Fig. 2, we provide an intuitive illustration of how latent random variables f, g are used to model observed count data y in our proposed zero-inflated Gaussian process with zero-inflated Poisson likelihood model.

Definition of expanded model with inducing points. Our inducing point expansion of the ZIGP-ZIP model in Eq. (3) includes as random variables the function evaluations $\mathbf{h}^G, \mathbf{h}^F$ at the M inducing points.

Here, we define the matrices A, B used in that model

$$A^G = K^G_{N,M}\left(K^G_{M,M}\right)^{-1}, \qquad B^G = K^G_{N,N} - K^G_{N,M}\left(K^G_{M,M}\right)^{-1}K^G_{M,N} \tag{8}$$

$$A^F = K^F_{N,M}\left(K^F_{M,M}\right)^{-1}, \qquad B^F = K^F_{N,N} - K^F_{N,M}\left(K^F_{M,M}\right)^{-1}K^F_{M,N} \tag{9}$$

where for function f, $K^F_{N,N}$ represents the $N \times N$ kernel matrix for all pairs in the training set, $K^F_{M,M}$ is the kernel matrix for all pairs in the inducing set, and $K^F_{M,N}$ is the rectangular kernel matrix where the value in entry i, j represents the kernel at the i-th inducing point and the j-th training points. K^G matrices are defined similarly for function g using its corresponding kernel function.

B Variational Methods

We've assumed a variational density family of the form

$$q(h^G, h^F, g, f) = q(h^G; m^G, S^G)q(h^F; m^F, S^F)p(g|h^G)p(f|g, h^F) \tag{10}$$

Following Hegde et al. (2018), we can apply the sum rule to this joint distribution to obtain the marginals that are also Normally-distributed

$$q(g) = \mathcal{N}(\mu^G, \Sigma^G), \tag{11}$$

$$q(f|g) = \mathcal{N}(\text{diag}(\Phi(g))\mu^F, \Phi(g)\Phi(g)^T \odot \Sigma^F), \tag{12}$$

$$\Sigma^G = K^G_{N,N} + A^G(S^G - K^G_{M,M})A^{G,T}, \qquad \mu^G = A^G m^G, \tag{13}$$

$$\Sigma^F = K^F_{N,N} + A^F(S^F - K^F_{M,M})A^{F,T}, \qquad \mu^F = A^F m^F, \tag{14}$$

Looking at specific training point n, which is all that is needed to evaluate the expected likelihood of y_n in Eq. (6), we thus have

$$q(g_n) = \mathcal{N}(c^G_n, \kappa^2_n), \qquad c^G_n = A^G_{n,:}m^G, \;\; \kappa^2_n = K^G_{n,n} + A^G_{n,:}(S^G - K^G_{1:M,1:M})A^{G,T}_{:,n}$$

$$q(f_n|g_n) = \mathcal{N}(\Phi(g_n)c^F_n, \Phi(g_n)^2\tau^2_n), \quad c^F_n = A^F_{n,:}m^F, \;\; \tau^2_n = K^F_{n,n} + A^F_{n,:}(S^F - K^F_{1:M,1:M})A^{F,T}_{:,n}$$

using the above, we can reparameterize to draw samples from $q(g_n)$ and $q(f_n|g_n)$ via

$$g_n(u) = \kappa_n u + c^G_n, \qquad u \sim \mathcal{N}(0,1) \tag{15}$$

$$f_n(u,w) = \Phi(g_n(u))\tau_n w + \Phi(g_n(u))c^F_n, \qquad w \sim \mathcal{N}(0,1) \tag{16}$$

C Applied Methods for Fatal Overdose Prediction

Here we provide more detail about the methods we compare

C.1 Simple baselines : no training required

We consider several very simple methods that involve no modeling whatsoever.

1. zero : Always guess zero: $\hat{y}_{s,T+1} = 0$
2. median : Guess each spatial tract's training set median: $\hat{y}_{s,T+1} = \text{median}(\mathbf{y}_{s,1:T})$
3. lastyear : Guess each spatial tract's previous year value: $\hat{y}_{s,T+1} = y_{s,T-3}$

These methods require no parameter fitting or hyperparameters whatsoever.

C.2 ML baselines

We further consider some baseline machine learning methods using implementations from sklearn.

1. GLM+Poisson : Generalized linear model with Poisson likelihood
2. RF+P : Random forest with Poisson

For the GLM with Poisson likelihood we used a grid search to select a regularization penalty of 1.0 and optimized using the L-BFGS solver until the gradients stopped changing with a tolerance of 1e-4.

For the Random forest with a Poisson loss, we conducted a grid search over maximum number of leaves in a tree, maximum depth, learning rate, and L2 regularization strength. We selected a maximum leaf count of 10, no maximum tree depth, a learning rate of 0.01 and no L2 regularization. The model was optimized using the L-BFGS solver until the gradients stopped changing with a tolerance of 1e-4.

C.3 Gaussian Process Methods

1. GP+N : Standard GP (no zero inflation) with normal likelihood
2. ZIGP+N : Zero-inflated GP with normal likelihood and exact inference Hegde et al. (2018)
3. ZIGP+ZIP : Proposed ZIGP with zero-inflated Poisson likelihood

For all GP methods, to define covariances we consider several kernels of squared-exponential form, including

$$\text{spatial} \qquad k_{\mathbb{S}}(s,t,s',t') = \alpha_{\mathbb{S}} \exp\left(-\frac{1}{2}\frac{||\ell_s - \ell_s||}{\lambda_{\mathbb{S}}}\right) \tag{17}$$

$$\text{temporal} \qquad k_{\mathbb{T}}(s,t,s',t') = \alpha_{\mathbb{T}} \exp\left(-\frac{1}{2}\frac{(\tau_t - \tau_{t'})^2}{\lambda_{\mathbb{T}}}\right) \tag{18}$$

$$\text{demographics/economics} \qquad k_{\mathbb{D}}(s,t,s',t') = \alpha_{\mathbb{D}} \exp\left(-\frac{1}{2}\frac{||\mathbf{x}_{s,t} - \mathbf{x}_{s',t'}||}{\lambda_{\mathbb{D}}}\right) \tag{19}$$