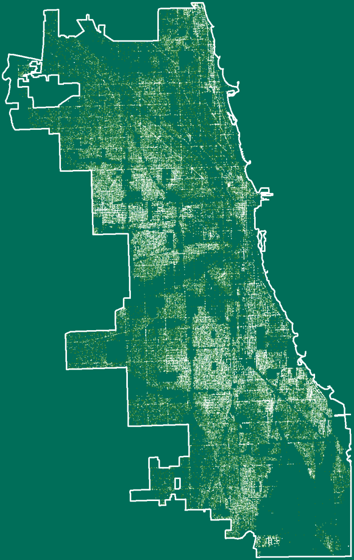


Distribution regression, ecological inference, encoding GP aggregates and the change-of-support problem

Seth Flaxman
Department of Computer Science
University of Oxford
March 2024

MACHINE LEARNING & GLOBAL HEALTH NETWORK

UNIVERSITY OF OXFORD

Plan for my talk

- Distribution regression for ecological inference
- More recent work on Gaussian process aggregation
- A theorem and some open questions

Kernel mean embeddings and distribution regression[1]

Individual-level data with group-level labels:

$$\left(\{x_1^j\}_{j=1}^{N_1}, y_1\right), \left(\{x_2^j\}_{j=1}^{N_2}, y_2\right), \ldots \left(\{x_n^j\}_{j=1}^{N_n}, y_n\right)$$

Learn a function:

$$f : \{x^j\}_{j=1}^{N} \to y$$

[1]Flaxman, Wang, Smola, "Who Supported Obama in 2012?: Ecological Inference through Distribution Regression," KDD 2015

Learning from distributions

- Previous work: Jebara et al, 2004; Hein and Bousquet, 2005; Muandet et al, 2012; Póczos et al, 2013, Szabó et al (2014), Lopez-Paz et al, 2015, Lopez-Paz (2016).
- Distribution regression / distribution classification relies on the kernel mean embedding [see Muandet et al 2017's survey]
- Given kernel $k(x, \cdot)$, RKHS $\mathcal{H}_k$, and corresponding embedding $\phi(x) \in \mathcal{H}_k$, consider a measure with $X \sim \mathcal{P}$. Then define:

$$\mu_{\mathcal{P}} := \mathrm{E}[\phi(X)] = \int_{\mathcal{X}} \phi(x) d\mathcal{P}(x) \tag{1}$$

  Obvious empirical estimator for samples $x_1, \ldots, x_n$:

$$\hat{\mu_P} := \frac{1}{n} \sum_i \phi(x_i) \tag{2}$$

- Learning: use any supervised learning method to learn a function $f(\mu_{\mathcal{P}})$.

Distribution embedding illustration

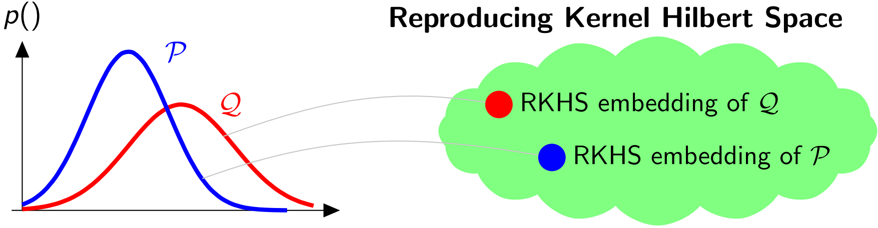
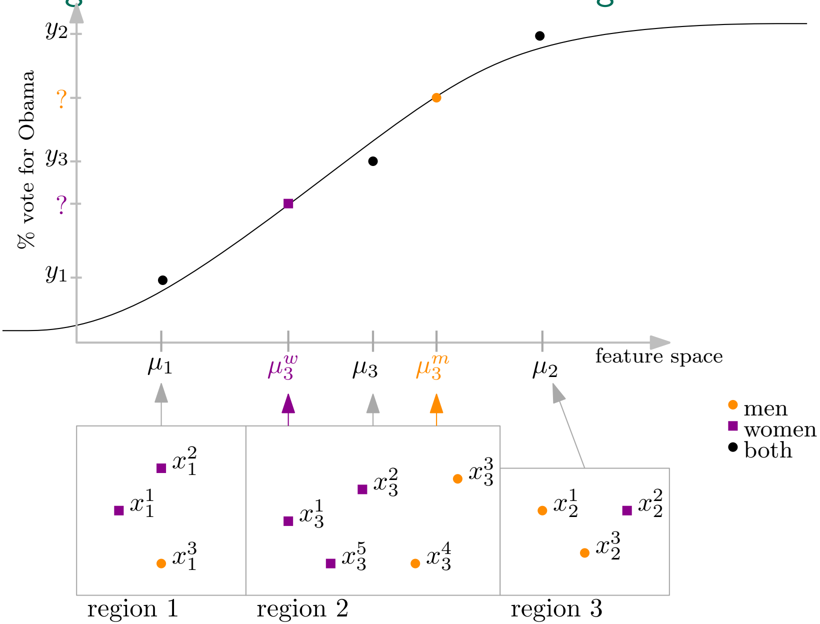


Figure: Each distribution is mapped into the reproducing kernel Hilbert space via an expectation operation. (Source: Muandet et al 2017)

Ecological inference with distribution regression



Bayesian distribution regression

- Estimate $\widehat{\mu_1}, \ldots, \widehat{\mu_n} \in \mathcal{R}^n$ using kernel embeddings:

$$\widehat{\mu}_i = \frac{1}{N} \sum_j k(x_i^j, \cdot) = \frac{1}{N} \sum_j \phi(x_i^j)$$

- Use GP logistic regression
- Additive kernels with a spatial component:

$$K_{ij} = \sigma_x^2 \langle \widehat{\mu}_i, \widehat{\mu}_j \rangle + k_s(s_i, s_j)$$

$$\boldsymbol{f} \sim \mathcal{GP}(0, \boldsymbol{K})$$

$$k_i | f_i \sim \text{Binomial}(n_i, \text{logit}^{-1}(f_i))$$

Obama received k_i out of n_i votes in region i.

- Make predictions for demographic subgroups:

$$\widehat{\boldsymbol{f}}(\mu_i^{\text{women}}, s_i)$$

Kernel details

- Demographic attributes (Gaussian RBF):
  - Standardize coordinates
  - Expand discrete attributes:
    (low, medium, high income) $\rightarrow$ $([1\ 0\ 0], [0\ 1\ 0], [0\ 0\ 1])$.
  - Use random Fourier features for speed:
    $k(x, x') = \langle \phi(x), \phi(x') \rangle \approx \langle \hat{\phi}(x), \hat{\phi}(x') \rangle$ with $\hat{\phi}(x) \in \mathcal{R}^{2048}$.
- Spatial attributes with Matérn-$\frac{3}{2}$:

$$k(s, s') = (1 + \rho \|s - s'\|) \exp(-\rho \|s - s'\|)$$

Millions of observations, but the covariance matrix is 843×843 for the 843 electoral regions.

Algorithm details

- One pass through census data to create mean embeddings:

$$\widehat{\mu_1} = \frac{\sum_j w_1^j \phi(x_1^j)}{\sum_j w_1^j}, \quad \ldots, \quad \widehat{\mu_n} = \frac{\sum_j w_n^j \phi(x_n^j)}{\sum_j w_n^j} \qquad (3)$$

- Setup GP regression:

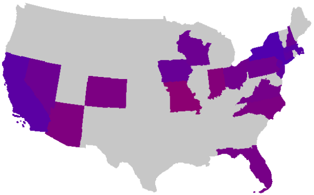
$$f \sim \mathcal{GP}(0, \sigma_x^2 K_x + \sigma_s^2 K_s)$$

$$k_i | f_i \sim \text{Binomial}(n_i, \text{logit}^{-1}(f_i))$$

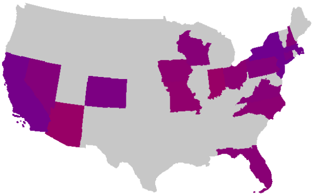
- Laplace approximation for hyperparameter learning $\theta = [\sigma_x, \sigma_s, \rho]$ w/ marginal likelihood
- Bayesian posterior inference to make predictions for latent f at new "locations":

$$p(f_*^{\text{men}} | y, \hat{\theta})$$

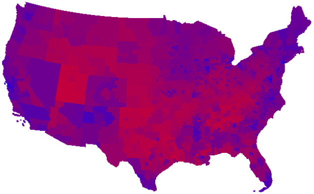
Experiments



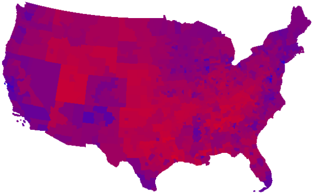
Exit poll women



Exit poll men

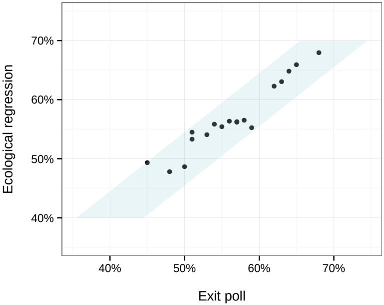


Ecological regression women

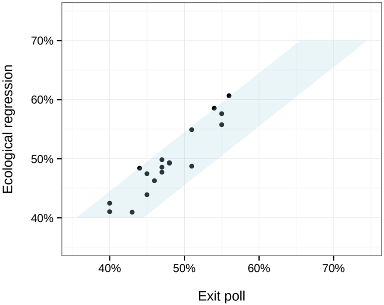


Ecological regression men

Experiments

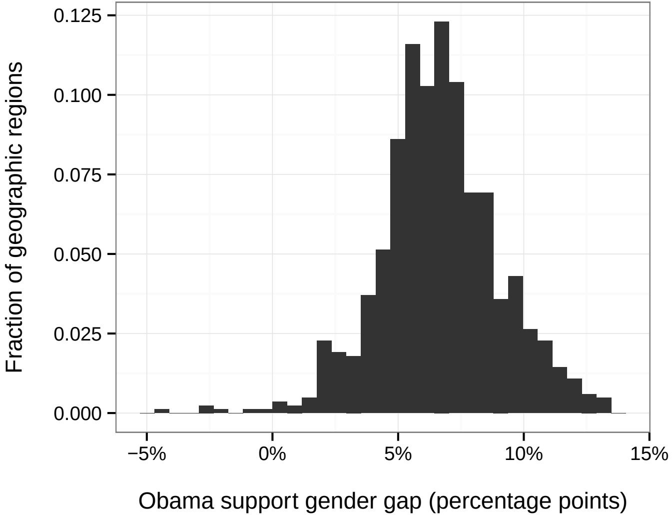


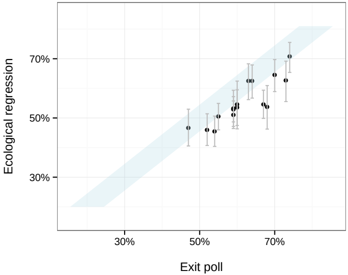
Women



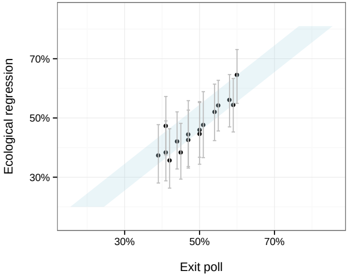
Men

Experiments

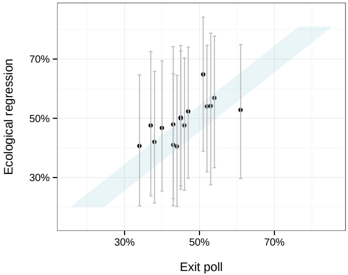




Low income



Medium income



High income

Refinements for 2016 election[2]

- Explicity model non-voters:

  $$_i = [\text{Clinton votes}, \text{Trump votes}, \text{Non-votes and third party votes}]^\top$$

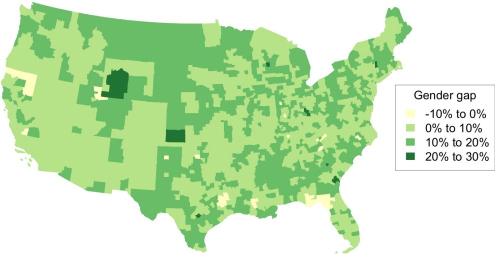
- Multinomial likelihood with softmax link, fit with penalized MLE with group lasso and L_2 penalty
- More interpretable / richer feature representation to allow for exploratory analysis / calculation of marginal effects:

  $$(x_i^j) := [\phi_1(x_{r1}^j), \ldots, \phi_d(x_{rd}^j)]^\top \quad (4)$$

- Incorporation of some exit polling data as extra set of labeled distributions

[2] arXiv:1611.03787

Results for 2016 Presidential Election



	Clinton	Trump	Frac. electorate	Participation rate
Men	0.45	0.55	0.47	0.50
Women	0.56	0.44	0.53	0.53
18–29 year olds	0.62	0.38	0.17	0.42
30–44	0.54	0.46	0.25	0.54
45–64	0.46	0.54	0.39	0.58
65 and older	0.45	0.55	0.18	0.47

Results for 2016 Presidential Election

	Clinton	Trump	Participation
Language other than English spoken at home	0.74	0.26	0.32
Mobility = lived here one year ago	0.45	0.55	0.55
Mobility = moved here from outside US and Puerto Rico	0.60	0.40	0.47
Mobility = moved here from inside US or Puerto Rico	0.56	0.44	0.48
Active duty military	0.45	0.55	0.56
Not enrolled in school	0.45	0.55	0.60
Enrolled in a public school or public college	0.61	0.39	0.39
Enrolled in private school, private college, or home school	0.66	0.34	0.53

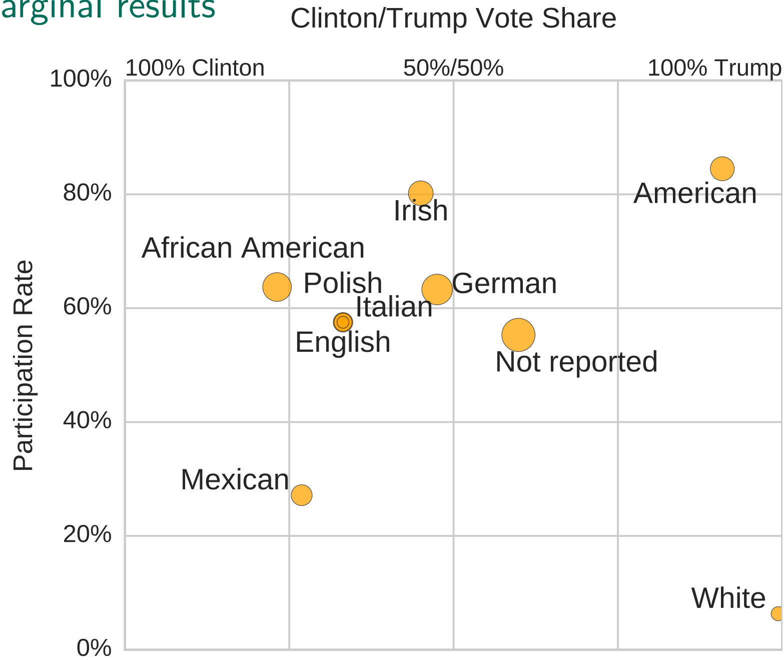
Results for 2016 Presidential Election

	Clinton	Trump	Frac	Participation
personal income $\leq$ 50000 & men	0.56	0.44	0.25	0.37
personal income $\leq$ 50000 & women	0.63	0.37	0.36	0.40
50000 $<$ personal income $\leq$ 100000 & men	0.40	0.60	0.15	0.67
50000 $<$ personal income $\leq <$ 100000 & women	0.53	0.47	0.13	0.84
personal income $>$ 100000 & men	0.49	0.51	0.08	0.70
personal income $>$ 100000 & women	0.62	0.38	0.03	0.80

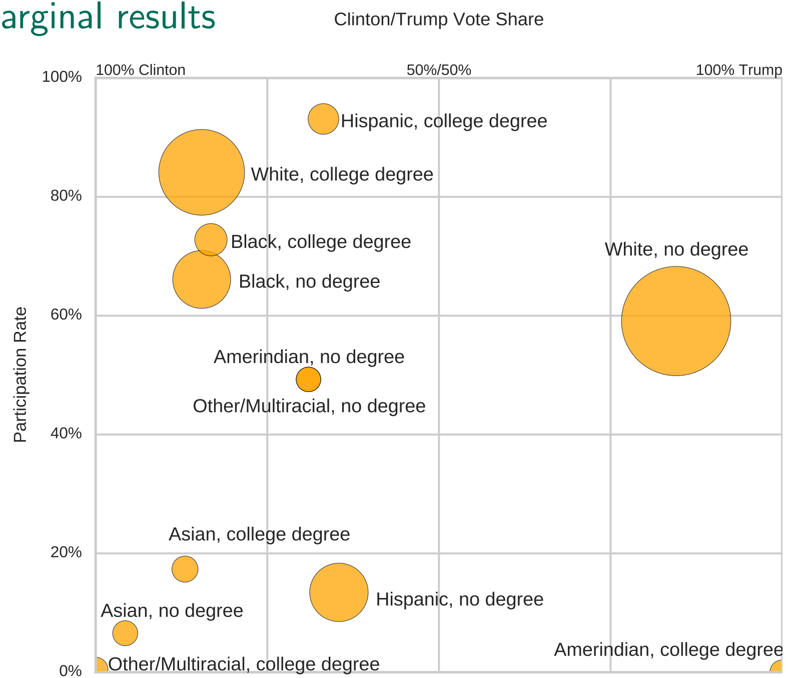
Exploratory results

	feature	deviance	frac.deviance
1	RAC3P - race coding	0.04	0.86
2	ethnicity interacted with has degree	0.04	0.74
3	schooling attainment	0.04	0.72
4	ANC2P - detailed ancestry	0.04	0.83
5	OCCP - occupation	0.04	0.75
6	COW - class of worker	0.04	0.67
7	ANC1P - detailed ancestry	0.05	0.77
8	NAICSP - industry code	0.05	0.71
9	RAC2P - race code	0.05	0.70
10	age interacted with usual hours worked per week (WKHP)	0.05	0.69
11	sex interacted with ethnicity	0.05	0.65
12	MSP - marital status	0.05	0.61
13	FOD1P - field of degree	0.05	0.61
14	ethnicity	0.06	0.57
15	RAC1P - recoded race	0.06	0.54
16	sex interacted with age	0.06	0.57
17	has degree interacted with age	0.06	0.55
18	age interacted with personal income	0.06	0.76
19	sex interacted with hours worked per week	0.06	0.62
20	personal income interacted with hours worked per week	0.06	0.69
21	personal income	0.06	0.59
22	RACSOR - single or multiple race	0.07	0.42
23	has degree interacted with hours worked per week	0.07	0.59
24	hispanic	0.07	0.56
25	sex interacted with personal income	0.07	0.57

Marginal results



Marginal results

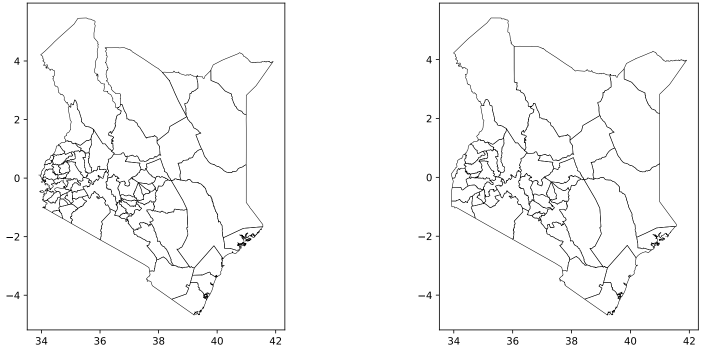


Conclusion: ecological inference

- New ecological inference method through Bayesian distribution regression
- Scalable to millions of observations through random features
- Good empirical results
- Realistic uncertainty intervals
- Simple method [off-the-shelf tools]
- Python package by Danica Sutherland and replication code
- Next steps (before Biden-Trump 2024!): fully Bayesian version of multinomial model, learning richer feature representations, validation on ground truth

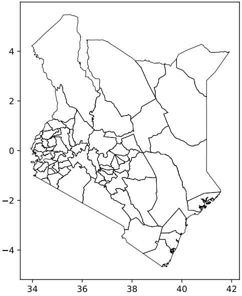
Encoding GP aggregates and change-of-support problem

Kenya: boundaries before and after 2010



aggVAE[3]: what are we solving?

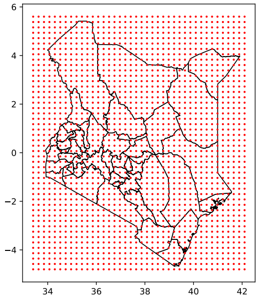
- Adjacency-based models assume heterogeneity.
- Changing boundaries: change-of-support.



[3]E Semenova, S Mishra, S Bhatt, S Flaxman, and HJT Unwin, "Deep learning and MCMC with aggVAE for shifting administrative boundaries: mapping malaria prevalence in Kenya", UAI 2023 workshop "Epistemic Uncertainty in Artificial Intelligence" Proceedings, Publisher: Springer, LNAI (Lecture Notes in Artificial Intelligence); https://arxiv.org/abs/2305.19779

Computational grid

- Create fine spatial grid $\{g_1, ...g_n\}$ over the domain of interest:

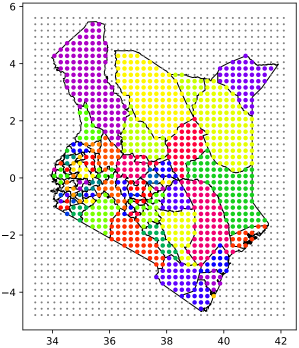


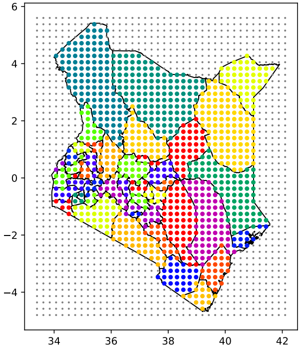
Computational grid

- Draw GP evaluations over the grid:

$$
\begin{aligned}
f &= \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} \sim \text{MVN}(0, \Sigma), \\
f_j &= f(g_j), \\
\Sigma_{jk} &= \sigma^2 \exp\left(-\frac{d_{jk}^2}{2l^2}\right), \\
d_{jk} &= ||g_j - g_k||
\end{aligned}
$$

Attribution of grid points over polygons





Computing GP aggregates over polygons

For each district (polygon) $p_i, i = 1, ..., K$, compute

$$f_{\text{aggGP}}^{p_i} = \int_{p_i} f(s)ds \approx c \sum_{g_j \in p_i} f_j = c\bar{f}_{\text{aggGP}}^{p_i}.$$

Spatial random effect:

$$f_{\text{aggGP}} = \begin{pmatrix} f_{\text{aggGP}}^{p_1} \\ \vdots \\ f_{\text{aggGP}}^{p_K} \end{pmatrix} = Mf \in \mathbb{R}^K,$$

$$M : \quad m_{ij} = I_{\{g_j \subset p_i\}}.$$

Joint encoding of priors

To tackle the the change-of-support problem, encode $\bar{f}^{\text{old}}_{\text{aggGP}}$ and $\bar{f}^{\text{new}}_{\text{aggGP}}$ jointly:

$$\bar{f}^{\text{joint}}_{\text{aggGP}} = \begin{pmatrix} \bar{f}^{p_1^{\text{old}}}_{\text{aggGP}} \\ \cdots \\ \bar{f}^{p_{K_1}^{\text{old}}}_{\text{aggGP}} \\ ----- \\ \bar{f}^{p_1^{\text{new}}}_{\text{aggGP}} \\ \bar{f}^{p_{K_2}^{\text{new}}}_{\text{aggGP}} \end{pmatrix} = \begin{pmatrix} M^{\text{old}} f \\ M^{\text{new}} f \end{pmatrix} \in \mathbb{R}^{K_1+K_2}.$$

'aggVAE' workflow

- Fix spatial structure of areal units as a collection of polygons $P = \{p_1, \dots, p_k\}$.
- Create an aritificial computational grid of sufficient granularity $G = \{g_1, \dots, g_n\}$.
- Pre-compute the matrix of indicators $M, \quad m_{ij} = I_{\{g_j \subset p_i\}}$.
- Draw GP evaluations over G using a selected kernel $k(.,.)$: $f = (f_1, \dots f_n)^T$.
- Compute GP aggregates at the level of P : $f_{\text{aggGP}} = cMf$
- Train PriorVAE on f_{aggGP} draws to obtain f_{aggVAE} priors.
- Use f_{aggVAE} at inference stage within MCMC.

Mapping malaria prevalence in Kenya

- **Model** Malaria prevalence $\theta_i, i \in 1, \ldots K$ is inferred using the Negative Binomial distribution

$$\begin{cases} n_i^{\text{pos}} & \sim \text{NegBin}(n_i^{\text{tests}}, \theta_i), \\ \text{logit}(\theta_i) & = b_0 + f_{\text{aggGP}}^{p_i}. \end{cases}$$

  where n_i^{tests} and n_i^{pos} are the number of total and positive RDT tests, correspondingly.

- **Inference.** Perform MCMC inference using f_{aggVAE} instead of f_{aggGP}.

Results

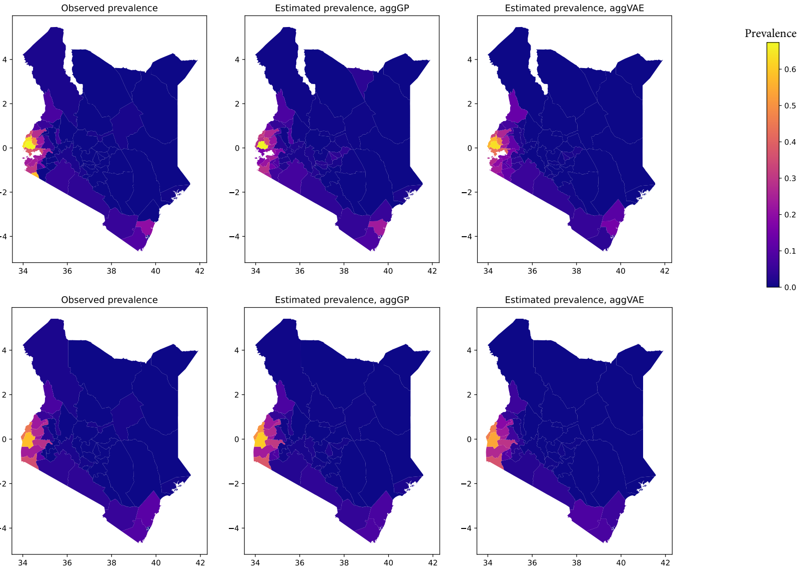
Comparison of MCMC for models with f_{aggGP} and f_{aggVAE} using 200 warm-up steps and 1000 iterations:

Model of the spatial random effect	Elapsed time	Average effective sample size of the random effects
aggGP	**15h***	129
aggVAE	**5s**	231

Table: Model comparison.

* aggGP model has not converged: $\hat{R} = 1.4$.

Results



Observed prevalence
Estimated prevalence, aggGP
Estimated prevalence, aggVAE
Prevalence
Observed prevalence
Estimated prevalence, aggGP
Estimated prevalence, aggVAE

From distribution regression to aggregated GPs[4]

Theorem. Consider a Gaussian process $g \sim \mathcal{GP}(0, \rho)$ with kernel $\rho(P, Q) = \langle \mu_P, \mu_Q \rangle_{\mathcal{H}_k}$ and $f \sim \mathcal{GP}(0, k)$.

Then for any $\Pi_1, \ldots, \Pi_n \in \mathcal{P}(\mathcal{X})$:

$$\left(\int f \mathrm{d}\Pi_1, \ldots, \int f \mathrm{d}\Pi_n \right) \quad \stackrel{d}{=} \quad (g(\Pi_1), \ldots, g(\Pi_n))$$

because $\rho(P, Q) = \int \int k(x, x') \mathrm{d}P(x) \mathrm{d}Q(x')$ for any $P, Q \in \mathcal{P}(\mathcal{X})$.

[4] See Zhu et al, "Aggregated Gaussian Processes with Multiresolution Earth Observation Covariates," https://arxiv.org/abs/2105.01460

From distribution regression to aggregated GPs

Theorem.

$$\left(\int f \,\mathrm{d}\Pi_1, \ldots, \int f \,\mathrm{d}\Pi_n\right) \quad \stackrel{d}{=} \quad (g(\Pi_1), \ldots, g(\Pi_n))$$

Remark. This justifies ecological inference aka disaggregation: for a single individual $x \in \mathcal{X}$, i.e. a point mass $\Pi = \delta_x$,

$$f(x) = \int f \,\mathrm{d}\Pi \stackrel{d}{=} g(\Pi) = g(\delta_x)$$

$\rightarrow$ we are justified in asking for **individual-level** predictions from a distribution regression / aggregated GP model!

Quiz. Does $g(P) = \langle f, \mu_P \rangle_{\mathcal{H}_k} = \int f dP$?

5

Quiz. Does $g(P) = \langle f, \mu_P \rangle_{\mathcal{H}_k} = \int f dP$?

No! f lies outside $\mathcal{H}_k$ almost surely[5]

[5] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. "Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences." arXiv:1807.02582

Quiz. Does $g(P) = \langle f, \mu_P \rangle_{\mathcal{H}_k} = \int f dP$?

No! f lies outside $\mathcal{H}_k$ almost surely[5]

Does it matter?

[5] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. "Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences." arXiv:1807.02582

Open questions

- What if $\rho(P, Q)$ is a nonlinear kernel, e.g.:

$$\rho(P, Q) = \exp(-\|\mu_P - \mu_Q\|^2)$$

- Can representation learning do better? Deep generative models?
- But what if we care about uncertainty? Fully Bayesian inference?
- Satellite imagery for disaggregation, see: Law, Sejdinovic, Cameron, Lucas, Flaxman, Battle, Fukumizu, "Variational Learning on Aggregate Outputs with Gaussian Processes" (NeurIPS 2018)
- Assessing sources of bias in survey data, see: Bradley, Kuriwaki, Isakov, Sejdinovic, Meng, and Flaxman, "Unrepresentative big surveys significantly overestimated US vaccine uptake" (Nature 2021)

Recap

- Distribution regression for ecological inference
- Encoding GP aggregates and change-of-support
- From distribution regression to aggregated GPs

Collaborators

Machine Learning & Global Health (MLGH) network

MACHINE LEARNING
& GLOBAL HEALTH NETWORK

- Juliette Unwin (Bristol)
- Elizaveta Semenova, Leonid Chindelevitch, Samir Bhatt (Imperial College London)
- Swapnil Mishra (National University of Singapore)

Thank you!

- www.sethrf.com