Data integration using constrained Gaussian process models with applications to nuclear physics

Shuang Zhou
Arizona State University
(Joint work with P. Ray, A. Bhattacharya and D. Pati)

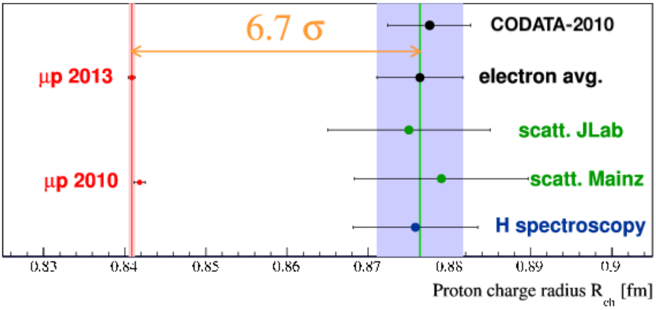
ISNET-9, Dept. of Physics, Washington University in St. Louis
May 25, 2023

Outline of the talk

- Motivating data: The proton radius puzzle
- Model: Hierarchical models for grouped responses; Incorporate shape constraints using Gaussian processes with a basis expansion
- Simulations & real applications

Motivating data

Motivation: Proton radius puzzle



RP et al., Nature 466, 213 (2010); Science 339, 417 (2013); ARNPS 63, 175 (2013).

- Old results from the electron scattering experiments have determined the proton radius to be $\sim$ 0.875 fm
- In 2010 high precision results from Muonic Lamb shift expt. estimated the proton radius as $\sim$ 0.844 fm; supported by $\sim$ 0.831 fm *(Nature, 2019)*, $\sim$ 0.845 fm *(Phys Rev C, 2019)* $\sim$ 0.848 fm *(Science, 2020)*

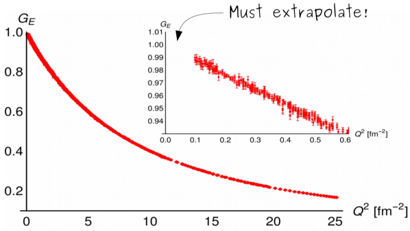
Electron scattering experiment

Electron scattering experiment

- Proton form factor G_E curve as a function of potential Q^2 is difficult to obtain analytically
- The proton radius r_p is related to the derivative of the G_E curve at $Q^2 = 0$,

$$r_p := \sqrt{-6 \frac{dG_E(Q^2)}{dQ^2}|_{Q^2=0}}$$

- Scattering experiment: noisy data obtained for G_E and Q^2



- Impossible to measure G_E for $Q^2 \approx 0$
- Puzzle lies in the extraction of the proton radius from the scattering data

More about the elec. scatt. experiment

- The electric form factor G_E as a function of potential Q^2 is "continuously monotone" with a fixed intercept

$$(-1)^n G_E^{(n)}(Q^2) > 0 \quad and \quad G_E(Q^2 = 0) = 1$$

- Data collected from $T = 34$ from difference sources (with known lables)

- Multiplicative uncertainties in measurements of form factor:

$$G_{E_t}^{obs} = n_{0t}\, G_{E_t}, \quad t = 1, \ldots T$$

where normalization parameters $\{n_0\}$ are unknown (close to 1), varying across difference sources

Existing methods and issues

- Existing methods: Parametric models such as *monopole, dipole, polynomial* (Robust OLS)
- Results can be sensitive to the particular parametric model used
- The error structure is still less understood
- New method: Flexible Bayesian semi-parametric model to incorporate the constraints and to detect the normalization parameters

Modeling with a basis representation

Model framework

- Grouped observation pairs $\{y_{ti}, x_{ti}\}$ from the source $t (t = 1, \ldots, T)$; y_{ti} observed G_E; x_{ti} scaled Q^2, $i = 1, \ldots, n_t$

- Model:

$$y_{ti} \;=\; (1 + \eta_t) f(x_{ti}) + \epsilon_{ti}, \quad \epsilon_{ti} \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad f \in \mathcal{C}_f.$$

- $\{\eta_t\}$ characterize unknown normalization factors

- The constraint set

$$\mathcal{C}_f \;=\; \{f : [0,1] \to \mathbb{R} : \; f(0) = 1, f'(x) < 0, f''(x) > 0, \; \forall\, x\},$$

- Our goal: Characterize the uncertainty in estimating the radius

A basis expansion approach

- For a twice continuously differentiable function

$$f(x) = f(0) + xf'(0) + \int_0^x \int_0^t f''(s)dsdt$$

- Given N equal-spaced knots $\{u_j\}$ and basis $h_j(x)$ (Maatouk & Bay, 2016),

$$f''(x) \approx \sum_{j=0}^{N} f''(u_j)\, h_j(x), \quad \phi_j(x) = \int_0^x \int_0^t h_j(s)dsdt$$

Illustration

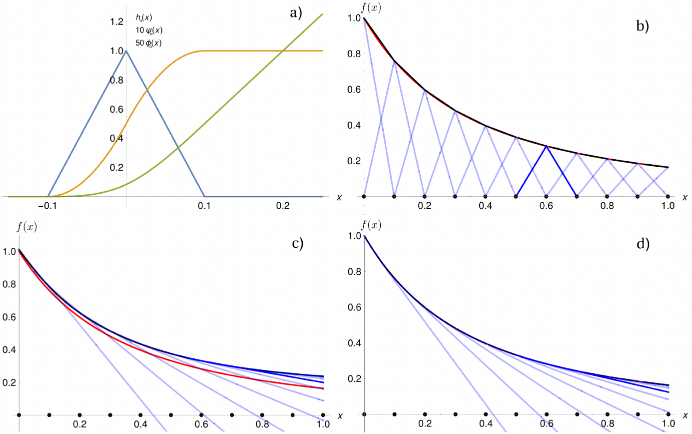


Figure: (a) Functions $h_0(x)$, $\psi_0(x)$ and $\phi_0(x)$; (b) Approximations (black) of the dipole function (Red) using the basis functions $h_j(x)$ on 11 gridpoints between 0 and 1 (black dots). (c)-(d) Approximation (black) of the same dipole function (red) using the basis functions ϕ_j.

A basis expansion approach

- Function approximation

$$f(x) \approx f(0) + xf'(0) + \sum_{j=0}^{N} f''(u_j)\, \phi_j(x)$$

- Re-parameterizing,

$$f_\theta(x) = \ \theta_1 + \theta_2\, x + \sum_{j=0}^{N} \theta_{j+3}\, \phi_j(x)$$

with unknown parameter $\theta = \{\theta_1, \ldots, \theta_{N+3}\}$.

Transferring the constraints

Find equivalent constraint set on coefficients θ:

Lemma

$f \in \mathcal{C}_f$ if and only if $\theta \in \mathcal{C}_\Theta$, where

$$\mathcal{C}_\Theta = \left\{ \theta_1 = 1, \quad \theta_2 + \sum_{j=0}^{N} \theta_{j+3}\, c_j < 0, \quad \theta_{j+3} > 0, \quad j = 0, \ldots, N. \right\}$$

where $c_j = \int_0^1 h_j(x)dx$ for $j = 0, \ldots, N$.

- Finite numbers of linear constraints on unknown coefficients. Easy to implement!

Prior choice and posterior inference

Prior choice

- A natural prior choice is a Gaussian process (GP) prior, $f'' \sim \mathrm{GP}(0, \tau^2 K)$, then

$$\theta_{[3:(N+3)]} = [f''(u_0), \ldots, f''(u_N)]^{\mathrm{T}} \sim \mathcal{N}(0, \tau^2 \Gamma)$$

- Univariate normal prior $\theta_2 \;\sim\; \mathcal{N}(\mu_0, \tau^2)$

- Set the prior distribution on θ as a truncated MVN

$$\theta_2, \theta_{[3:(N+3)]} \mid \tau^2 \quad \sim \quad \Pi\,(\theta_2)\; \Pi(\theta_{[3:(N+3)]})\, \mathbb{1}_{\mathcal{C}_\Theta}\, \big(\theta_2, \theta_{[3:(N+3)]}\big)$$

- Centered normal prior on $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\eta^2)$

Hyperparameter choices

- K: stationary Matérn kernel with smoothness parameter $\nu = 0.5$
- Inverse-gamma priors on τ^2, σ_η^2
- Gamma prior on σ^2
- Consider various choices of $\{N, \ell\}$ for model comparison ($I = \#(N) \times \#(\ell)$)

Posterior inference

- Posterior computation: MCMC algorithm using Gibbs sampling (Elliptical slice sampling to sample from the truncated posterior)

- Model averaging according to model comparison: Use Watanabe-Akaike Information Criterion values WAIC_i under different combinations of $\{N, \ell\}$

- Averaging the estimates with the weights

$$w_i = \frac{\exp(-\mathrm{WAIC}_i/2)}{\sum_{j=1}^{I} \exp(-\mathrm{WAIC}_j/2)}, \quad i = 1, \ldots, I.$$

- The final estimate of the proton radius

$$\tilde{r}_p = \sum_{i=1}^{I} w_i \, \hat{r}_{pi}, \quad \hat{r}_{pi} = S^{-1} \sum_{s=1}^{S} \sqrt{-6\, \theta_2^{(s)} / Q^2_{\max}}$$

Simulation results

Simulation: Data generation

- Set the true radius $r_p = 0.85$ fm
- Synthetic G_E values y_{it}^* from the data-generator (Yan et al., 2018) using Q^2s in the Mainz data
- Generate normalization parameters

$$\eta_t^* \overset{i.i.d.}{\sim} \text{Unif}\,[1 - \delta_0, 1 + \delta_0]$$

- Additive normal errors $\epsilon_{it} \overset{i.i.d.}{\sim} N(0, \sigma_0^2)$
- Observed responses:

$$y_{ti} = (1 + \eta_t^*)\, y_{it}^* + \epsilon_{it}, \quad i = 1, \ldots, n_t, \quad t = 1, \ldots, T.$$

Data separation

Table: Data separation

	n_t	Q^2_{low}	Q^2_{upp}
Group 1	106	0.005	0.0168
Group 2	41	0.0132	0.0249
Group 3	102	0.0147	0.086
Group 4	19	0.0249	0.0386
Group 5	38	0.055	0.0967
Group 6	17	0.0967	0.109
Group 7	104	0.0145	0.0638
Group 8	38	0.0561	0.1817
Group 9	40	0.0626	0.1882
Group 10	62	0.1473	0.2783
Group 11	77	0.0199	0.0747
Group 12	52	0.0747	0.1535
Group 13	42	0.0765	0.3478
Group 14	17	0.0769	0.1112
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Error set-ups

Case I: Large multiplicative errors and small additive errors

- Fix $\sigma_0 = 0.001$ and set $\delta_0 \in \{0.001, 0.003, 0.005\}$

Case II: Small multiplicative error and large additive error

- Fix $\sigma_0 = 0.001$ and set $\delta_0 \in \{0.0001, 0.0005, 0.001\}$

In cases I,II:

- Generate response observations in two scenarios, by taking the first 14 groups (low regime) and the first 28 groups (high regime) of data
- Replicate 50 data sets and fit the model

Results (Case I, low regime)

Table: Posterior (mean) estimates and 95% credible intervals (CI) of radius of the proton over 50 replicated data sets in low regime

	δ_0	0.001	0.003	0.005
N=25	$\hat{r}_p$	0.848	0.849	0.847
	r_{CI}	(0.846,0.851)	(0.842,0.859)	(0.837,0.857)
N=50	$\hat{r}_p$	0.85	0.850	0.855
	r_{CI}	(0.849,0.852)	(0.842,0.859)	(0.842, 0.869)
N=100	$\hat{r}_p$	0.850	0.853	0.851
	r_{CI}	(0.843,0.855)	(0.842,0.871)	(0.844,0.858)
WAIC-wt	$\hat{r}_p$	0.849	0.851	0.851
	r_{CI}	(0.848,0.851)	(0.842,0.862)	(0.844,0.862)

Results (Case I, high regime)

Table: Posterior (mean) estimates and 95% credible intervals (CI) of radius of proton over 50 replicated data sets

	δ_0	0.001	0.003	0.005
N=25	$\hat{r}_p$	0.848	0.847	0.845
	r_{CI}	(0.847,0.850)	(0.844,0.849)	(0.837,0.852)
N=50	$\hat{r}_p$	0.85	0.850	0.847
	r_{CI}	(0.848,851)	(0.846,0.853)	(0.840,0.853)
N=100	$\hat{r}_p$	0.85	0.845	0.847
	r_{CI}	(0.847,0.852)	(0.842,0.850)	(0.840,0.853)
WAIC-wt	$\hat{r}_p$	0.85	0.847	0.847
	r_{CI}	(0.848,0.852)	(0.843,0.851)	(0.839,0.853)

Results (Case II, low regime)

Table: Posterior (mean) estimates and 95% credible intervals (CI) of radius of proton over 50 replicated data sets

	δ_0	0.0001	0.0005	0.001
N=25	$\hat{r}_p$	0.848	0.849	0.848
	r_{CI}	(0.843,0.857)	(0.843,0.858)	(0.841,0.857)
N=50	$\hat{r}_p$	0.852	0.851	0.850
	r_{CI}	(0.843,0.860)	(0.845,0.858)	(0.845,0.858)
N=100	$\hat{r}_p$	0.849	0.855	0.855
	r_{CI}	(0.845,0.855)	(0.849,0.865)	(0.845,0.867)
WAIC-wt	$\hat{r}_p$	0.850	0.852	0.851
	r_{CI}	(0.848,0.852)	(0.844,0.860)	(0.844,0.863)

Results (Case II, high regime)

Table: Posterior (mean) estimates and 95% credible intervals (CI) of radius of proton over 50 replicated data sets

	δ_0	0.0001	0.0005	0.001
N=25	$\hat{r}_p$	0.850	0.849	0.850
	r_{CI}	(0.847,0.853)	(0.846,0.851)	(0.846,0.855)
N=50	$\hat{r}_p$	0.851	0.849	0.849
	r_{CI}	(0.848,0.854)	(0.846,0.852)	(0.846,0.852)
N=100	$\hat{r}_p$	0.851	0.848	0.849
	r_{CI}	(0.847,0.854)	(0.844,0.851)	(0.843,0.854)
WAIC-wt	$\hat{r}_p$	0.851	0.849	0.849
	r_{CI}	(0.847,0.854)	(0.845,0.851)	(0.844,0.854)

WAIC-weighted estimate of η_t^* under $\delta_0 = 0.003$ in case I (low regime)

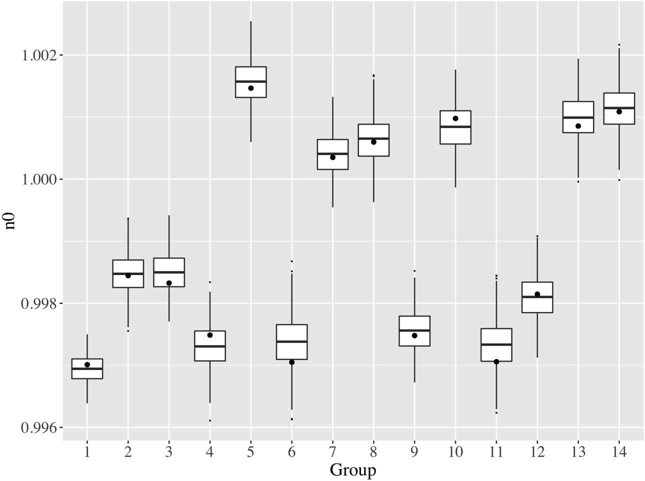


Figure: Box-plot of weighted estimates of normalization parameter per group. Black dots: true; black stars: outliers.

WAIC-weighted estimate of η_t^* under $\delta_0 = 0.005$ in case II (high regime)

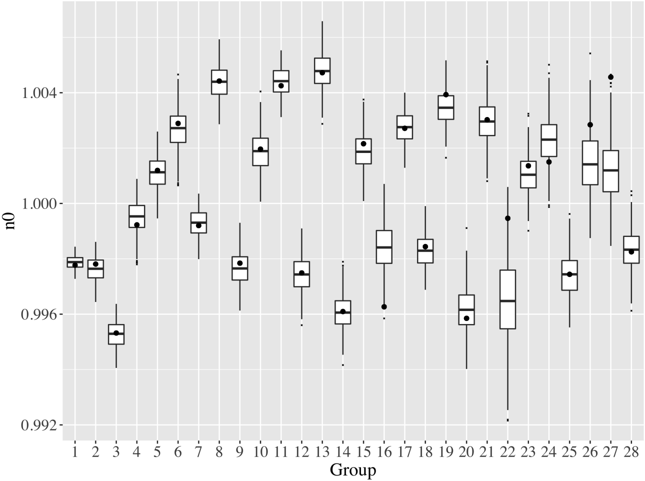


Figure: Box-plot of weighted estimates of normalization parameter per group. Black dots: true; black stars: outliers.

Robustness check under $\delta_0 = 0.003$ (low regime)

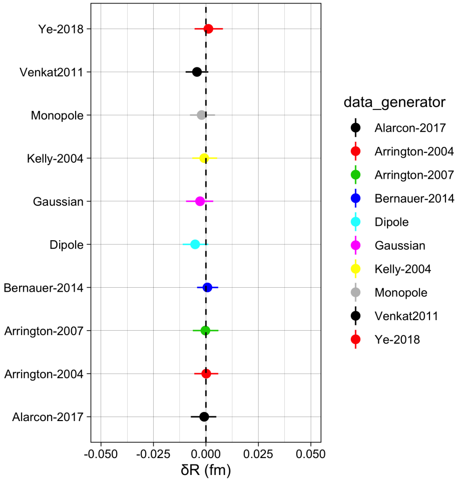
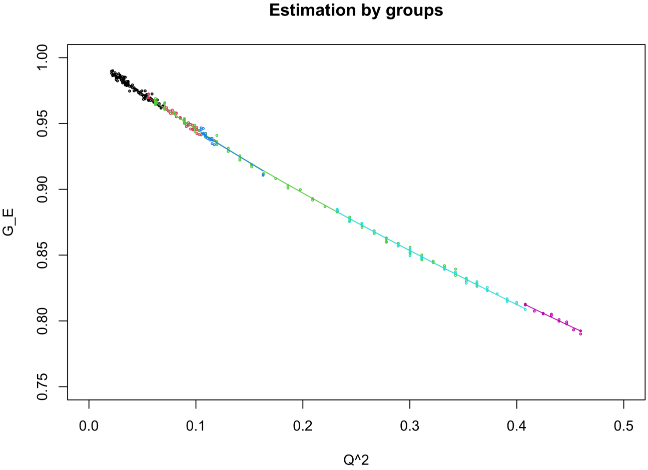


Figure: Posterior estimate with 95% error bar under different data generators.

Real data analysis

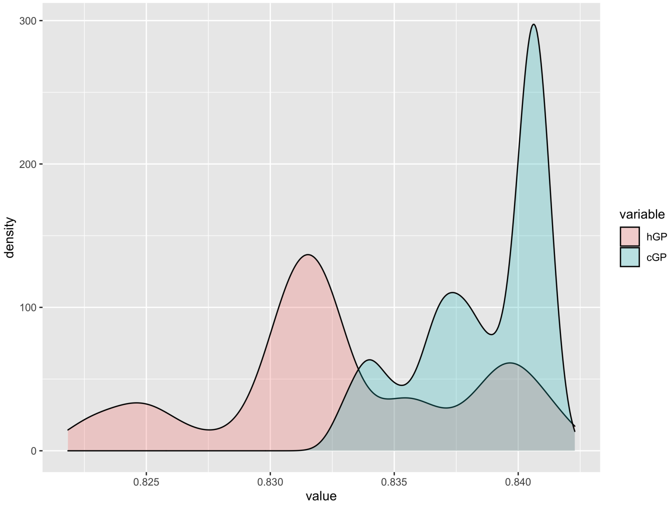
Real data analysis (preliminary)

- CODATA-2010: low regime. Model fit accommodated by recovered normalization parameters:



Real data analysis (preliminary)

- Posterior density plot of the proton radius obtained by the hierarchical model (hGP) and constrained GP (treat normalization parameters universally)



Conclusion & Future work

Summary:

- Develop a hierarchical constrained GP model
- Provide reasonable estimates of the proton radius
- Recover the true normalization parameter of synthetic data

To-dos:

- Update the hyperparameters, make the model more robust
- Model exploration with different choices of basis functions
- Model under heteroscedastic cases
- Extension to multidimensional models

Collaborators

- Palavi Ray (Eli Lily)
- Debdeep Pati (TAMU)
- Anirban Bhattacharya (TAMU)

References

- Revisiting the proton-radius problem using constrained Gaussian processes, *Physical Review C, 2019* (S. Zhou, P. Giulani, J. Piekarewicz, A. Bhattacharya, and D. Pati)
- Robust Gaussian process models for extrapolation of electronic proton radius (SZ, PR, DP, AB)
- Data integration with hierarchical Gaussian processes under constraints (SZ, AB, DP)
- Code: https://github.com/szh0u/Constrained-GP

Thank you!