A short overview of preference learning with Gaussian process based approaches

Dario Azzimonti

Dalle Molle Institute for Artificial Intelligence

IDSIA USI SUPSI

MASCOT-NUM 2024

3rd April 2024 - Giens Peninsula, Hyères, France

Who

The works presented here are the results of collaborations with

- Alessio Benavoli, School of Computer Science and Statistics. Trinity College, Dublin, Ireland.

- Dario Piga, IDSIA, Lugano, Switzerland.

- Diego Moranda, IDSIA, Lugano, Switzerland.

The following paper provides an extensive tutorial on the subject:

Benavoli, A., & Azzimonti, D. 2024. A tutorial on learning from preferences and choices with Gaussian Processes. *arXiv:2403.11782*.

All algorithms are implemented in Python and available at

prefGP: A Python package for preference and choice learning with Gaussian processes.

Outline

How to model users' preferences?

Often, we express preferences and not measurable ratings.

Examples:

- Buying a new laptop;
- Understanding travel preferences of a group of users;
- Learning which is the best gaming platform for a user.

Types of preference data

We can broadly divide preference data into three types

1) Object preferences

   - pairwise comparison between objects;
   - features associated with the items;
   - *Example:* buying a new laptop.

2) Label preferences

   - pairwise comparison between labels associated with objects;
   - features associated with the users;
   - *Example:* Learning the best gaming platform for a user.

3) Choice data

   - a user selects items from a fixed set of options

We will see how to model those with Gaussian processes.

What is a preference?

Given a finite set $\mathcal{X}$, a (strict) preference is a *binary relation* $\succ$ which is

- *symmetric:* $\forall x, y \in \mathcal{X}$ if $x \succ y$ then not $y \succ x$
- *negatively transitive:* if $x \not\succ z$ and $z \not\succ y$ then $x \not\succ y$

Representation via utility functions

For any set $\mathcal{X}$ and preference relation $\succ$ on $\mathcal{X}$, the *utility* function $u : \mathcal{X} \rightarrow \mathbb{R}$ represents $\succ$ if

$$x \succ y \quad \textit{iff} \quad u(x) > u(y)$$

This representation is *not unique*: for any increasing function g

$$x \succ y \;\; \textit{iff} \;\; u(x) > u(y) \;\; \textit{iff} \;\; g(u(x)) > g(u(y)).$$

How to learn from preference data?

Objective of preference learning: learn the *latent* utility function.

There are many models available:

- Random utility models (McFadden, 1974; Hunter, 2004)
- Generalised linear models (Critchlow and Fligner, 1991)
- Support vector machines (Maldonado et al., 2015)

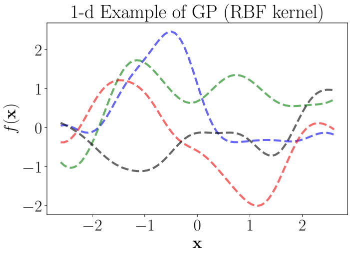
Here we focus on **Gaussian process based models**.

What is a Gaussian process?

A GP is a stochastic process $(f(\mathbf{x}))_{\mathbf{x}\in\mathcal{X}} \sim GP(\mu(\mathbf{x}), k(\mathbf{x},\mathbf{x}'))$ defined by its mean (μ often $\mu(\mathbf{x}) \equiv 0$) and kernel ($k(\mathbf{x},\mathbf{x}')$) functions.

- For a fixed $\mathbf{x}$, $f(\mathbf{x})$ is a Gaussian r.v.
- For a fixed ω, $f(\omega, \mathbf{x})$ is a function of $\mathbf{x}$.
- k controls the properties of f.

More



Realization of $f \sim GP(0, k)$, RBF kernel.

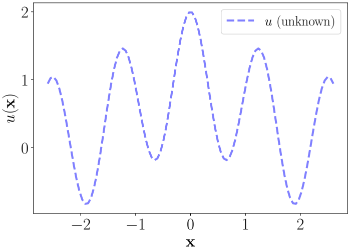
We use GPs to define a prior distribution over functions.

Rasmussen & Williams. 2006. Gaussian processes for machine learning. *MIT Press.*

What is a Gaussian process model?

A Bayesian, non-parametric model to fit unknown functions.

Problem: fit unknown function.



Start: Prior on functions $p(f)$.

Observe: the function at few values

$\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n, \quad \mathbf{x}_i \in D \subset \mathbb{R}^d$

$y_i = u(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_n^2)$

What is a Gaussian process model?

A Bayesian, non-parametric model to fit unknown functions.

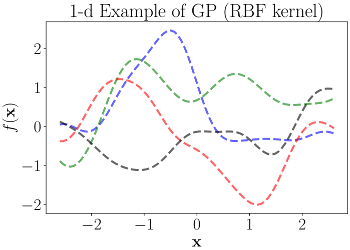
Problem: fit unknown function.

Start: Prior on functions $p(f)$.

Observe: the function at few values

$\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n, \quad \mathbf{x}_i \in D \subset \mathbb{R}^d$

$y_i = u(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_n^2)$



What is a Gaussian process model?

A Bayesian, non-parametric model to fit unknown functions.

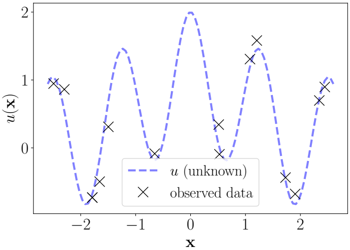
Problem: fit unknown function.

Start: Prior on functions $p(f)$.

Observe: the function at few values

$\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n, \quad \mathbf{x}_i \in D \subset \mathbb{R}^d$

$y_i = u(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_n^2)$



What is a Gaussian process model?

A Bayesian, non-parametric model to fit unknown functions.

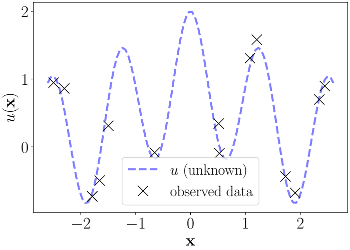
Problem: fit unknown function.

Start: Prior on functions $p(f)$.

Observe: the function at few values

$\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n, \quad \mathbf{x}_i \in D \subset \mathbb{R}^d$

$y_i = u(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_n^2)$



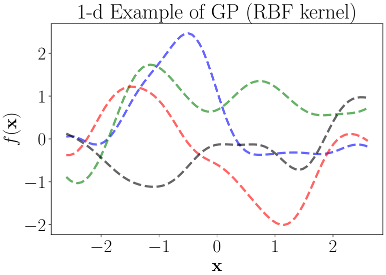
Update the prior include data knowledge with Bayes theorem!

$$p(f|\mathbf{y}) = \frac{p(\mathbf{y}|f)p(f)}{p(\mathbf{y})}$$

How to incorporate data?

Bayes theorem! $p(f|\mathbf{y}) = \frac{p(\mathbf{y}|f)p(f)}{p(\mathbf{y})}$

Prior: $p(f) = GP(0, k)$



Observe $y_i = f(\mathbf{x}_i) + \epsilon_i$, $i = 1, \ldots, n$ with $\epsilon_i \sim N(0, \sigma^2)$

$\mathbf{y}_n = (y_1, \ldots, y_n)$ and

$\mathbf{f}_n = f(X_n) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$

Likelihood: $p(\mathbf{y}_n|\mathbf{f}_n) = N_n(\mathbf{f}_n, \sigma_n^2 I_n)$

Posterior: $p(f(\mathbf{x})|\mathbf{y}_n) = GP(\mu_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}))$

How to incorporate data?

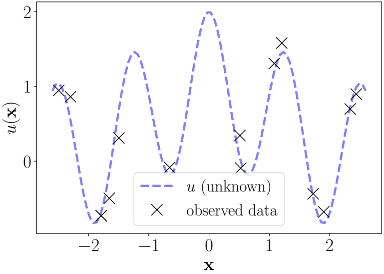
Bayes theorem! $p(f|\mathbf{y}) = \frac{p(\mathbf{y}|f)p(f)}{p(\mathbf{y})}$

Prior: $p(f) = GP(0, k)$

Observe $y_i = f(\mathbf{x}_i) + \epsilon_i$, $i = 1, \ldots, n$ with $\epsilon_i \sim N(0, \sigma^2)$

$\mathbf{y}_n = (y_1, \ldots, y_n)$ and

$\mathbf{f}_n = f(X_n) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$



Likelihood: $p(\mathbf{y}_n|\mathbf{f}_n) = N_n(\mathbf{f}_n, \sigma_n^2 I_n)$

Posterior: $p(f(\mathbf{x})|\mathbf{y}_n) = GP(\mu_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}))$

How to incorporate data?

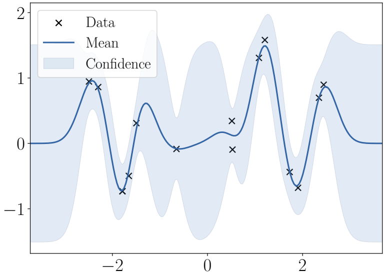
Bayes theorem! $p(f|\mathbf{y}) = \frac{p(\mathbf{y}|f)p(f)}{p(\mathbf{y})}$

Prior: $p(f) = GP(0, k)$

Observe $y_i = f(\mathbf{x}_i) + \epsilon_i$,
$i = 1, \ldots, n$ with $\epsilon_i \sim N(0, \sigma^2)$

$\mathbf{y}_n = (y_1, \ldots, y_n)$ and

$\mathbf{f}_n = f(X_n) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$



Likelihood: $p(\mathbf{y}_n|\mathbf{f}_n) = N_n(\mathbf{f}_n, \sigma_n^2 I_n)$

Posterior: $p(f(\mathbf{x})|\mathbf{y}_n) = GP(\mu_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}))$

Why use Gaussian processes?

- Bayesian model $\Rightarrow$ easy to update
- Non-parametric model $\Rightarrow$ flexible
- The posterior is analytical for (Gaussian) regression!

$$\mu_n(\mathbf{x}) = k(\mathbf{x}, X_n)(k(X_n, X_n) + \sigma_n^2 I_n)^{-1}\mathbf{y}_n$$
$$k_n(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, X_n)(k(X_n, X_n) + \sigma_n^2 I_n)^{-1} k(X_n, \mathbf{x}')$$

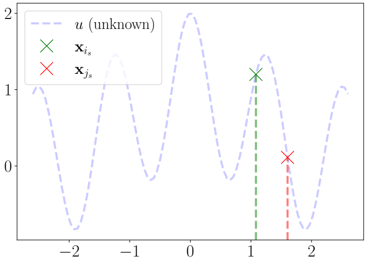
What if we **do not observe** output values $\mathbf{y}_n$?

A model for consistent object preferences

Preference: $\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s}$ for two items $\mathbf{x}_{i_s} \neq \mathbf{x}_{j_s} \in \mathcal{X}$;

Assumption: underlying hidden function u such that

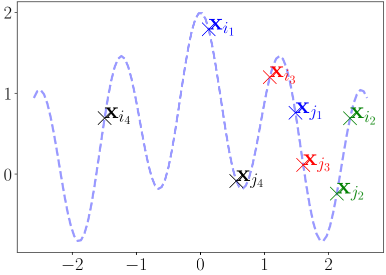
$$u(\mathbf{x}_{i_s}) \geq u(\mathbf{x}_{j_s}) \quad \text{if} \quad \mathbf{x}_{i_s} \succ \mathbf{x}_{j_s}$$



A model for consistent object preferences

Data: $m = 50$ pairwise preferences

$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$



Prior: $u \sim GP(0, k)$;

Likelihood:

$$p(\mathcal{D}_m | u(X)) = \prod_{s=1}^{m} I_{\{u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s}) > 0\}}(u(X)) = I_{\{Wu(X) > 0\}}(u(X)).$$

A model for consistent object preferences

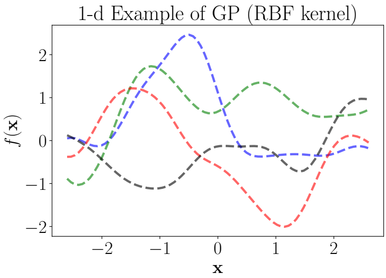
Data: $m = 50$ pairwise preferences

$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$

Prior: $u \sim GP(0, k)$;

Likelihood:

$$p(\mathcal{D}_m | u(X)) = \prod_{s=1}^{m} I_{\{u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s}) > 0\}}(u(X)) = I_{\{Wu(X) > 0\}}(u(X)).$$



A model for consistent object preferences

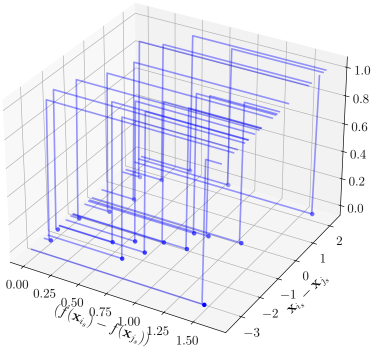
Data: $m = 50$ pairwise preferences

$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$

Prior: $u \sim GP(0, k)$;

Likelihood:

$$p(\mathcal{D}_m | u(X)) = \prod_{s=1}^{m} I_{\{u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s}) > 0\}}(u(X)) = I_{\{Wu(X) > 0\}}(u(X)).$$



A model for consistent object preferences

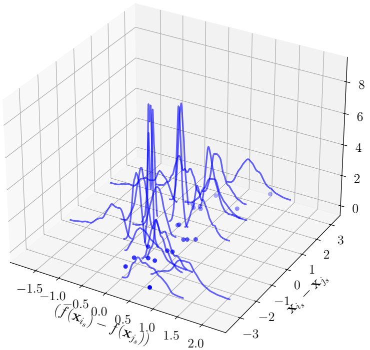
Data: $m = 50$ pairwise preferences

$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$

Prior: $u \sim GP(0, k)$;

Likelihood:

$$p(\mathcal{D}_m | u(X)) = \prod_{s=1}^{m} I_{\{u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s}) > 0\}}(u(X)) = I_{\{Wu(X) > 0\}}(u(X)).$$



Posterior: $p(u(X)|\mathcal{D}_m) = TN_{\{Wu(X) \geq 0\}}(u(X); \mathbf{0}, K_{\boldsymbol{\theta}}(X, X))$

Note: $TN_{\{Wu(X) \geq 0\}}(u(X))$, truncated normal in $\{Wu(X) \geq 0\}$

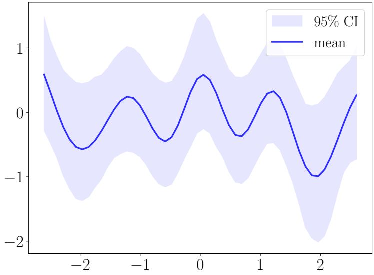
A model for consistent object preferences

Predictive posterior:

$$p(u(X^*)|\mathcal{D}_m) = \int p(u(X^*)|u(X))\; TN_{\{Wu(X)\geq 0\}}(u(X); \mathbf{0}, K_{\boldsymbol{\theta}}(X,X))du(X),$$

where:

- W matrix of preferences
  $W_{i_s} = 1,\; W_{j_s} = -1$ for $\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s}$
- $p(u(X^*)|u(X))$ conditional Gaussian distribution



A model for consistent object preferences

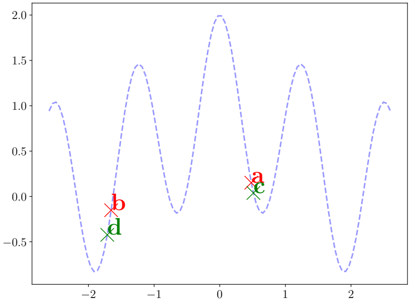
Consistent object preferences: underlying hidden function u such that

$$u(\mathbf{x}_{i_s}) \geq u(\mathbf{x}_{j_s}) \quad \text{if} \quad \mathbf{x}_{i_s} \succ \mathbf{x}_{j_s}$$

What if we have this data?

$\mathbf{a} \succ \mathbf{b}$

$\mathbf{d} \succ \mathbf{c}$

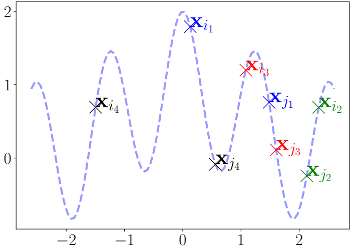


Issues: what if preferences are not consistent?

A model for erroneous preferences

Data: m pairwise preferences

$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$



Prior: $u \sim GP(0, k)$;

Likelihood: $p(\mathcal{D}_m | u(X)) =$

$$\prod_{s=1}^{m} \Phi\left(u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s})\right) =$$

$$\Phi_m(Wu(X)).$$

Posterior: $p(u(X)|\mathcal{D}_m) = \textbf{?}$

Chu, W. and Ghahramani, Z. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning* (ICML '05).

A model for erroneous preferences

Data: m pairwise preferences

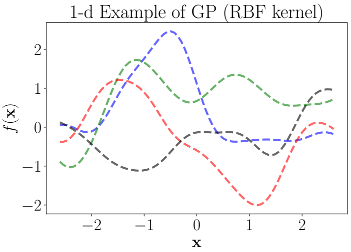
$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$

Prior: $u \sim GP(0, k)$;

Likelihood: $p(\mathcal{D}_m | u(X)) =$

$$\prod_{s=1}^{m} \Phi\left(u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s})\right) =$$

$$\Phi_m(Wu(X)).$$



Posterior: $p(u(X)|\mathcal{D}_m) = $ **?**

Chu, W. and Ghahramani, Z. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning* (ICML '05).

A model for erroneous preferences

Data: m pairwise preferences

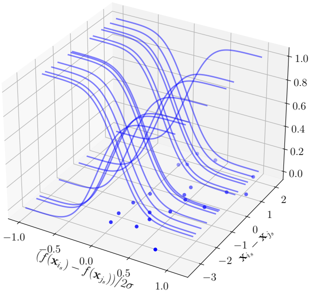
$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$

Prior: $u \sim GP(0, k)$;

Likelihood: $p(\mathcal{D}_m | u(X)) =$

$$\prod_{s=1}^{m} \Phi\left(u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s})\right) =$$

$$\Phi_m(Wu(X)).$$



Posterior: $p(u(X)|\mathcal{D}_m) = ?$

Chu, W. and Ghahramani, Z. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning* (ICML '05).

A model for erroneous preferences

Data: m pairwise preferences

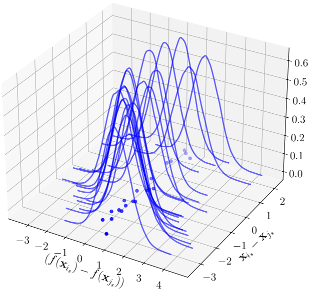
$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$

Prior: $u \sim GP(0, k)$;

Likelihood: $p(\mathcal{D}_m | u(X)) =$

$$\prod_{s=1}^{m} \Phi\left(u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s})\right) =$$

$$\Phi_m(Wu(X)).$$



Posterior: $p(u(X)|\mathcal{D}_m)$ = **?**

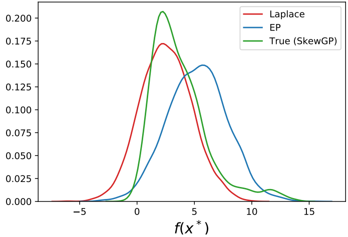
Chu, W. and Ghahramani, Z. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning* (ICML '05).

Let's zoom in, what does the posterior look like?

Let's compute the posterior with Monte Carlo simulations

$$p(f(X)|\mathcal{D}) \propto p(\mathcal{D}|f(X))p(f)$$

- GP prior, preference data
- compute posterior with MC, EP and Laplace approximations
- plot posterior (sample) distribution at one point



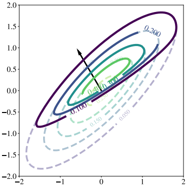
The posterior is a skewed distribution! How do we model it?

H. Nickisch and C. E. Rasmussen. 2008. Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research* 9 2035-2078

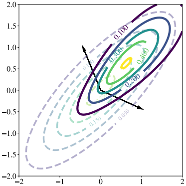
SkewGPs - what is a skew normal?

A vector $\mathbf{z} \in \mathbb{R}^p \sim SUN_{p,s}(\boldsymbol{\xi}, \Omega, \Delta, \gamma, \Gamma)$ if its density[1] is

$$p(\mathbf{z}) = \phi_p(\mathbf{z} - \boldsymbol{\xi}; \Omega) \frac{\Phi_s\left(\gamma + \Delta^T \Omega^{-1}(\mathbf{z} - \boldsymbol{\xi}); \Gamma - \Delta^T \Omega^{-1} \Delta\right)}{\Phi_s\left(\gamma; \Gamma\right)},$$



$s = 1,\ \Gamma = 1,\ \Delta = [0.8, 0.3]^T$



$s = 2,\ \Gamma_{1,2} = -0.3,\ \Delta = \begin{bmatrix} 0.3 & 0.8 \\ 0.8 & 0.3 \end{bmatrix}$

The parameters $\gamma \in \mathbb{R}^s, \Gamma \in \mathbb{R}^{s \times s}, \Delta^{p \times s}$ control the skewness. More

1. See Azzalini (2013) for a complete reference on skew-normal distributions.

SkewGPs - definition

$$f \sim \text{SkewGP}_s(\xi, \Omega, \Delta, \gamma, \Gamma)$$

if, for any sequence of n points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, the vector $[f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)] \in \mathbb{R}^n \sim SUN(\xi(X), \Omega(X,X), \Delta(X), \gamma, \Gamma)$ given by

$$\xi(X) := \begin{bmatrix} \xi(\mathbf{x}_1) \\ \xi(\mathbf{x}_2) \\ \vdots \\ \xi(\mathbf{x}_n) \end{bmatrix}, \quad \Omega(X,X) := \begin{bmatrix} \Omega(\mathbf{x}_1,\mathbf{x}_1) & \Omega(\mathbf{x}_1,\mathbf{x}_2) & \dots & \Omega(\mathbf{x}_1,\mathbf{x}_n) \\ \Omega(\mathbf{x}_2,\mathbf{x}_1) & \Omega(\mathbf{x}_2,\mathbf{x}_2) & \dots & \Omega(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & \dots & \vdots \\ \Omega(\mathbf{x}_n,\mathbf{x}_1) & \Omega(\mathbf{x}_n,\mathbf{x}_2) & \dots & \Omega(\mathbf{x}_n,\mathbf{x}_n) \end{bmatrix},$$

$$\Delta(X) := \begin{bmatrix} \Delta(\mathbf{x}_1) & \Delta(\mathbf{x}_2) & \dots & \Delta(\mathbf{x}_n) \end{bmatrix}.$$

- Well defined if $M = \begin{bmatrix} \Gamma & \Delta(X) \\ \Delta(X)^T & \Omega(X,X) \end{bmatrix}$ is **positive definite** for all X.
- If $\Delta \equiv 0$ or the latent dimension $s = 0$, then SkewGP reduces to a GP.

SkewGPs and probit likelihoods

Prior $u(\mathbf{x}) \sim GP(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'))$;

Likelihood $p(W \mid u(X)) = \Phi_m(Wu(X))$ where $W \in \mathbb{R}^{m \times n}$.

The **posterior distribution** of $u(X)$ is a SUN:

$$p(u(X)|W) = \text{SUN}_{n,m}(\tilde{\xi}, \tilde{\Omega}, \tilde{\Delta}, \tilde{\gamma}, \tilde{\Gamma}) \quad \text{with}$$

$$\tilde{\xi} = \boldsymbol{\xi}, \quad \tilde{\Omega} = \Omega, \quad \tilde{\Delta} = \Omega W^T, \quad \tilde{\gamma} = W\xi, \quad \tilde{\Gamma} = W\Omega W^T + I_m,$$

Durante, D. 2019. Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4), 765–779.

Alessio Benavoli, D. A., Dario Piga. 2020. Skew Gaussian Processes for Classification. *Machine Learning*, 109(9):1877–1902.

Preferential learning with SkewGPs - Model

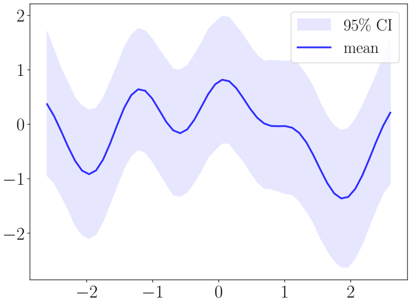
Data: $m = 50$ pairwise preferences

$$\mathcal{D}_m = \{\mathbf{x}_{i_s} \succ \mathbf{x}_{j_s} : s = 1, \ldots, m\}$$

Prior: $u \sim GP(0, k)$;

Likelihood: $p(\mathcal{D}_m | u(X)) =$

$$\prod_{s=1}^{m} \Phi\left(\frac{u(\mathbf{x}_{i_s}) - u(\mathbf{x}_{j_s})}{\sqrt{2}\sigma}\right)$$



Posterior: $p(u(X)|\mathcal{D}_m) = SkewGP(\tilde{\xi}, \tilde{\Omega}, \tilde{\Delta}, \tilde{\gamma}, \tilde{\Gamma})$

Predictive posterior: rejection-free sampling method (*lin-ess*) More

Transportation mode preference

Objective: predict transportation mode for a subject on a journey.

Four transportation modes: air, bus, car and train

Information about the journey:

- `dist` the distance of the trip;
- `cost` the monetary cost;
- `ivt` in-vehicle-time;
- `ovt` out-of-vehicle time;
- `freq` frequency.

Only subjects with the same individual features (income and type of trip).

Transportation mode preference

Example of data

case	alt	choice	dist	cost	ivt	ovt	freq
129	train	0	387	58.25	316	59	2
129	air	0	387	145.80	56	85	9
129	bus	0	387	26.67	301	58	8
129	car	1	387	73.53	251	0	0

Preferences encoded as pairwise comparisons

$$\mathbf{x}_{car} \succ \mathbf{x}_{air}, \;\; \mathbf{x}_{car} \succ \mathbf{x}_{bus}, \;\; \mathbf{x}_{car} \succ \mathbf{x}_{train}$$

Transportation mode preference - Linear model

Dataset: 679 cases, 5 input features

Prior: GP $u \sim GP(0, k)$ with $k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{5} \sigma_j^2 x_j x_j'$ (linear kernel)

Likelihood: erroneous preference likelihood $p(\mathcal{D}_m | u(X)) = \Phi_m(Wu(X))$

Posterior: SkewGP with known analytical parameters.

Results: 0.84 Accuracy over 10 repetition (70-30 train/test split)

Note: this is equivalent to a Bayesian linear model

$$u(\mathbf{x}) = \beta_1 \texttt{dist} + \beta_2 \texttt{cost} + \beta_3 \texttt{ivt} + \beta_4 \texttt{ovt} + \beta_5 \texttt{freq}$$

Transportation mode preference - RBF model

Dataset: 679 cases, 5 input features

Prior: GP $u \sim GP(0, k)$ with RBF kernel

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\sum_{j=1}^{5} \frac{(x_j - x_j')^2}{2\ell_j^2} \right)$$

Likelihood: erroneous preference likelihood $p(\mathcal{D}_m | u(X)) = \Phi_m(Wu(X))$

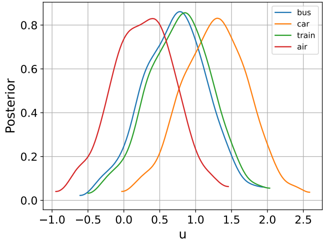
Posterior: SkewGP with known analytical parameters.

Results: 0.90 Accuracy over 10 repetition (70-30 train/test split)

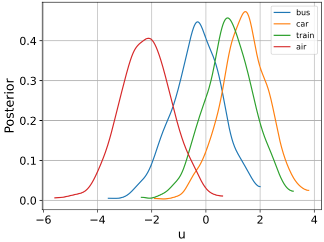
Transportation mode preference - predictive distributions

alt	dist	cost	ivt	ovt	freq
bus	175	18.43	119	75	4
car	175	33.25	115	0	0
train	175	38.15	109	65	5
air	175	155.90	60	118	6

Test example: four transportation modes.



Linear kernel



RBF kernel

GPs for object preferences

Coherent preferences model

- Requires coherent observations (no user errors);
- Likelihood based on indicator functions;
- Truncated normal posterior distribution.

Erroneous preferences model

- Accounts for possible user's errors;
- Probit likelihood;
- SkewGP posterior distribution.

More models (limit of discernibility, two-argument function) available in
Benavoli, A., & Azzimonti, D. 2024. A tutorial on learning from preferences and choices with Gaussian Processes. *arXiv:2403.11782*.

A model for label preferences

Label preferences:

- Predefined set of labels $\mathcal{C} = \{c_1, \ldots, c_d\}$;
- A subject is asked to express preference relations over the label set;
- Features associated with the subject (not the object)
- $c_i \succ_{\mathbf{x}} c_j$ means "for the object **x**, the label c_i is preferred to c_j".

Example

- Survey to students about ranking of gaming platforms
- Preference observations for student **x** are

$$c_{\texttt{GameBoy}} \succ_{\mathbf{x}} c_{\texttt{PC}} \succ_{\mathbf{x}} c_{\texttt{PlayStation}} \succ_{\mathbf{x}} c_{\texttt{PSPortable}} \succ_{\mathbf{x}} c_{\texttt{Xbox}}$$

Plackett-Luce distribution

Setup: for each instance $\mathbf{x}_k$, an *ordering* is

$$c_{\pi_1^{(k)}} \succ_{\mathbf{x}^{(k)}} c_{\pi_2^{(k)}} \succ_{\mathbf{x}^{(k)}} c_{\pi_3^{(k)}} \succ_{\mathbf{x}^{(k)}} \cdots \succ_{\mathbf{x}^{(k)}} c_{\pi_d^{(k)}},$$

where $[\pi_1^{(k)}, \ldots, \pi_d^{(k)}]$ is a permutation of $\{c_1, \ldots, c_d\}$.

Plackett-Luce distribution: consider score variables $\mathbf{a} \in \mathbb{R}^d$ with $a_i > 0$, we define a distribution over the permutations

$$p(\boldsymbol{\pi}^{(k)}|\mathbf{a}) = \prod_{i=1}^{d-1} \frac{a_{\pi_i^{(k)}}}{\sum_{j=i}^{d} a_{\pi_j^{(k)}}}.$$

Luce, R. D. 1959. Individual Choice Behavior: A Theoretical Analysis. *Wiley, New York*.

Plackett, R. L. 1975. The analysis of permutations. *JRSS-C*, 24(2):193-202.

Plackett-Luce distribution

What is the idea behind the Plackett-Luce distribution?

Luce's axiom of choice

The odds of choosing an item over another do not depend on the set of items from which the choice is made.

For any choice $\pi^{(k)}$ and for any two alternatives i_1 and i_2

$$\frac{\frac{a_{\pi^{(k)}_{i_1}}}{\sum_{j=i}^{d} a_{\pi^{(k)}_{j}}}}{\frac{a_{\pi^{(k)}_{i_2}}}{\sum_{j=i}^{d} a_{\pi^{(k)}_{j}}}} = \frac{a_{\pi^{(k)}_{i_1}}}{a_{\pi^{(k)}_{i_2}}},$$

only depends on i_1 and i_2.

We can intepret a ranking as a sequence of independent $d-1$ choices.

Plackett-Luce model

Idea: use the PL distribution as a likelihood for a GP-based model.

Training data: $\{c_1, \ldots, c_d\}$ labels, X set of subjects.

$$\mathcal{D}_m = \{c_{\pi_1^{(k)}} \succ_{\mathbf{x}^{(k)}} \cdots \succ_{\mathbf{x}^{(k)}} c_{\pi_d^{(k)}} : \quad \boldsymbol{\pi}^{(k)} \in \Pi_d, \mathbf{x}^{(k)} \in X, \; k = 1, \ldots, m\}.$$

Likelihood: link the ranking with d utility functions

$$p(\mathcal{D}_m | \mathbf{u}(X)) = \prod_{k=1}^{m} \prod_{i=1}^{d-1} \frac{e^{u_{\pi_i^{(k)}}(\mathbf{x}^{(k)})}}{\sum_{j=i}^{d} e^{u_{\pi_j^{(k)}}(\mathbf{x}^{(k)})}}.$$

Note: the Plackett-Luce distribution becomes our likelihood.
(We use functions $u_i(\mathbf{x})$ in place of the score variables a_i)

Plackett-Luce model

How to model the utility function?

GP prior: independent GP prior for each utility function:

$$u_i \sim \mathrm{GP}(u; 0, k_i), \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix} \sim GP\left(\mathbf{0}, \begin{bmatrix} k_1 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & k_d \end{bmatrix}\right),$$

same kernel $k_i = k_j$ for each utility, different hyper-parameters.

Posterior distribution:

- no analytical form;
- variational Bayes approximation.

Plackett-Luce model - Inference

Approximate $p(\mathbf{u}(X)|\mathcal{D}_m)$ with **variational inference**.

Variational density: $q(\mathbf{u}(X)) \sim N(\boldsymbol{\mu}, S)$, with S block-diagonal.

Find the q and the model parameters (lengthscales and σ) by maximizing

$$ELBO(q) = \underbrace{\int q(\mathbf{u}(X)) \log p(\mathcal{D}_m|\mathbf{u}(X)) d\mathbf{u}}_{\text{likelihood term}} - \underbrace{KL[q(\mathbf{u}(X))||p(\mathbf{u}(X))]}_{\text{KL between priors}}$$

We approximate the likelihood term with Monte Carlo integration.

Predictive posterior:

$$p(\mathbf{u}(\mathbf{x}^*)|\mathcal{D}_m) = \int p(\mathbf{u}(\mathbf{x}^*)|\mathbf{u}(X)) q(\mathbf{u}(X)|\mathcal{D}_m) d\mathbf{u}(X)$$

Ranking gaming platforms

Survey among 91 Dutch students (details in Fok et al. 2012)

Students were required to rank six options: Xbox, PlayStation, GameCube, PlayStation Portable, Game Boy, PC

We know features of the students:

- age;
- which platforms the student owns;
- average number of hours gaming each week;

Ranking gaming platforms - Data example

age	hours	platform	choice	own	choceid
33	2.00	GameBoy	6	0	1
33	2.00	GameCube	5	0	1
33	2.00	PC	4	1	1
33	2.00	PlayStation	1	1	1
33	2.00	PSPortable	3	0	1
33	2.00	Xbox	2	0	1

The individual with covariates age=33, hours=2, ownPlayStation=1 and ownPC=1, expressed the ranking

$$c_{\texttt{PlayStation}} \succ_{\texttt{x}} c_{\texttt{Xbox}} \succ_{\texttt{x}} c_{\texttt{PSPortable}} \succ_{\texttt{x}} c_{\texttt{PC}} \succ_{\texttt{x}} c_{\texttt{GameCube}} \succ_{\texttt{x}} c_{\texttt{GameBoy}}$$

Ranking gaming platforms - Plackett-Luce model

Data: $d = 6,\ m = 92,$

$$\mathcal{D}_m = \{c_{\pi_1^{(k)}} \succ_{\mathbf{x}^{(k)}} \cdots \succ_{\mathbf{x}^{(k)}} c_{\pi_d^{(k)}} : \quad \boldsymbol{\pi}^{(k)} \in \Pi_d, \mathbf{x}^{(k)} \in X,\ k = 1, \ldots, m\}.$$

Prior: each latent $u_i(\mathbf{x})$ modelled independent GP:

$$u_i(\mathbf{x}) \sim \mathrm{GP}_i(0, k_i(\mathbf{x}, \mathbf{x}')), \qquad i = 1, 2, \ldots, d.$$

Likelihood: Plackett-Luce likelihood

$$p(\mathcal{D}_m | \mathbf{u}(X)) = \prod_{k=1}^{m} \prod_{i=1}^{d-1} \frac{e^{u_{\pi_i^{(k)}}(\mathbf{x}^{(k)})}}{\sum_{j=i}^{d} e^{u_{\pi_j^{(k)}}(\mathbf{x}^{(k)})}}.$$

Posterior: computed with variational inference.

Ranking gaming platforms - Prediction example

Consider the student with covariates
`ownPlayStation`=1, `ownPC`=1, `age`=20 and `hours`=15.

The model predicts

	Xbox	PlayStation	PSPortable	GameCube	GameBoy	PC
predicted	3	2	4	5	6	1
observed	3	2	5	4	6	1

Evaluate the predictions with scaled *Kendall-*τ

$$\tau' = \left(\frac{n_c - n_d}{n_0} + 1 \right) /2$$

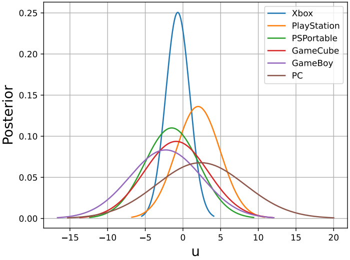
- $n_0 = n(n-1)/2$ possible pairs for n elements;
- n_c concordant pairs, n_d discordant pairs.

Example above: $n_0 = 15$, $n_c = 14$, $n_d = 1$, $\tau' = 0.93$

Ranking gaming platforms - Prediction example

Student with covariates
`ownPlayStation`=1, `ownPC`=1, `age`=20 and `hours`=15.

We return the posterior distribution for $u_i(\mathbf{x}^*)$, $i = 1, \ldots, 6$.



Average τ' over whole dataset with 10-fold cross-validation is 0.69.

GPs for label preference

Plackett-Luce label preference model

- Probabilistic learning from ranking data;
- Plackett-Luce likelihood;
- Posterior requires variational approximation.

More models (Thurston model, paired comparisons) in
Benavoli, A., & Azzimonti, D. 2024. A tutorial on learning from preferences and choices with Gaussian Processes. *arXiv:2403.11782*.

Choice function

Consider

- $\mathcal{X}$, a finite set of items;

$\mathcal{X} = \{$  ,  ,  ,  ,  $\}$

- $\mathcal{Q}$, the set of all (finite) subsets of $\mathcal{X}$.

$\mathcal{Q} = \{ \{$  $\}, \ldots, \{$  ,  $\}, \ldots, \{$  ,  ,  ,  ,  $\} \}$

A choice function is a map $C : \mathcal{Q} \to \mathcal{Q}$ such that

$$C : A \in \mathcal{Q} \mapsto C(A) \in \mathcal{Q}$$

$C : A = \{$   $\} \in \mathcal{Q} \mapsto C(A) = \{$  $\} \in \mathcal{Q}$

Learning user's behavior $\Rightarrow$ learning choice function from history.

Choice function - a few considerations

We model each item in A with a vector $\mathbf{x} \in \mathbb{R}^{n_x}$ containing its features.

Set of rejected items: $R(A) = A \setminus C(A)$, for any $A \in \mathcal{Q}$.

- If $\mathbf{x}_j \in R(A)$, there is at least one object in $C(A)$ better than $\mathbf{x}_j$

Incomparability: if $\{\mathbf{x}_j, \mathbf{x}_k\} \subseteq C(A)$ then $\mathbf{x}_j$ and $\mathbf{x}_k$ are incomparable

- The user may have multiple utilities
- Lack of knowledge

How to model choice functions?

Vector of utility functions $\mathbf{u} = [u_1(\mathbf{x}), \ldots, u_d(\mathbf{x})]^T$.

Pareto-dominant option $\mathbf{x_1}$ Pareto-dominates $\mathbf{x_2}$ ($\mathbf{x_1} \succ \mathbf{x_2}$) if

i) for all $j = 1, \ldots, d$, $u_j(\mathbf{x_1}) \geq u_j(\mathbf{x_2})$

ii) $\exists j \in \{1, \ldots, d\}$ s.t. $u_j(\mathbf{x_1}) > u_j(\mathbf{x_2})$

Non-dominated Pareto set Given $A = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, the set of non-dominated options is

$$A' = \{\mathbf{x} \in A : \nexists \mathbf{x}' \in A \text{ s.t. } \mathbf{x}' \succ \mathbf{x}\}$$

Pareto-rationalisable choice functions

u describes the choice function C if, for each $A \subset \mathcal{X}$,

- $C(A)$ is the *non-dominated set* in the *strong Pareto sense* for **u**;
- $R(A)$ is the set of dominated objects.

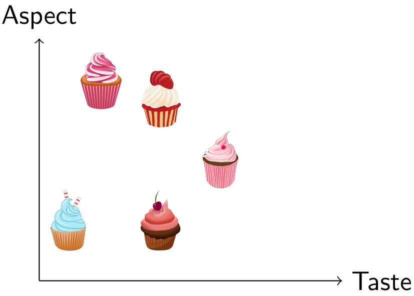
Note: not all choice functions are Pareto-rationalisable.

Pareto-rationalisable choice functions: example

Example: choose best cake

- utilities: taste, aspect
-  dominates 
- set of non-dominated options

$A' = \{$  ,  ,  $\}$



We do not observe the vector of utility functions

An exact link between choices and utilities

Assume: latent vector of utilities $\mathbf{u} = [u_1(\mathbf{x}), \dots, u_d(\mathbf{x})]^T$.

For each $A \subset \mathcal{X}$, we can link choices and utilities with

$$\neg \left(\min_{i \in \{1,\dots,d\}} (u_i(\mathbf{o}) - u_i(\mathbf{v})) \leq 0, \ \forall \mathbf{o} \in C(A) \right), \forall \mathbf{v} \in R(A), \qquad (1)$$

(For each $\mathbf{v} \in R(A)$, there is at least a object in $C(A)$ not worse than $\mathbf{v}$)

$$\min_{i \in \{1,\dots,d\}} (u_i(\mathbf{o}) - u_i(\mathbf{v})) \leq 0, \ \forall \mathbf{o}, \mathbf{v} \in C(A), \ \mathbf{o} \neq \mathbf{v}. \qquad (2)$$

(For each object in $C(A)$, there is no better object in $C(A)$)

Using the link to build a likelihood

Given a choice dataset

$$\mathcal{D}_m = \{(C(A_s), A_s) : \text{ for } s = 1, \ldots, m\}, \quad A_s \subset \mathcal{X} \text{ for each } s,$$

$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t]^\top$ features associated with t objects in $\mathcal{X}$

We define a likelihood

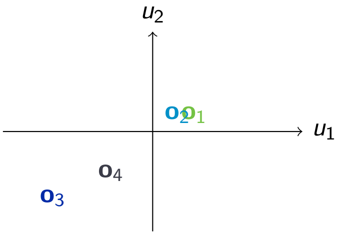
$$p_{exact}(\mathcal{D}_m | \mathbf{u}(X)) = \prod_{k=1}^{m} p_{exact}(C(A_k), A_k | \mathbf{u}(X)),$$

where $\mathbf{u}(X) = [\mathbf{u}(\mathbf{x}_1), \mathbf{u}(\mathbf{x}_2), \ldots, \mathbf{u}(\mathbf{x}_t)]^\top$ and

$$p_{exact}(C(A_k), A_k | \mathbf{u}(X)) = \begin{cases} 1 & \text{if both conditions are satisfied} \\ 0 & \text{otherwise} \end{cases}$$

Case of non-Pareto rational choices

Consider $\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4$ with



Assume we are given the following choices:

$$C(\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}) = \{\mathbf{o}_1, \mathbf{o}_2\}, \quad C(\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_4\}) = \{\mathbf{o}_1\},$$

These choices are not Pareto rational, $p_{exact}(\mathcal{D}_m | \mathbf{u}(X))$ is zero.

Likelihood accounting for errors

On the choice dataset $\mathcal{D}_m$ we define the likelihood

$$p(\mathcal{D}_m|\mathbf{u}(X)) = \prod_{k=1}^{m} p(C(A_k), A_k|\mathbf{u}(X))$$

$$= \prod_{k=1}^{m} \prod_{\{\mathbf{o},\mathbf{v}\}\in C_\sharp(A_k)} \left(1 - \prod_{i=1}^{d} \Phi\left(\frac{u_i(\mathbf{o}) - u_i(\mathbf{v})}{\sigma}\right) - \prod_{i=1}^{d} \Phi\left(\frac{u_i(\mathbf{v}) - u_i(\mathbf{o})}{\sigma}\right)\right)$$

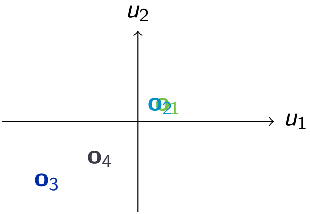
$$\prod_{\mathbf{v}\in R(A_k)} \left(1 - \prod_{\mathbf{o}\in C(A_k)} \left(1 - \prod_{i=1}^{d} \Phi\left(\frac{u_i(\mathbf{o}) - u_i(\mathbf{v})}{\sigma}\right)\right)\right)$$

The blue part is a probabilistic relaxation of the first condition
(For each $\mathbf{v} \in R(A)$, there is at least a object in $C(A)$ not worse than $\mathbf{v}$)

The green part is a probabilistic relaxation of the second condition
(For each object in $C(A)$, there is no better object in $C(A)$)

Non-Pareto rational choices

Consider the same objects $\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4$



With the following choices:
$C(\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}) = \{\mathbf{o}_1, \mathbf{o}_2\}$,
$C(\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_4\}) = \{\mathbf{o}_1\}$

Probabilistic relaxation likelihood
$p(\{\mathbf{o}_1, \mathbf{o}_2\}, \{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\} | \mathbf{u}(X)) \approx 0.48$

$p(\{\mathbf{o}_1\}, \{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_4\} | \mathbf{u}(X)) \approx 0.12$

$p(\mathcal{D}_m | \mathbf{u}(X)) \approx 0.48 \cdot 0.12 = 0.057$

ChoiceGP

Data: $\mathcal{D}_m = \{(C(A_s), A_s) : \text{ for } s = 1, \ldots, m\}, \quad A_s \subset \mathcal{X}$ for each s;

Prior: each latent utility function in $\mathbf{u}(\mathbf{x}) = [u_1(\mathbf{x}), \ldots, u_d(\mathbf{x})]^\top$ is modelled as an independent GP:

$$u_i(\mathbf{x}) \sim \mathrm{GP}_i(0, k_i(\mathbf{x}, \mathbf{x}')), \qquad i = 1, 2, \ldots, d.$$

Likelihood: Error accounting likelihood

$$p(\mathcal{D}_m | \mathbf{u}(X)) = \prod_{k=1}^{m} p(C(A_k), A_k | \mathbf{u}(X))$$

ChoiceGP - Inference

Posterior: $$p(\mathbf{u}(X)|\mathcal{D}_m) = \frac{p(\mathbf{u}(X))}{p(\mathcal{D}_m)} \prod_{k=1}^{m} p(C(A_k), A_k|\mathbf{u}(X))$$

Model parameters:

- ARD lengthscales of each $k_i(\cdot,\cdot)$;
- scale parameter σ in the likelihood.

The posterior is not a GP $\Rightarrow$ computation with variational inference.

ChoiceGP - Prediction

For a new vector of p objects $X^* = \{\mathbf{x}_1^*, \ldots, \mathbf{x}_p^*\}$

$$p(\mathbf{u}(X^*)|\mathcal{D}_m) = \int p(\mathbf{u}(X^*)|\mathbf{u}(X))q(\mathbf{u}(X)|\mathcal{D}_m)d\mathbf{u}(X)$$

where $q(\mathbf{u}|\mathcal{D}_m)$ is the approximate VI posterior.

For a new set A^* and a possible choice $C(A^*)$

$$p(C(A^*), A^*|\mathcal{D}_m) = \int p(C(A^*), A^*|\mathbf{u}(X^*))p(\mathbf{u}(X^*)|\mathcal{D}_m)d\mathbf{u}(X^*),$$

computed via Monte Carlo sampling from $p(\mathbf{u}(X^*)|\mathcal{D}_m)$.

Benavoli, A., Azzimonti, D., and Piga, D. 2023. Learning Choice Functions with Gaussian Processes *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 141-151.

Revisiting transportation modes

Objective: predict transportation mode for a subject on a journey.

Example of data

case	alt	choice	dist	cost	ivt	ovt	freq
129	train	0	387	58.25	316	59	2
129	air	0	387	145.80	56	85	9
129	bus	0	387	26.67	301	58	8
129	car	1	387	73.53	251	0	0

Data encoded as choices: $A_k = \{\mathbf{x}_{air}, \mathbf{x}_{bus}, \mathbf{x}_{car}, \mathbf{x}_{train}\}$ $C(A_k) = \{\mathbf{x}_{car}\}$

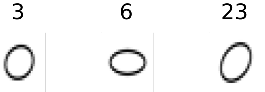
Results: 0.90 Accuracy over 10 repetition (70-30 train/test split)

Note: here $C(A)$ always contains one element
$\Rightarrow$ choices are equivalent to pairwise preferences.

Choosing ellipses

Task:

- a student is shown images of three ellipses;
- they are asked to select "best ellipses" based on
  (i) small eccentricity (being close to a circular shape);
  (ii) alignment of their axes with the Cartesian axes



Example: three ellipses images.

Data: Task repeated 160 times generating 160 choice sets.
Example: $A = \{3, 6, 23\}$ $C(A) = \{6\}$.

Choosing ellipses

Dataset:

- 160 choices data;
- for each ellipse, the (25×25) pixel values of the image.

Model: ChoiceGP with erroneous likelihood,
2 latent dimensions, RBF kernels k_i with ARD lengthscales.

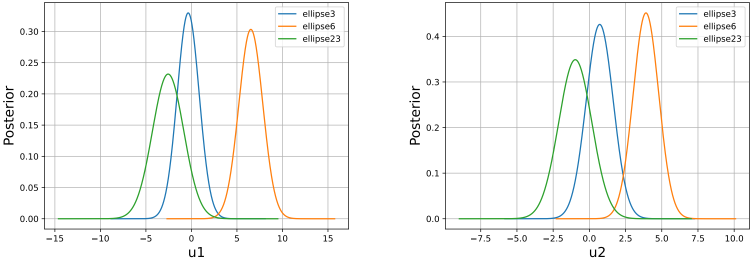
Results: On a $90-10$ train-test split, accuracy 0.84.

Accuracy is computed as $acc = \frac{1}{|A|}\left(\sum_{o\in C(A)} I_{\hat{C}(A)}(o) + \sum_{v\in R(A)} I_{\hat{R}(A)}(v)\right)$

Choosing ellipses - example

3 6 23

Three ellipses images: the student selected 6.

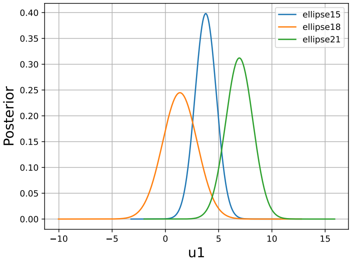


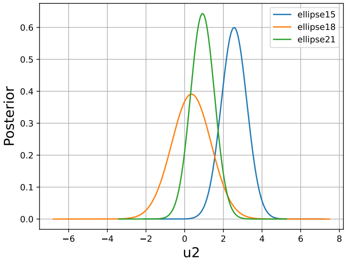
Estimated latent functions.

Choosing ellipses - example

15 18 21

Three ellipses images: the student selected $\{15, 21\}$.





Estimated latent functions.

Conclusions

- Preference learning with Gaussian processes;
- Three types of preference encoding
    - Object preferences (features of the object);
    - Label preferences (features of the subject, rankings);
    - Choice data (generalized preferences).

Where to use this?

- learning rankings/preferences;
- Bayesian optimization (use learned model to suggest better items)

References (1)

Benavoli, A., & Azzimonti, D. 2024. A tutorial on learning from preferences and choices with Gaussian Processes. *arXiv:2403.11782*. arXiv.

McFadden, D. 1974. The measurement of urban travel demand. *Journal of Public Economics*, 3(4): 303-328.

Hunter, D. R. 2004. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1): 384-406.

Critchlow, D. E., & Fligner, M. A. 1991. Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, 56(3): 517-533.

Maldonado, S., Montoya, R., & Weber, R. 2015. Advanced conjoint analysis using feature selection via support vector machines. *European Journal of Operational Research*, 241(2): 564-574.

Chu, W. and Ghahramani, Z. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany) (ICML '05) 137-144.

References (2)

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. 2011. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745 (2011).

Benavoli, A., Azzimonti, D., and Piga, D. 2020. Skew Gaussian Processes for Classification. *Machine Learning*, 109(9):1877-1902.

Benavoli, A., Azzimonti, D., and Piga, D. 2021. Preferential Bayesian optimisation with skew Gaussian processes. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO '21)*, 1842–1850.

Benavoli, A., Azzimonti, D., and Piga, D. 2021. A unified framework for closed-form nonparametric regression, classification, preference and mixed problems with Skew Gaussian Processes. *Machine Learning* 110, 3095–3133.

References (3)

González, J., Dai, Z., Damianou, A. and Lawrence, N. D. 2017. Preferential Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning*. Volume 70. JMLR. 1282-1291.

Nickisch, H., and Rasmussen, C. E. 2008. Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research* 9 2035-2078

Durante, D. 2019. Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4), 765-779.

Azzalini, A. 2013. The Skew-Normal and related families (Vol. 3). Cambridge: Cambridge University Press.

Gessner, A., Kanjilal, O., and Hennig, P. 2020. Integrals over Gaussians under linear domain constraints. In *Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics* (AISTATS) 2020, Palermo, Italy.

References (4)

Croissant, Yves. 2020. Estimation of random utility models in R: the mlogit package. *Journal of Statistical Software*, 95:1-41.

Chau, S. L., Gonzalez, J., and Sejdinovic, D. 2022. Learning Inconsistent Preferences with Gaussian Processes. In *25th International Conference on Artificial Intelligence and Statistics*, 2266-2281. PMLR.

Moulin, H. 1985. Choice Functions Over a Finite Set: A Summary. *Social Choice and Welfare* 2(2): 147-60.

Benavoli, A., Azzimonti, D., and Piga, D. 2023. Learning Choice Functions with Gaussian Processes *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 141-151.

K. Pfannschmidt and E. Hüllermeier. 2020. Learning Choice Functions via Pareto-Embeddings. In *German Conference on Artificial Intelligence. Springer.*

References (5)

A. Vehtari, A. Gelman, and J. Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5):1413-1432.

Benavoli, A., Azzimonti, D., and Piga, D.. 2023. Bayesian Optimization For Choice Data. *In GECCO '23. Association for Computing Machinery.*

Daulton, S., Balandat, M., and Bakshy, E. 2020. Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization. In *Advances in Neural Information Processing Systems* Vol. 33. Curran Associates, Inc., 9851-9864.

Thompson, William R. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika 25, 3/4, 285.*