HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

FCAI Finnish Center for Artificial Intelligence

Better priors for everyone

Arto Klami

Department of Computer Science

University of Helsinki

Building on work of many others:
M. Hartmann, P. Bürkner, G. Agiashvili,
E. de Souza da Silva, T. Kuśmierczyk, O. Martin, …

Bayesian statistics

Specify a **statistical model** as joint distribution over data and parameters

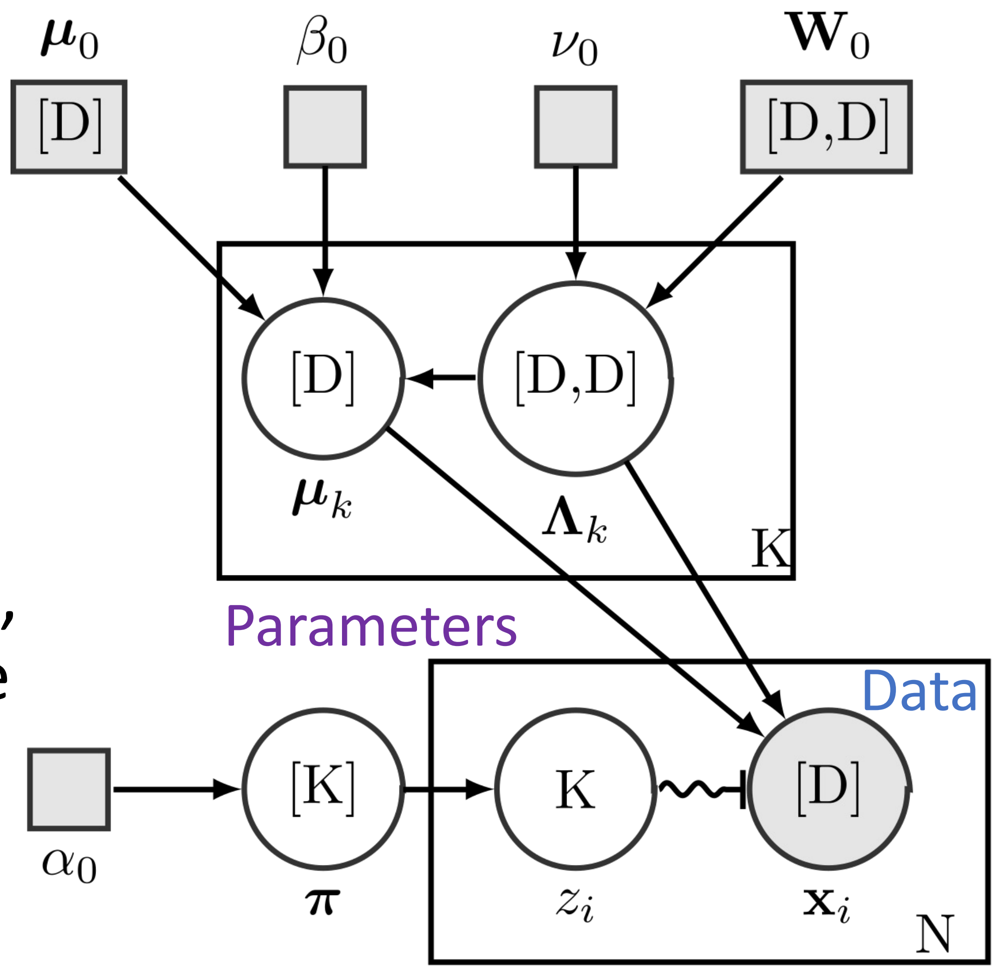
$\mathcal{D}$: Data $\qquad \theta$: Parameters

$p(\mathcal{D}, \theta | \lambda)$: Model

Given the model and some observations, infer a distribution of possible values the model's parameters could take

$$p(\theta | \mathcal{D}, \lambda) = \frac{p(\mathcal{D}, \theta | \lambda)}{p(\mathcal{D} | \lambda)}$$

Hyperparameters



Choice of prior

The priors are part of the model specification, just as e.g. neural network architecture would be

Choice of prior means

- Choosing the form (parameteric family), often factorized over the parameters
- Choosing the hyperparameters that specify the prior itself

$$p(\theta_1, \theta_2 | \lambda_1, \lambda_2) = \mathcal{N}(\theta_1 | \lambda_1) \mathrm{Gamma}(\theta_2 | \lambda_2)$$

Model parameters

Hyperparameters

All kinds of models need priors

Statistical modelling

Often

- Proper model of a phenomenom
- Relatively few parameters
- Priors encode subjective knowledge
- Implemented e.g. in Stan

Example: Cognitive theory
Disease transmission

Machine learning

Often

- General-purpose model
- Huge number of parameters
- Priors encode desired properties or heuristic inductive biases
- Implemented e.g. in PyTorch

Example: Neural network
Recommender engine

How to choose the priors

Form of prior

- Computational convenience
- Literature
- Domain knowledge

Hyperparameters

- Domain knowledge
- Statistical expertise

Form of prior

- Computational convenience
- Desired properties (e.g. sparsity)
- Whatever previous authors used

Hyperparameters

- Default values and heuristics
- Cross-validation

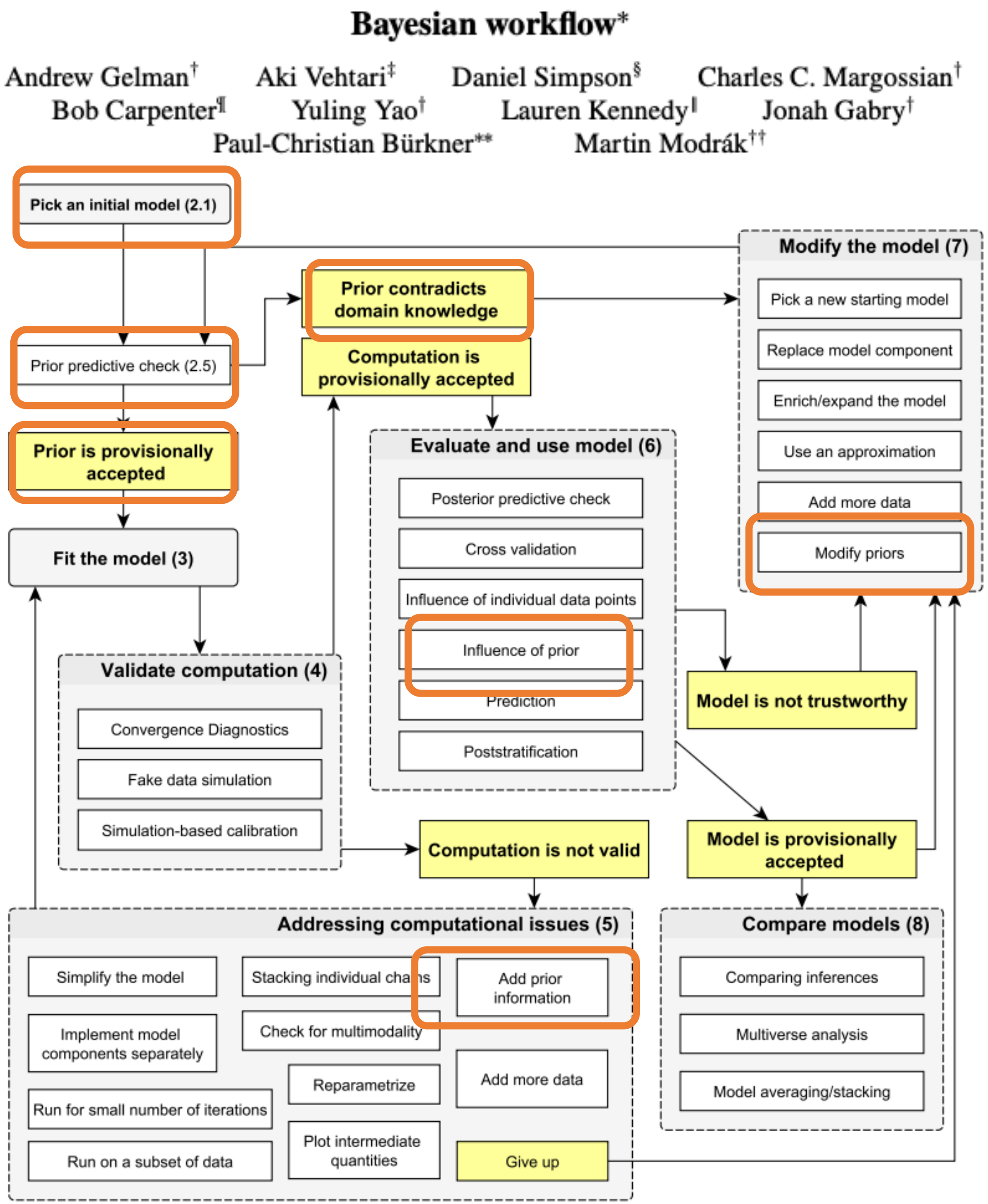
Domain knowledge

Near impossible in practice

- Statistician does not have domain knowledge
- Domain expert does not know enough statistics

In practice

- Highly iterative workflow
- Quality depends a lot on the individual



Example

Ice cream shop

- $\alpha, \beta \sim \mathcal{N}(0, 100)$
- $\mu_t \sim \mathcal{N}(15, 2)$
- $t_i \sim \mathcal{N}(\mu_t, 2)$
- $s_{t,i} | \{t_i\} \sim \mathcal{N}(t_i, 1)$
- $s_i | \{t_i, \alpha, \beta\} \sim \text{Poisson}(\text{rate} = \alpha + \beta t_i)$

What is the effect of changing these?

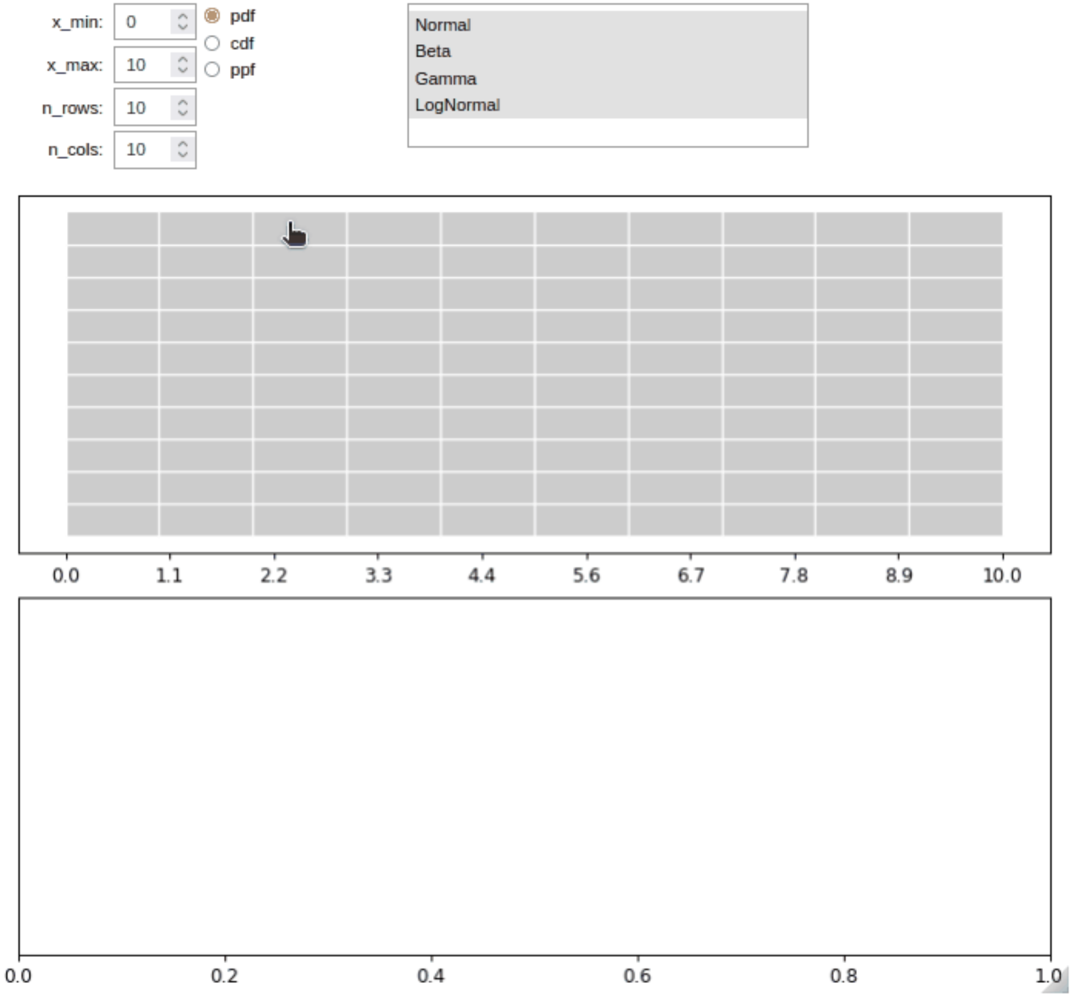
Prior elicitation

Let's help the user

Facilitator: Statistician, asks questions

Expert: Domain knowledge, provides answers (via graphical interface)

Goal: Transform tacit knowledge into proper prior distributions, without requiring their direct specification



https://github.com/arviz-devs/preliz

Why don't we use it?

Open Access

2023

Prior Knowledge Elicitation: The Past, Present, and Future

Petrus Mikkola, Osvaldo A. Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, Paul-Christian Bürkner, Arto Klami

Author Affiliations +

Bayesian Anal. Advance Publication 1-33 (2023). DOI: 10.1214/23-BA1381

+30 page Supplement

Why don’t we use it?

1. Methods are model-specific
   - Only helps if you use that exact model
   - No support for general probabilistic programs
2. Lack of software support
   - Nothing that integrates with PP tools (Stan etc)
   - No robust and general implementations
3. Lack of high profile examples
   - Why risk using poor software or methods for your most important studies?
   - No tools available for your model

Open Access

2023

Prior Knowledge Elicitation: The Past, Present, and Future

Petrus Mikkola, Osvaldo A. Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, Paul-Christian Bürkner, Arto Klami

Author Affiliations +

Bayesian Anal. Advance Publication 1-33 (2023). DOI: 10.1214/23-BA1381

Suggestions for

- General algorithms
- Evaluation
- Software
- …

Why is it hard?

See Kadane and Wolfson (1998) for more

Priors are over the parameter space, but

- Not all parameters have interpretation
- There may be complex dependencies even with univariate priors
- Requires significant understanding of statistics

Experts often know the observed data better:

"On a hot day I sell 1000-2000 ice creams" vs *"Std of alpha is 100"*

By asking expert information about the data space we make it easier for them, but need to solve a harder computational task

Prior predictive distribution

All probabilistic models define a prior predictive distribution (PPD)

$$p(x|\lambda) = \int p(x|\theta)p(\theta|\lambda)d\theta$$

This is the basis of prior predictive check

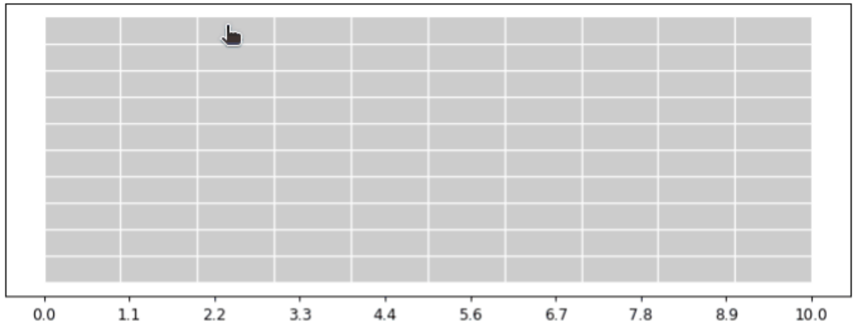
1. Sample imaginary data from the prior
2. Plot the imaginary data
3. Modify the prior if the imaginary data is really weird

Football example from https://mc-stan.org/docs/stan-users-guide/example-of-prior-predictive-checks.html

1. Poisson distribution for the number of goals for each team
2. Jeffrey's uninformative prior for rates

Average number of goals per team is around 50,000!!!

Prior predictive elicitation: Idea



1. Ask the expert what they expect from data
2. For any λ PPD defines what kind of data is likely under the model
3. Solve λ so that the PPD matches the expert's answers

$$p(x^*) \approx p(x|\lambda) = \int p(x|\theta)p(\theta|\lambda)d\theta$$

Expert information

PPD

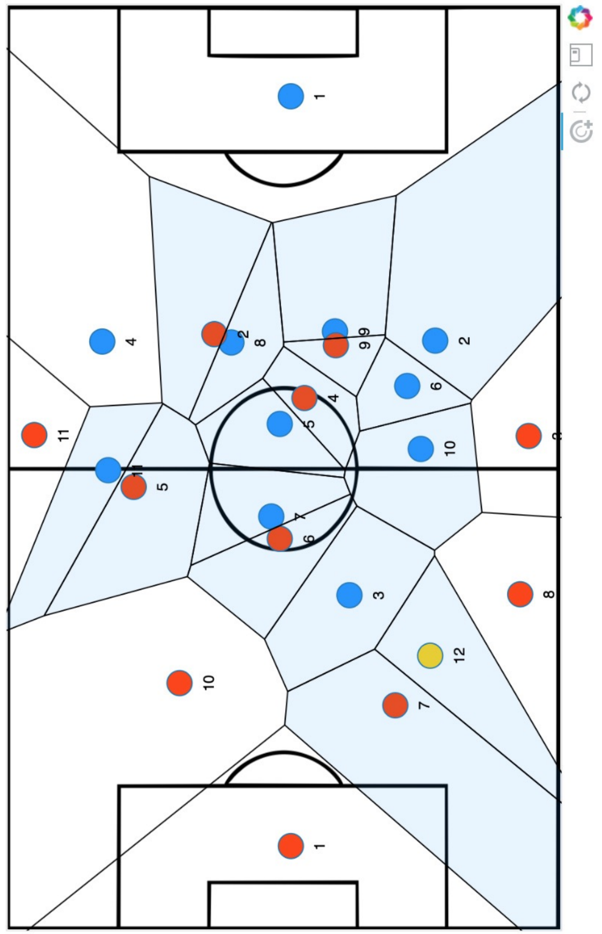
Prior predictive elicitation

Hartmann et al. **Flexible prior elicitation via the prior predictive distribution**, UAI 2020

1. Partition observation space arbitrarily
2. Expert provides expected probability for each part
3. Treat expert annotations as noisy realizations of the PPD

$$\pi(\mathbf{p}) \sim \mathcal{D}(\mathbf{p}\,|\,\alpha, \mathbb{P}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{n} \Gamma(\alpha\,\mathbb{P})} \prod_{i=1}^{n} \mathrm{p}_i^{\alpha\,\mathbb{P}-1}$$

4. Maximize their likelihood wrt to the prior hyperparameters



Example: Human (male) growth rate

See Agiashvili: **Prior Predictive Elicitation** (2021) for more details

Model: Lawless (2011)

$$Y_t | \boldsymbol{\theta}, b \sim \mathcal{W}(h(t;\boldsymbol{\theta}), b)$$

Likelihood

$$b \sim \mathcal{G}(a_0, b_0)$$

$$\theta_d \overset{i.i.d}{\sim} \mathcal{LN}(a_d, b_d)$$

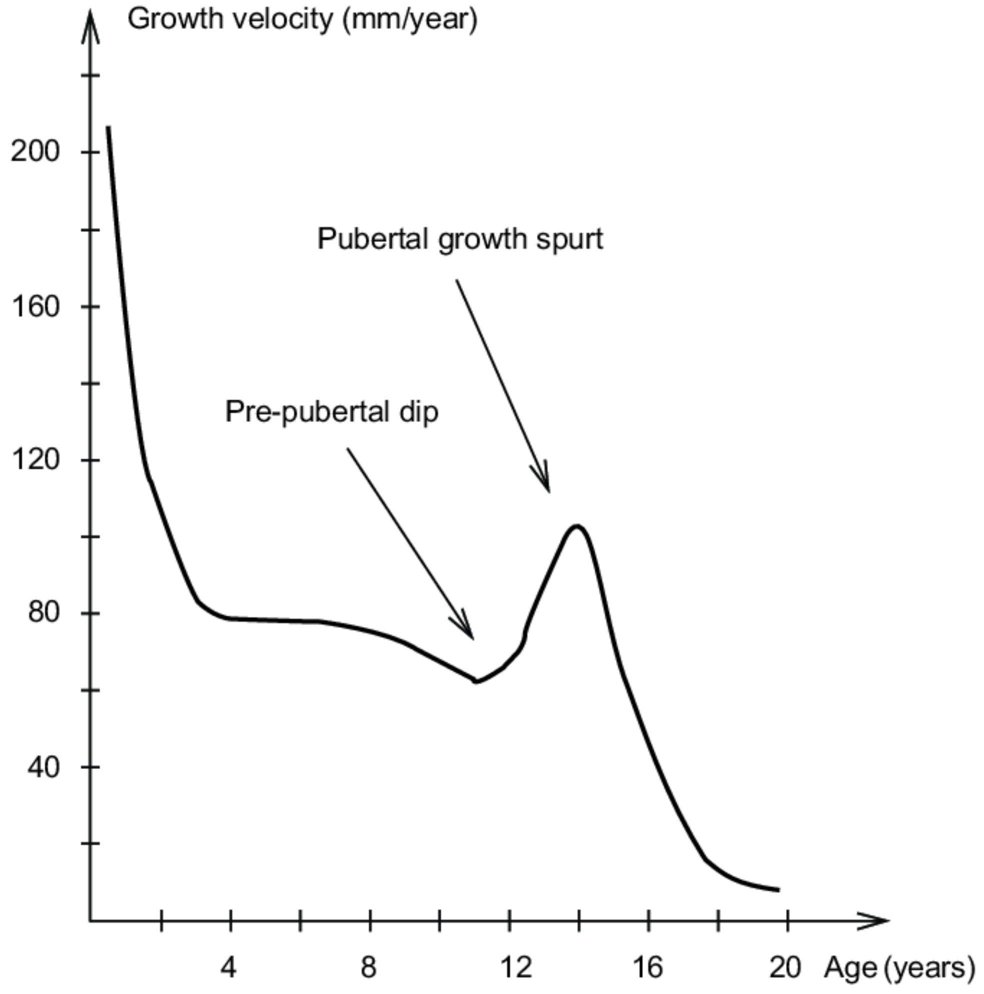
Prior

$$h(t;\boldsymbol{\theta}) = h_1 - \frac{2(h_1 - h_{t_*})}{\exp[s_0\ (t - t_*)] + \exp[s_1\ (t - t_*)]}.$$

Parameters: 6

- Average adult height
- Height and age of growth spurth
- And some others

Prior hyperparameters: 12



Example: Human (male) growth rate

5 statisticians set priors using two alternative strategies but with the same graphical interface, providing quantile probabilities

Structural

- What values are likely for each parameter

Predictive

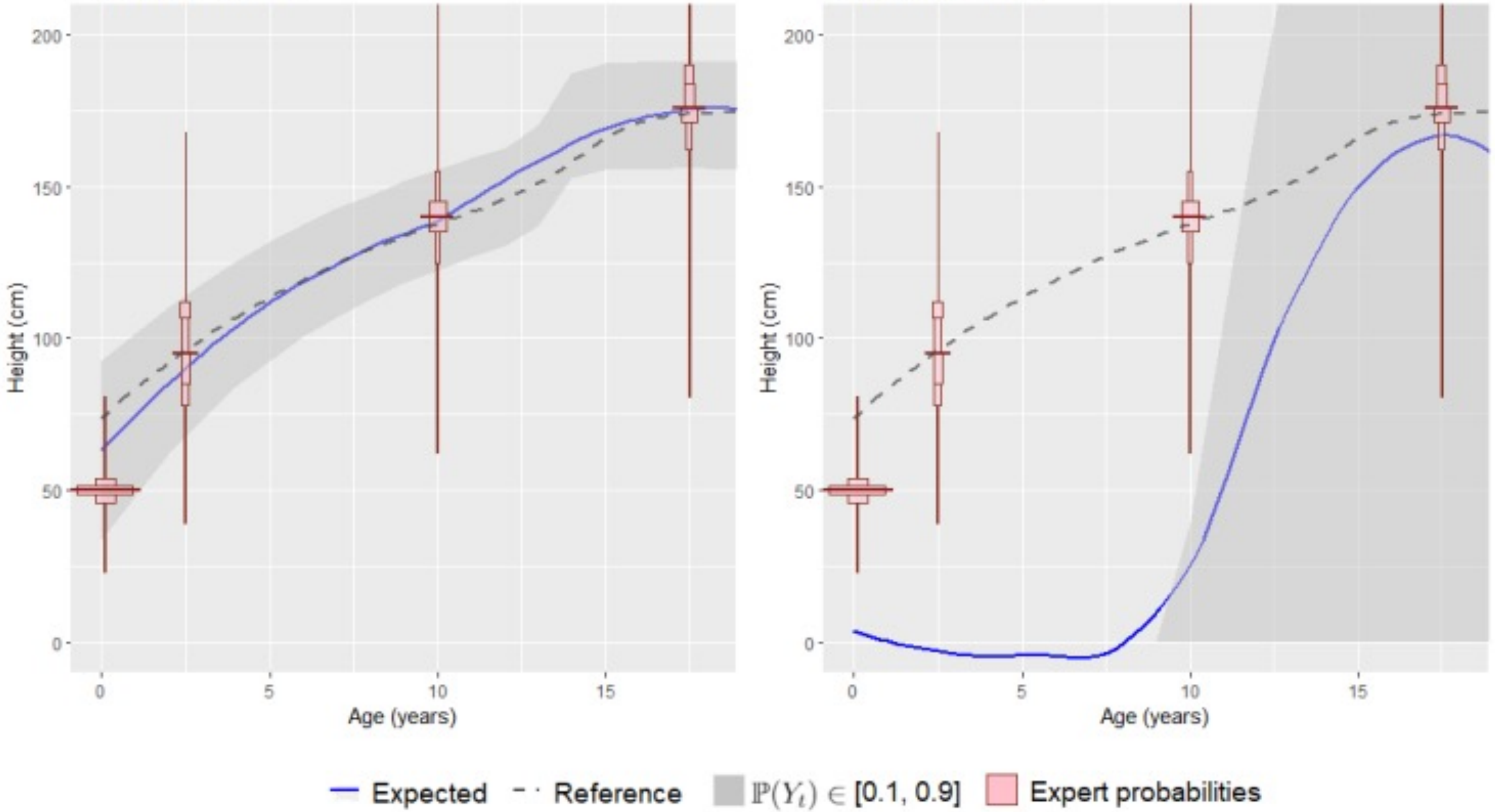
- How tall are men at certain ages
- Note: Need to choose the ages

		Predictive		**Structural**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	174.5	0.8	176.2	105.3
h_{t_*}	162.9	162.8	4.2	129.1	33.6
s_0	0.1	0.1	< 0.1	1.2	1.1
s_1	1.2	3.3	0.2	1.2	1.1
t_*	14.6	13.4	0.01	12.5	0.6
b	–	15.8	12.9	2.0	4.6
α	–	6.9	–	1.2	–

Example: Human (male) growth rate

Predictive

Structural



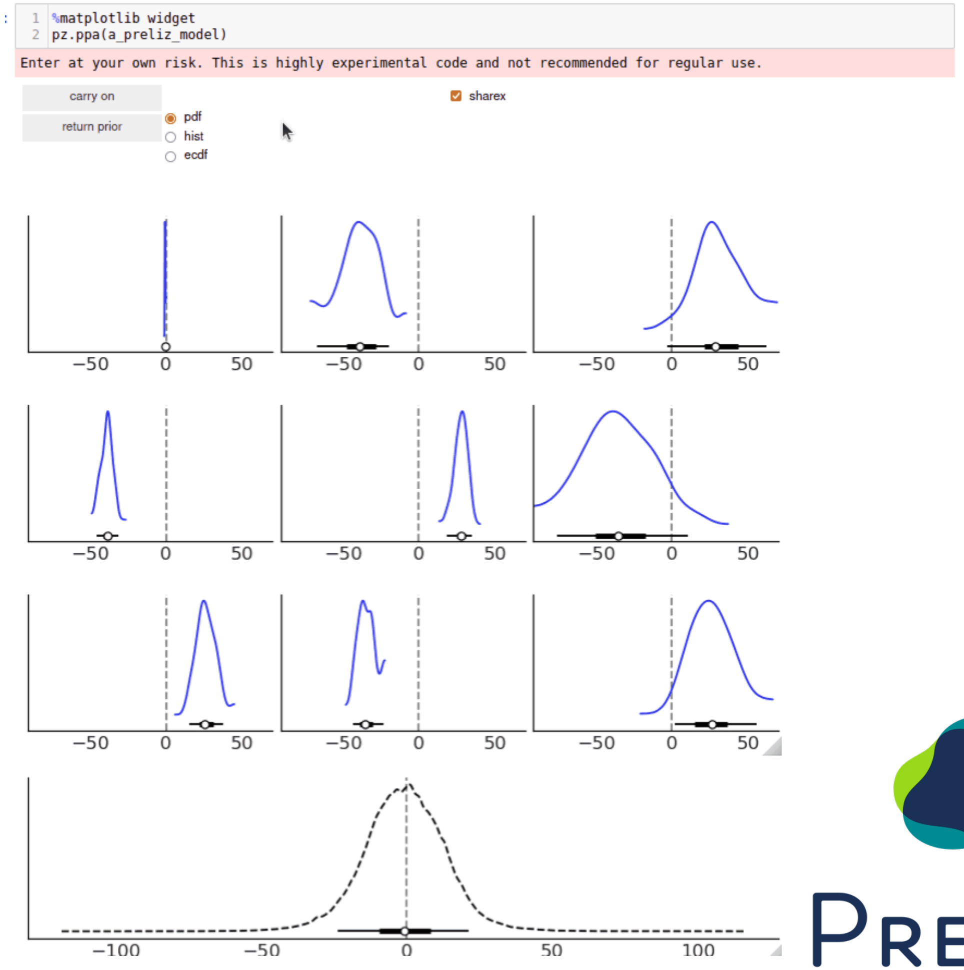
Needs for elicitation software

Needs to be be modular and applicable to arbitrary models

- Connects to typical PP languages and inference engines
- Visualization and interaction
- Elicitation algorithms
- Evaluation

PreliZ is one ongoing attempt

https://github.com/arviz-devs/preliz



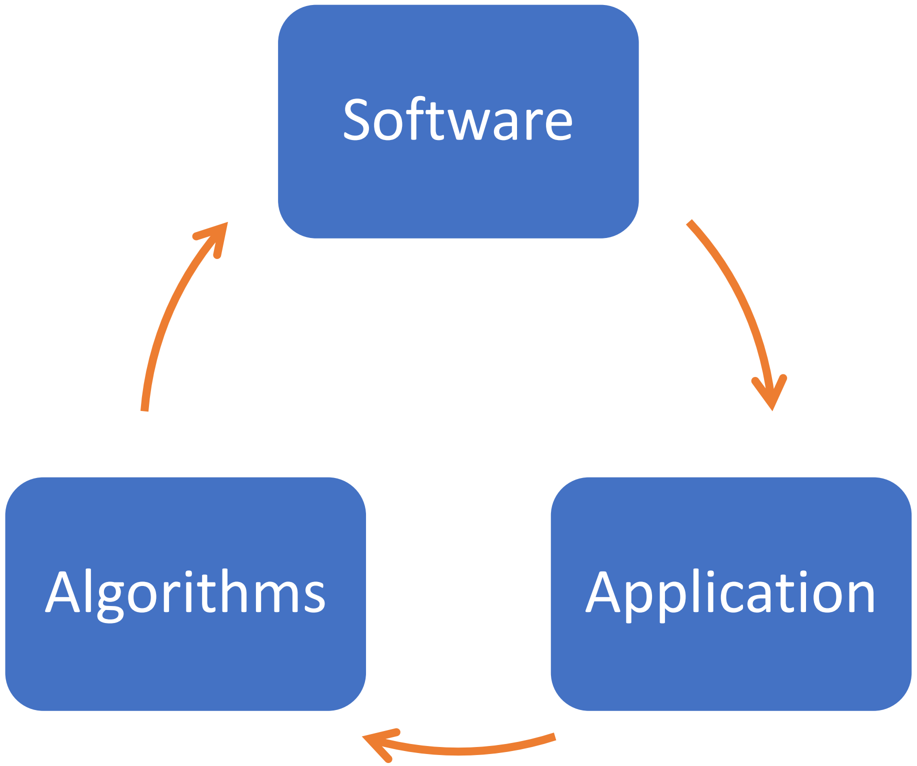
Virtuous cycle

Once we have good algorithms and implementations, people will use them

First strong applications encourage further algorithm/software development

Leads to new examples and benchmarks

You should use some elicitation method when writing your next Nature paper!



How about machine learning?

Parameters of most flexible models have no interpretation so subjective prior knowledge is out of question

BAD

Lazy ML researcher: "*We set this to 0.1, but other values could be used*"

ML engineer: "*Let's do cross-validation, trying out lots of lambdas*"

- Kind 1: Grid search is good enough
- Kind 2: Bayesian optimization is better

NEEDS INFERENCE

Automatic prior specification

de Souza da Solva et al. **Prior specification for Bayesian matrix factorization via prior predictive matching.** JMLR, 2023

PPD helped in prior predictive elicitation and is defined for all generative models, also the large ML models

$$p(x|\lambda) = \int p(x|\theta)p(\theta|\lambda)d\theta$$

Let's use that to create method for automatic choice of hyperparameters that does not need repeated inference

We demonstrated the general principle for matrix factorization models

Gist

1. Choose a few statistics (mean, variance, …) to match
2. Choose a distance measure between the statistics
3. Compute target statistics from the observed data
4. Solve for optimal hyperparameters

$$\lambda^* = \arg\min_{\lambda} d(\underbrace{T(x|\lambda)}_{\text{PPD statistics}}, \underbrace{T^*}_{\text{Target/data statistics}})$$

Gist

1. Choose a few statistics (mean, variance, …) to match
2. Choose a distance measure between the statistics
3. Compute target statistics from the observed data
4. Solve for optimal hyperparameters

Cheating? Not when using simple summary statistics, but you can also

- Estimate the target statistics from separate validation data
- Use subjective knowledge (*“around 2 goals per team on average”*)

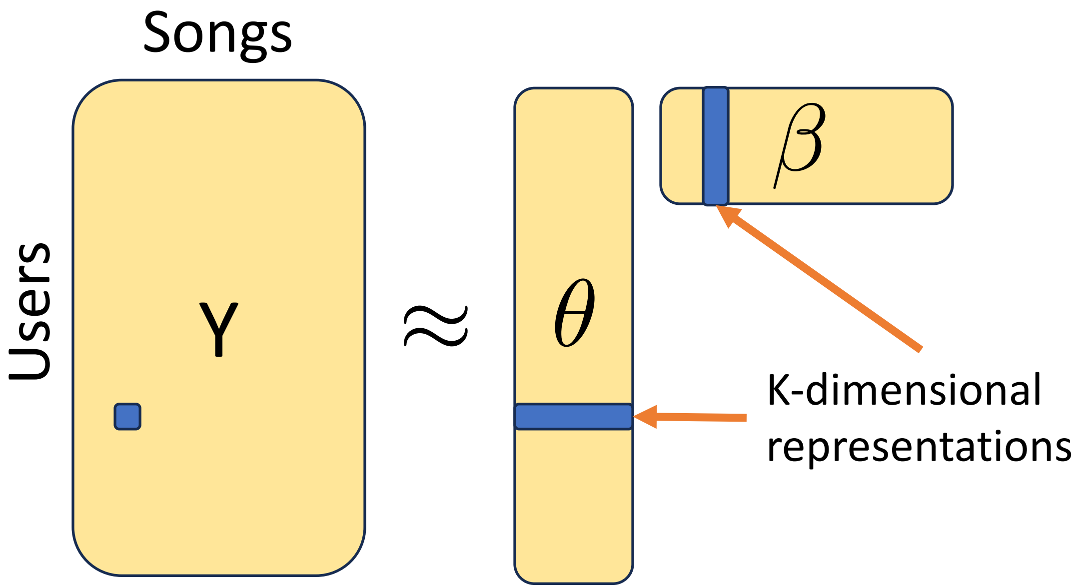
Example: Matrix factorization

Poisson matrix factorization (PMF)

- 5 hyperparameters (one is K)
- Lots and lots of parameters

$$\theta_{ik} \overset{\text{iid}}{\sim} F(\mu_\theta, \sigma_\theta^2), \quad \beta_{jk} \overset{\text{iid}}{\sim} F(\mu_\beta, \sigma_\beta^2),$$

$$Y_{ij} \overset{\text{iid}}{\sim} \text{Poisson}\left(\sum_{k=1}^{K} \theta_{ik}\beta_{jk}\right).$$



Note: Not assuming conjugate priors here

Method 1: Human labor

For PMF we can compute expectations of PPD analytically and solve for the optimal hyperparameters analytically as well

$$\lambda^* = \arg\min_{\lambda} d(T(x|\lambda), T^*)$$

$$\Downarrow$$

$$T(x|\lambda^*) = T^*$$

Often quite involved derivations, but worth it for important models

Method 1: Human labor

For PMF we get

$$\mathbb{E}[Y_{ij}] = K\mu_\theta\mu_\beta$$

$$\begin{aligned}\mathbb{V}[Y_{ij}] = K[&\mu_\theta\mu_\beta + (\mu_\beta\sigma_\theta)^2 \\ &+ (\mu_\theta\sigma_\beta)^2 + (\sigma_\theta\sigma_\beta)^2]\end{aligned}$$

$$\rho_1 = \frac{K(\mu_\beta\sigma_\theta)^2}{\mathbb{V}[Y_{ij}]}$$

$$\rho_2 = \frac{K(\mu_\theta\sigma_\beta)^2}{\mathbb{V}[Y_{ij}]}$$

Method 1: Human labor

For PMF we get

$$\mathbb{E}[Y_{ij}] = K\mu_\theta\mu_\beta$$

$$\begin{aligned}\mathbb{V}[Y_{ij}] = K[&\mu_\theta\mu_\beta + (\mu_\beta\sigma_\theta)^2 \\ &+ (\mu_\theta\sigma_\beta)^2 + (\sigma_\theta\sigma_\beta)^2]\end{aligned}$$

$$\rho_1 = \frac{K(\mu_\beta\sigma_\theta)^2}{\mathbb{V}[Y_{ij}]}$$

$$\rho_2 = \frac{K(\mu_\theta\sigma_\beta)^2}{\mathbb{V}[Y_{ij}]}$$

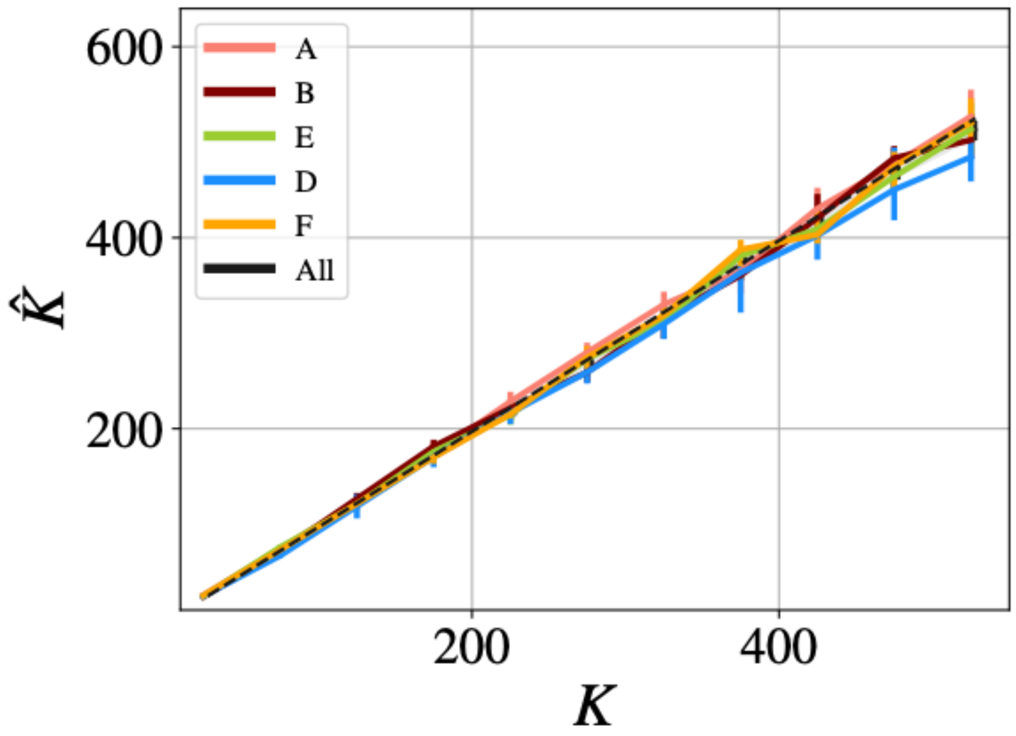
Gives, e.g. automatic choice of number of components

$$K = \frac{\tau\,\mathbb{V}[Y_{ij}] - \mathbb{E}[Y_{ij}]}{\rho_1\rho_2}\left(\frac{\mathbb{E}[Y_{ij}]}{\mathbb{V}[Y_{ij}]}\right)^2$$

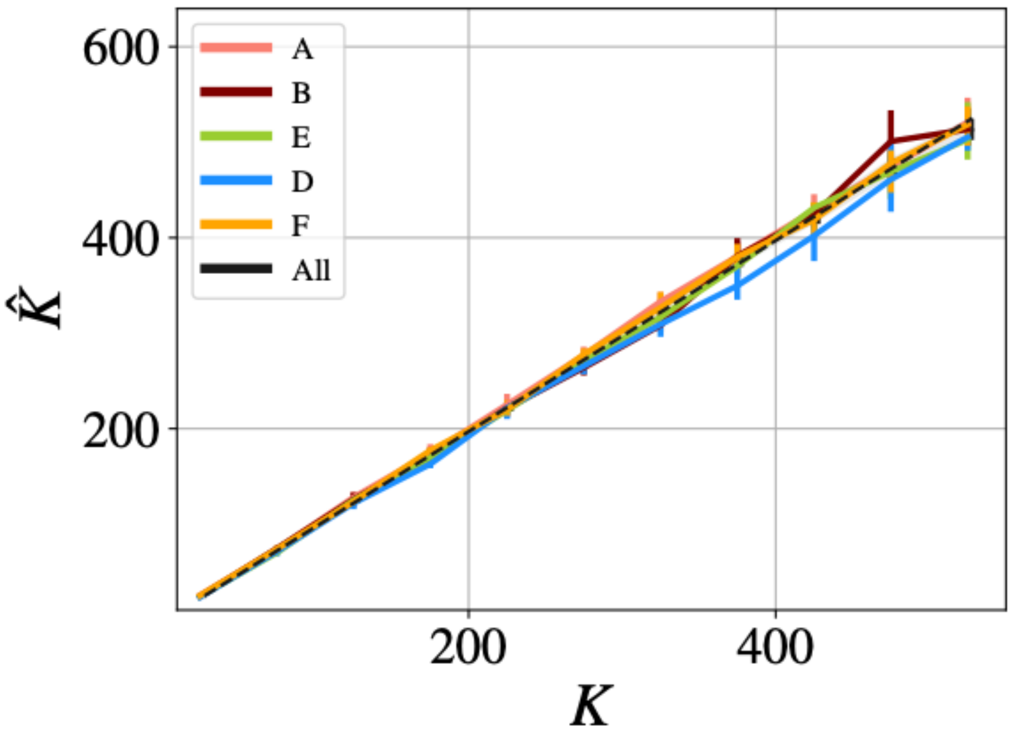
(Similar results for other MF models)

Method 1: Human labor

Finds right K for PMF



…and Compound PMF



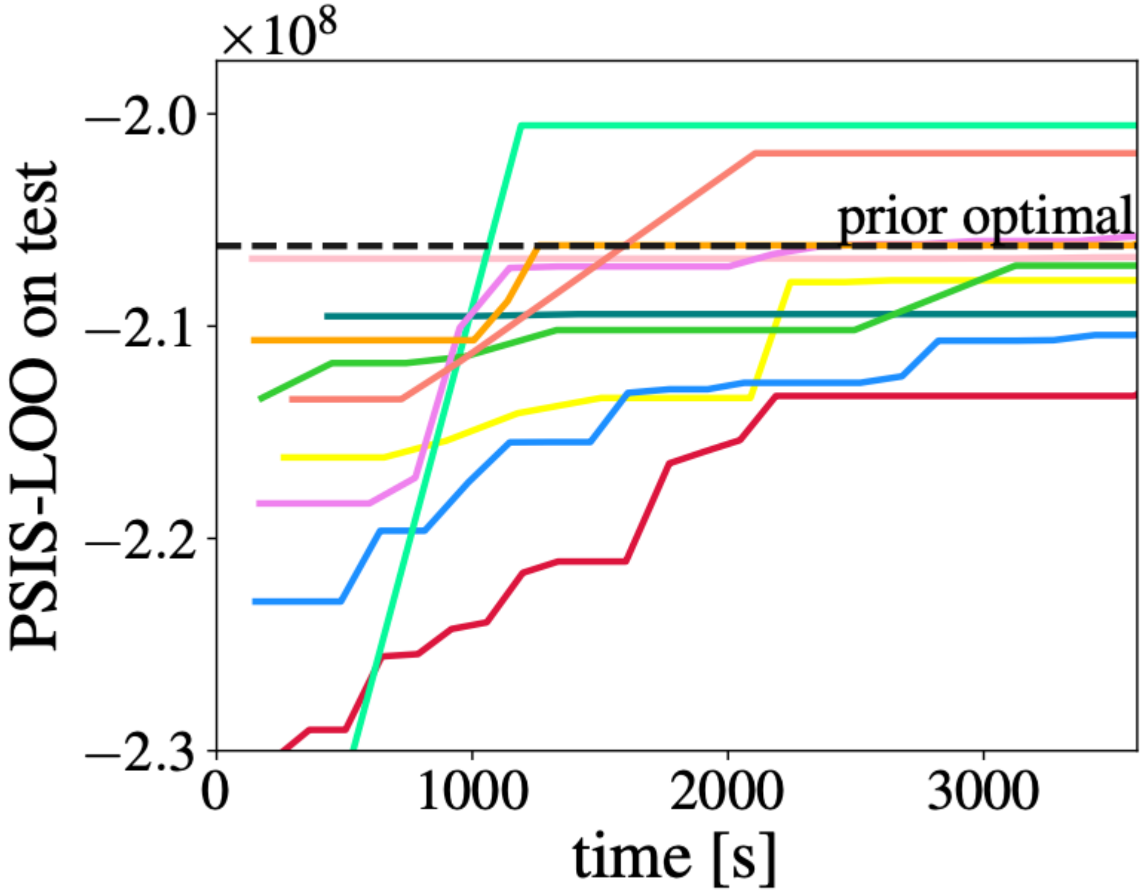
See the JMLR paper for many more examples, e.g. sensitivity to model mismatch and other MF models

What really matters

Bayesian optimization eventually leads to better prior, but is slow even for fast models

What if the model was truly large (foundation models etc)?

Also: Works as initialization for BO



PMF with VI on lastfm data

Method 2: SGD

For models with no analytic statistics we need Monte Carlo and iterative optimization

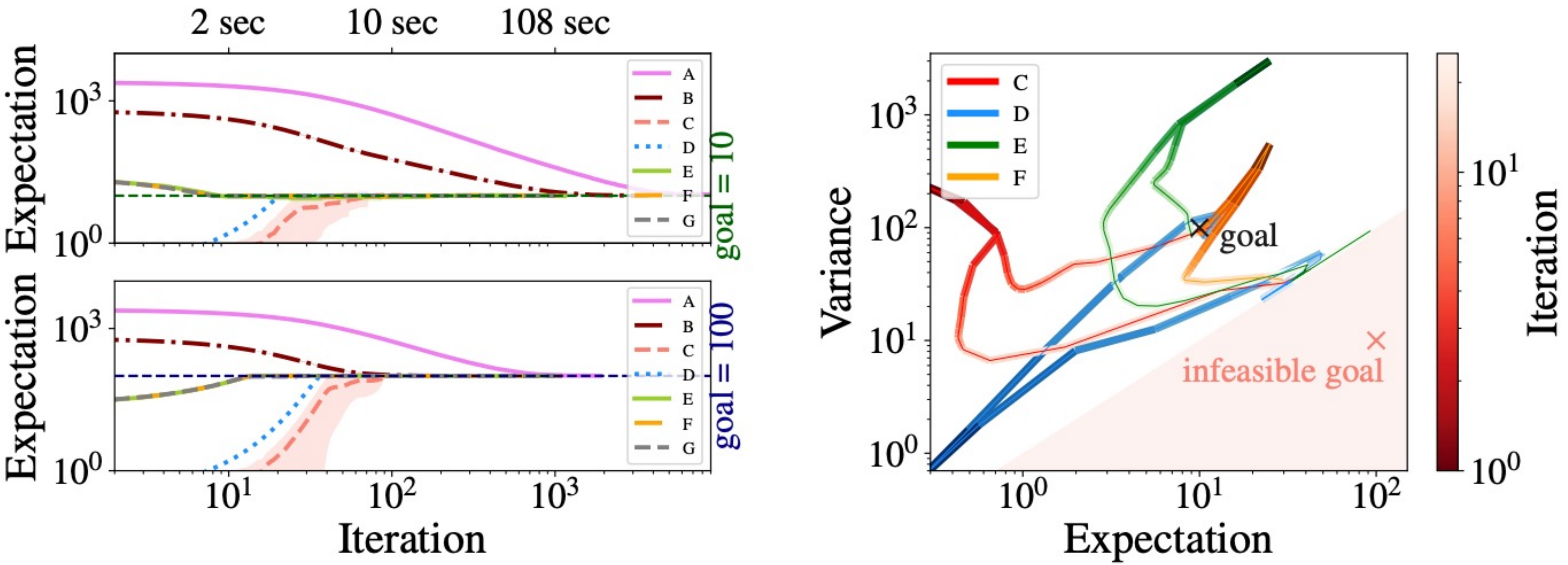
Standard stochastic gradient descent

- Reparameterization for propagating gradients through sampling

$$\lambda^* = \arg\min_{\lambda} d(T(x|\lambda), T^*)$$

No longer immediate analytic solution, but still no inference needed

Method 2: SGD



Fast enough and can detect infeasible targets

Summary

How to choose the priors (amongst chosen prior family)

Statistical models

- Direct specification near impossible for domain experts
- Prior elicitation in (low-dimensional) observation space truly helps
- We still need better algorithms, software and killer applications

Machine learning

- We can skip cross-validation for Bayesian models
- Immediate or very fast solution, without inference
- More work needed for discriminative models or high-dimensional outputs

You can start the virtuous cycle for prior elicitation:

Make new general algorithms easily available or use one in your Science paper

References

1. Agiashvili. **Probabilistic predictive elicitation.** MSc thesis, University of Helsinki, 2021.
2. Gelman et al. **Bayesian workflow.** arXiv:2011.01808, 2020.
3. Hartmann, Agiashvili, Bürkner, Klami. **Flexible prior elicitation via the prior predictive distribution.** UAI, 2020.
4. Kadane and Wolfson. **Experiences in elicitation**, JRSS:D, 1998.
5. Lawless, J. **Statistical Models and Methods for Lifetime Data.** Wiley Series in Probability and Statistics, 2011.
6. Mikkola et al. **Prior knowledge elicitation: The past, present, and future**. Bayesian Analysis, DOI: 10.1214/23-BA1381, 2023.
7. de Souza da Silva et al. **Prior specification for Bayesian matrix factorization via prior predictive matching.** JMLR, 2023.