Constrained Data-Fitters

an inspiration from machine learning

Larry Samuelson, Jakub Steiner

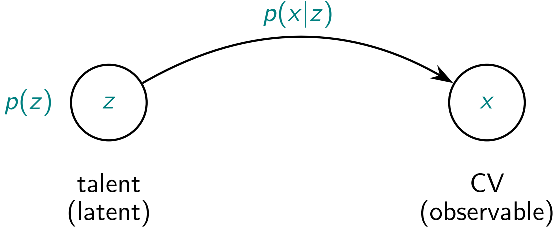
ESA Columbus 2024

The Map is Not the Territory

Alfred Korzybski

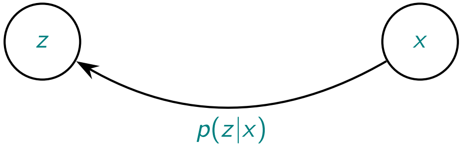
Statistical Model

a "map"



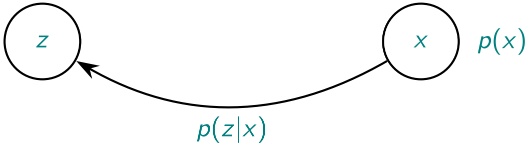
Using the Model

a posteriori optimal choice



Posterior Approach

same model, different perspective



How are the Models Chosen?

estimation

unknown data-generating process $q(y)$

maximum-likelihood estimate

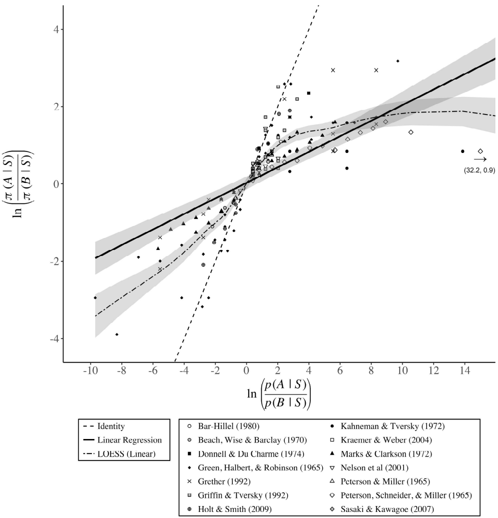
$$\arg\max_{p \in \mathcal{P}} p\text{-likelihood}(y_1, \ldots y_n)$$

Wald'49:

well-specified agent learns the true distribution: $p_n \to q$

People Aren't Bayesian

Benjamin'19, Ortoleva'24



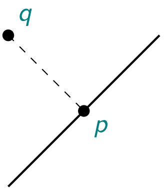
People are Misspecified

true process is not included in the set of hypotheses

Berk'66, White'82:

as the sample expands, the estimate converges to the least wrong model

$$\arg\min_{p \in \mathcal{P}} \mathrm{KL}\left(q \parallel p\right)$$



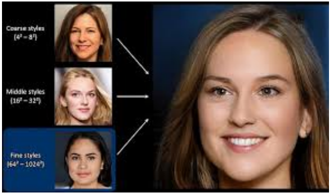
Blei et al. '17:

> *One of the core problems of modern statistics is to approximate difficult-to-compute probability densities. This problem is especially important in Bayesian statistics, which frames all inference [. . .] involving the posterior density.*

the true posterior is projected on a set of tractable distributions

Machines Are Misspecified Too

generative task



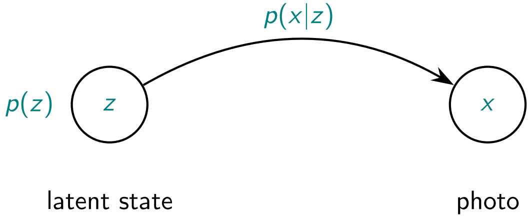
how it's done:

- estimate $q(x)$ from the training sample
- draw x_{n+1} from the estimated distribution

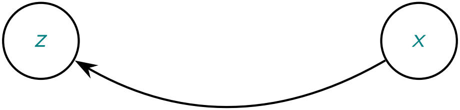
two frictions:

- misspecification
- fit is difficult to evaluate

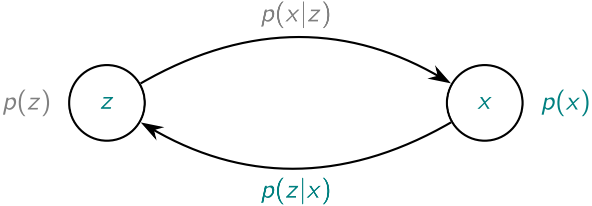
Variational Autoencoders



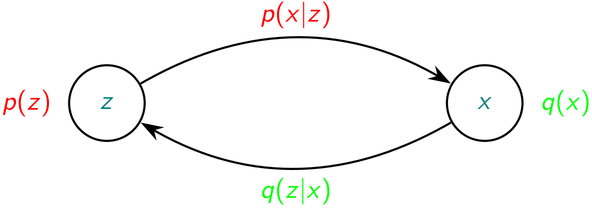
Variational Autoencoders



Variational Autoencoders

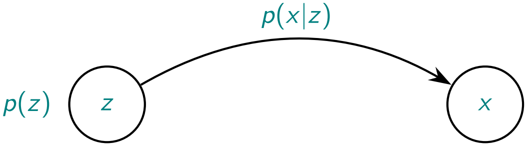


Variational Autoencoders

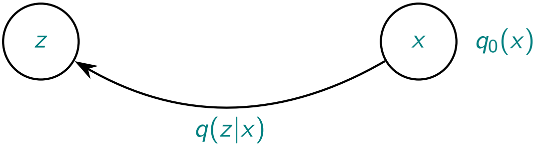


generative model vs recognition model

Two Models



generative model $p(x, z) = p(z)p(x \mid z)$



recognition model $q(x, z) = q_0(x)q(z \mid x)$

Example

generative model:

- $p(z)$ prior distribution of talent in population
- $p(x \mid z)$ stochastic CV of each talent type

recognition model:

- large sample of the job applicants' CVs
- $q_0(x)$ – empirical distribution of the sample
- belief $q(z \mid x)$ about talent z of a candidate with CV x

Choice of the Two Models

let's proceed backwards

given the generative model, choose the recognition model

- that is most consistent with the generative model
- subject to a cognitive constraint

choose the generative model with the best subjective fit to data

- accounting for own cognitive constraint during updating

Choice of the Recognition model

variational Bayes' methods, Jordan et al.'99

$$\min_{\tilde{q}(x,z)} \quad \mathrm{KL}\left(\tilde{q}(x,z) \,\|\, p(x,z)\right)$$

$$\text{s.t.} \quad \tilde{q}(x,z) \in \mathcal{Q}$$

$$\tilde{q}(x) = q_0(x)$$

Choice of the Recognition model

variational Bayes' methods, Jordan et al.'99

$$\min_{\tilde{q}(x,z)} \quad \mathrm{KL}\left(\tilde{q}(x,z) \,\|\, p(x,z)\right)$$

$$\text{s.t.} \quad \tilde{q}(x,z) \in \mathcal{Q}$$

$$\tilde{q}(x) = q_0(x)$$

empirical constraint

Choice of the Recognition model

variational Bayes' methods, Jordan et al.'99

$$\min_{\tilde{q}(x,z)} \quad \mathrm{KL}\left(\tilde{q}(x,z) \,\|\, p(x,z)\right)$$

$$\text{s.t.} \quad \tilde{q}(x,z) \in \mathcal{Q}$$

$$\tilde{q}(x) = q_0(x)$$

updating constraint

Choice of the Recognition model

variational Bayes' methods, Jordan et al.'99

$$\begin{aligned} \min_{\tilde{q}(x,z)} \quad & \mathrm{KL}\left(\tilde{q}(x,z) \,\|\, p(x,z)\right) \\ \text{s.t.} \quad & \tilde{q}(x,z) \in \mathcal{Q} \\ & \tilde{q}(x) = q_0(x) \end{aligned}$$

subjective fit

Choice of the Generative Model

variational autoencoder, Kingma&Welling'13

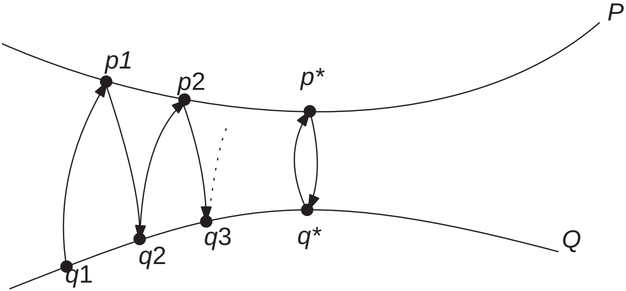
$$\min_{\tilde{p}(x,z),\tilde{q}(x,z)} \quad \mathrm{KL}\left(\tilde{q}(x,z) \parallel \tilde{p}(x,z)\right)$$

$$\text{s.t.} \quad \tilde{p}(x,z) \in \mathcal{P}$$

$$\tilde{q}(x,z) \in \mathcal{Q}$$

$$\tilde{q}(x) = q_0(x)$$

Information Geometry



Some Updating Constraints

no constraint:

- $\mathcal{Q} = \Delta(X \times Z)$

$\Rightarrow$ back to homo economicus:

- Bayesian updates: $q(z \mid x) = p(z \mid x)$
- standard likelihood

Some Updating Constraints

computational constraint:

- $\mathcal{Q}$ is a family of tractable distributions

relevant in machine learning

Some Updating Constraints

causal constraint; e.g.:

- $x = \text{CV}$, $z = (\text{aptitude, grit})$
- q must comply with a causal graph, e.g.:

$$\text{aptitude} \to \text{CV} \leftarrow \text{grit}$$

- $\Leftrightarrow$ a factorization constraint on $q(\text{CV, aptitude, grit})$

see Spiegler'20 for review

Updating As Constrained Optimization

Kullback'59 principle of minimum discrimination information

- posterior minimizes KL-divergence from prior s.t. new information
- originates in Laplace's principle of insufficient reason

Dominiak, Kovach & Tserenjigmid'21

- axiomatization and extensions

In Machine Learning

machine evaluates fit of a model $p(x, z)$ to observable data x

- intractable marginalization $p(x) = \sum_z p(x, z)$

instead, the machine evaluates evidence lower bound

Sanov's Theorem

draw a large sample from p

suppose the empirical distribution q is in $\mathcal{Q}$

then, it is

$$\begin{aligned} \arg\min_{\tilde{q}} \quad & \mathrm{KL}(\tilde{q} \parallel p) \\ \text{s.t.} \quad & \tilde{q} \in \mathcal{Q} \end{aligned}$$

our agent reasons about the sample of (x, z) s.t. the constraints

Example

$x = \text{CV}$, $z = (\text{aptitude}, \text{grit})$

generative model:

- any $p(\text{aptitude}, \text{grit})$
- restrictions imposed on $p(\text{CV} \mid \text{aptitude}, \text{grit})$

recognition model:

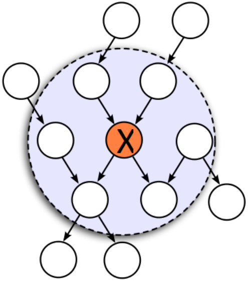
- $\text{CV} \rightarrow \text{aptitude} \rightarrow \text{grit}$

deterministic collapse

The agent models grit as a deterministic function of aptitude.

grit has no explanatory power at the recognition stage
$\Rightarrow$ it is not used in the generative stage

Generalization



deterministic collapse

The agent models variables from outside the Markov Boundary as a deterministic function of the variables from within the boundary.

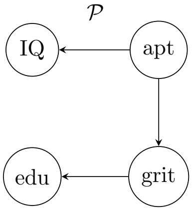
Two Frictions

		$\mathcal{P}$	
		Well-specified	**Miss-specified**
$\mathcal{Q}$	**Bayes' Rationality**	Wald'49	Berk'66
	Updating Friction	model-fitting	model-fitting

Correlation Neglect

$x = (\text{IQ score}, \ \text{education})$ and $z = (\text{aptitude}, \ \text{grit})$

the true process $q_0(\text{IQ}, \text{edu})$ exhibits correlation



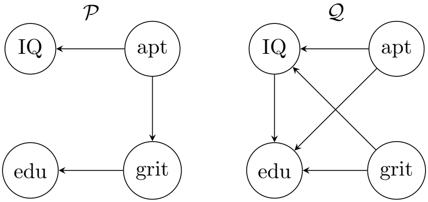
agent is well-specified

- $\Rightarrow$ she learns the true process if updating is unconstrained

Correlation Neglect

$x = (\text{IQ score}, \ \text{education})$ and $z = (\text{aptitude}, \ \text{grit})$

the true process $q_0(\text{IQ}, \text{edu})$ exhibits correlation



the updating constraint

- $\Rightarrow$ optimal correlation neglect $p(\text{IQ}, \text{edu}) = p(\text{IQ})p(\text{edu})$

definition

The agent has rational expectations if she isn't systematically surprised:

$$p(z) = \mathrm{E}_{q_0(x)}\, q(z \mid x).$$

in general, our agent does not have RE

Rational Expectations

result

proposition

An agent who can conceive any $p(z)$ forms rational expectations.

standard RE is forced by the Bayes' law

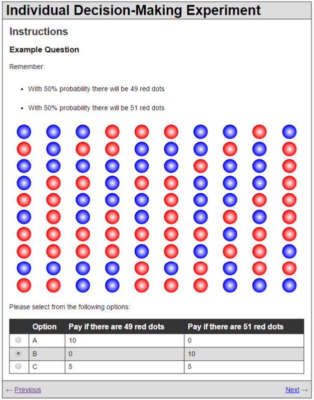
- RE can fail in our framework, but hold at the optimum

a popular non-Bayesian intuition in support of RE:

- systematically surprised agent should adjust her prior ✓

Dean & Neligh'23:

observable information structures



Individual Decision-Making Experiment

Instructions

Example Question

Remember:

- With 50% probability there will be 49 red dots
- With 50% probability there will be 51 red dots

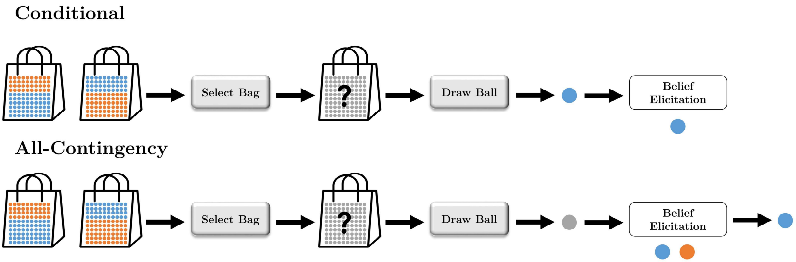
Please select from the following options:

	Option	Pay if there are 49 red dots	Pay if there are 51 red dots
○	A	10	0
◉	B	0	10
○	C	5	5

← Previous Next →

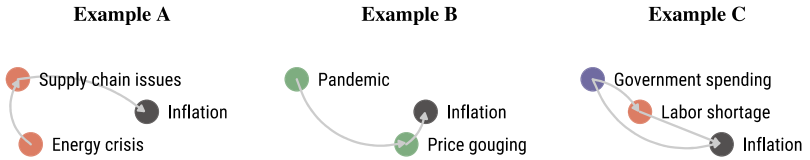
Aina, Amelio & Brütt'23

- is it misspecification? Bohren & Hauser'23
- failure of Bayesian reasoning?



Narratives

Andre, Haaland, Roth & Wohlfart'23



Machines and Humans

machines

- variational autoencoders

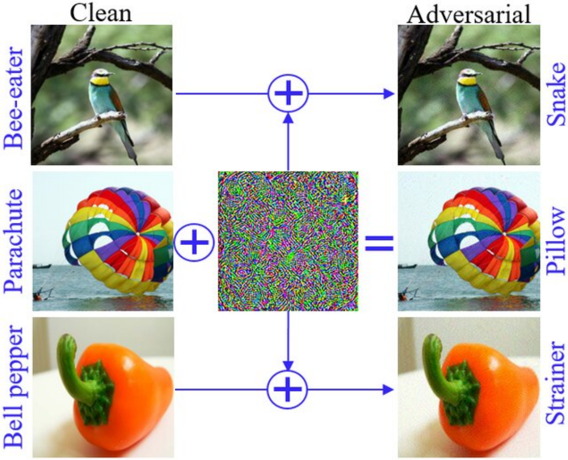
humans

- predictive coding

machine as a model of humans?

- perfect fMRI: observable latent representation
- flexible design: you can deceive machine
- twin studies: copies of the same machine

Experimental Methods in Machine Learning
Akhtar et al. '21



this informs about the geometry of the latent representation

Literature

non-Bayesian updating: Dominiak, Kovach & Tserenjigmid '21; Jakobsen '21; Ortoleva '12; Zhao '22

machine learning: Caplin, Martin & Marx '23; Zhao, Ke, Wang & Hsieh '20; Aridor, da Silveira & Woodford '24

Bayesian networks: Spiegler '16,'20; Sloman '05; Pearl '88; Ambuehl & Thysen '24; Andre, Haaland, Roth & Wohlfart '23

belief inconsistencies: Aina, Amelio & Brütt '23; Bohren & Hauser '23

misspecified learning: Esponda & Pouzo '16; Fudenberg, Lanzani & Strack '21; Frick, Iijima, & Ishii '23