Variational Bayesian Methods (in Neuroscience)

Tyler Benster & Aaron Andalman
Deisseroth (Tyler & Aaron) and Druckmann (Tyler) Labs

Computational Neuroscience Journal Club
Stanford University

July 10, 2019

2019-07-15

Variational Bayesian Methods (in Neuroscience)
Tyler Benster & Aaron Andalman
Deisseroth (Tyler & Aaron) and Druckmann (Tyler) Labs
Computational Neuroscience Journal Club
Stanford University
July 10, 2019

Roadmap

Introduction: Why care about the distribution of data?

Problem: Analyzing high dimensional data is hard

Solution: Variational Methods

Discussion: Neuroscience applications

Roadmap

Introduction: Why care about the distribution of data?

Problem: Analyzing high dimensional data is hard

Solution: Variational Methods

Discussion: Neuroscience applications

Introduction: Why care about the distribution of data?

Introduction: Why care about the distribution of data?

Problem: Analyzing high dimensional data is hard

Solution: Variational Methods

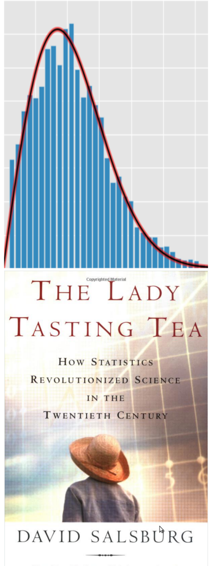
Discussion: Neuroscience applications

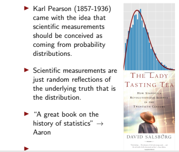
2019-07-15

Introduction: Why care about the distribution of data?
└Introduction: Why care about the distribution of data?
└Introduction: Why care about the distribution of data?

The probability distribution revolution

- Karl Pearson (1857-1936) came with the idea that scientific measurements should be conceived as coming from probability distributions.
- Scientific measurements are just random reflections of the underlying truth that is the distribution.
- "A great book on the history of statistics" → Aaron





Lets start with a bit of history. The idea that scientific measurements are best understood as reflecting underlying probabilities distributions is a relatively new idea.
It was a now famous thinker and scientist, Karl Pearson (of the Pearson correlation coefficient) who conceived the idea in late 18 hundreds.
He realized randomness was inherent part of nature and of scientific measure, and he formulated the idea that all measurments should be conceived of as coming from an underlying probability distributions.
In other words, the underlying truth is the distrubutions, and the measurements are just random reflections of this truth. At the time, this was a revolutionary idea.

The power of probability distributions

Distributions allow scientists to:

- Understand scientific measurement
- Predict the probability of specific data
- Test specific hypothesis (p-values)
- Produce generative models
- Better conceptual understanding data.

Distributions allow scientists to:
- Understand scientific measurement
- Predict the probability of specific data
- Test specific hypothesis (p-values)
- Produce generative models
- Better conceptual understanding data.

And it was an idea that revolutized science.
It allowed scientists to:

- Better understand their measurements.
- To make predictions about what data they should expect to observe.
- To test scientific hypotheses in a mathematically rigorous way.
- To build generative models of their data, and to test how well those models explain the observed measurments.
- And in general to have a better conceptual understanding of the data they generated.

Estimating distributions from data

Low-Dimensional:

- Great tools to fit and understand the underlying probability distribution of data.

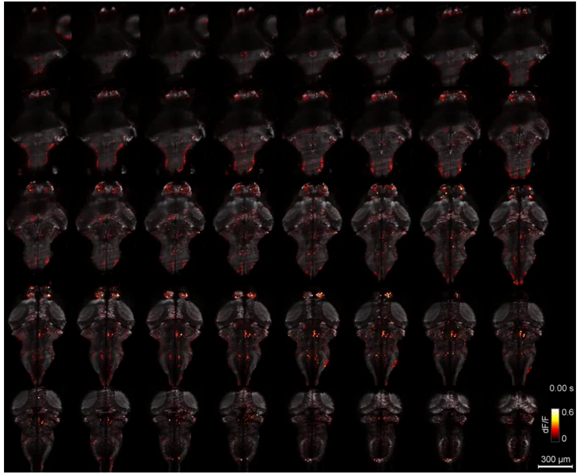
High-Dimensional:

- In some cases, classical statistical tools are insufficient.
- Problematic for modern neuroscience:
  - Thousands of electrodes.
  - Millions of voxels.

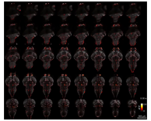
Low-Dimensional:
- Great tools to fit and understand the underlying probability distribution of data.
High-Dimensional:
- In some cases, classical statistical tools are insufficient.
- Problematic for modern neuroscience:
  - Thousands of electrodes.
  - Millions of voxels.

Since Pearson's early work, scientists and statisticians have devised an enormous number of related tools.
For example they've defined many many distributions, and they've created tools for working with those distrubitons (calculating likelihoods and fitting them).
One class of tools aim to estimate the underlying distribution that generated an observed empirical measurment.
These tools are highly effective when data is low dimensional, but they are sometimes insufficient when data is high dimensional.

How can we build statistical distributions for neuroscience datasets?



https://www.youtube.com/watch?v=CXYp9xCUhe0



https://www.youtube.com/watch?v=CXYp9xCUhe0

For example, consider whole brain imaging data from the zebrafish. This data can have millions of dimensions, which makes estimating the understanding joint probability distribution difficult. The number of possible states is enormous. The voxels are not independent. You can't simply make a histogram or a heat-map.

Variational Bayesian Methods

So this brings me to the topic I want to introduce today. Variational Bayesian Methods.

Varitional Bayesian methods are powerful statistical approach for estimating the underlying probability distrution of high dimensional data.

Variational Bayesian Methods

Estimate the probability distribution of high-dimensional neural data.

- Compute the probability of observing a particular neural state.
- Sample neural states/trajectory from the estimate.
- Generate statistics / test hypotheses
- Estimate latent factors/states that drive the observations.
- Reduce dimensionality (with advantages over other methods, e.g. PCA, T-SNE)

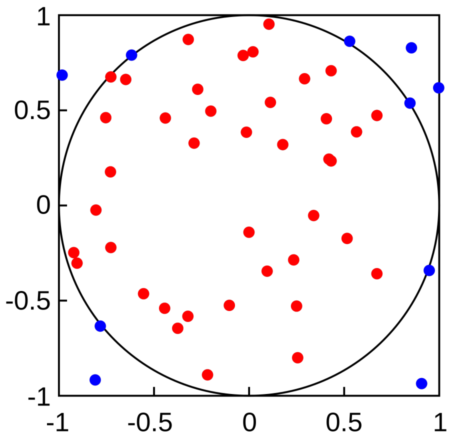
Estimate the probability distribution of high-dimensional neural data.
- Compute the probability of observing a particular neural state.
- Sample neural states/trajectory from the estimate.
- Generate statistics / test hypotheses
- Estimate latent factors/states that drive the observations.
- Reduce dimensionality (with advantages over other methods, e.g. PCA, T-SNE)

And this has several important possible uses in neuroscience. By estimating the probability distrubiton of, say, whole brain calcium imaging data, one could: . . .

Problem: Analyzing high dimensional data is hard

Introduction: Why care about the distribution of data?
Problem: Analyzing high dimensional data is hard
Solution: Variational Methods
Discussion: Neuroscience applications

Approximating the area of a circle with Monte Carlo



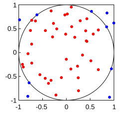
$40/50$ samples in circle gives area of $4 * 0.8 = 3.2 \approx \pi$. $\sim 2\%$ error

2019-07-15

Problem: Analyzing high dimensional data is hard

└─Problem: Analyzing high dimensional data is hard

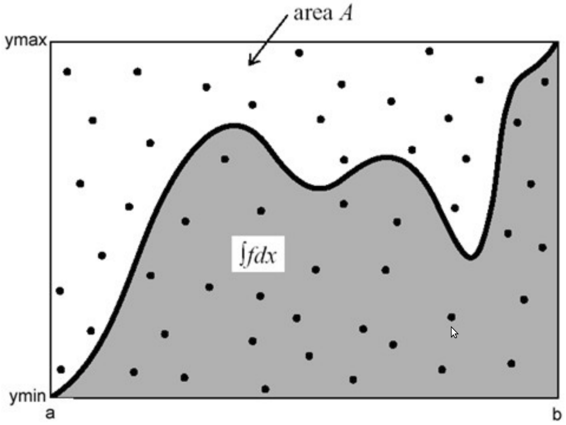
└─Approximating the area of a circle with Monte Carlo



40/50 samples in circle gives area of $4 * 0.8 = 3.2 \approx \pi$. $\sim 2\%$ error

CC0 1.0, wikimedia, MonteCarloIntegrationCircle.svg

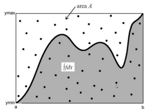
Useful for analytically intractable integrals



Robert Lin - Monte Carlo Integration

2019-07-15

└Useful for analytically intractable integrals



Robert Lin - Monte Carlo Integration

Extension to unit hypercube

We can approximate a high-dimensional integral using a Monte Carlo approximation:

$$\int_0^1 \cdots \int_0^1 g(x_1, \ldots, x_n) dx_1, \ldots, dx_n \approx \frac{1}{N} \sum_{j=1}^{N} g(\bar{x}_j)$$

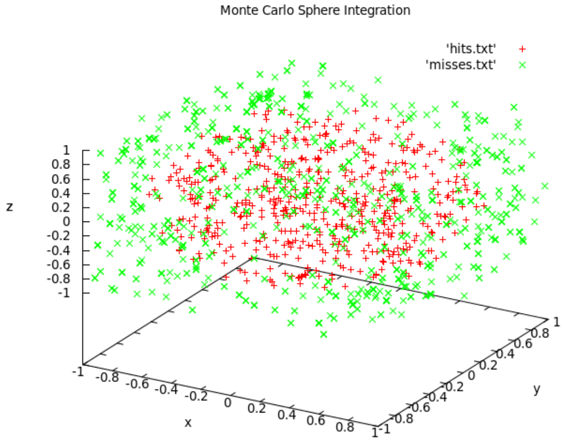
where $\bar{x}_1, \ldots, \bar{x}_N \sim \mathcal{U}(0,1)$ is the i^{th} random sample

We can approximate a high-dimensional integral using a Monte Carlo approximation:

$$\int_0^1 \cdots \int_0^1 g(x_1, \ldots, x_n) dx_1, \ldots, dx_n \approx \frac{1}{N} \sum_{j=1}^{N} g(\bar{x}_j)$$

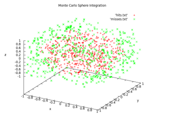
where $\bar{x}_1, \ldots, \bar{x}_N \sim \mathcal{U}(0,1)$ is the i^{th} random sample

Performance scales poorly with number of dimensions



1000 samples, estimate $= 4.072$, actual $= 4.189$, $\sim 3\%$ error

http://www2.hawaii.edu/ yuxian/phys305/a6/


1000 samples, estimate = 4.072, actual = 4.189, ~ 3% error
http://www2.hawaii.edu/ yuxian/phys305/a6/

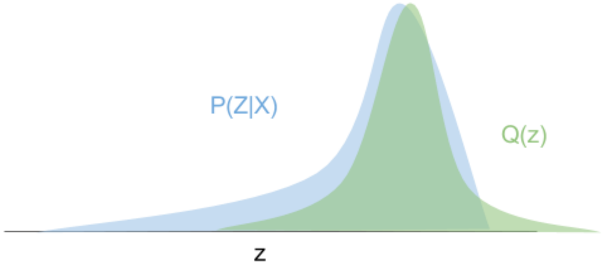
2019-07-15

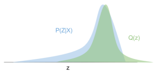
Solution: Variational Methods
└Solution: Variational Methods
└Solution: Variational Methods

Introduction: Why care about the distribution of data?
Problem: Analyzing high dimensional data is hard
Solution: Variational Methods
Discussion: Neuroscience applications

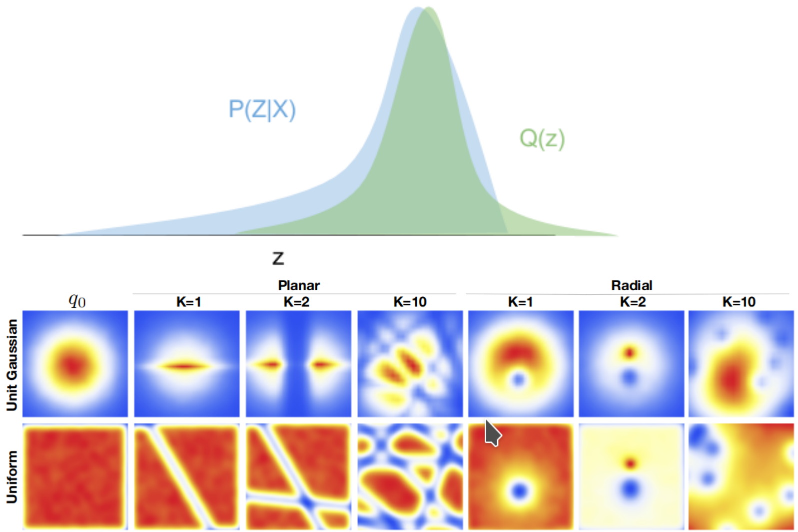
Alternate approach: Variational Bayes

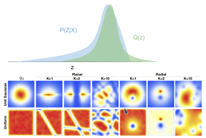
Variational auto-encoders and normalizing flows



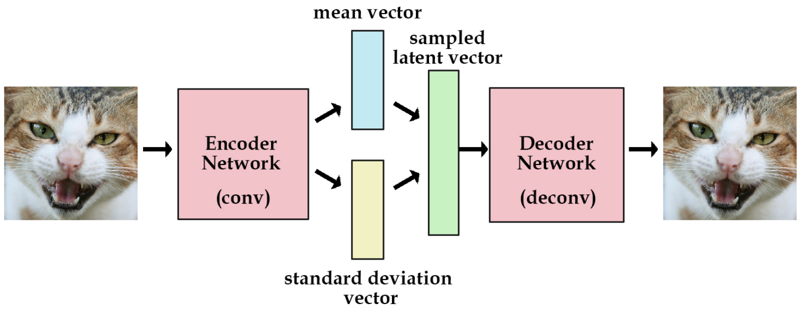


Alternate approach: Variational Bayes

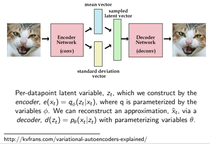




Variational auto-encoder



Per-datapoint latent variable, z_t, which we construct by the *encoder*, $e(x_t) = q_\phi(z_t|x_t)$, where q is parameterized by the variables ϕ. We can reconstruct an approximation, $\hat{x}_t$, via a *decoder*, $d(z_t) = p_\theta(x_t|z_t)$ with parameterizing variables θ.



Derive lower bound → optimize!

Starting with the log probability of x, we derive (5), the Evidence lower bound (ELBO):

$$
\begin{aligned}
\log p_\theta(x) &= \log \int_Z p_\theta(x,z) && (1)\\
&= \log \int_Z p_\theta(x,z)\frac{q_\phi(z|x)}{q_\phi(z|x)} && (2)\\
&= \log E_{q(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] && (3)\\
&\geq E_{q(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] && (4)\\
&= E_{q(z|x)}\left[\log p_\theta(x,z)\right] - H(q_\phi(z|x)) = L && (5)
\end{aligned}
$$

Kingma & Welling 2013

Starting with the log probability of x, we derive (5), the Evidence lower bound (ELBO):

$$
\begin{aligned}
\log p_\theta(x) &= \log \int_Z p_\theta(x,z) && (1)\\
&= \log \int_Z p_\theta(x,z)\frac{q_\phi(z|x)}{q_\phi(z|x)} && (2)\\
&= \log E_{q(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] && (3)\\
&\geq E_{q(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] && (4)\\
&= E_{q(z|x)}\left[\log p_\theta(x,z)\right] - H(q_\phi(z|x)) = L && (5)
\end{aligned}
$$

Kingma & Welling 2013

Note that 4 is true thanks to Jensen's inequality: $\varphi(E[X]) \geq E[\varphi(X)]$ for convex function φ. Intuitively, we want to improve our approximation of $p(x_t)$ by optimizing θ, ϕ to maximize the ELBO.

Understanding when the bound is tight

An alternate derivation gives insight for when this bound is tight:

$$\mathrm{KL}(q_\phi(z|x)||p(z|x)) = \int_Z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} \tag{6}$$

$$= \int_Z q_\phi(z|x) \log \frac{q_\phi(z|x)p_\theta(x)}{p_\theta(x,z)} \tag{7}$$

$$= H(q_\phi(z|x)) + \log p_\theta(x) \int_Z q_\phi(z|x) - E_{q_\phi(z|x)}[\log p_\theta(x,z)] \tag{8}$$

$$L = \log p_\theta(x) - \mathrm{KL}(q_\phi(z|x)||p_\theta(z|x)) \tag{9}$$

An alternate derivation gives insight for when this bound is tight:

$$\mathrm{KL}(q_\phi(z|x)||p(z|x)) = \int_Z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} \tag{6}$$

$$= \int_Z q_\phi(z|x) \log \frac{q_\phi(z|x)p_\theta(x)}{p_\theta(x,z)} \tag{7}$$

$$= H(q_\phi(z|x)) + \log p_\theta(x) \int_Z q_\phi(z|x) - E_{q_\phi(z|x)}[\log p_\theta(x,z)] \tag{8}$$

$$L = \log p_\theta(x) - \mathrm{KL}(q_\phi(z|x)||p_\theta(z|x)) \tag{9}$$

Demo time!

https://bit.ly/2LcEhow

https://bit.ly/2LcEhow

Summary

- Problem: Analyzing high dimensional data is hard
  -
- Solution: Variational Methods
  -
- Example: variational autoencoder
  -

- Problem: Analyzing high dimensional data is hard
  -
- Solution: Variational Methods
  -
- Example: variational autoencoder
  -

└Discussion: Neuroscience applications

Introduction: Why care about the distribution of data?
Problem: Analyzing high dimensional data is hard
Solution: Variational Methods
Discussion: Neuroscience applications

Example: LFADS