Statistics in Astronomy

A View Through the Looking Glass

Harrison B. Prosper

Kirby W. Kemper Endowed Professor of Physics

Department of Physics, Florida State University
Seminar, Queen Mary University, London

October 22, 2020

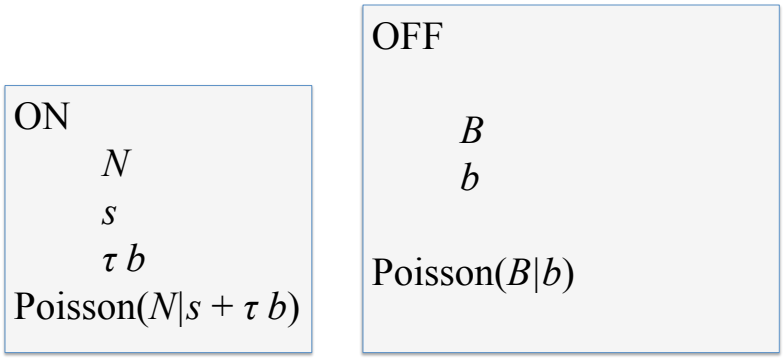
Outline

Outline

Astronomy and particle physics share common statistical problems of which the On/Off problem is one of the most important.

The ON/OFF problem An observation is made in one region called the ON-source (or signal) region and an independent observation is made in another region called the OFF-source (or background) region.

The On/Off problem[1] in astronomy and particle physics[2]:



Let τ be the ratio of ON-source to OFF-source observation times of a telescope, then for N and B photon counts in the two observation regions, respectively, the likelihood is given by

$$p(D \mid s, b) = \text{Poisson}(N, s + \tau b)\,\text{Poisson}(B, b).$$

Question: What is the statistical significance of the signal s?

---

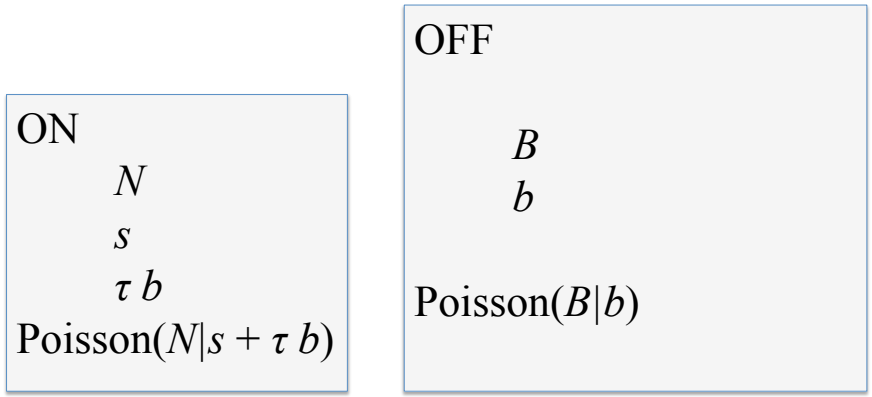
[1] T. P. Li and Y. Q. Ma, Analysis method for results in gamma-ray astronomy, *Astrophs. J.* **272**, 313 (1983).

[2] J. T. Linnemann, Measures of Significance in HEP and Astrophysics, PHYSTAT2003, SLAC, Stanford CA, September 8-11, 2003; R. D. Cousins, J. T. Linnemann, J. Tucker, Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process, NIM A **595** 480-501 (2008).

The frequentist solution is well-known and widely used. One computes the profile likelihood

$$L_p(s) = p(D \mid s, \hat{b}(s)),$$

where $\hat{b}(s)$ is the conditional maximum likelihood estimate (CMLE) of b, that is, the maximum likelihood estimate (MLE) of b for a given value of s.

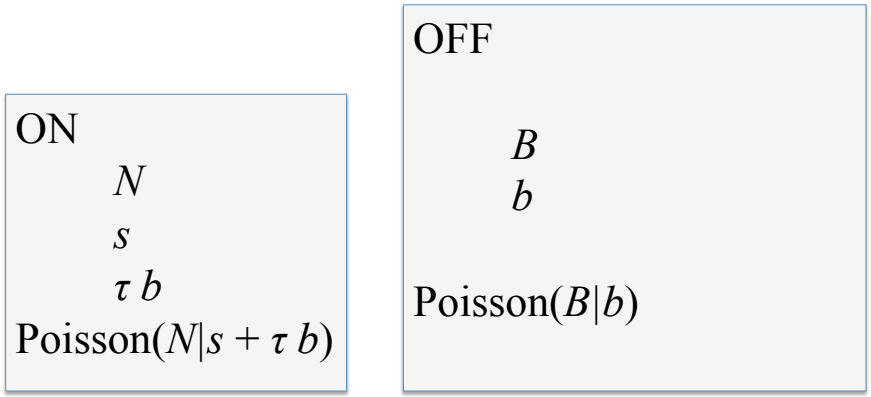


According to Wilks' theorem[3], when the counts are large the probability density of the quantity $F = -2 \log \Lambda(0)$, where $\Lambda(s) = L_p(s)/L_p(\hat{s})$, is $p(\chi^2, \text{ndf} = 1)$ if the hypothesis $s = 0$ is true.

Since, $F \approx \chi^2$, it follows that $Z = \sqrt{F}$ indicates a Z-standard deviation observation away from $s = 0$.

[3] See for example, G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur.Phys.J.C71:1554,2011.

The Bayesian solution to the On/Off problem requires the probability $p(H_1 \mid D)$ of the hypothesis $s > 0$, which we denote by H_1.



This probability is given by

$$p(H_1 \mid D) = \frac{B_{10}\, p(H_1)}{B_{10}\, p(H_1) + p(H_0)},$$

where H_0 denotes the hypothesis $s = 0$.

The ratio $B_{10} = p(D \mid H_1) \,/\, p(D \mid H_0)$ is called the Bayes factor. It is the amount by which the probability of hypothesis H_1 has *changed* relative to that of hypothesis H_0, given the observations.

$p(H_1)$ and $p(H_0)$ are the prior probabilities you assign to the respective hypothesis. It is uncontroversial to argue that $p(H_1) = p(H_0)$ assigns equal weight to each!

What *is* controversial is computing

$$p(D \mid H_1) = \int ds \int db\, p(D \mid s, b)\, \pi(s, b),$$
$$= \int \left[\int db\, p(D \mid s, b)\, \pi(b) \right] \pi(s|b)\, ds,$$

where we have used $\pi(s, b) = \pi(s|b)\, \pi(b)$. Note also that

$$p(D \mid H_0) = \int db\, p(D \mid s = 0, b)\, \pi(b)\, db.$$

Scientists can usually agree on the form of the evidence-based prior $\pi(b)$. But, the choice of the prior $\pi(s|b)$ is controversial and therefore problematic. The point is that in order for the Bayes factor $B_{10} = p(D \mid H_1) \,/\, p(D \mid H_0)$ to be well-defined, that is, not scaled by an arbitrary constant, the prior $\pi(s|b)$ must be proper, that is, it must satisfy

$$\int \pi(s|b)\, ds = 1.$$

If your friendly neighborhood astrophysicist makes a precise prediction for the signal, say $s = s_0$, this prediction could be encoded in the (proper) prior as follows

$$\pi(s|b) = \delta(s - s_0),$$

whereupon the probability of the hypothesis $s = s_0$ can be computed using the procedure on the previous slide. But, what if you don't want to do that?

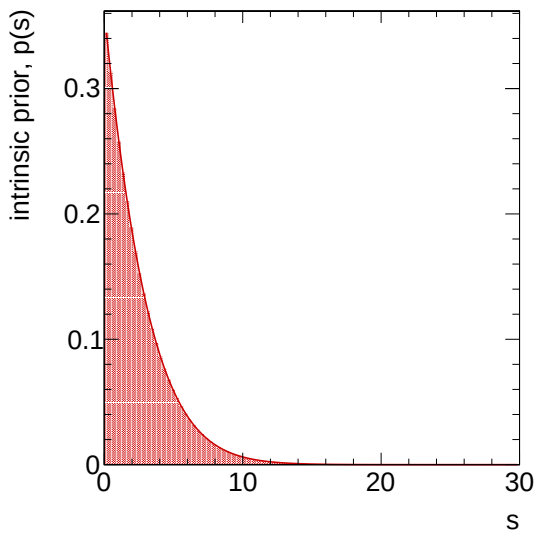
At a conference in 2008 (PHYSTAT 2008), American statistician Jim Berger introduced us to the *intrinsic prior* construction.

1. Compute $p_0(s \,|\, D) = p(D \,|\, s)\, \pi_0(s)/p_0(D)$, making sure that a prior is used for which $p_0(D) = \int p(D|s)\, \pi_0(s)\, ds < \infty$.
2. Before observations, the data data D are unknown. Therefore, following standard Bayesian practice we marginalize (that is, integrate) over unknowns. Assuming that the signal is negligible, we compute $\pi_I(s) =< p_0(s \,|\, D) >$, where we average with respect to $p(D \,|\, s = 0)$.

Example (Particle Physics: The Top Quark Discovery (1995, CDF, D0))

ON observation of $N = 17$ events with an effective OFF observation of $B = 40.1$ events with a scale factor $\tau = 0.0947$. At a signal of exactly $s = 14$ events, $p(D \mid H_1) = 9.3 \times 10^{-2}$, while $p(D \mid H_0) = 3.0 \times 10^{-6}$. Therefore, $B_{10} = 3.1 \times 10^4$, or $Z_{BF} = \sqrt{2 \log B_{10}} = 4.5$, which is a Bayesian analog of a frequentist $Z = 4.5\sigma$ effect.

However, if one prefers not to specify a specific signal hypothesis, it would be necessary to perform the integral $p(D \mid H_1) = \int p(D \mid s)\, \pi_I(s)\, ds$ over the intrinsic prior, $\pi_I(s)$.



But, what if you find Jim Berger's reasoning unpersuasive?

Outline

Question: How can one construct a prior such that $\pi(\theta)d\theta = \pi(\lambda)d\lambda$?

Consider the separation between two densities p and q from the same family, measured by the Kullback-Leibler (K-L) divergence

$$D(p, q) = \mathbb{E}_x[\log p(x|\theta)/q(x|\theta)]$$

between them. Notice that $D(p, q) \neq D(q, p)$; therefore, $D(p, q)$ cannot be interpreted as a distance. But, when $q = p(x|\theta + d\theta)$ it follows that

$$2D(p, q) = \sum_i \sum_j g_{ij} d\theta_i d\theta_j,$$

where g_{ij} is called the Fisher Information matrix, which is given by

$$g_{ij} = \mathbb{E}_x \left[\frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j} \right].$$

All averages are with respect to $p(x|\theta)$.

From the theory of differential geometry, it follows that

$$dV = \sqrt{\det g} \prod_i d\theta_i \equiv \sqrt{\det g}\, d\theta,$$

is an infinitesimal invariant volume within the parameter space, by which we mean

$$dV(\theta) = dV(\lambda)$$

holds for different parameterizations θ and λ of the probability density p.

In the 1930s, Cambridge physicist Sir Harold Jeffreys suggested the choice $\pi(\theta)d\theta = dV(\theta)$ for the prior whenever the only available information is the form of the probability function $p(x|\theta)$ and the domain of its parameters θ.

This prior, known as the Jeffreys prior, satisfies the desired property

$$\pi(\theta)d\theta = \pi(\lambda)d\lambda.$$

Myung *et al.*[4] provide a beautiful interpretation of the

Jeffreys Prior

$$\pi(\theta)\, d\theta = \sqrt{\det g}\, d\theta,$$

$$g_{ij} = \mathbb{E}_x \left[\frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j} \right] = -\mathbb{E}_x \left[\frac{\partial^2 \log p}{\partial \theta_i \partial \theta_j} \right]$$

for probability function $p(x|\theta)$,

namely, it is the prior that assigns equal weight to *probability densities* indexed by θ. That is, the Jeffreys prior is the *flat prior* in the space of probability densities and is independent of the parameterization.

For 1-dimensional parameter spaces, as in the Poisson problem, this invariant prior is widely accepted as *the* solution.

[4] I. J. Myung, V. Balasubramanian, and M. A. Pitt, *Counting probability distributions: Differential geometry and model selection*, PNAS, vol. **97**, 11171 (2000).

Alas, for the 2-parameter Gaussian $p(x|\mu,\sigma) = e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}/(\sigma\sqrt{2\pi})$, the Jeffreys prior is

$$\pi(\mu,\sigma)d\mu d\sigma = \frac{1}{\sigma^2}d\mu d\sigma.$$

Why "Alas"? Because this prior can yield very bad results. Indeed, the recommended prior, which follows from work by the statisticians Bernardo and Berger[5], is

$$\pi(\mu,\sigma)d\mu d\sigma = \frac{1}{\sigma}d\mu d\sigma.$$

Does this mean that the beautiful geometrical reasoning regarding priors is ultimately worthless?

Not necessarily!

[5] For a physicist's introduction see L. Demortier, S. Jain, HBP, *Reference priors for high energy physics*, Phys. Rev. D **82**, 034002 (2010).

What the difference between $\pi(\mu, \sigma) = 1/\sigma^2$ and $\pi(\mu, \sigma) = 1/\sigma$ is telling us is that in more than one dimension, assigning equal weight to every probability density is *not* necessarily sensible. We may be forced to weight these densities differently. For the Gaussian density, the prior should be

$$\pi(\mu, \sigma) d\mu d\sigma = w(\mu, \sigma) dV(\mu, \sigma) = w(\mu, \sigma) \frac{d\mu d\sigma}{\sigma^2},$$

where $w(\mu, \sigma)$ is the weight assigned to each density. If we set $w(\mu, \sigma) \propto \sigma$ we obtain the prior that statisticians prefer.

Remarkably, the principle of maximum entropy of physicist Edwin Jaynes[6] implies $w(\mu, \sigma) \propto e^S = e^{\log(\sigma/\sigma_0)}$, where S is the entropy of the Gaussian.

In summary The general form of an invariant prior is

$$\boxed{\pi(\theta) d\theta = w(\theta) dV(\theta)}$$

.

[6] E.T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. **106** (4) 620–630 (1957); *Information theory and statistical mechanics*, Phys. Rev. **108** (2) 171–190 (1957).

Outline

Consider the task of estimating the parameters a, b, c, and d, in the following problem

$$y = a + b \sin(cx + d).$$

This problem has 4 parameters. However, suppose that the amount of data is such that it is difficult to estimate the coefficient b. In that case, the effective number of parameters is closer to one.

A model in which some parameters cannot be identified uniquely, even with perfect data, is said to be *structurally non-identifiable*. A structurally identifiable model can become *practically non-identifiable* when using real data, which are always noisy.

Complex non-linear models, such as the first-order coupled differential equation models used in epidemiology, can be plagued[7] with non-identifiable parameters. Similar problems occur in astronomy and particle physics.

[7] No pun intended!

In a 2002 paper, Spiegelhalter et al.[8], suggested the following measure of the effective number of parameters

$$P = < \ell(\theta) > - \ell(\hat{\theta}),$$

where,

$$\ell(\theta) = -2 \log p(D \mid \theta),$$

and the average is with respect to the posterior density $p(\theta \mid D)$.

Note that $\ell(\hat{\theta}) \approx \chi^2_K + C$, where the number of degrees of freedom K is lower by the effective number of parameters P. If we knew the true value of $\ell(\theta)$ and, therefore, the number of degrees of freedom K_0 prior to fitting, then we could estimate the effective number of parameters using $P = K_0 - K$. However, we do not know $\ell(\theta)$. But, it can be estimated by averaging $\ell(\theta)$ over all possible values of θ weighted by $p(\theta|D)$.

[8] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. van der Linde, Bayesian measures of model complexity and fit, J. R. Statist. Soc.B **64** Part 4, 583-639 (2002)

Counting the effective number of parameters in a model is a useful diagnostic for flagging non-identifiability. We illustrate this using the Union 2.1 Compilation of Type 1a supernova data by the Supernova Cosmology Project[9]. The data comprise the redshift z and the distance modulus μ, and associated uncertainty, for 580 supernovae.

Example (Fitting ΛCDM model to Type 1a Sn Data)

The matter/energy density for the ΛCDM model as a function of the scale factor a of the universe is given by

$$\Omega(a) = \frac{\Omega_M}{a^3} + \frac{1 - \Omega_M - \Omega_\Lambda}{a^2} + \Omega_\Lambda,$$

where Ω_M, Ω_Λ, and H_0, the matter, vacuum energy, and Hubble constant, respectively, are the free parameters of the model.

[9] http://supernova.lbl.gov

Example (Fitting ΛCDM model to Type 1a Sn Data)

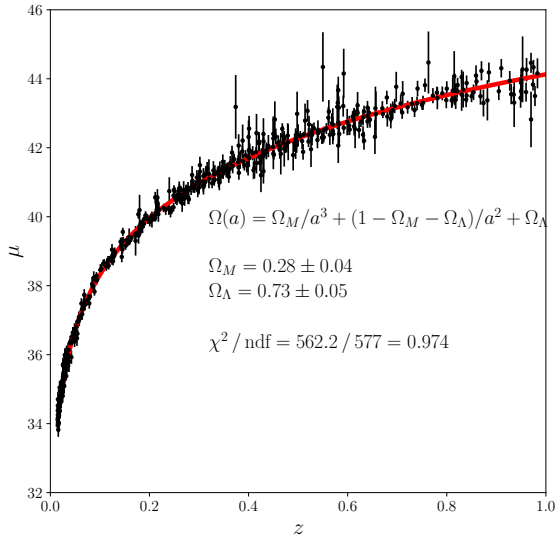
Distance modulus:

$$\mu = 5\log_{10}[(1+z)\,\sin(\sqrt{-(1-\Omega_M-\Omega_\Lambda)}\,u(z))\,/\,\sqrt{-(1-\Omega_M-\Omega_\Lambda)}] \\ -5\log_{10}(H_0) + 5\log_{10}(c) + 25,$$

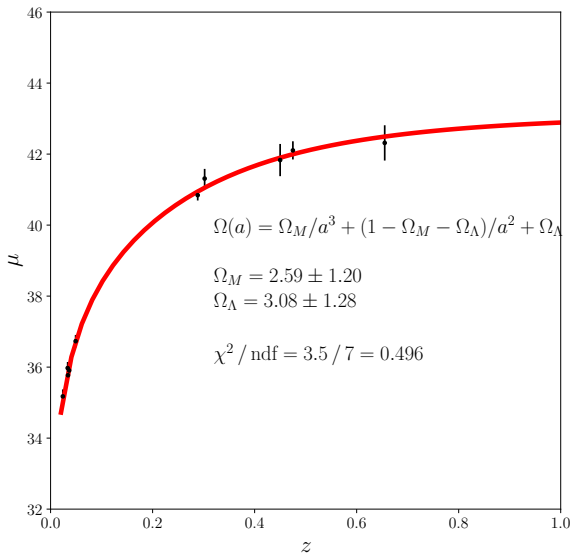
where

$$u(z) \equiv \int_{1/(1+z)}^{1} \frac{da}{a^2\sqrt{\Omega(a)}}.$$

We perform two fits: one with all the data points and another with only 10 data points. And we compute the effective number of parameters.



$P = 3.0$



$P = 2.6$

The effective number of parameters P behaves as expected. However, it would be useful to study the stability of the Spiegelhalter et al. measure when the sample size is small. It would also be of interest to understand how it is related to model selection[10].

[10] S. I. Vrieze, Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), Psychol Methods. 2012 June; 17(2): 228-243. doi:10.1037/a0027127; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3366160/.

Outline

Machine learning (ML) has been used in astronomy and particle physics for decades. However, the exponential rise in the development and use of ML can be traced to a breakthrough that occurred in 2006.

- In 2006, the field of machine learning suddenly became HOT when Hinton, Osindero and Teh[11] succeeded in training a deep neural network (DNN) by initializing its parameters sequentially, layer by layer. Each layer was trained to produce a representation of its inputs that served as the training data for the next layer. Then the network was tweaked using stochastic gradient descent.

- This breakthrough was viewed as compelling evidence that the training of DNNs requires careful initialization of parameters and sophisticated training algorithms.

[11] G. E. Hinton, S. Osindero, and Y. Teh, *A fast learning algorithm for deep belief nets*, Neural Computation **18**, 1527-1554.

- But, in 2010, a surprising counter example to the conventional wisdom was demonstrated by Cireșan *et al.*[12].

- The authors trained deep neural networks to classify the handwritten digits in the MNIST[13] data set, which comprises 60,000 $28 \times 28 = 784$ pixel images for training and 10,000 images for testing.

- Their model, with structure (784, 2500, 2000, 1500, 1000, 500, 10), outperformed all other methods that had been applied to the MNIST data set as of 2010. The error rate of this ∼12 million parameter DNN was 35 images out of 10,000.

The lessons drawn were: 1) very deep models are useful, 2) huge amounts of data are, however, needed to fit them, and 4) huge amounts of computing is a must.

[12] Cireșan DC, Meier U, Gambardella LM, Schmidhuber J. ,*Deep, big, simple neural nets for handwritten digit recognition.* Neural Comput. 2010 Dec; 22 (12): 3207-20.

[13] `http://yann.lecun.com/exdb/mnist/`

ARTICLE

doi:10.1038/nature24270

Mastering the game of Go without human knowledge

David Silver[1]*, Julian Schrittwieser[1]*, Karen Simonyan[1]*, Ioannis Antonoglou[1], Aja Huang[1], Arthur Guez[1], Thomas Hubert[1], Lucas Baker[1], Matthew Lai[1], Adrian Bolton[1], Yutian Chen[1], Timothy Lillicrap[1], Fan Hui[1], Laurent Sifre[1], George van den Driessche[1], Thore Graepel[1] & Demis Hassabis[1]

A long-standing goal of artificial intelligence is an algorithm that learns, *tabula rasa*, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. Here we introduce an algorithm based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules. AlphaGo becomes its own teacher: a neural network is trained to predict AlphaGo's own move selections and also the winner of AlphaGo's games. This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration. Starting *tabula rasa*, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo.

David Silver *et al.* Nature, vol. **550**, 19 October 2017.

2.08×10^{170}

Almost every ML training algorithm can be cast as an optimization problem whose goal is to minimize the average

$$R(f) = \frac{1}{N}\sum_{i=1}^{N} L(y, f(x_i, \theta)) + C(\theta)$$

of a suitable loss function $L(y, f)$ subject to some constraint $C(\theta)$.

The key point to note is that this sum approximates the *functional*

$$\begin{aligned} R[f] &= \int \left[\int dy\, L(y, f(x, \theta)\, p(y, x) \right] dx, \\ &\equiv \int G(f)\, dx, \end{aligned}$$

where $p(y, x)$ is the probability density of the targets y and features x of which the training data are a sample.

Example (Quadratic Loss: $L(y, f) = (y - f)^2$)

For the quadratic loss,

$$G(f) = \int (y - f)^2 \, p(y, x) \, dy$$

$$= p(x) \int (y - f)^2 \, p(y|x) \, dy,$$

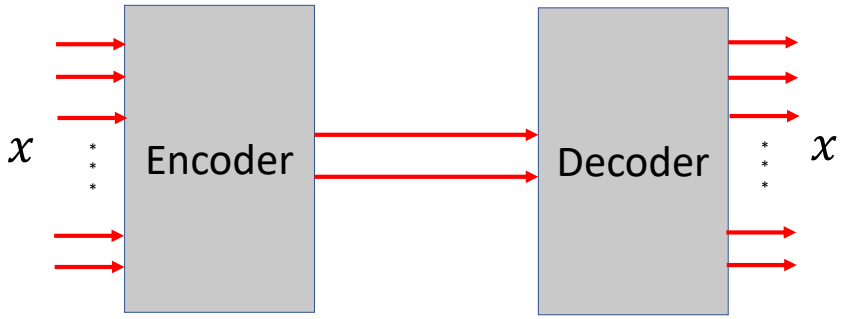
$$\frac{\delta G}{\delta f} = -2p(x) \int (y - f) \, p(y|x) \, dy = 0,$$

which implies $\boxed{f(x, \theta) = \int y \, p(y|x) \, dy, \text{ for some value of } \theta.}$

Conclusion If 1) the training data are sufficient and 2) $f(x, \theta)$ is sufficiently flexible (i.e., $\exists$ an $f(x, \theta)$ such that the functional derivative $\delta G / \delta f$ reaches zero) and 3) we use the quadratic loss then $f(x, \theta)$ will approximate the *mean* of the conditional density $p(y|x)$.

Example (Quadratic Loss: Mapping Gaia Variable Star Data to 2-D)

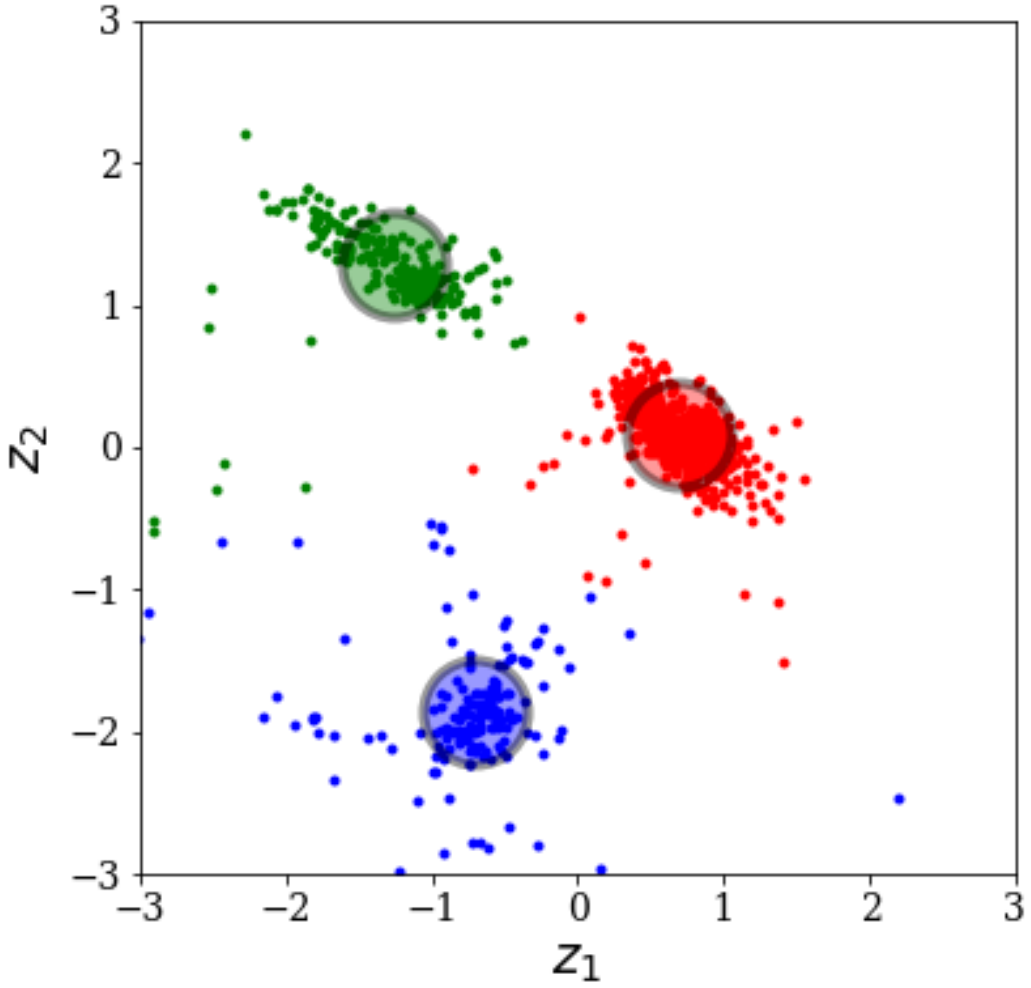
Fit an autoencoder



by minimizing the quadratic loss between the input to the encoder and the output of the decoder in order to map the Gaia Data Release 1 data $x = \nu, \Delta m, \overline{m}, A_2/A_1, \phi_2 - 2\phi_1$ for about 3,000 Cepheid and RRLyrae stars to a 2-D space $(z_1, z_2) \in R^2$. (See jupyter notebook for details.)

For the quadratic loss, $f(x, \theta)$ will approximate the integral $\int y\, p(y \,|\, x)\, dx = \int y\, \delta(y - x)\, dx = x$ provided that the conditions on the previous slide are met, *irrespective of the details of the autoencoder* $f(x, \theta)$. (Note, by the way, that for autoencoders, the quadratic loss yields functions that are *unbiased* estimates of the data x).

Example (Quadratic Loss: Mapping Gaia Variable Star Data to 2-D)



The blue dots are identified in the Gaia database as Cepheids, while the red and green dots are identified as RRLyrae stars.

The three clusters were found automatically.

There is a huge, and rapidly growing, number of ML models on the market, and many many variations on stochastic gradient descent for minimizing average loss functions.

However, these models, are ultimately approximating the same small set of mathematical quantities, which is determined by the loss function.

The quality of the approximation, however, is determined both my the flexibility of the model and the effectiveness of the minimization algorithm. This observation is the main motivation for the frenetic search for better ML models and algorithms.

Unfortunately, however, there seems to be very little work on the following inverse problem: given the mathematical quantity to be approximated, what average loss function should one minimize?

Outline

- I think it is helpful, from time to time, to see what's happening in a related field.
- In astronomy, just as in particle physics, there has been an explosion of developments in statistics over the past two decades and especially since about 2010.
- But, in spite of the wide variety of models and creative developments, the underlying principles of statistics remain intact and many fundamental problems have yet to be fully resolved. There is still a lot of difficult work to do.

Thank you!

1. T. P. Li and Y. Q. Ma, *Analysis method for results in gamma-ray astronomy*, Astrophys. J. **272**, 313 (1983).

2. J. T. Linnemann, *Measures of Significance in HEP and Astrophysics*, PHYSTAT2003, SLAC, Stanford CA, September 8-11, 2003; R. D. Cousins, J. T. Linnemann, J. Tucker, *Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process*, NIM A **595** 480-501 (2008).

3. G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur.Phys.J.C71:1554, 2011.

4. K. Maisinger, M. P. Hobson and A. N. Lasenby, *Maximum-entropy image reconstruction using wavelets*, Mon. Not. R. Astron. Soc. **347**, 339-354 (2004).

5. I. J. Myung, V. Balasubramanian, and M. A. Pitt, *Counting probability distributions: Differential geometry and model selection*, PNAS, vol. **97**, 11171 (2000).

6. L. Demortier, S. Jain, H. B. Prosper, *Reference priors for high energy physics*, Phys. Rev. D **82**, 034002 (2010).

7. E.T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. **106** (4) 620–630 (1957); *Information theory and statistical mechanics*, Phys. Rev. **108** (2) 171–190 (1957).
8. D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. van der Linde, *Bayesian measures of model complexity and fit*, J. R. Statist. Soc.B **64** Part 4, 583-639 (2002).
9. Supernova Cosmology Project, `http://supernova.lbl.gov`.
10. S. I. Vrieze, *Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)*, Psychol Methods. 2012 June; 17(2): 228-243. doi:10.1037/a0027127; `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3366160/`.
11. G. E. Hinton, S. Osindero,. and Y. Teh, *A fast learning algorithm for deep belief nets*, Neural Computation **18**, 1527-1554..
12. D. C. Cireşan, U. Meier, L. M. Gambardella, J. Schmidhuber, *Deep, big, simple neural nets for handwritten digit recognition.* Neural Comput. 2010 Dec; 22 (12): 3207-20.
13. MNIST, `http://yann.lecun.com/exdb/mnist/`.