Introduction to Bayesian Statistics - 8

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

1. Bayesian classification

data X, classes C

this likelihood is defined by training data

$$P(C|X) = \frac{P(X|C)}{P(X)} P(C)$$

the prior is also defined by training data

we can use the prior learning to assign a class to new data

$$C_k = \arg\max_{C_k} \frac{P(X|C_k)}{P(X)} P(C_k) = \arg\max_{C_k} P(X|C_k) P(C_k)$$

Consider a vector of *N* attributes given as Boolean variables **x** = {x_i} and classify the data vectors with a single Boolean variable.

The learning procedure must yield:

$$P(y)$$

it is easy to obtain it as an empirical distribution from an histogram of training class data: y is Boolean, the histogram has just two bins, and a hundred examples suffice to determine the empirical distribution to better than 10%.

$$P(\mathbf{x}|y)$$

there is a bigger problem here: the arguments have 2^{N+1} different values, and we must estimate $2(2^N-1)$ parameters ... for instance, with N = 30 there are more than 2 billion parameters!

How can we reduce the huge complexity of learning?

we assume the conditional independence of the *x_n’s*:
naive Bayesian learning

for instance, with just two attributes

$$P\left(x_1, x_2 \middle| y\right) = P\left(x_1 \middle| x_2, y\right) P\left(x_2 \middle| y\right) = P\left(x_1 \middle| y\right) P\left(x_2 \middle| y\right)$$

conditional independence assumption

with more than 2 attributes

$$P\left(\mathbf{x} \middle| y\right) \approx \prod_{k=1}^{N} P\left(x_k \middle| y\right)$$

Therefore:

$$P(y_k|\mathbf{x}) = \frac{P(\mathbf{x}|y_k)}{P(\mathbf{x})} P(y_k) = \frac{P(\mathbf{x}|y_k)}{\sum_j P(\mathbf{x}|y_j) P(y_j)} P(y_k)$$

$$\approx \frac{\prod_{n=1}^{N} P(x_n|y_k)}{\sum_j P(y_j) \prod_{n=1}^{N} P(x_n|y_j)} P(y_k)$$

and we assign the class according to the rule (MAP)

$$\boxed{y = \arg\max_{y_k} \frac{\prod_{n=1}^{N} P(x_n|y_k)}{\sum_j P(y_j) \prod_{n=1}^{N} P(x_n|y_j)} P(y_k)}$$

More general discrete inputs

If any of the *N x* variables has *J* different values, e if there are *K* classes, then we must estimate in all *NK*(*J*-1) free parameters with the Naive Bayes Classifier (this includes normalization) (compare this with the $K(J^N\text{-}1)$ parameters needed by a complete classifier)

Continuous inputs and discrete classes – the Gaussian case

$$P\left(x_n \middle| y_k\right) = \frac{1}{\sqrt{2\pi\sigma_{nk}^2}} \exp\left[-\frac{\left(x_n - \mu_{nk}\right)^2}{2\sigma_{nk}^2} \right]$$

here we must estimate 2*NK* parameters + the shape of the distribution *P*(*y*) (this adds up to another *K*-1 parameters)

Gaussian special case with class-independent variance and Boolean classification (two classes only):

$$P\left(y=0\middle|\mathbf{x}\right)=\frac{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)}{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)+P\left(\mathbf{x}\middle|y=1\right)P\left(y=1\right)}$$

$$P\left(x_n\middle|y=0\right)=\frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left[-\frac{\left(x_n-\mu_{n0}\right)^2}{2\sigma_n^2}\right]$$

$$P\left(x_n\middle|y=1\right)=\frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left[-\frac{\left(x_n-\mu_{n1}\right)^2}{2\sigma_n^2}\right]$$

$$
\begin{aligned}
P\left(y=0\middle|\mathbf{x}\right) &= \frac{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)}{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)+P\left(\mathbf{x}\middle|y=1\right)P\left(y=1\right)} \\
&= \frac{1}{1+\dfrac{P\left(\mathbf{x}\middle|y=1\right)P\left(y=1\right)}{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)}} \\
&= \frac{1}{1+\dfrac{P\left(y=1\right)}{P\left(y=0\right)}\displaystyle\prod_{n=1}^{N}\exp\left[-\frac{\left(x_n-\mu_{n1}\right)^2}{2\sigma_n^2}+\frac{\left(x_n-\mu_{n0}\right)^2}{2\sigma_n^2}\right]} \\
&= \frac{1}{1+\exp\left\{\ln\left(\dfrac{P\left(y=1\right)}{P\left(y=0\right)}\right)+\displaystyle\sum_{n=1}^{N}\left[\frac{\left(\mu_{n1}-\mu_{n0}\right)x_n}{\sigma_n^2}+\frac{\mu_{n0}^2-\mu_{n1}^2}{2\sigma_n^2}\right]\right\}}
\end{aligned}
$$

$$w_0 = \ln\left(\frac{P(y=1)}{P(y=0)}\right) + \sum_{n=1}^{N}\left[\frac{\mu_{n0}^2 - \mu_{n1}^2}{2\sigma_n^2}\right]$$

$$w_n = \frac{(\mu_{n1} - \mu_{n0})}{\sigma_n^2}$$

logistic shape

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$P(y=1|\mathbf{x}) = 1 - P(y=0|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

Finally an input vector belongs to class y = 0 if

$$\frac{P\left(y=0|\mathbf{x}\right)}{P\left(y=1|\mathbf{x}\right)} > 1$$

$$\Downarrow$$

$$P\left(y=0|\mathbf{x}\right) = \frac{1}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$P\left(y=1|\mathbf{x}\right) = \frac{\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$\Rightarrow \quad \exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right) < 1$$

$$\Rightarrow \quad w_0 + \sum_{n=1}^{N} w_n x_n < 0$$

2. Logistic regression (logit regression)

The odds ratio

$$\frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \frac{p}{1-p}$$

with the exponential expansion that we just found (the logistic expression for *p*) gives

$$\ln \frac{p}{1-p} = w_0 + \sum_{n=1}^{N} w_n x_n$$

For a given set of *K* class determinations where the fraction of assignments to class 1 is $p^{(k)}$ for a parameter vector of $\{x_n^{(k)}\}_{k=1,K}$ this log odds ratio becomes

$$\ln \frac{p^{(k)}}{1-p^{(k)}} = w_0 + \sum_{n=1}^{N} w_n x_n^{(k)}$$

The expression

$$\ln \frac{p^{(k)}}{1-p^{(k)}} = w_0 + \sum_{n=1}^{N} w_n x_n^{(k)}$$

is the basis for a generalized linear regression, to determine the *w* parameters.

This can be done with the least squares method, where one minizes

$$S = \sum_{k=1}^{K} \left[\ln \frac{p^{(k)}}{1-p^{(k)}} - \left(w_0 + \sum_{n=1}^{N} w_n x_n^{(k)} \right) \right]^2$$

This logit regression is often used in classification problems.

3. Model selection

The generic purpose of a model selection statistic is to set up a tension between the predictiveness of a model (for instance indicated by the number of free parameters) and its ability to fit observational data. Oversimplistic models offering a poor fit should of course be thrown out, but so should more complex models that offer poor predictive power.

There are two main types of model selection statistic that have been used in the literature so far. Information criteria look at the best-fitting parameter values and attach a penalty for the number of parameters; they are essentially a technical formulation of "chi-squared per degrees of freedom" arguments. By contrast, the Bayesian evidence applies the same type of likelihood analysis familiar from parameter estimation, but at the level of models rather than parameters. It depends on goodness of fit across the entire model parameter space.

(Liddle & al., 2006)

Akaike Information Criterion (AIC).

This was derived by Hirotugu Akaike in 1974, and takes the form

$$\mathrm{AIC} = -2\ln\mathcal{L}_{\max} + 2k$$

where k is the number of parameters in the model. The subscript “max” indicates that one should find the parameter values yielding the highest possible likelihood within the model. ***This second term acts as a kind of “Occam factor”; initially, as parameters are added, the fit to data improves rapidly until a reasonable fit is achieved, but further parameters then add little and the penalty term 2k takes over.*** *The generic shape of the AIC as a function of number of parameters is a rapid fall, a minimum, and then a rise. The preferred model sits at the minimum.*

The AIC was derived from information-theoretic considerations, specifically an approximate minimization of the Kullback–Leibler information entropy which measures the distance between two probability distributions.

(Liddle & al., 2006)

Bayesian Information Criterion (BIC).

This was derived by Gideon Schwarz in 1978, and strongly resembles the AIC. It is given by

$$\mathrm{BIC} = -2 \ln \mathcal{L}_{\max} + k \ln N$$

where N is the number of datapoints. Since a typical dataset will have lnN > 2, the BIC imposes a stricter penalty against extra parameters than the AIC.

It was derived as an approximation to the Bayesian evidence*, to be discussed next, but* ***the assumptions required are very restrictive and unlikely to hold in practice, rendering the approximation quite crude.***

(Liddle & al., 2006)

Bayesian evidence

Model selection aims to determine which theoretical models are most plausible given some data, without necessarily considering preferred values of model parameters.

Ideally, we would like to estimate posterior probabilities on the set of all competing models using Bayes' theorem:

$$P(M_i|D,I) = \frac{P(D|M_i,I)P(M_i|I)}{\sum_k P(D|M_k,I)P(M_k|I)}$$

and select the best model using the odds ratio

$$\mathcal{O}_{i,j} = \frac{P(M_i|D,I)}{P(M_i|D,I)} = \frac{P(D|M_i,I)P(M_i|I)}{P(D|M_j,I)P(M_j|I)}$$

or the Bayes factor, if we assume equal prior probabilities for the different models:

$$B_{i,j} = \frac{P(D|M_i,I)}{P(D|M_j,I)}$$

Thus we see that the Bayes factor is a ratio of evidences

$$B_{i,j} = \frac{P(D|M_i, I)}{P(D|M_j, I)}$$

As usual, each evidence is obtained by marginalizing the likelihood with respect to the (potentially different) parameters:

$$P(D|M_i, I) = \int_{\Theta_i} P(D|\boldsymbol{\theta}_i, M_i, I) p(\boldsymbol{\theta}_i|M_i, I) d\boldsymbol{\theta}_i$$

The evidence of a model is thus the average likelihood of the model in the prior.

Unlike the AIC and BIC, it does not focus on the best-fitting parameters of the model, but asks "of all the parameter values you thought were viable before the data came along, how well on average did they fit the data?". Literally, it is the likelihood of the model given the data.

The evidence rewards predictability of models, provided they give a good fit to the data, and hence gives an axiomatic realization of Occam's razor.

A model with little parameter freedom is likely to fit data over much of its parameter space, whereas a model that could match pretty much any data that might have cropped up will give a better fit to the actual data but only in a small region of its larger parameter space, pulling the average likelihood down.

(Liddle & al., 2006)

Which statistics?

Of these statistics, we would advocate using – wherever possible – the Bayesian evidence, which is a full implementation of Bayesian inference and can be directly interpreted in terms of model probabilities. It is computationally challenging to compute, being a highly peaked multidimensional integral, but recent algorithm development has made it feasible in cosmological contexts.

If the Bayesian evidence cannot be computed, the BIC can be deployed as a substitute. *It is much simpler to compute as one need only find the point of maximum likelihood for each model.* ***However, interpreting it can be difficult. Its main usefulness is as an approximation to the evidence, but this holds only for gaussian likelihoods and provided the datapoints are independent and identically distributed.*** *The latter condition holds poorly for the current global cosmological dataset, though it can potentially be improved by binning of the data, hence decreasing the N in the penalty term.*

The AIC has been widely used outside astrophysics, but is of debatable utility. ***It has been shown to be "dimensionally inconsistent", meaning that it is not guaranteed to give the right result even in the limit of infinite unbiased data.*** *It may be useful for checking the robustness of conclusions drawn using the BIC.* ***The evidence and BIC are dimensionally consistent.***

(Liddle & al., 2006)

4. The EM algorithm (Dempster, Laird & Rubin, 1977)

Recall the max. likelihood principle:

uniform distribution
(usually an improper prior)

$$P(\boldsymbol{\theta} \mid \mathbf{d}, I) = \frac{P(\mathbf{d} \mid \boldsymbol{\theta}, I)}{P(\mathbf{d} \mid I)} \cdot P(\boldsymbol{\theta} \mid I)$$

likelihood

$$= \frac{\mathcal{L}(\mathbf{d}, \boldsymbol{\theta})}{P(\mathbf{d} \mid I)} \cdot P(\boldsymbol{\theta} \mid I) \propto \mathcal{L}(\mathbf{d}, \boldsymbol{\theta})$$

evidence

in this (approximate) setting, the MAP estimate coincides with the ML estimate.

when data are independent and identically distributed (i.i.d.) we find the following likelihood function

$$\mathcal{L}(\mathbf{d},\boldsymbol{\theta}) = \prod_i p\left(d_i \middle| \boldsymbol{\theta}\right)$$

and we estimate the parameters by maximizing the likelihood function

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{d},\boldsymbol{\theta})$$

or, equivalently, its logarithm

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left[\log \mathcal{L}(\mathbf{d},\boldsymbol{\theta})\right]$$

(in real life, this procedure is often complex and almost invariably it requires a numerical solution)

The EM algorithm is used to maximize likelihood with incomplete information, and it has two main steps that are iterated until convergence:

E. expectation of the log-likelihood, averaged with respect to missing data:

parameters (with respect to which we want to maximize the expression

measured data

missing data

previous parameter estimate (constant values)

likelihood

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(n-1)}\right)=E_{\mathbf{y}}\left[\log p\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\theta}\right)\middle|\mathbf{x},\boldsymbol{\theta}^{(n-1)}\right]$$

$$=\int_{Y}\left[\log p\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\theta}\right)\right]p\left(\mathbf{y}\middle|\mathbf{x},\boldsymbol{\theta}^{(n-1)}\right)d\mathbf{y}$$

M. maximization of the averaged log-likelihood with respect to parameters:

$$\boldsymbol{\theta}^{(n)}=\arg\max_{\boldsymbol{\theta}}Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(n-1)}\right)$$

Example: an experiment with an exponential model (Flury and Zoppè)

Light bulbs fail following an exponential distribution with mean failure time θ

To estimate the mean two experiments are performed

1. *n* light bulbs are tested, all failure times u_i are recorded
2. *m* light bulbs are tested, only the total number *r* of bulbs failed at time *t* are recorded

1. $$\mathcal{L} = \prod_{i=1}^{n} \frac{1}{\theta} \exp\left(-\frac{u_i}{\theta} \right) = \frac{1}{\theta^n} \exp\left(-\frac{\sum_i u_i}{\theta} \right) = \frac{1}{\theta^n} \exp\left(-\frac{n \langle u \rangle}{\theta} \right)$$

2. $$\mathcal{L} = \prod_{i=1}^{m} \frac{1}{\theta} \exp\left(-\frac{v_i}{\theta} \right)$$

missing data!

combined likelihood

$$\frac{1}{\theta^n}\exp\left(-\frac{n\langle u\rangle}{\theta}\right)\cdot\prod_{i=1}^{m}\frac{1}{\theta}\exp\left(-\frac{v_i}{\theta}\right)$$

log-likelihood

$$-n\ln\theta-\frac{n\langle u\rangle}{\theta}-\sum_{i=1}^{m}\left(\ln\theta+\frac{v_i}{\theta}\right)$$

expected failure time for a bulb that is still burning at time t

$$t+\theta$$

expected failure time for a bulb that is not burning at time t

$$\theta-\frac{t\exp\left(-t/\theta\right)}{1-\exp\left(-t/\theta\right)}$$

Note on mean failure time for a bulb that is not burning at time *t*

$$p(t') \propto \frac{1}{\theta} e^{-t'/\theta} \qquad 0 \le t' \le t$$

$$\text{normalization} = \int_0^t p(t')\,dt' = \int_0^t \frac{dt'}{\theta} e^{-t'/\theta} = 1 - e^{-t/\theta}$$

$$\text{mean failure time} = \int_0^t t' p(t')\,dt' = \frac{1}{1-e^{-t/\theta}} \int_0^t t' e^{-t'/\theta} \frac{dt'}{\theta}$$

$$= \frac{\theta}{1-e^{-t/\theta}} \left[1 - e^{-t/\theta} - (t/\theta) e^{-t/\theta} \right]$$

$$= \theta - \frac{t e^{-t/\theta}}{1-e^{-t/\theta}}$$

average log-likelihood

$$Q = E\left[-n\ln\theta - \frac{n\langle u\rangle}{\theta} + \sum_{i=1}^{m}\left(-\ln\theta - \frac{v_i}{\theta}\right)\right]$$

$$= -(n+m)\ln\theta - \frac{n\langle u\rangle}{\theta} - \frac{r}{\theta}\left(\theta - \frac{t\exp(-t/\theta)}{1-\exp(-t/\theta)}\right) - \frac{(m-r)}{\theta}(\theta+t)$$

this ends the expectation step

the max of the mean likelihood

$$Q = -(n+m)\ln\theta - \frac{1}{\theta}\left[n\langle u\rangle + r\left(\theta - \frac{t\exp(-t/\theta)}{1-\exp(-t/\theta)}\right) + (m-r)(\theta+t)\right]$$

can be found by maximizing the approximate expression

$$Q \approx -(n+m)\ln\theta - \frac{1}{\theta}\left[n\langle u\rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)}\right) + (m-r)\left(\theta^{(k)}+t\right)\right]$$

$$\Downarrow$$

$$\frac{dQ}{d\theta} \approx -(n+m)\frac{1}{\theta} + \frac{1}{\theta^2}\left[n\langle u\rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)}\right) + (m-r)\left(\theta^{(k)}+t\right)\right] = 0$$

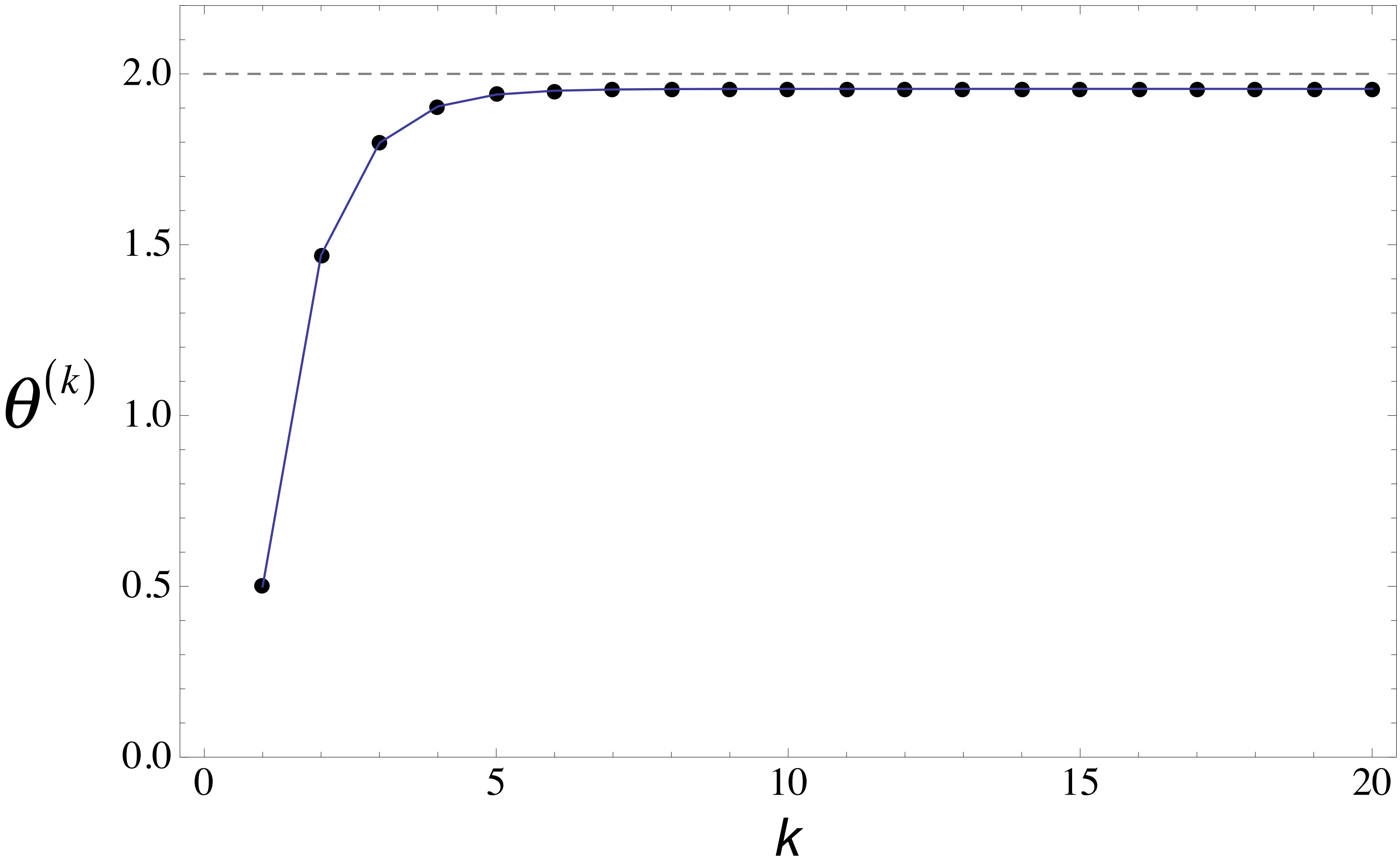
$$\frac{dQ}{d\theta} \approx -\left(n+m\right)\frac{1}{\theta} + \frac{1}{\theta^{2}}\left[n\langle u\rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)} \right) + \left(m-r\right)\left(\theta^{(k)}+t\right) \right] = 0$$

⇩

$$\theta^{(k+1)} = \frac{1}{n+m}\left[n\langle u\rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)} \right) + \left(m-r\right)\left(\theta^{(k)}+t\right) \right]$$

this formula summarizes expectation and maximization: therefore, the recipe is to iterate this until convergence ...

Example with mean failure time = 2 (a.u.), and randomly generated data (n = 100; m = 100). In this example r = 36.

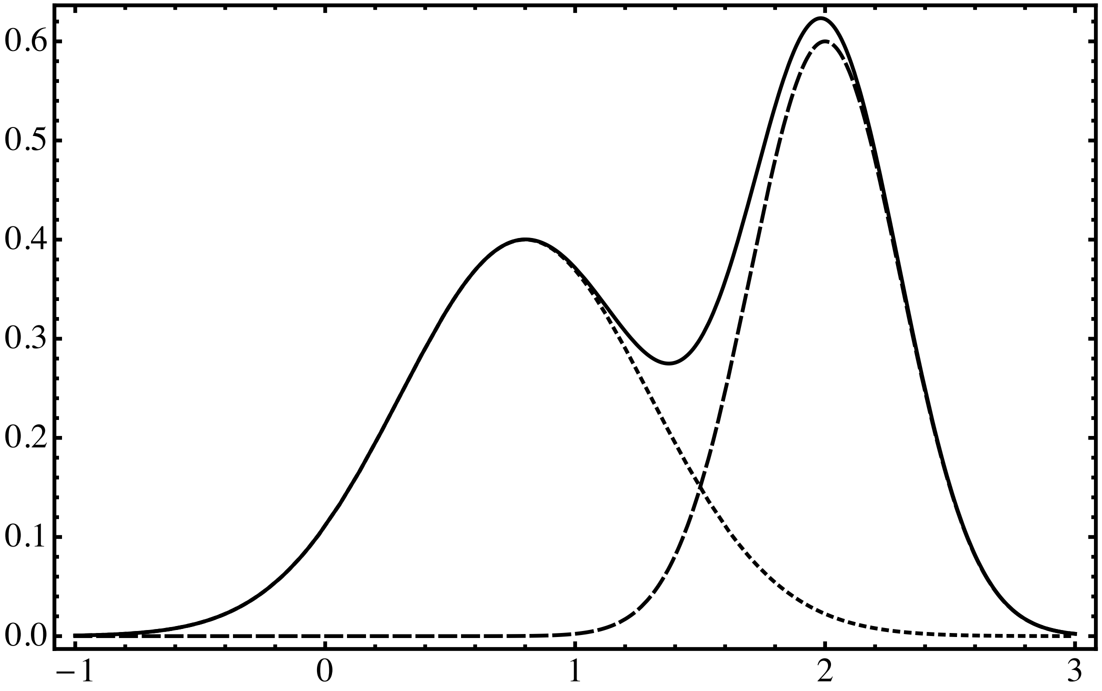


The EM method is often used to estimate the parameters of "mixture models".

$$p\left(x_n \middle| \boldsymbol{\theta}\right) = \sum_{i=1}^{M} \alpha_i p_i\left(x_n \middle| \boldsymbol{\theta}_i\right)$$

$$\boldsymbol{\theta} = \left(\alpha_1, \ldots, \alpha_M; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\right)$$

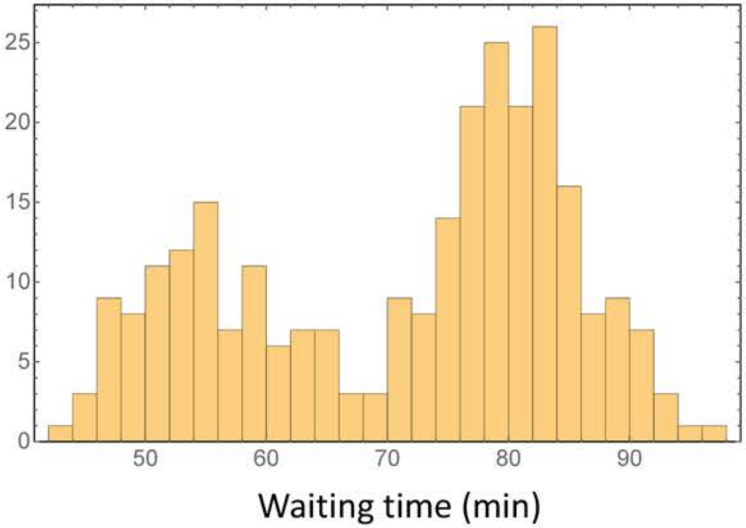
$$\sum_{i=1}^{M} \alpha_i = 1$$



Example: a Gaussian mixture model (M=2)

Easy-to-understand example: waiting times between eruptions of the Old Faithful Geiser (Yellowstone National Park – Wyoming)

Here we analyze the waiting times assuming a 2-Gaussian mixture model for the waiting time distribution
(data taken from an R example)



In this case, the mixture model has two Gaussian components

$$p(w|\boldsymbol{\theta}) = \alpha N(w;\mu_1,\sigma_1) + (1-\alpha)N(w;\mu_2,\sigma_2)$$

where the vector of parameters is $\boldsymbol{\theta} = (\alpha,\mu_1,\mu_2,\sigma_1,\sigma_2)$

The resulting log likelihood with *n* waiting times w_i is

$$\ln\mathcal{L} = \sum_i \ln\left[\alpha N(w_i;\mu_1,\sigma_1) + (1-\alpha)N(w_i;\mu_2,\sigma_2)\right]$$

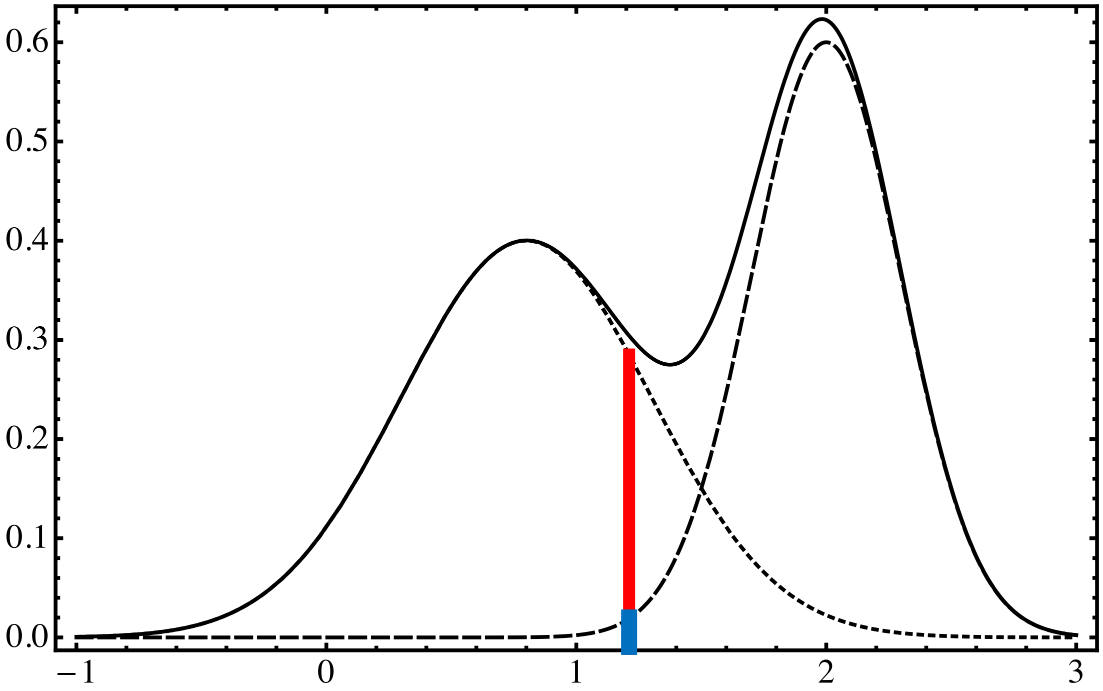
We substitute the likelihood with the new one, containing unobserved data:

$$\mathcal{L} = \prod_i \alpha^{y_i} (1-\alpha)^{1-y_i} [N(w_i;\mu_1,\sigma_1)]^{y_i} [N(w_i;\mu_2,\sigma_2)]^{1-y_i}$$

where the new, unobserved data y_i are indicator variables that select extraction from the first (y_i = 1) or the second (y_i = 0) Gaussian (i.e., *two classes*).

Then

$$\ln \mathcal{L} = \sum_i \left[y_i \ln \alpha + (1-y_i)\ln(1-\alpha) + y_i \left(-\frac{1}{2}\ln(2\pi\sigma_1) - \frac{(w_i-\mu_1)^2}{2\sigma_1^2} \right) \right.$$
$$\left. +(1-y_i)\left(-\frac{1}{2}\ln(2\pi\sigma_2) - \frac{(w_i-\mu_2)^2}{2\sigma_2^2} \right) \right]$$



The probability that a given time interval belongs to the first Gaussian is

this probability is also equal to the mean value of the indicator variable

$$p_i = \frac{\alpha \times N(w_i; \mu_1, \sigma_1)}{\alpha \times N(w_i; \mu_1, \sigma_1) + (1-\alpha) \times N(w_i; \mu_2, \sigma_2)}$$

$$= \frac{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2/2(\sigma_1^{(k)})^2]/\sqrt{2\pi(\sigma_1^{(k)})^2}}{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2/2(\sigma_1^{(k)})^2]/\sqrt{2\pi(\sigma_1^{(k)})^2} + (1-\alpha^{(k)}) \exp[-(w_i - \mu_2^{(k)})^2/2(\sigma_2^{(k)})^2]/\sqrt{2\pi(\sigma_2^{(k)})^2}}$$

Now, **averaging the log likelihood with respect to the unobserved data** we find

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \sum_i \left[p_i^{(k)} \ln \alpha + (1 - p_i^{(k)}) \ln(1-\alpha) + p_i^{(k)} \left(-\frac{1}{2} \ln(2\pi\sigma_1^2) - \frac{(w_i - \mu_1)^2}{2\sigma_1^2} \right) \right.$$
$$\left. + (1 - p_i^{(k)}) \left(-\frac{1}{2} \ln(2\pi\sigma_2^2) - \frac{(w_i - \mu_2)^2}{2\sigma_2^2} \right) \right]$$

(the mean value of the indicator variable is equal to the current estimate of the probability α)

Next we maximize with respect to all the remaining parameters, and we find:

$$\alpha^{(k+1)} = \frac{1}{N} \sum_i p_i^{(k)}$$

$$\left(\sigma_1^{(k+1)}\right)^2 = \frac{\sum_i p_i^{(k)} (w_i - \mu_1^{(k)})^2}{\sum_i p_i^{(k)}}; \qquad \mu_1^{(k+1)} = \frac{\sum_i p_i^{(k)} w_i}{\sum_i p_i^{(k)}}$$

$$\left(\sigma_2^{(k+1)}\right)^2 = \frac{\sum_i (1 - p_i^{(k)}) (w_i - \mu_2^{(k)})^2}{\sum_i (1 - p_i^{(k)})}; \qquad \mu_2^{(k+1)} = \frac{\sum_i (1 - p_i^{(k)}) w_i}{\sum_i (1 - p_i^{(k)})}$$

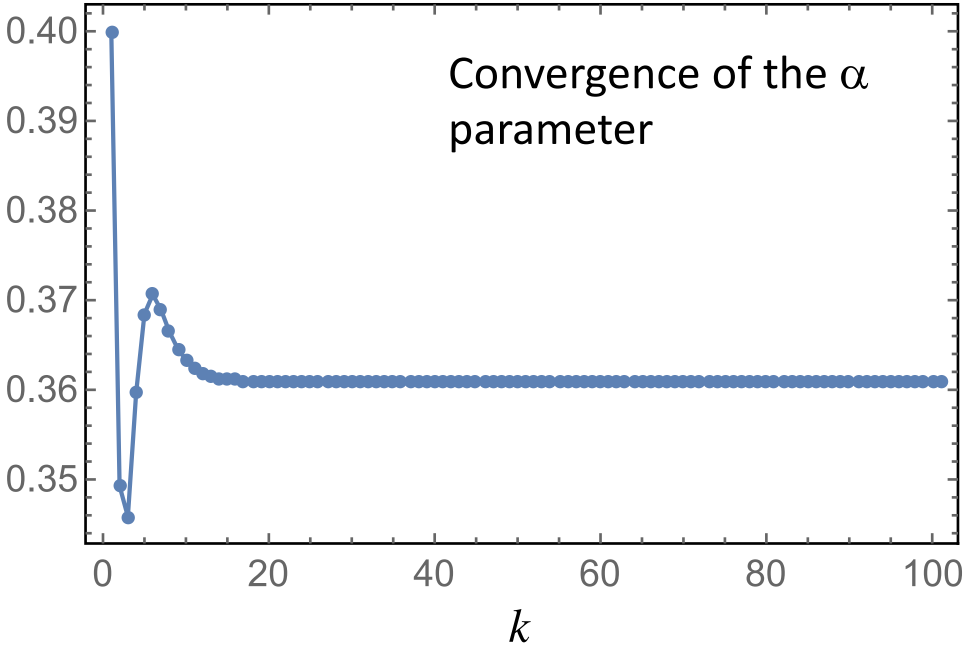
Finally we have the following set of equations:

$$p_i^{(k)} = \frac{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2/2(\sigma_1^{(k)})^2]/\sqrt{2\pi(\sigma_1^{(k)})^2}}{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2/2(\sigma_1^{(k)})^2]/\sqrt{2\pi(\sigma_1^{(k)})^2} + (1-\alpha^{(k)}) \exp[-(w_i - \mu_2^{(k)})^2/2(\sigma_2^{(k)})^2]/\sqrt{2\pi(\sigma_2^{(k)})^2}}$$

$$\alpha^{(k+1)} = \frac{1}{N}\sum_i p_i^{(k)}$$

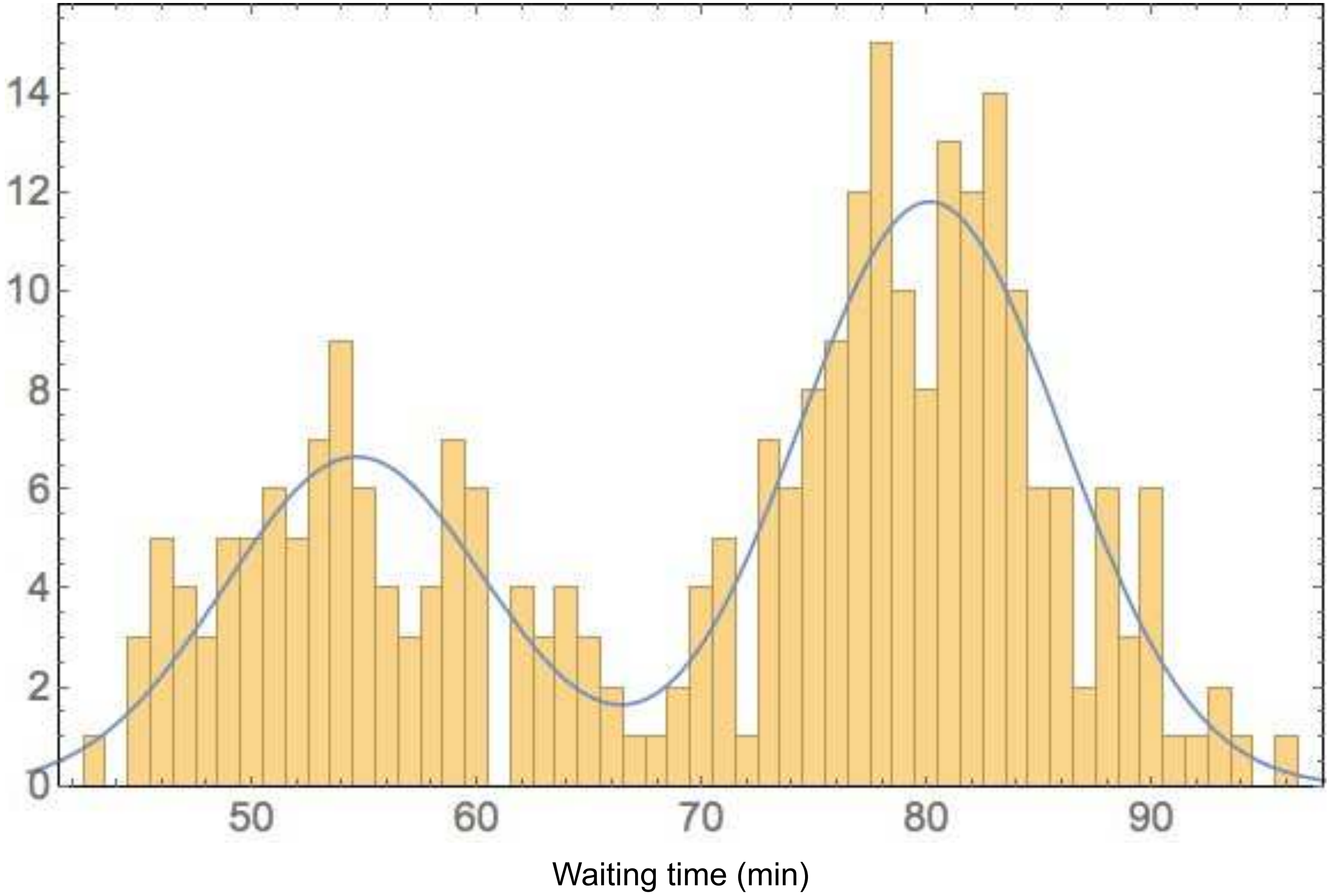
$$\left(\sigma_1^{(k+1)}\right)^2 = \frac{\sum_i p_i^{(k)}(w_i - \mu_1^{(k)})^2}{\sum_i p_i^{(k)}}; \qquad \mu_1^{(k+1)} = \frac{\sum_i p_i^{(k)} w_i}{\sum_i p_i^{(k)}}$$

$$\left(\sigma_2^{(k+1)}\right)^2 = \frac{\sum_i (1 - p_i^{(k)})(w_i - \mu_2^{(k)})^2}{\sum_i (1 - p_i^{(k)})}; \qquad \mu_2^{(k+1)} = \frac{\sum_i (1 - p_i^{(k)}) w_i}{\sum_i (1 - p_i^{(k)})}$$



Convergence of the α parameter

Comparison of the original data with the mixture model obtained with the EM algorithm



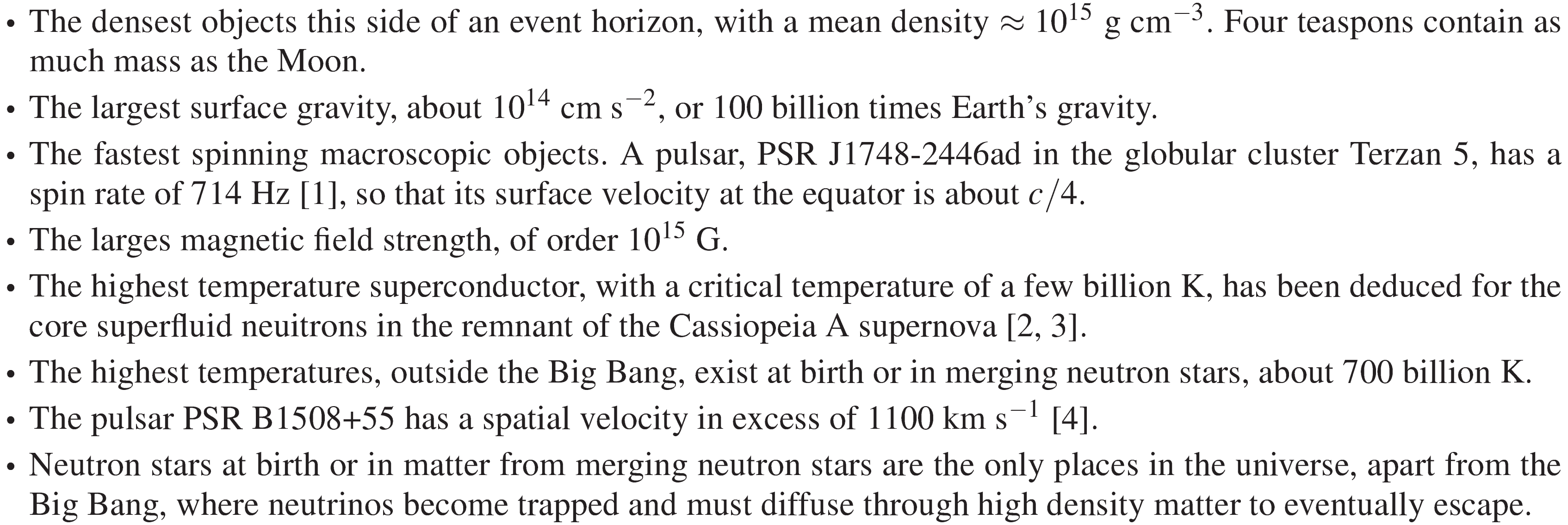
Notes:

1. the approach outlined in this 2-component Gaussian mixture model can be generalized to more than two components
2. the assignment to different components (hidden index variable) amounts to the solution of a classification problem

5. Neutron star mass range

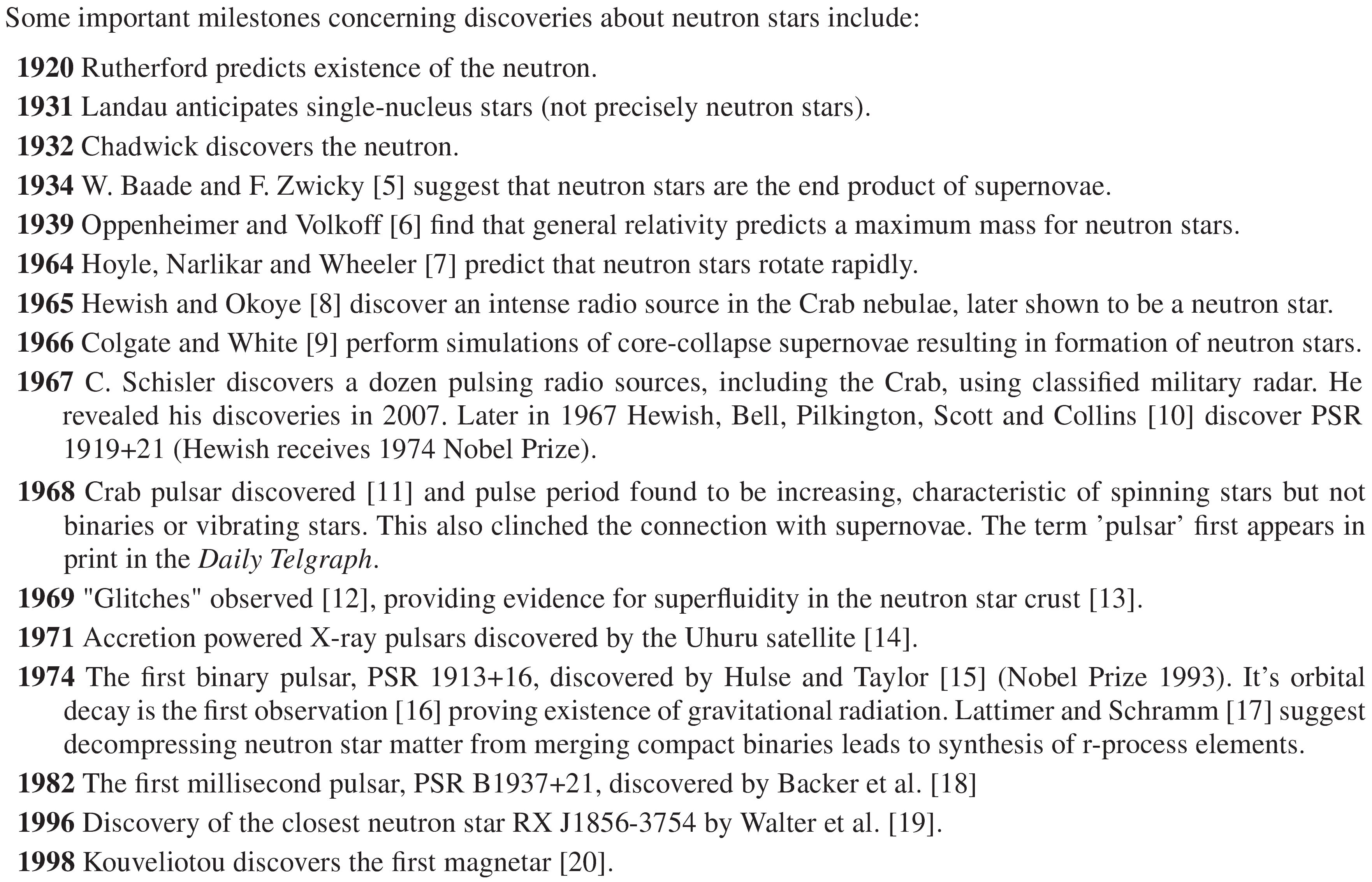
(Finn, PRL 73 (1994) 1878)

Neutron stars:



- The densest objects this side of an event horizon, with a mean density $\approx 10^{15}$ g cm^{-3}. Four teaspons contain as much mass as the Moon.
- The largest surface gravity, about 10^{14} cm s^{-2}, or 100 billion times Earth's gravity.
- The fastest spinning macroscopic objects. A pulsar, PSR J1748-2446ad in the globular cluster Terzan 5, has a spin rate of 714 Hz [1], so that its surface velocity at the equator is about $c/4$.
- The larges magnetic field strength, of order 10^{15} G.
- The highest temperature superconductor, with a critical temperature of a few billion K, has been deduced for the core superfluid neuitrons in the remnant of the Cassiopeia A supernova [2, 3].
- The highest temperatures, outside the Big Bang, exist at birth or in merging neutron stars, about 700 billion K.
- The pulsar PSR B1508+55 has a spatial velocity in excess of 1100 km s^{-1} [4].
- Neutron stars at birth or in matter from merging neutron stars are the only places in the universe, apart from the Big Bang, where neutrinos become trapped and must diffuse through high density matter to eventually escape.

from J. Lattimer: "Introduction to neutron stars", AIP Conference Proceedings 1645, 61 (2015); https://doi.org/10.1063/1.4909560



Some important milestones concerning discoveries about neutron stars include:

1920 Rutherford predicts existence of the neutron.

1931 Landau anticipates single-nucleus stars (not precisely neutron stars).

1932 Chadwick discovers the neutron.

1934 W. Baade and F. Zwicky [5] suggest that neutron stars are the end product of supernovae.

1939 Oppenheimer and Volkoff [6] find that general relativity predicts a maximum mass for neutron stars.

1964 Hoyle, Narlikar and Wheeler [7] predict that neutron stars rotate rapidly.

1965 Hewish and Okoye [8] discover an intense radio source in the Crab nebulae, later shown to be a neutron star.

1966 Colgate and White [9] perform simulations of core-collapse supernovae resulting in formation of neutron stars.

1967 C. Schisler discovers a dozen pulsing radio sources, including the Crab, using classified military radar. He revealed his discoveries in 2007. Later in 1967 Hewish, Bell, Pilkington, Scott and Collins [10] discover PSR 1919+21 (Hewish receives 1974 Nobel Prize).

1968 Crab pulsar discovered [11] and pulse period found to be increasing, characteristic of spinning stars but not binaries or vibrating stars. This also clinched the connection with supernovae. The term ’pulsar’ first appears in print in the *Daily Telgraph*.

1969 "Glitches" observed [12], providing evidence for superfluidity in the neutron star crust [13].

1971 Accretion powered X-ray pulsars discovered by the Uhuru satellite [14].

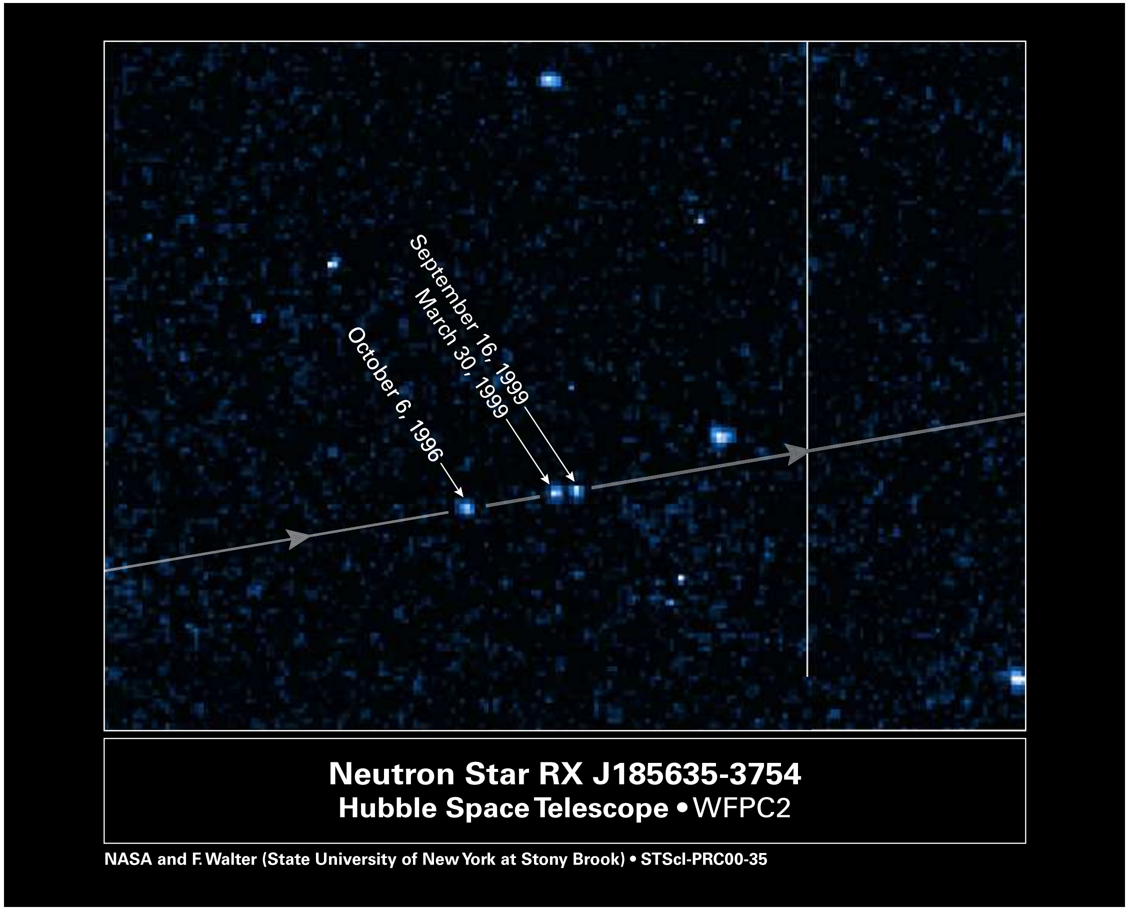
1974 The first binary pulsar, PSR 1913+16, discovered by Hulse and Taylor [15] (Nobel Prize 1993). It’s orbital decay is the first observation [16] proving existence of gravitational radiation. Lattimer and Schramm [17] suggest decompressing neutron star matter from merging compact binaries leads to synthesis of r-process elements.

1982 The first millisecond pulsar, PSR B1937+21, discovered by Backer et al. [18]

1996 Discovery of the closest neutron star RX J1856-3754 by Walter et al. [19].

1998 Kouveliotou discovers the first magnetar [20].

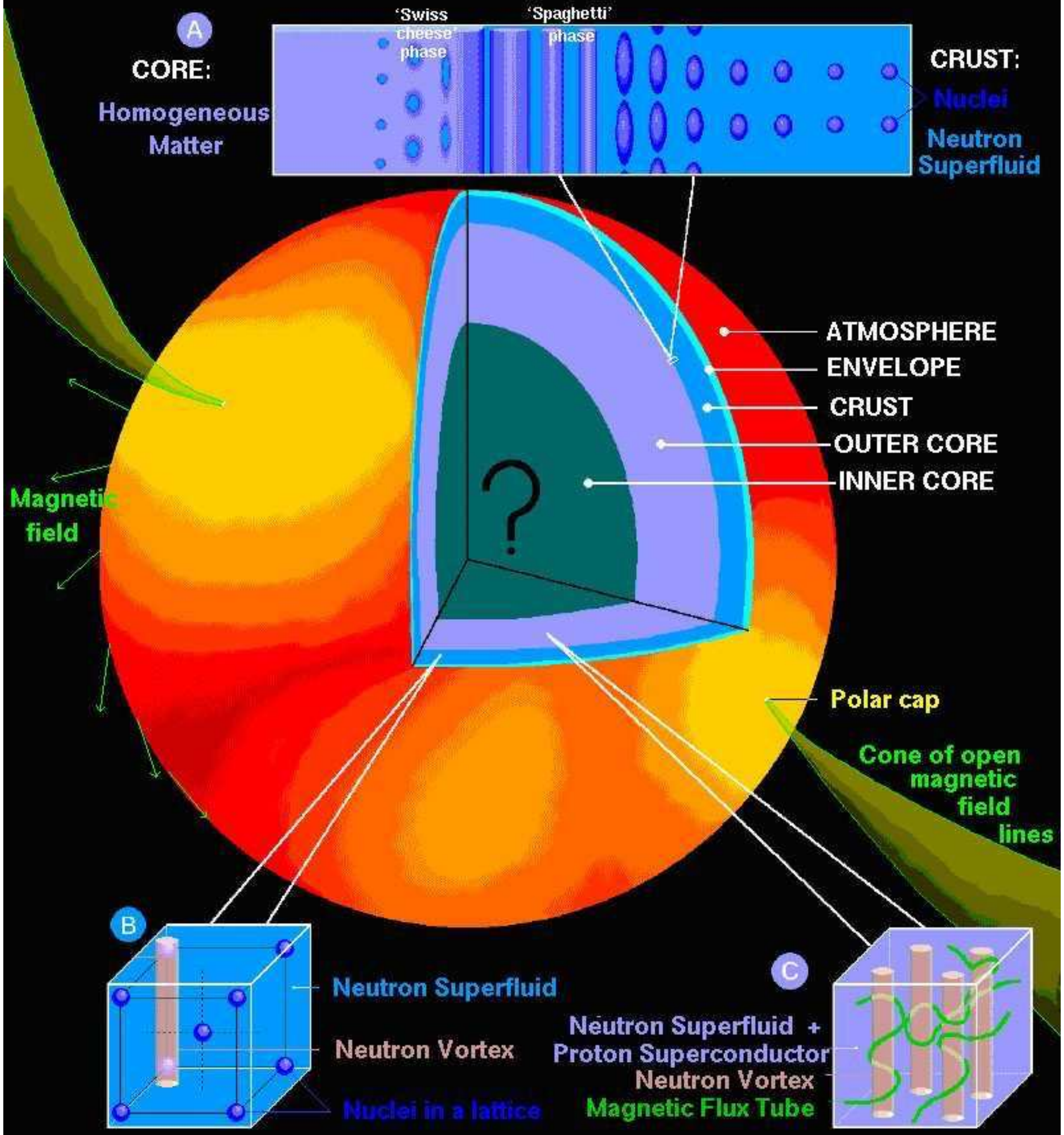
from J. Lattimer: "Introduction to neutron stars", AIP Conference Proceedings 1645, 61 (2015); https://doi.org/10.1063/1.4909560


September 16, 1999
March 30, 1999
October 6, 1996
Neutron Star RX J185635-3754
Hubble Space Telescope • WFPC2
NASA and F. Walter (State University of New York at Stony Brook) • STScI-PRC00-35

The Motion of RX J185635-3754 - The Nearest Neutron Star to Earth

This photograph is the sum of three Hubble Space Telescope images. North is down, east is to the right. The image, taken by the Wide Field and Planetary Camera 2, is 8.8 arc seconds across (west to east), and 6.6 arc seconds top-to-bottom (south to north).
All stars line up in this composite picture, except the neutron star, which moves across the image in a direction 10 degrees south of east. The three images of the neutron star are labeled by date. The proper motion is 1/3 of an arc second per year. The small wobble caused by parallax (not visible in the image) has a size of 0.016 arc seconds, giving a distance of 200 light-years.

A NEUTRON STAR: SURFACE and INTERIOR



Simple exercise: derive the Newtonian version of the Tolman-Oppenheimer-Volkov equation for pressure and mass of a neutron star

$$\frac{dp}{dr} = -\frac{G\rho(r)\mathcal{M}(r)}{r^2} = -\frac{G\epsilon(r)\mathcal{M}(r)}{c^2 r^2}$$

$$\mathcal{M}(r) = 4\pi \int_0^r \rho(r')r'^2 dr' = \frac{4\pi}{c^2}\int_0^r \epsilon(r')r'^2 dr'$$

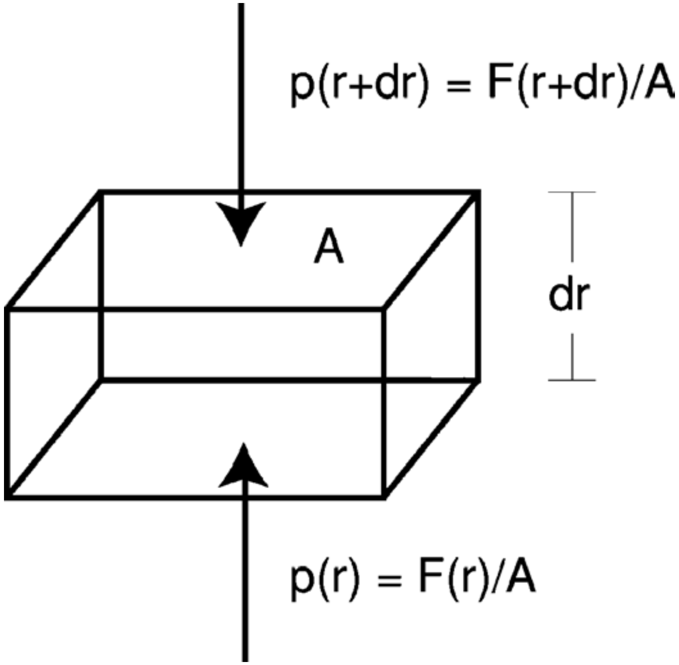
energy density

$$Ap(r) = Ap(r+dr) + \frac{G[\rho(r)Adr]\mathcal{M}(r)}{r^2}$$

$$\Downarrow$$

$$\frac{dp}{dr} = -\frac{G\rho(r)\mathcal{M}(r)}{r^2} = -\frac{G\epsilon(r)\mathcal{M}(r)}{c^2 r^2}$$

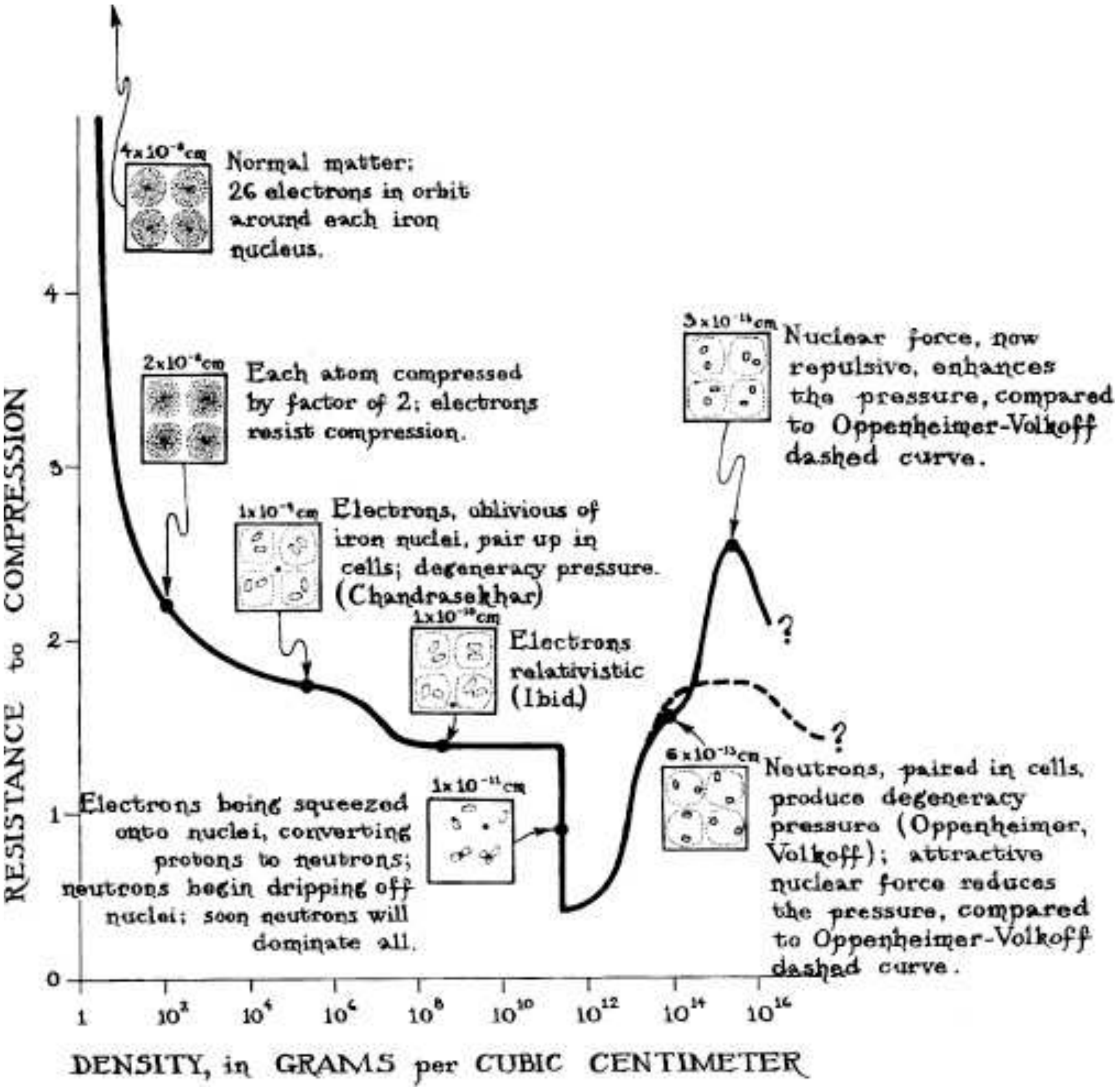
$$\mathcal{M}(r) = 4\pi \int_0^r \rho(r')r'^2 dr' = \frac{4\pi}{c^2}\int_0^r \epsilon(r')r'^2 dr'$$



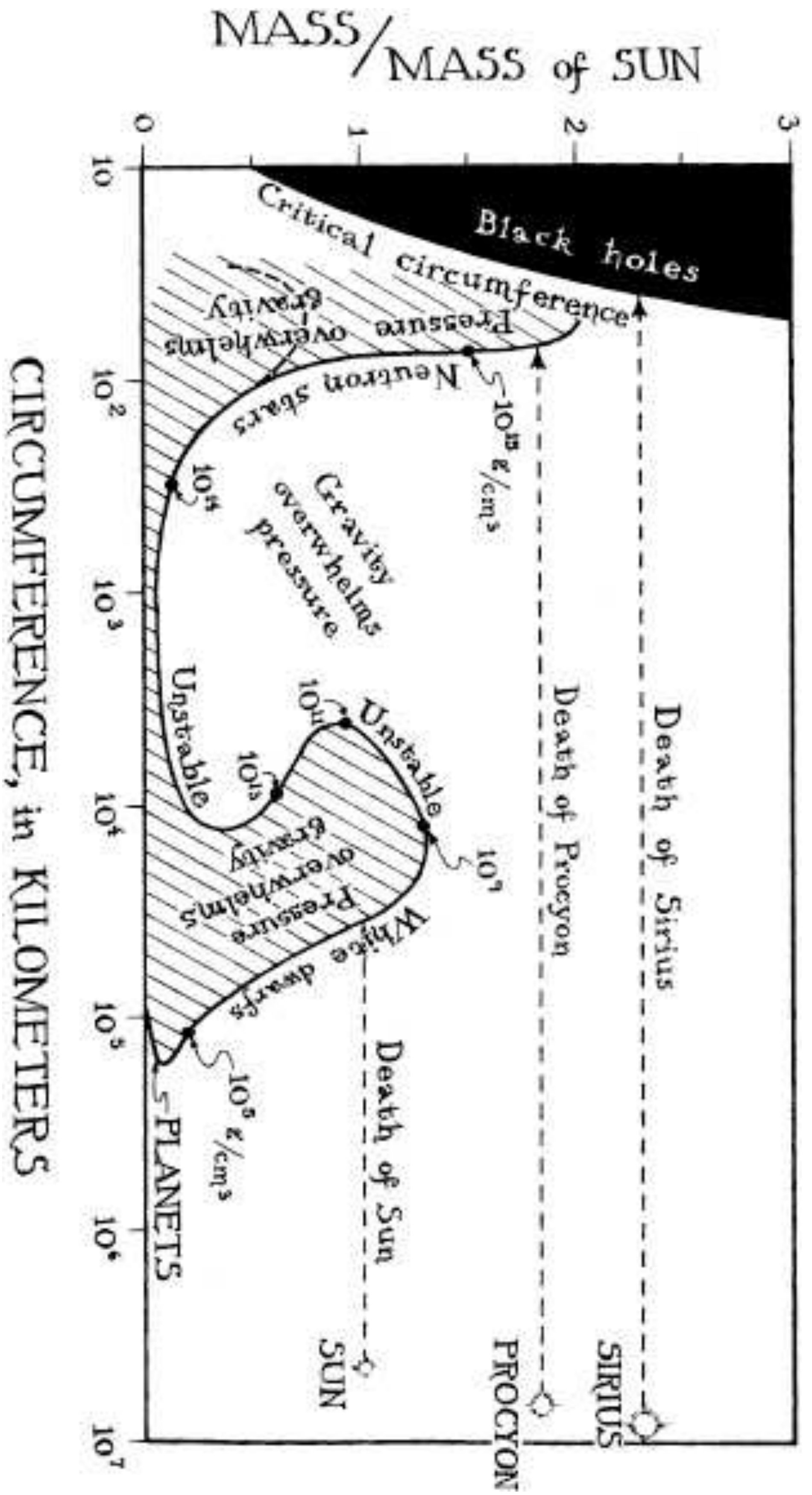
The complete, relativistic equation, contains corrections that involve the mass-energy density (the energy density and pressure are connected by the Equation Of State, EOS).

$$\frac{dp}{dr} = -\frac{G}{c^2}\frac{(m+4\pi r^3 p/c^2)(\varepsilon + p)}{r(r-2GM/c^2)}, \qquad \frac{dm}{dr} = 4\pi\frac{\varepsilon}{c^2}r^2$$

The Harrison–Wheeler Equation of State for Cold, Dead Matter

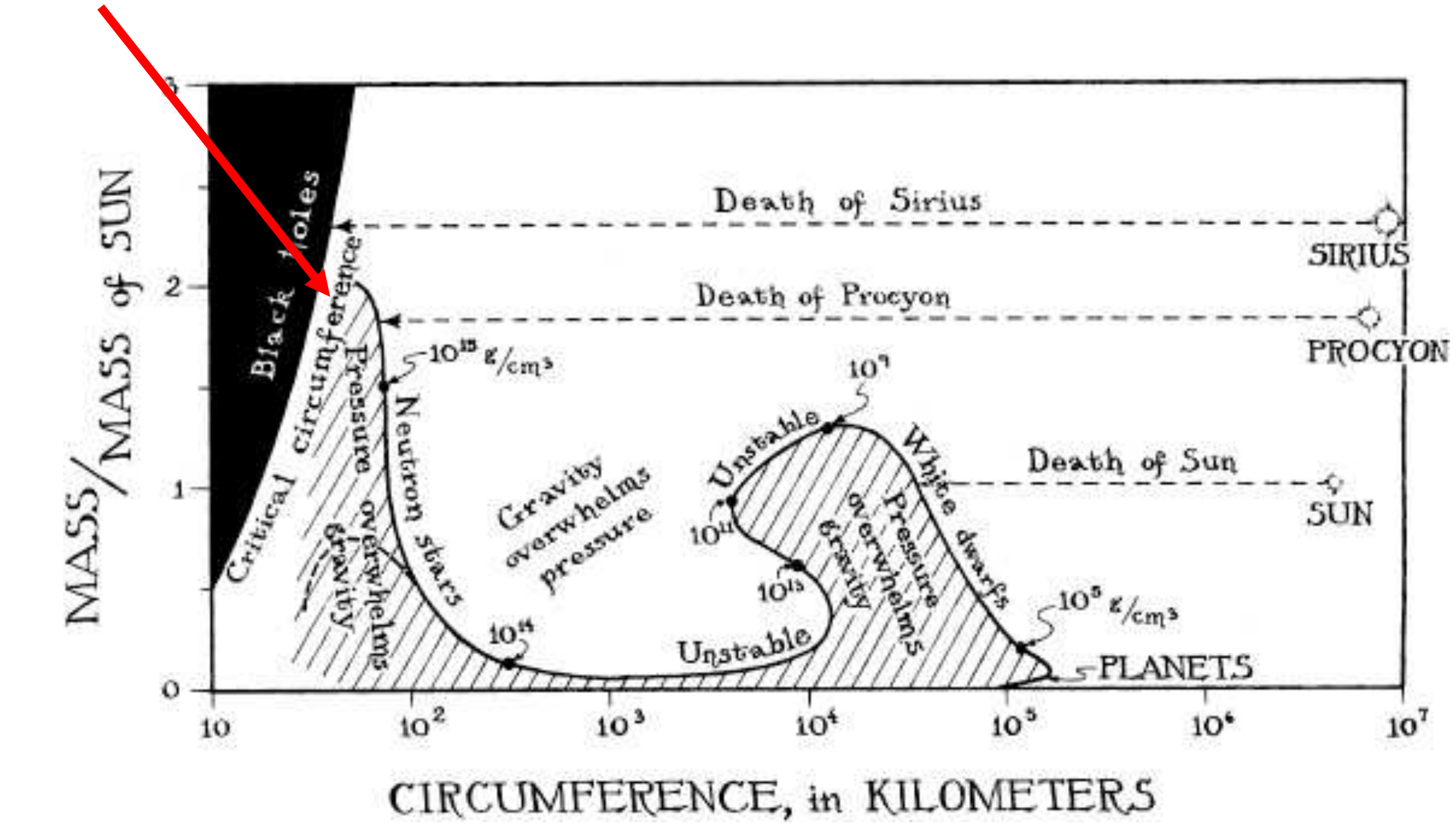


from K. S. Thorne: "Black Holes and Time Warps", Norton (1994)



from K. S. Thorne: "Black Holes and Time Warps", Norton (1994)

quark stars?
exotic stars?



from K. S. Thorne: "Black Holes and Time Warps", Norton (1994)

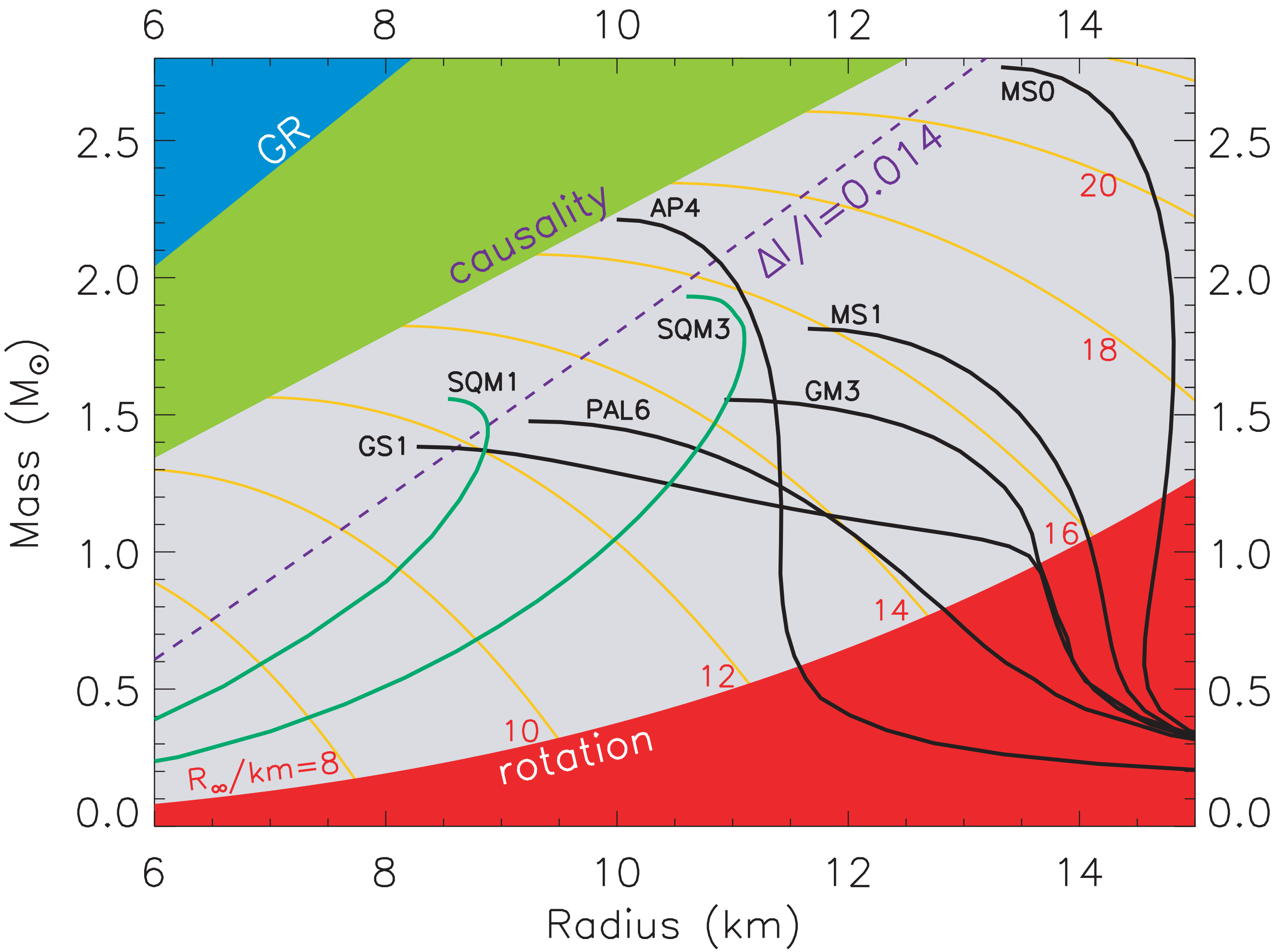


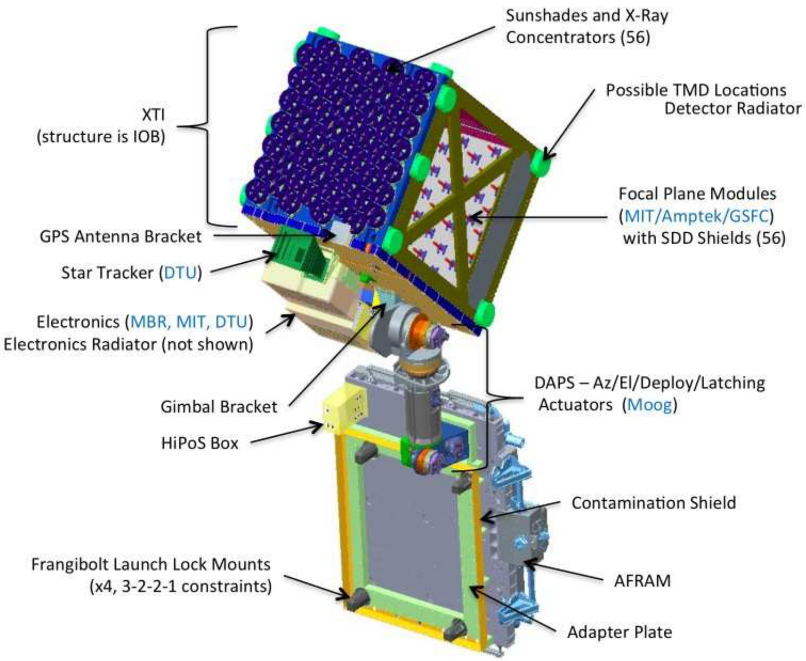
Fig. 2. Mass-radius diagram for neutron stars. Black (green) curves are for normal matter (SQM) equations of state [for definitions of the labels, see (*27*)]. Regions excluded by general relativity (GR), causality, and rotation constraints are indicated. Contours of radiation radii R_∞ are given by the orange curves. The dashed line labeled $\Delta I/I = 0.014$ is a radius limit estimated from Vela pulsar glitches (*27*).

from J. M. Lattimer and M. Prakash: "The Physics of Neutron Stars",
Science 304 (2004) 536

The mass-radius relation is important to pin down the EOS, and hence the composition of the neutron star nucleus.

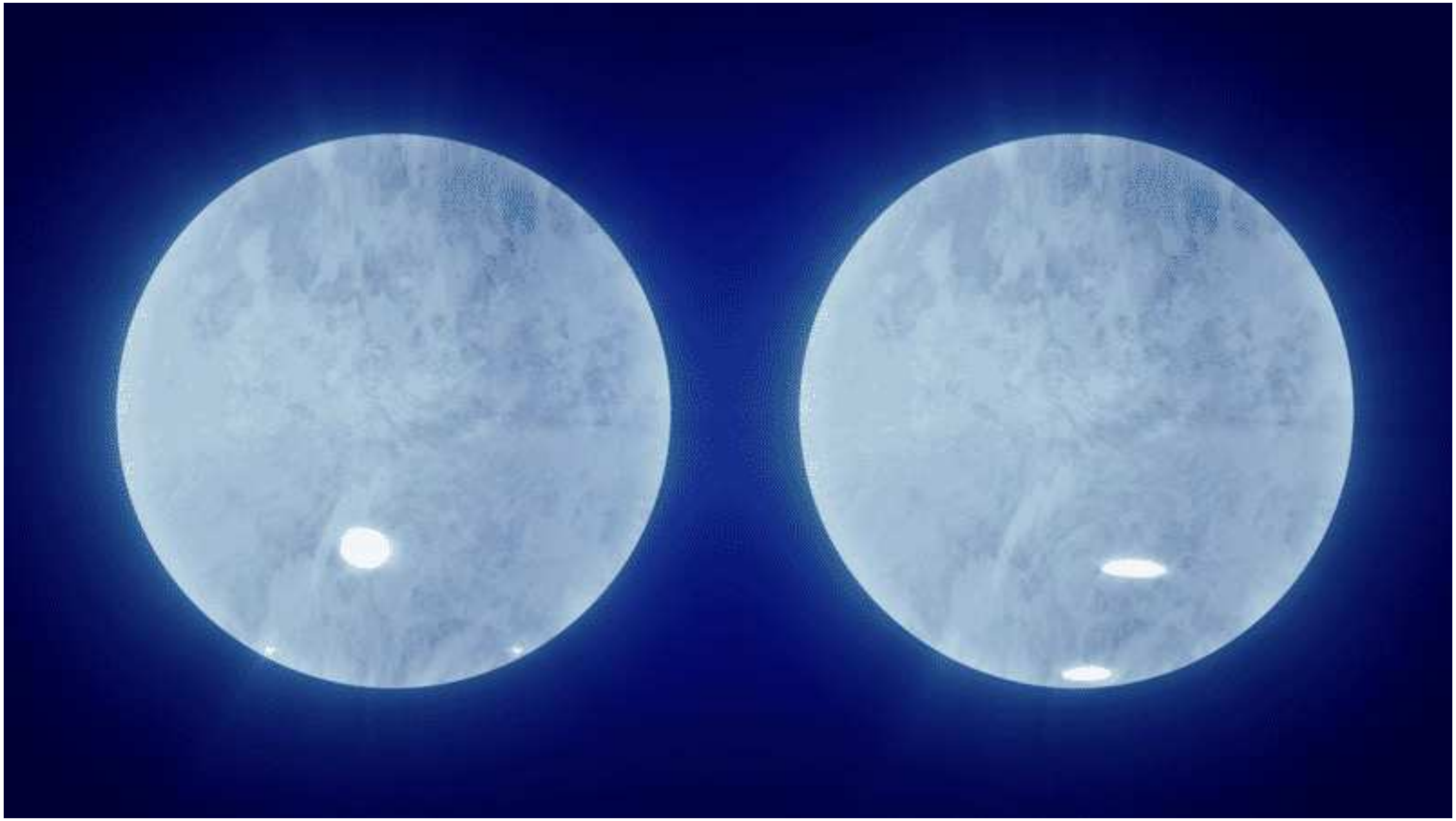
The NICER (Neutron star Interior Composition ExploreR) mission is steadily providing new informations on neutron stars (resolution obtained on radius ~ 0.5 km)

The NICER instrument onboard the ISS

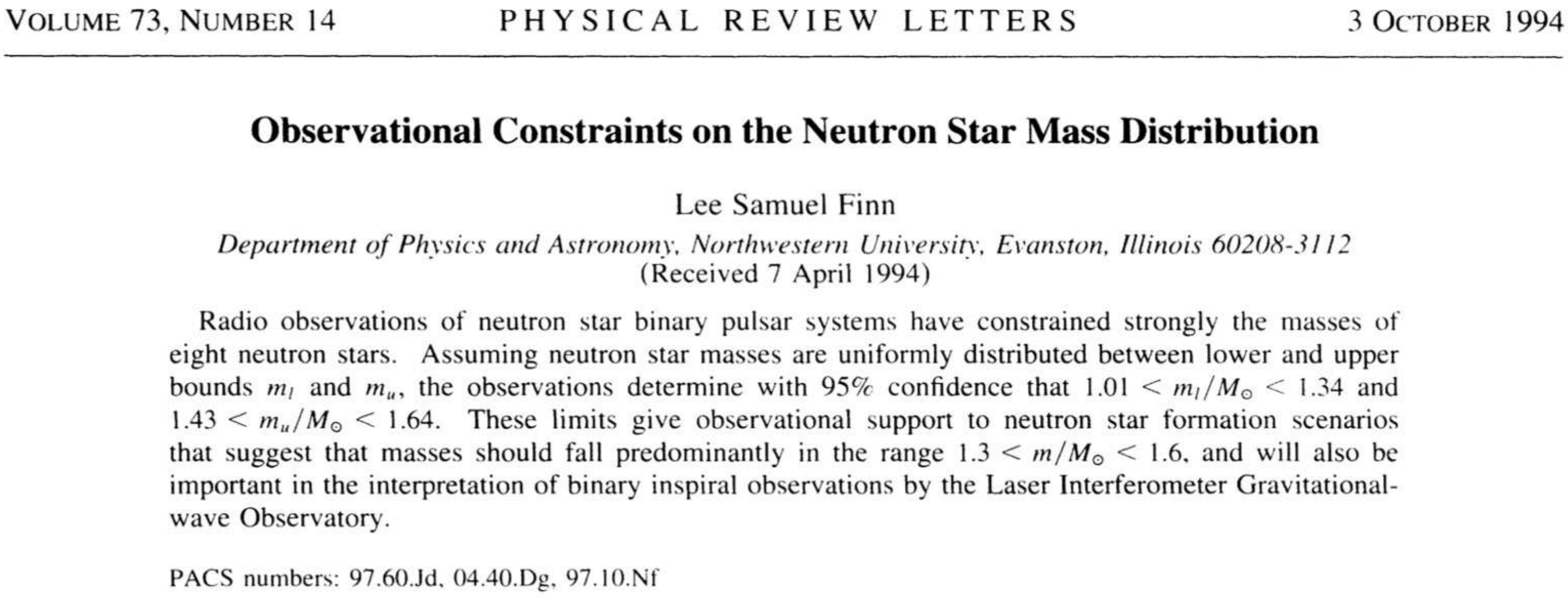

Sunshades and X-Ray Concentrators (56)
Possible TMD Locations
Detector Radiator
XTI (structure is IOB)
Focal Plane Modules (MIT/Amptek/GSFC) with SDD Shields (56)
GPS Antenna Bracket
Star Tracker (DTU)
Electronics (MBR, MIT, DTU)
Electronics Radiator (not shown)
DAPS – Az/El/Deploy/Latching Actuators (Moog)
Gimbal Bracket
HiPoS Box
Contamination Shield
Frangibolt Launch Lock Mounts (x4, 3-2-2-1 constraints)
AFRAM
Adapter Plate

The X-ray Timing Instrument (XTI) consists of an array of 56 X-ray “concentrator” optics and matching silicon detectors, which record the times of arrival (100 ns resolution) and energies of individual X-ray photons (0.2-12 keV). The payload uses an on-board GPS receiver to register photon detections to precise GPS time and position, while a star-tracker camera guides the pointing system, which uses gimbaled actuators to track targets with the XTI.

J0030+0451, is an isolated pulsar that spins roughly 200 times per second and is 337 parsecs (1,100 light years) from Earth, in the constellation Pisces. M ≈ 1.3 – 1.4 Mo; radius ≈ 13 km



Hotspots rotate in two scenarios for the pulsar J0030+0451, based on analysis of NICER data.Credit: NASA’s Goddard Space Flight Center/CI Lab



VOLUME 73, NUMBER 14 PHYSICAL REVIEW LETTERS 3 OCTOBER 1994

Observational Constraints on the Neutron Star Mass Distribution

Lee Samuel Finn

Department of Physics and Astronomy, Northwestern University, Evanston, Illinois 60208-3112

(Received 7 April 1994)

Radio observations of neutron star binary pulsar systems have constrained strongly the masses of eight neutron stars. Assuming neutron star masses are uniformly distributed between lower and upper bounds m_l and m_u, the observations determine with 95% confidence that $1.01 < m_l/M_\odot < 1.34$ and $1.43 < m_u/M_\odot < 1.64$. These limits give observational support to neutron star formation scenarios that suggest that masses should fall predominantly in the range $1.3 < m/M_\odot < 1.6$, and will also be important in the interpretation of binary inspiral observations by the Laser Interferometer Gravitational-wave Observatory.

PACS numbers: 97.60.Jd, 04.40.Dg, 97.10.Nf

Two sets of data

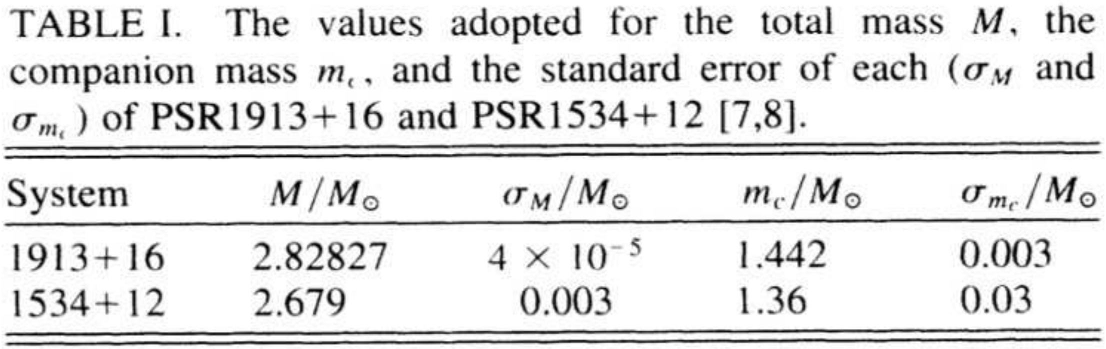


TABLE I. The values adopted for the total mass M, the companion mass m_c, and the standard error of each (σ_M and σ_{m_c}) of PSR1913+16 and PSR1534+12 [7,8].

System	$M/M_\odot$	$\sigma_M/M_\odot$	$m_c/M_\odot$	$\sigma_{m_c}/M_\odot$
1913+16	2.82827	4×10^{-5}	1.442	0.003
1534+12	2.679	0.003	1.36	0.03

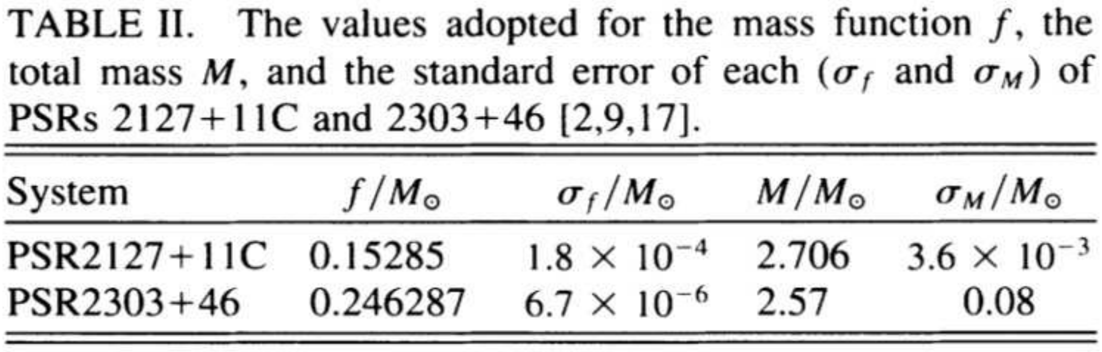


TABLE II. The values adopted for the mass function f, the total mass M, and the standard error of each (σ_f and σ_M) of PSRs 2127+11C and 2303+46 [2,9,17].

System	$f/M_\odot$	$\sigma_f/M_\odot$	$M/M_\odot$	$\sigma_M/M_\odot$
PSR2127+11C	0.15285	1.8×10^{-4}	2.706	3.6×10^{-3}
PSR2303+46	0.246287	6.7×10^{-6}	2.57	0.08

mass function

$$f = \frac{(m_2 \sin i)^3}{(m_1 + m_2)^2}$$

$$f,\ M,\ m_c \ \to\ \underbrace{\hat{f},\ \hat{M},\ \hat{m}_c}_{\text{measured values}}$$

mass function

total mass

mass of companion

Gaussian likelihoods:

$$P(x|\hat{x}, I) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{(x-\hat{x})^2}{2\sigma_x^2}\right]$$

any of the variables listed above

Our target: determine the upper and lower bound for the NS mass distribution.

The prior distribution is determined from very general considerations

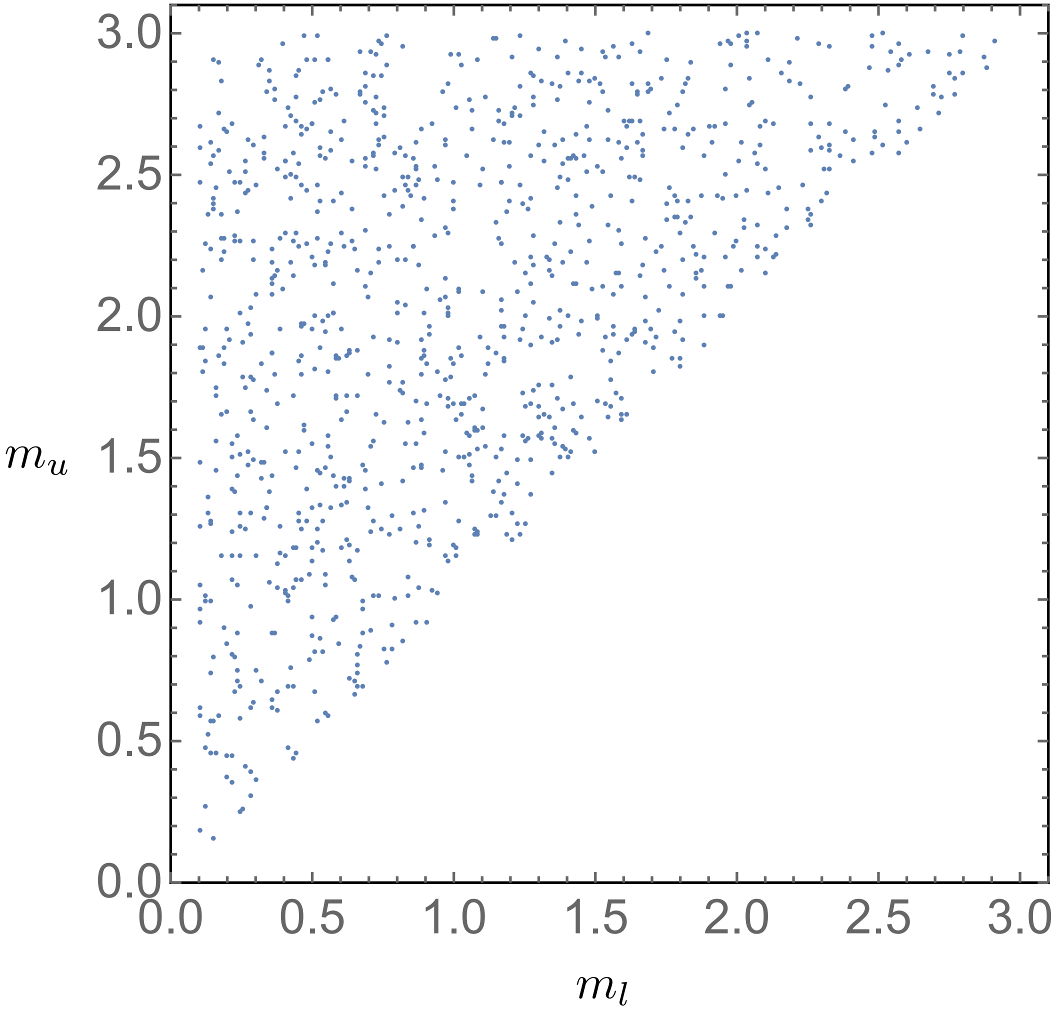
$$M_u > m_u > m_l > M_l$$

$$M_u \approx 3M_\odot$$

from causality and general relativity

$$M_l \approx 0.1M_\odot$$

from our understanding of the EOS



Uniform prior distribution
for the mass bounds

$$P(m_l, m_u|I) = \frac{2}{(M_u - M_l)^2}$$

Posterior distribution for the mass bounds

$$P(m_l, m_u|D, I) = \frac{P(D|m_l, m_u, I)}{P(D|I)} P(m_l, m_u|I)$$

Thanks to the independence of individual measurements

$$P(D|m_l, m_u, I) = \prod_n P(D_n|m_l, m_u, I)$$

Next we have to evaluate the different contributions to the individual likelihoods

$$P(\hat{m}_c, \hat{M}|m_l, m_u, I) = \int P(\hat{m}_c, \hat{M}|m_c, M, I)P(m_c, M|m_l, m_u, I)dm_c dM$$

we assume that the masses in the binary are independent

⇒

$$P(m_c, M|m_l, m_u, I) = P(m_c, M - m_c|m_l, m_u, I) = \frac{1}{(m_u - m_l)^2}$$

Similarly

$$P(\hat{f}, \hat{M}|m_l, m_u, I) = \int P(\hat{f}, \hat{M}|f, M, I)P(f, M|m_l, m_u, I)df dM$$

however, here

$$P(f, M|m_l, m_u, I) = P(f|M, m_l, m_u, I)P(M|m_l, m_u, I)$$

$$P(f, M|m_l, m_u, I) = P(f|M, m_l, m_u, I)P(M|m_l, m_u, I)$$

The two distributions on the r.h.s. can be evaluated separately

$$P(\hat{M}|m_l, m_u, I) = \frac{\max[0, \min(\hat{M} - m_l, m_u) - \max(\hat{M} - m_u, m_l)]}{(m_u - m_l)^2}$$

$$P(\hat{f}|\hat{M}, m_l, m_u, I) = \frac{1}{3} \frac{(\arcsin x_1 - \arcsin x_0)\,(\hat{M}/\hat{f})^{2/3}}{\min(\hat{M} - m_l, m_u) - \max(\hat{M} - m_u, m_l)} \, ,$$

where

$$x_0^2 = \max\left[0, 1 - \frac{(\hat{f}\hat{M}^2)^{2/3}}{\min[m_u, \max(\hat{M} - m_u, m_l)]^2}\right]$$

$$x_1^2 = \max\left[0, 1 - \frac{(\hat{f}\hat{M}^2)^{2/3}}{\max[m_l, \min(\hat{M} - m_l, m_u)]^2}\right]$$

Results

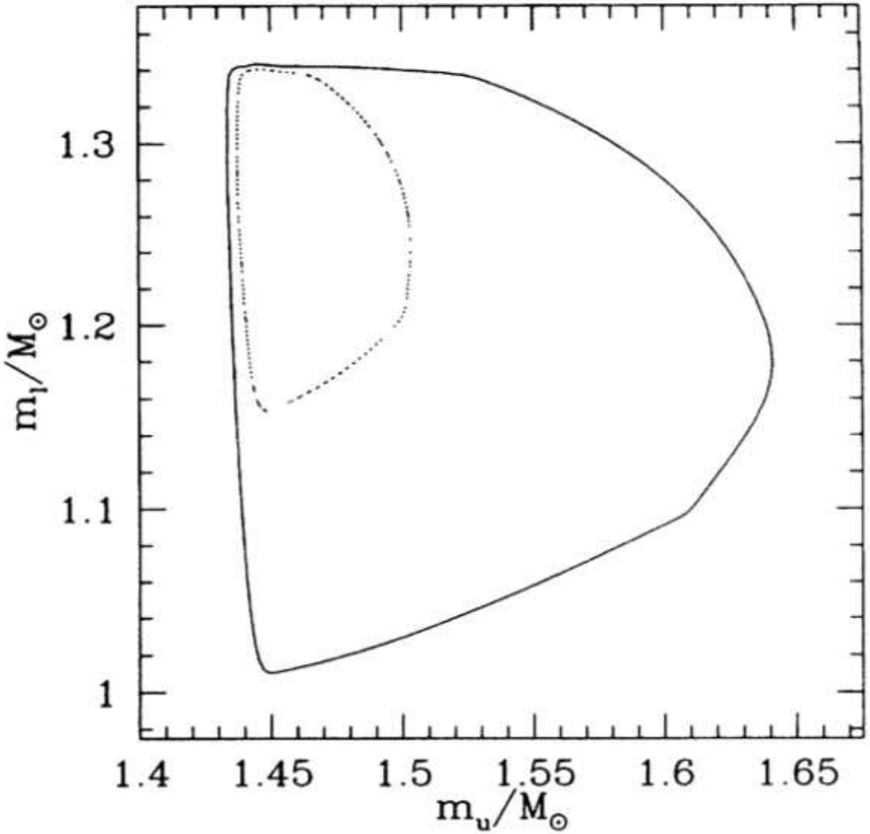


FIG. 1. Assuming ns masses are uniformly distributed between m_l and m_u, observations of PSRs 1534+12, 1913+16, 2127+11C, and 2303+46 determine the joint probability distribution for m_l and m_u. Shown here are contours enclosing regions of 68% (dotted) and 95% (solid) of this distribution.

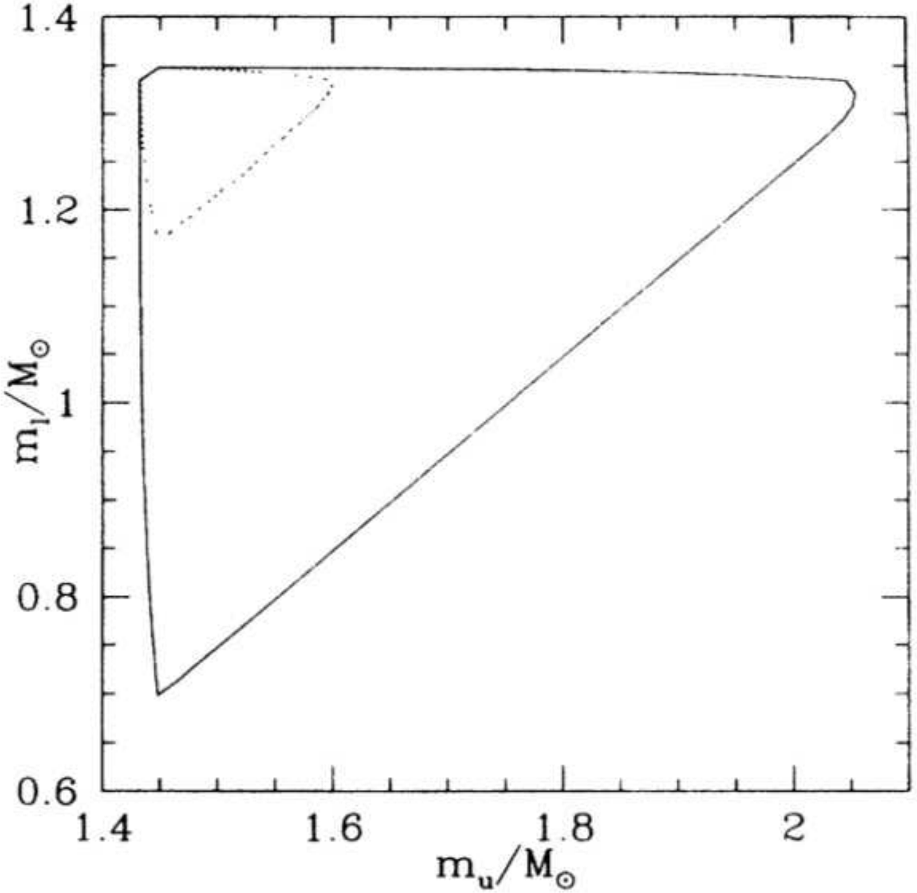


FIG. 2. As in Fig. 1, except that the contours are based on the constraints provided by observations of PSRs 1534+12 and 1913+16.

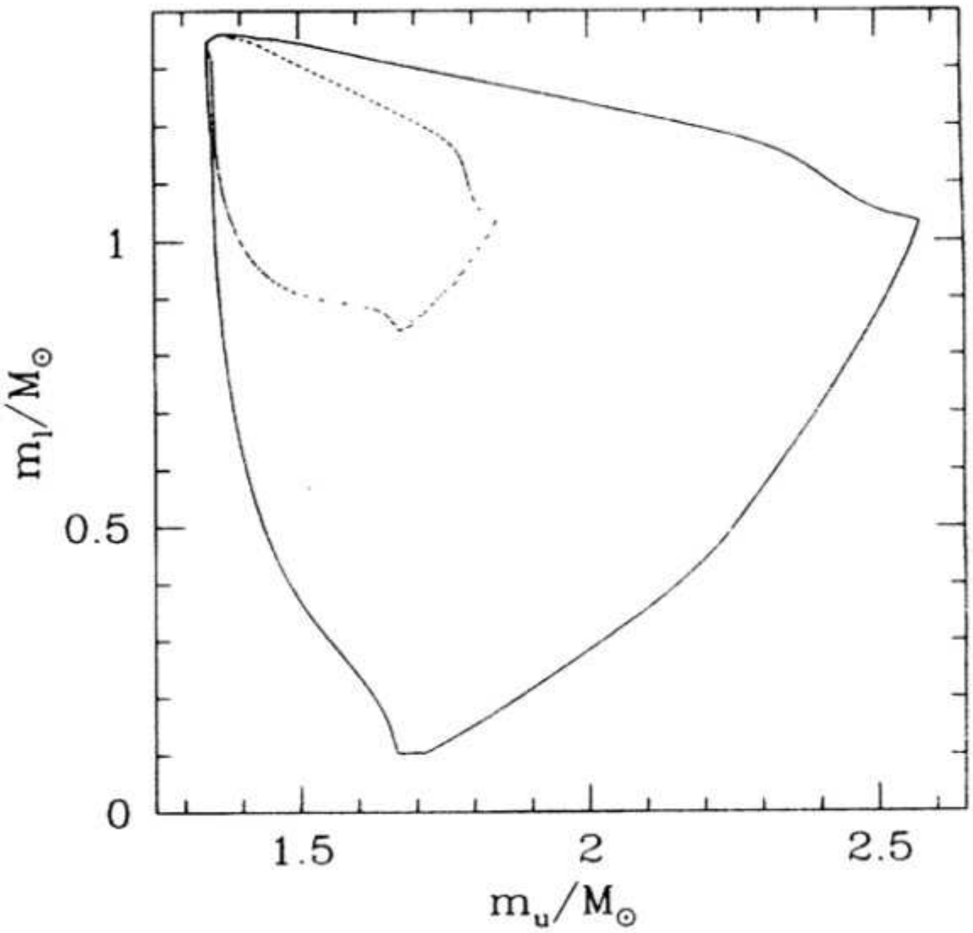


FIG. 3. As in Fig. 1, except that the contours are based on the constraints provided by observations of PSRs 2127+11C and 2303+46.

Additional references:

Model selection

A. Liddle, P. Mukherjee, D. Parkinson: "Model selection in cosmology", Astronomy & Geophysics 47 (2006) 4.30, https://academic.oup.com/astrogeo/article/47/4/4.30/206812

EM algorithm

- A. P. Dempster, N. M. Laird, and D. B. Rubin: “Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society series B, **39** (1977) 1
- J. Bilmes: “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”, ICSI preprint TR-97-021 (1998)
- B. Flury and A. Zoppé, “Exercises in EM” American Statistician, **54** (2000) 207.