

Introduction to Bayesian Statistics - 8

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

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

Recall the max. likelihood principle:

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

uniform distribution
(usually an improper prior)

likelihood

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

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

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

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

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

(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

likelihood

previous parameter estimate (constant values)

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

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

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

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

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

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

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

← 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

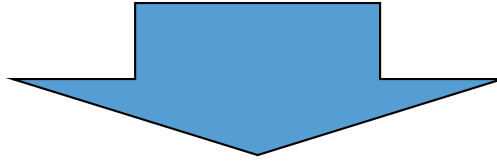
$$\theta - \frac{t \exp(-t/\theta)}{1 - \exp(-t/\theta)}$$

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

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

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

$$\begin{aligned} \text{mean failure time} &= \int_0^t t' p(t') dt' = \frac{1}{1 - e^{-t/\theta}} \int_0^t t' e^{-t'/\theta} \frac{dt'}{\theta} \\ &= \frac{\theta}{1 - e^{-t/\theta}} \left[1 - e^{-t/\theta} - (t/\theta) e^{-t/\theta} \right] \\ &= \theta - \frac{te^{-t/\theta}}{1 - e^{-t/\theta}} \end{aligned}$$



average log-likelihood

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

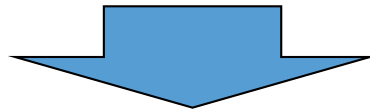
this ends the expectation step

the max of the mean likelihood

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

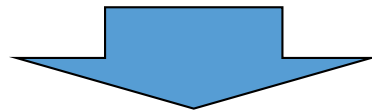
can be found by maximizing the approximate expression

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



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

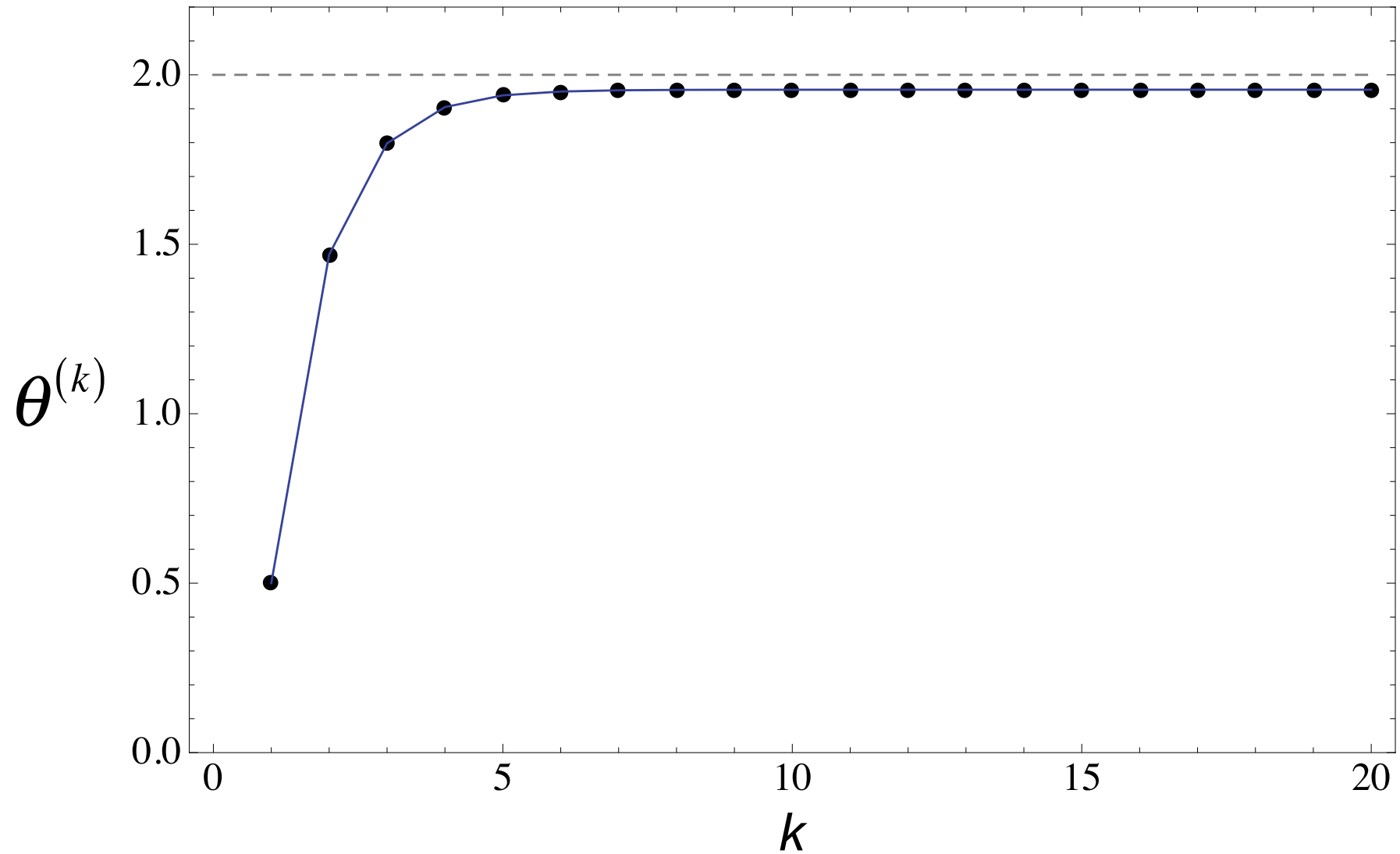
$$\frac{dQ}{d\theta} \approx -(n+m)\frac{1}{\theta} + \frac{1}{\theta^2} \left[n\langle u \rangle + r \left(\theta^{(k)} - \frac{t \exp(-t/\theta^{(k)})}{1 - \exp(-t/\theta^{(k)})} \right) + (m-r)(\theta^{(k)} + t) \right] = 0$$



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

iterate this until convergence ...

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

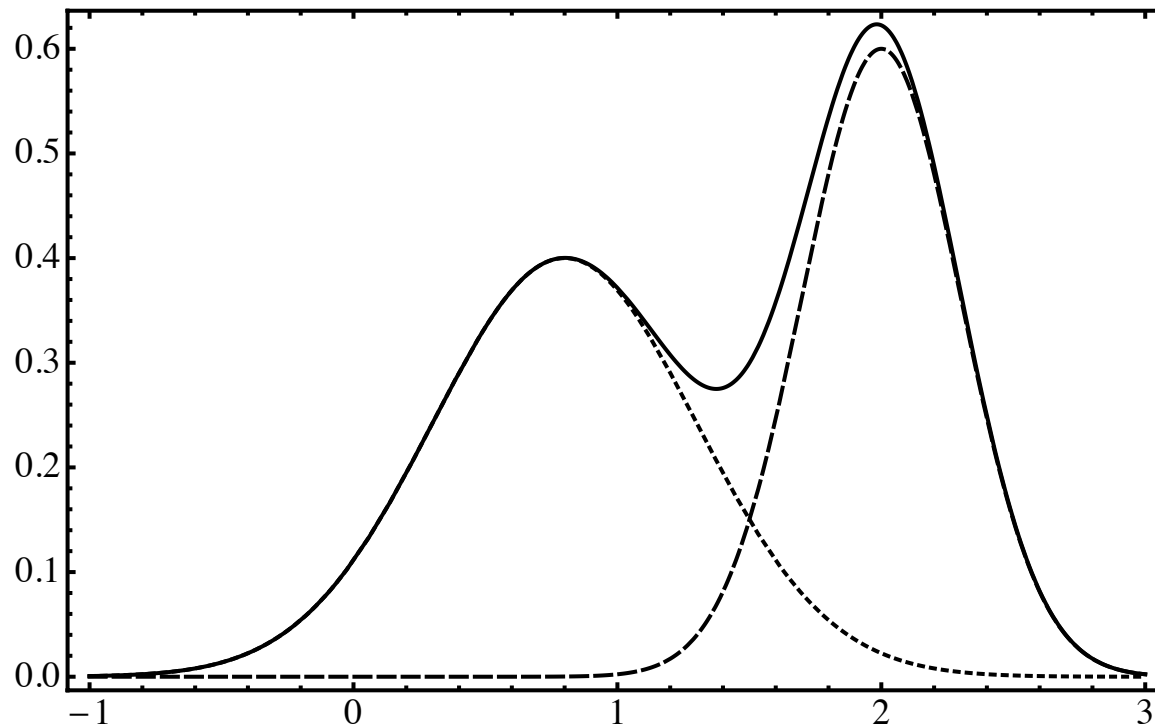


Important application of the EM method: parameters of “mixture models”.

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

$$\boldsymbol{\theta} = (\alpha_1, \dots, \alpha_M; \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M)$$

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



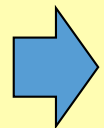
Example: a Gaussian mixture model (M=2)

direct maximization of log likelihood

$$\begin{aligned}\log \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}) &= \log \prod_n p(x_n | \boldsymbol{\theta}) = \sum_n \log p(x_n | \boldsymbol{\theta}) \\ &= \sum_n \log \left[\sum_{i=1}^M \alpha_i p_i(x_n | \boldsymbol{\theta}_i) \right]\end{aligned}$$

difficult numerical treatment ... however we can manage with a reinterpretation of the mixture model parameters ...

α_k = probability of drawing the k -th component of the mixture model



new (hidden) variable: y = index of component (integer values only)

thus we must redefine data and parameters

new likelihood which includes the hidden variables

$$\begin{aligned}\log \mathcal{L}'(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) &= \log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \\ &= \log \prod_n p(x_n, y_n | \boldsymbol{\theta}) \\ &= \sum_n \log \left[p(x_n | y_n, \boldsymbol{\theta}) p(y_n | \boldsymbol{\theta}) \right] \\ &= \sum_n \log \left[\alpha_{y_n} p_{y_n} \left(x_n | \boldsymbol{\theta}_{y_n} \right) \right]\end{aligned}$$

($\boldsymbol{\theta}_i$ are the parameters restricted to the i-th component)

The structure is simpler now, there is no sum in the argument of the logarithm, however there is a new hidden variable y .

Now we proceed by averaging the likelihood
(Expectation step)

new parameter
estimate

previous parameter
estimate

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) &= E_{\mathbf{y}} \left[\log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \mid \mathbf{x}, \boldsymbol{\theta}^{(i-1)} \right] \\ &= \int_Y \left[\log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \right] p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}^{(i-1)}) d\mathbf{y} \\ &\rightarrow \sum_{\mathbf{y}} \left[\log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \right] p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}^{(i-1)}) \end{aligned}$$

sum instead of integral, because the
y variate is discrete

prior probabilities in the expression of the averaged log-likelihood

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) = \sum_{\mathbf{y}} [\log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta})] p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}^{(i-1)})$$

and now we use Bayes:

$$p(y_n | x_n, \boldsymbol{\theta}) = \frac{p(x_n | y_n, \boldsymbol{\theta}) p(y_n | \boldsymbol{\theta})}{p(x_n | \boldsymbol{\theta})} = \frac{\alpha_{y_n} p_{y_n}(x_n | \boldsymbol{\theta}_{y_n})}{\sum_{k=1}^M \alpha_k p_k(x_n | \boldsymbol{\theta}_k)}$$

$$p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) = \prod_{n=1}^N p(y_n | x_n, \boldsymbol{\theta}) = \prod_{n=1}^N \frac{\alpha_{y_n} p_{y_n}(x_n | \boldsymbol{\theta}_{y_n})}{\sum_{k=1}^M \alpha_k p_k(x_n | \boldsymbol{\theta}_k)}$$

Therefore, using $\log \mathcal{L}'(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \sum_n \log \left[\alpha_{y_n} p_{y_n} \left(x_n \mid \boldsymbol{\theta}_{y_n} \right) \right]$

and $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) = \prod_{n=1}^N p(y_n \mid x_n, \boldsymbol{\theta})$

we find

$$\begin{aligned}
 Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) &= \sum_{\mathbf{y}} \left[\log p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta}) \right] p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}^{(i-1)}) \\
 &= \sum_{\mathbf{y}} \sum_{k=1}^N \log \left[\alpha_{y_k} p_{y_k} \left(x_k \mid \boldsymbol{\theta}_{y_k} \right) \right] \prod_{j=1}^N p(y_j \mid x_j, \boldsymbol{\theta}^{(i-1)}) \\
 &= \sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \sum_{k=1}^N \log \left[\alpha_{y_k} p_{y_k} \left(x_k \mid \boldsymbol{\theta}_{y_k} \right) \right] \prod_{j=1}^N p(y_j \mid x_j, \boldsymbol{\theta}^{(i-1)})
 \end{aligned}$$

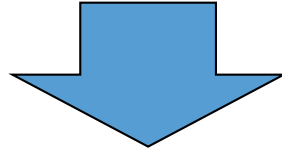
$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) &= \sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \sum_{k=1}^N \log \left[\alpha_{y_k} p_{y_k} (x_k | \boldsymbol{\theta}_{y_k}) \right] \prod_{j=1}^N p(y_j | x_j, \boldsymbol{\theta}^{(i-1)}) \\
&= \sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \sum_{k=1}^N \sum_{\ell=1}^M \delta_{\ell, y_k} \log \left[\alpha_{\ell} p_{\ell} (x_k | \boldsymbol{\theta}_{\ell}) \right] \prod_{j=1}^N p(y_j | x_j, \boldsymbol{\theta}^{(i-1)})
\end{aligned}$$

to decouple the variables, we add one sum and one Kronecker's delta...


after the decoupling, we can use the normalization of conditional probabilities

$$\sum_{y_j=1}^M p(y_j | x_j, \boldsymbol{\theta}^{(i-1)}) = 1$$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) = \sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \sum_{k=1}^N \sum_{\ell=1}^M \delta_{\ell, y_k} \log [\alpha_{\ell} p_{\ell}(x_k | \boldsymbol{\theta}_{\ell})] \prod_{j=1}^N p(y_j | x_j, \boldsymbol{\theta}^{(i-1)})$$



$$\begin{aligned}
 Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) &= \sum_{\ell=1}^M \sum_{k=1}^N \log [\alpha_{\ell} p_{\ell}(x_k | \boldsymbol{\theta}_{\ell})] \sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \delta_{\ell, y_k} \prod_{n=1}^N p(y_n | x_n, \boldsymbol{\theta}^{(i-1)}) \\
 &= \sum_{\ell=1}^M \sum_{k=1}^N \log [\alpha_{\ell} p_{\ell}(x_k | \boldsymbol{\theta}_{\ell})] \left\{ \sum_{y_1=1}^M \dots \sum_{y_{k-1}=1}^M \sum_{y_{k+1}=1}^M \dots \sum_{y_N=1}^M \prod_{\substack{j=1 \\ j \neq k}}^N p(y_j | x_j, \boldsymbol{\theta}^{(i-1)}) \right\} p(\ell | x_k, \boldsymbol{\theta}^{(i-1)}) \\
 &= \sum_{\ell=1}^M \sum_{k=1}^N \log [\alpha_{\ell} p_{\ell}(x_k | \boldsymbol{\theta}_{\ell})] \left\{ \prod_{\substack{j=1 \\ j \neq k}}^N \sum_{y_j=1}^M p(y_j | x_j, \boldsymbol{\theta}^{(i-1)}) \right\} p(\ell | x_k, \boldsymbol{\theta}^{(i-1)}) \\
 &= \sum_{\ell=1}^M \sum_{k=1}^N \log [\alpha_{\ell} p_{\ell}(x_k | \boldsymbol{\theta}_{\ell})] p(\ell | x_k, \boldsymbol{\theta}^{(i-1)})
 \end{aligned}$$


 these sums all add to 1 (normalization of conditional probabilities)

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) &= \sum_{\ell=1}^M \sum_{k=1}^N \ln [\alpha_{\ell} p(\ell | x_k, \boldsymbol{\theta})] p_{\ell}(x_k, \boldsymbol{\theta}^{(i-1)}) \\
&= \sum_{\ell=1}^M \sum_{k=1}^N \ln \alpha_{\ell} p_{\ell}(x_k, \boldsymbol{\theta}^{(i-1)}) + \sum_{\ell=1}^M \sum_{k=1}^N \ln p(\ell | x_k, \boldsymbol{\theta}) p_{\ell}(x_k, \boldsymbol{\theta}^{(i-1)})
\end{aligned}$$



this depends only on the α parameters

this term depends on the parameters of the component distributions

Thus there are two terms that can be maximized separately. Moreover, the first term must be maximized with the normalization constraint, i.e.

$$\frac{\partial}{\partial \alpha_m} \left[\sum_{\ell=1}^M \sum_{k=1}^N \log \alpha_{\ell} p(\ell | x_k, \boldsymbol{\theta}^{(i-1)}) + \lambda \left(\sum_{\ell=1}^M \alpha_{\ell} - 1 \right) \right] = 0$$

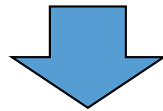


$$\sum_{k=1}^N \frac{1}{\alpha_m} p(m | x_k, \boldsymbol{\theta}^{(i-1)}) + \lambda = 0$$

$$\sum_{k=1}^N \frac{1}{\alpha_m} p(m|x_k, \boldsymbol{\theta}^{(i-1)}) + \lambda = 0$$



$$\sum_{k=1}^N p(m|x_k, \boldsymbol{\theta}^{(i-1)}) = -\lambda \alpha_m$$



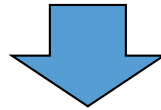
$$\sum_{m=1}^M \sum_{k=1}^N p(m|x_k, \boldsymbol{\theta}^{(i-1)}) = -\lambda \sum_{m=1}^M \alpha_m$$



$$\lambda = -N \quad \rightarrow \quad \alpha_m = \frac{1}{N} \sum_{k=1}^N p(m|x_k, \boldsymbol{\theta}^{(i-1)})$$

This is as far as we can go without introducing an explicit form for the component distributions: to evaluate the other term we explicitly consider the 1D Gaussian mixture model:

$$p_\ell(x|\mu_\ell, \sigma_\ell) = \frac{1}{\sqrt{2\pi\sigma_\ell^2}} \exp\left(-\frac{(x - \mu_\ell)^2}{2\sigma_\ell^2}\right)$$



$$\sum_{\ell=1}^M \sum_{k=1}^N \ln p_\ell(x_k, \boldsymbol{\theta}) p(\ell|x_k, \boldsymbol{\theta}^{(i-1)}) = \sum_{\ell=1}^M \sum_{k=1}^N \left[-\frac{1}{2} \ln(2\pi\sigma_\ell^2) - \frac{(x_k - \mu_\ell)^2}{2\sigma_\ell^2} \right] p(\ell|x_k, \mu_\ell^{(i-1)}, \sigma_\ell^{(i-1)})$$



$$\frac{\partial}{\partial \mu_m} \sum_{\ell=1}^M \sum_{k=1}^N \ln p_\ell(x_k, \boldsymbol{\theta}) p(\ell|x_k, \boldsymbol{\theta}^{(i-1)}) = -2 \sum_{k=1}^N \frac{(x_k - \mu_m)}{2\sigma_m^2} p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = 0$$

$$\frac{\partial}{\partial \mu_m} \sum_{\ell=1}^M \sum_{k=1}^N \ln p_{\ell}(x_k, \boldsymbol{\theta}) p(\ell|x_k, \boldsymbol{\theta}^{(i-1)}) = -2 \sum_{k=1}^N \frac{(x_k - \mu_m)}{2\sigma_m^2} p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = 0$$



$$\mu_m = \frac{\sum_{k=1}^N x_k p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})}{\sum_{k=1}^N p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})}$$

moreover, if we let $c_m = 1/\sigma_m^2$

$$\begin{aligned} \frac{\partial}{\partial c_m} \sum_{\ell=1}^M \sum_{k=1}^N \ln p_{\ell}(x_k, \boldsymbol{\theta}) p(\ell|x_k, \boldsymbol{\theta}^{(i-1)}) &= \frac{\partial}{\partial c_m} \sum_{\ell=1}^M \sum_{k=1}^N \left[-\frac{1}{2} \ln(2\pi\sigma_{\ell}^2) - \frac{(x_k - \mu_{\ell})^2}{2\sigma_{\ell}^2} \right] p(\ell|x_k, \mu_{\ell}^{(i-1)}, \sigma_{\ell}^{(i-1)}) \\ &= \sum_{k=1}^N \left[\frac{1}{2c_m} - \frac{1}{2}(x_k - \mu_m)^2 \right] p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) \\ &= \sum_{k=1}^N \left[\frac{\sigma_m^2}{2} - \frac{1}{2}(x_k - \mu_m)^2 \right] p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = 0 \end{aligned}$$

$$\frac{\partial}{\partial c_m} \sum_{\ell=1}^M \sum_{k=1}^N \ln p_{\ell}(x_k, \boldsymbol{\theta}) p(\ell|x_k, \boldsymbol{\theta}^{(i-1)}) = \sum_{k=1}^N \left[\frac{\sigma_m^2}{2} - \frac{1}{2}(x_k - \mu_m)^2 \right] p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = 0$$



$$\sigma_m^2 = \frac{\sum_{k=1}^N (x_k - \mu_m)^2 p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})}{\sum_{k=1}^N p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})}$$

Finally we find the following set of recursive formulas, that combine the E and M steps:

$$p_m(x|\mu_m, \sigma_m) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{(x - \mu_m)^2}{2\sigma_m^2}\right)$$

$$p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = \frac{\alpha_m^{(i-1)} p_m(x_k|\mu_m^{(i-1)}, \sigma_m^{(i-1)})}{\sum_{k=1}^M \alpha_m^{(i-1)} p_m(x_k|\mu_m^{(i-1)}, \sigma_m^{(i-1)})}$$

$$\alpha_m^{(i)} = \frac{1}{N} \sum_{k=1}^N p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})$$

$$\mu_m^{(i)} = \frac{\sum_{k=1}^N x_k p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})}{\sum_{k=1}^N p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})}$$

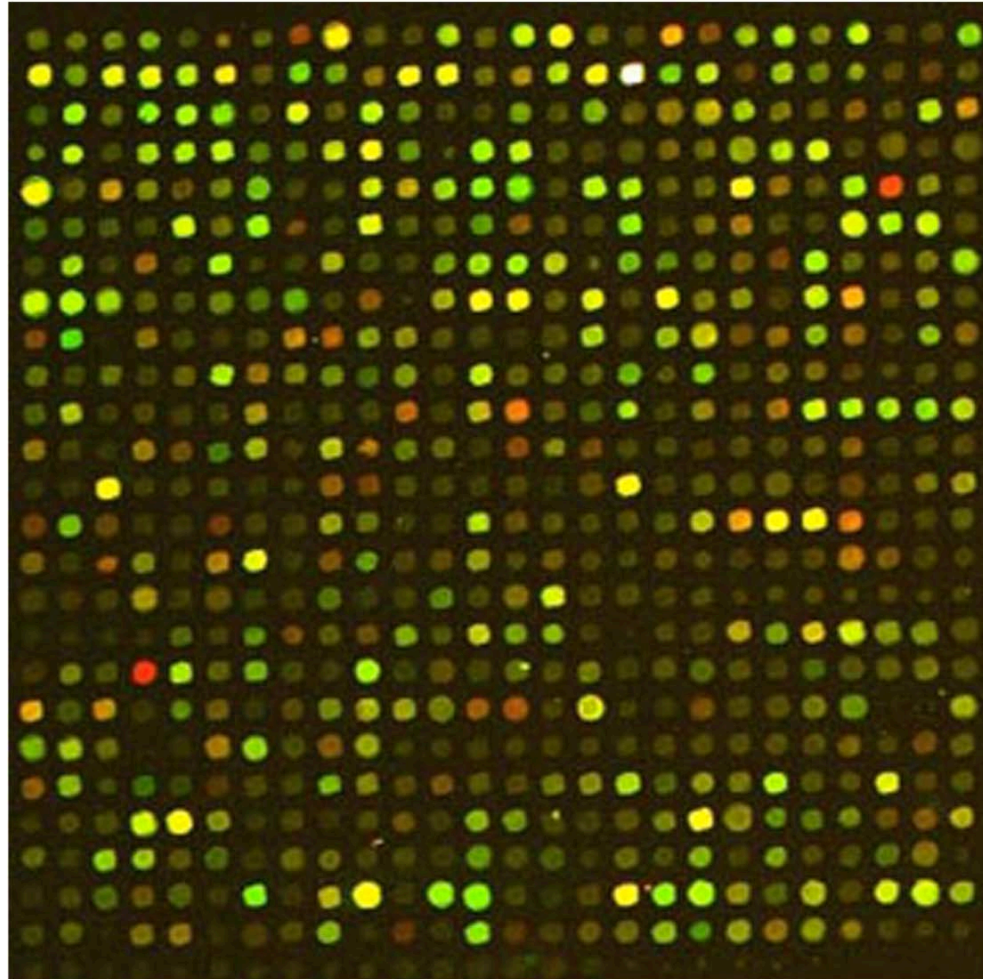
$$(\sigma_m^{(i)})^2 = \frac{\sum_{k=1}^N (x_k - \mu_m^{(i)})^2 p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})}{\sum_{k=1}^N p(m|x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})}$$

We remark that the probabilities

$$p(y_n | x_n, \boldsymbol{\theta}) = \frac{\alpha_{y_n} p_{y_n}(x_n | \boldsymbol{\theta}_{y_n})}{\sum_{k=1}^M \alpha_k p_k(x_n | \boldsymbol{\theta}_k)}$$

are an estimate of the frequencies of the y_n using the observed data x_n , and this amounts to a classification (selection of one of the component distributions).

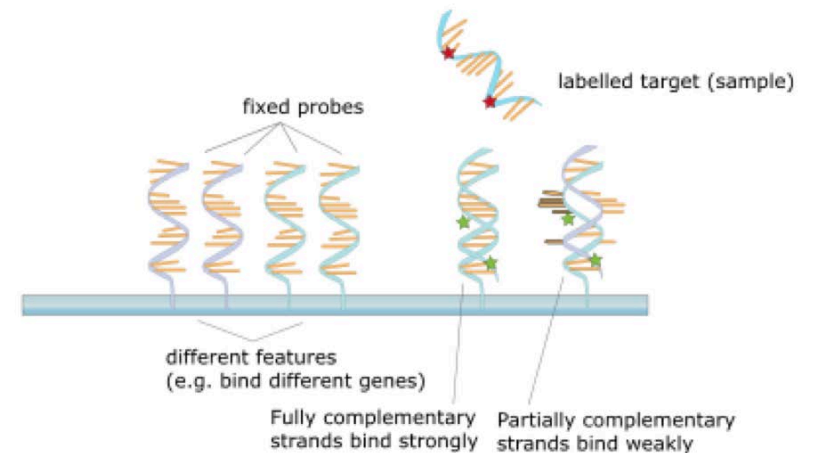
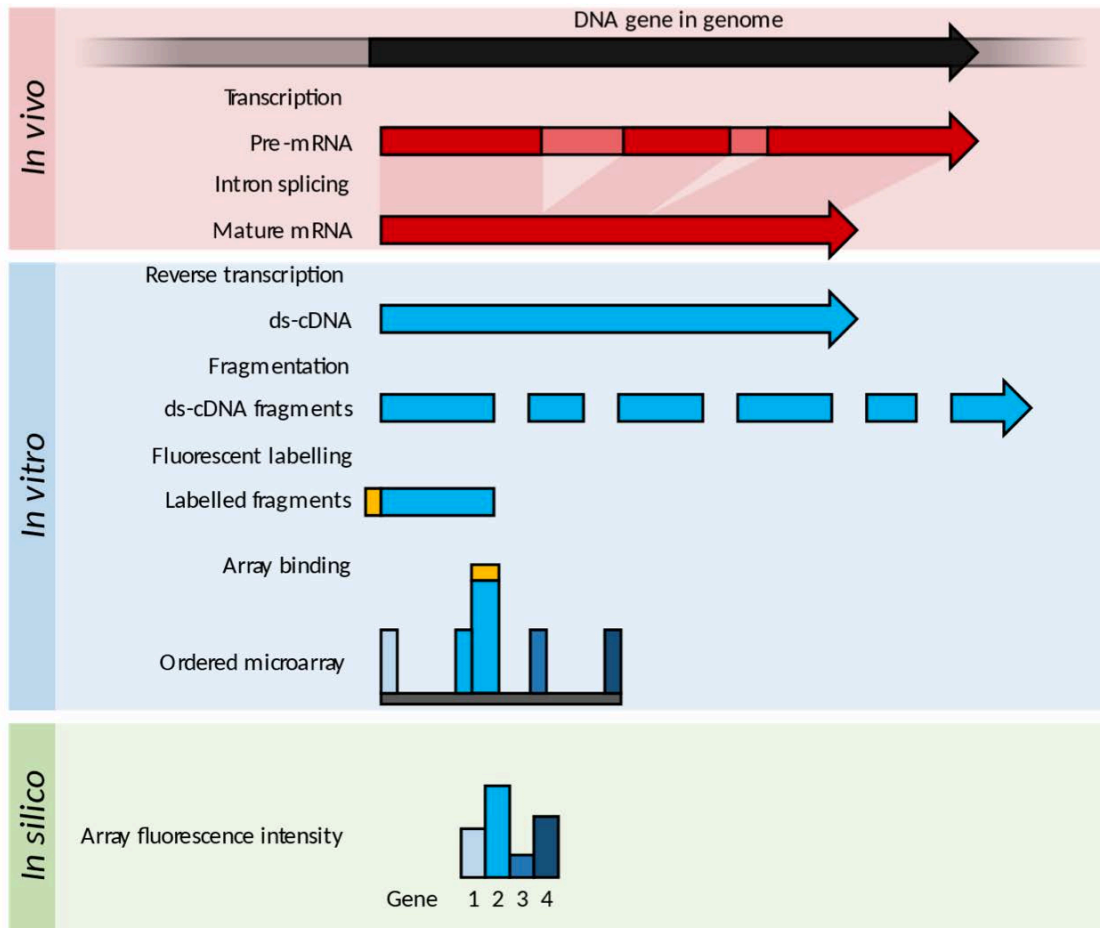
Example: classification of response of DNA microarrays.



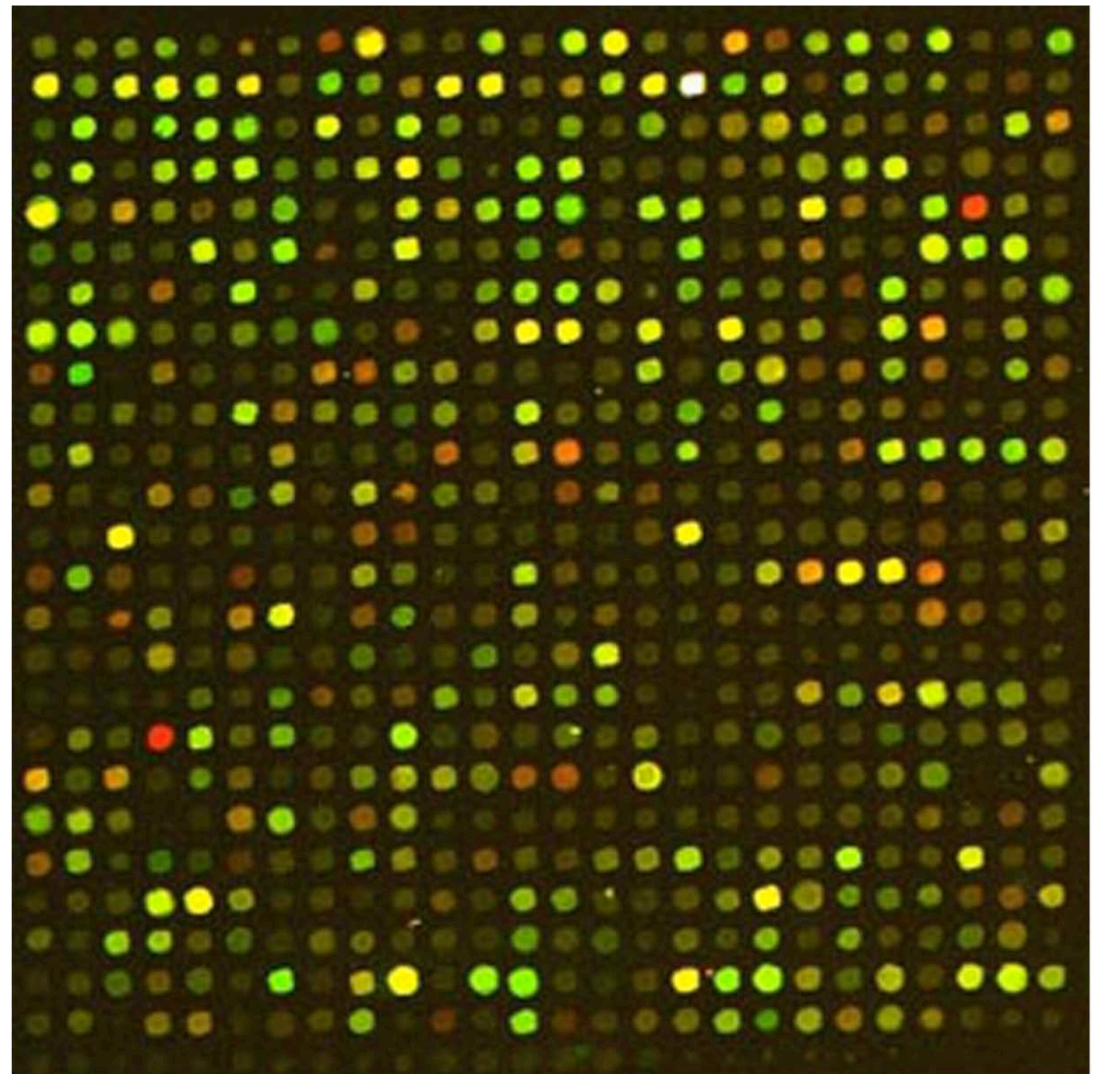
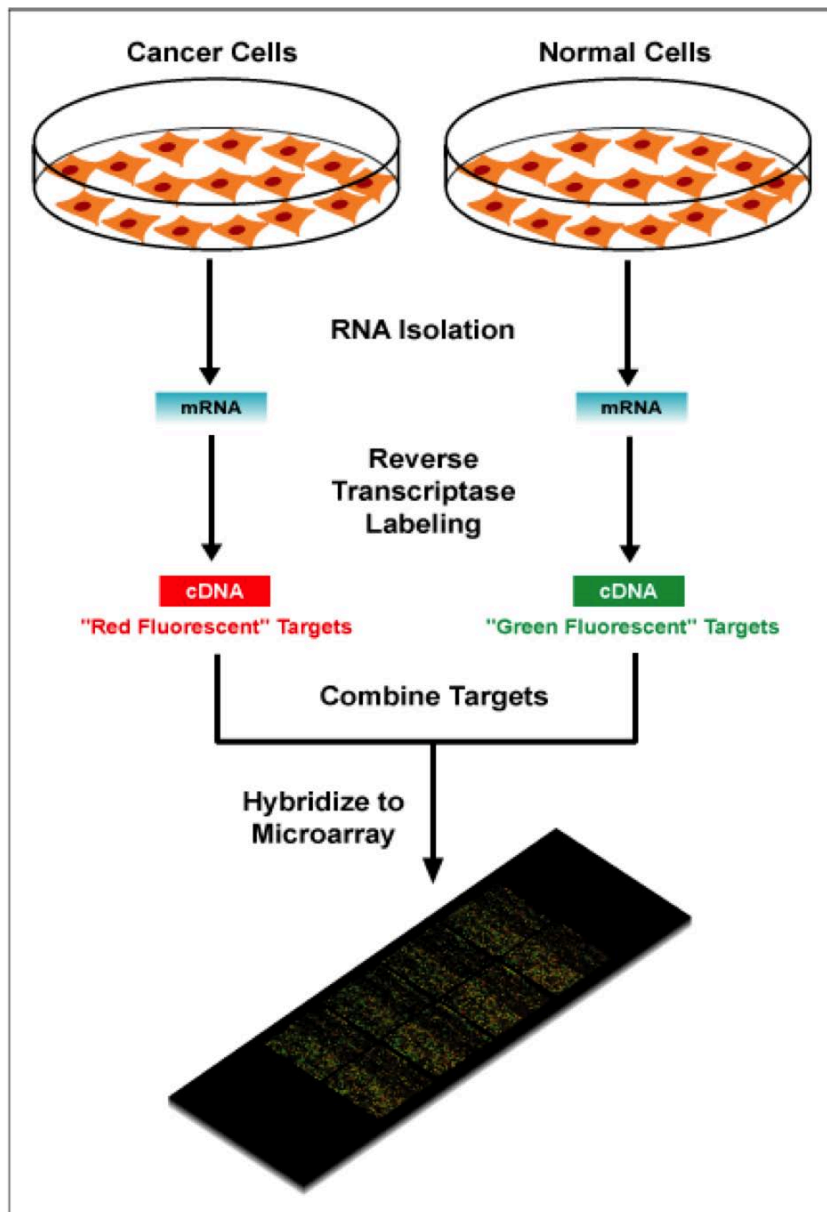
Microarray image
from: http://www.wormbook.org/chapters/www_germlinogenomics/germlinogenomics.html

Microarray: lab on a chip





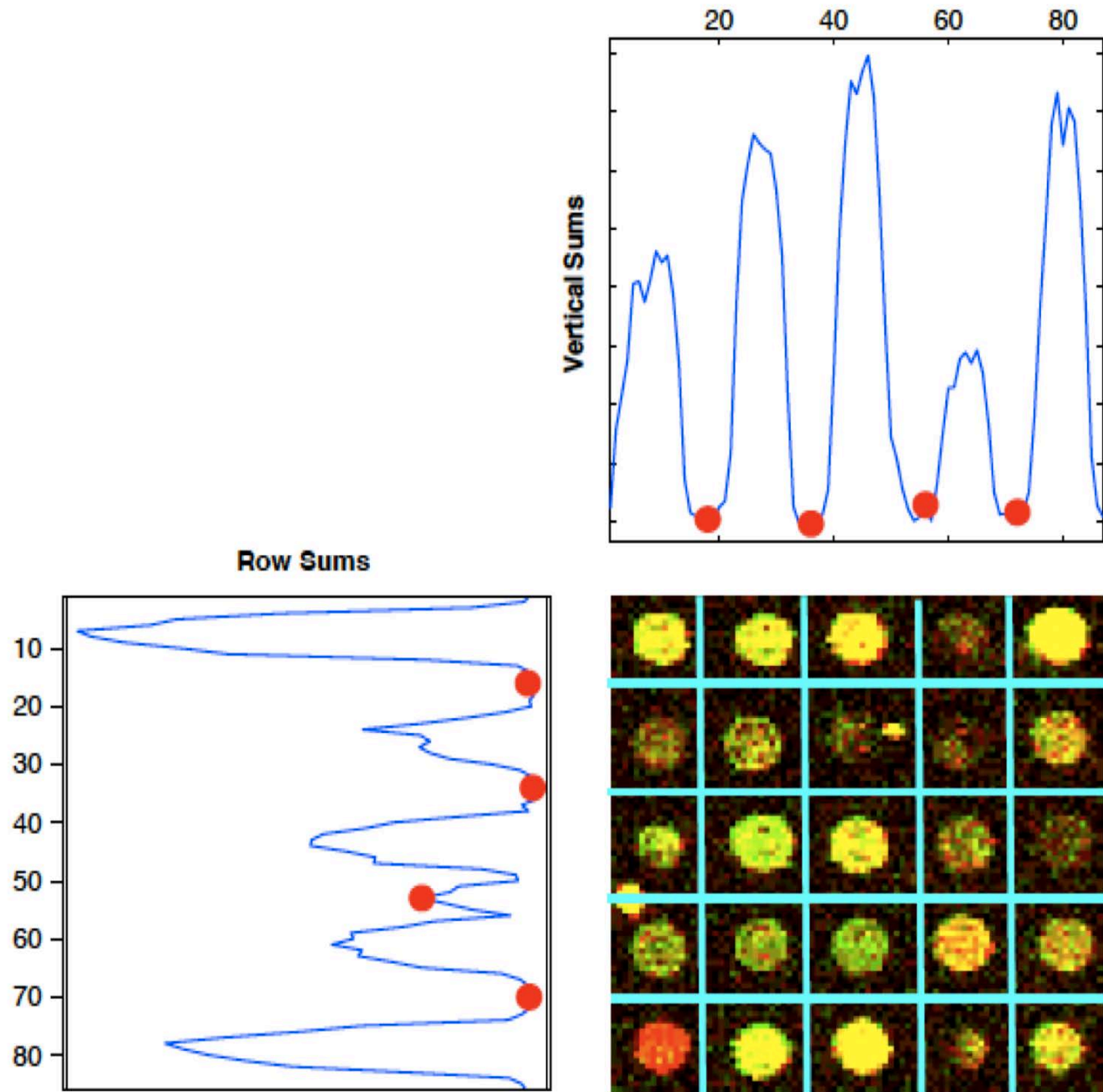
Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism and reverse transcriptase is used to copy the mRNA into stable cDNA (blue). In microarrays, the cDNA is fragmented and fluorescently labelled (orange). The labelled fragments bind to an ordered array of complementary oligonucleotides and measurement of fluorescent intensity across the array indicates the abundance of a predetermined set of sequences. These sequences are typically specifically chosen to report on genes of interest within the organism's genome. (from https://en.wikipedia.org/wiki/File:Summary_of_RNA_Microarray.svg)



Microarray image from: http://www.wormbook.org/chapters/www_germlinegenomics/germlinegenomics.html

Further informations on DNA microarrays: <http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html>

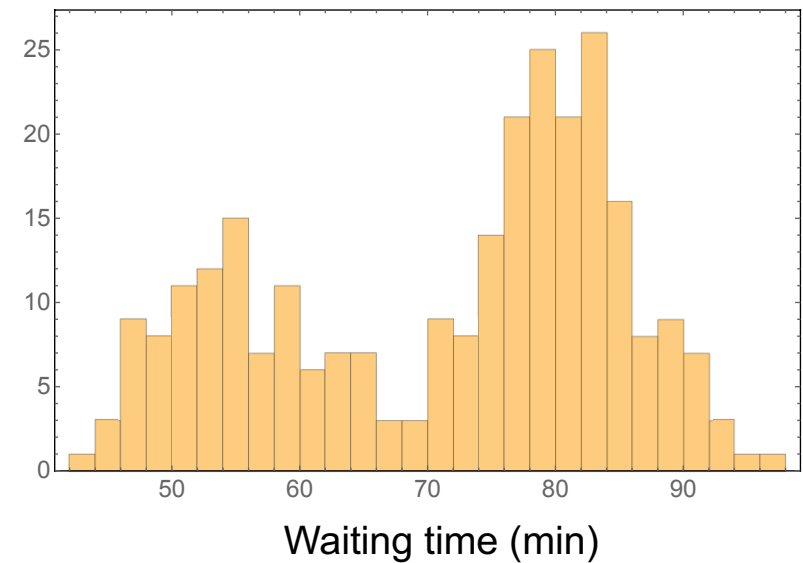
From Blekas et al., "Mixture Model Analysis of DNA Microarray Images", IEEE Trans. on Medical Imaging 24 (2005) 901



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



Gaussian mixture model for waiting time distribution (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

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

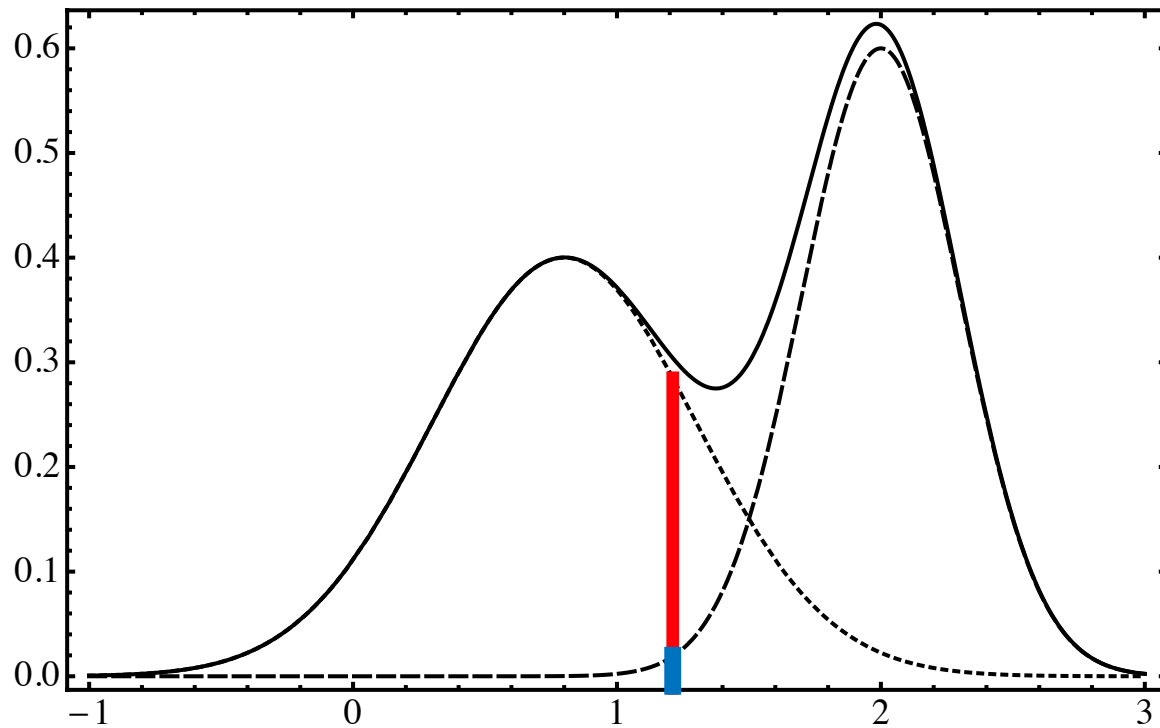
Again, we substitute the likelihood with the new one

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

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.

Then

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



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

$$\begin{aligned}
 p_i &= \frac{\alpha \times N(w_i; \mu_1, \sigma_1)}{\alpha \times N(w_i; \mu_1, \sigma_1) + (1 - \alpha) \times N(w_i; \mu_2, \sigma_2)} \\
 &= \frac{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2 / 2(\sigma_1^{(k)})^2] / \sqrt{2\pi(\sigma_1^{(k)})^2}}{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2 / 2(\sigma_1^{(k)})^2] / \sqrt{2\pi(\sigma_1^{(k)})^2} + (1 - \alpha^{(k)}) \exp[-(w_i - \mu_2^{(k)})^2 / 2(\sigma_2^{(k)})^2] / \sqrt{2\pi(\sigma_2^{(k)})^2}}
 \end{aligned}$$

Now, averaging the log likelihood with respect to the missing 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 probability α)

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

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

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

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

Finally we have the following set of equations:

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

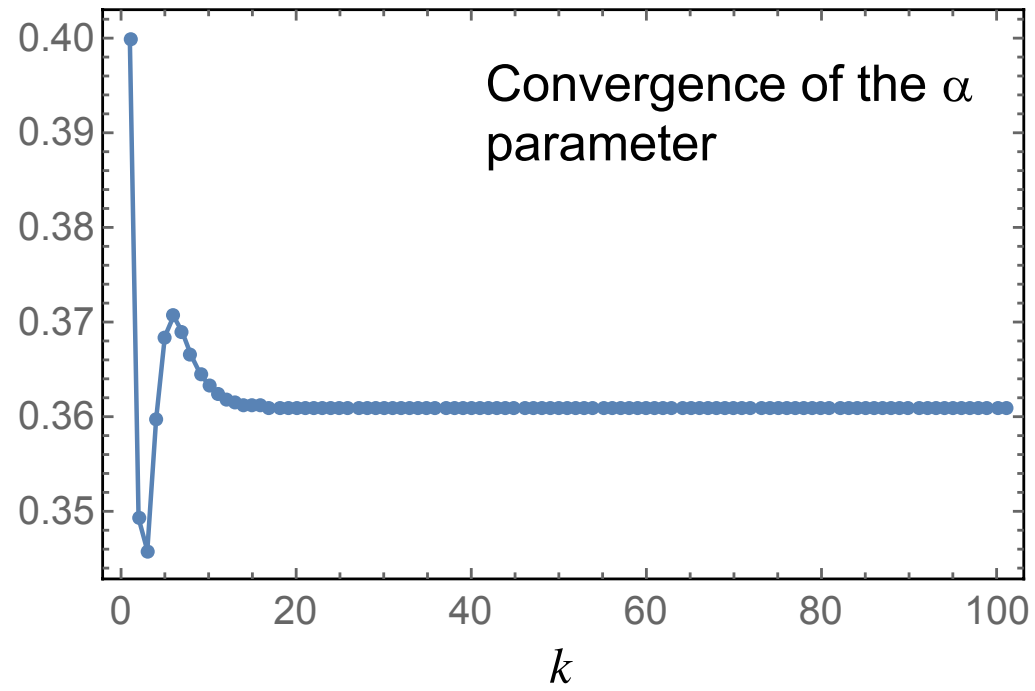
$$\alpha^{(k+1)} = \frac{1}{N} \sum_i p_i^{(k)}$$

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

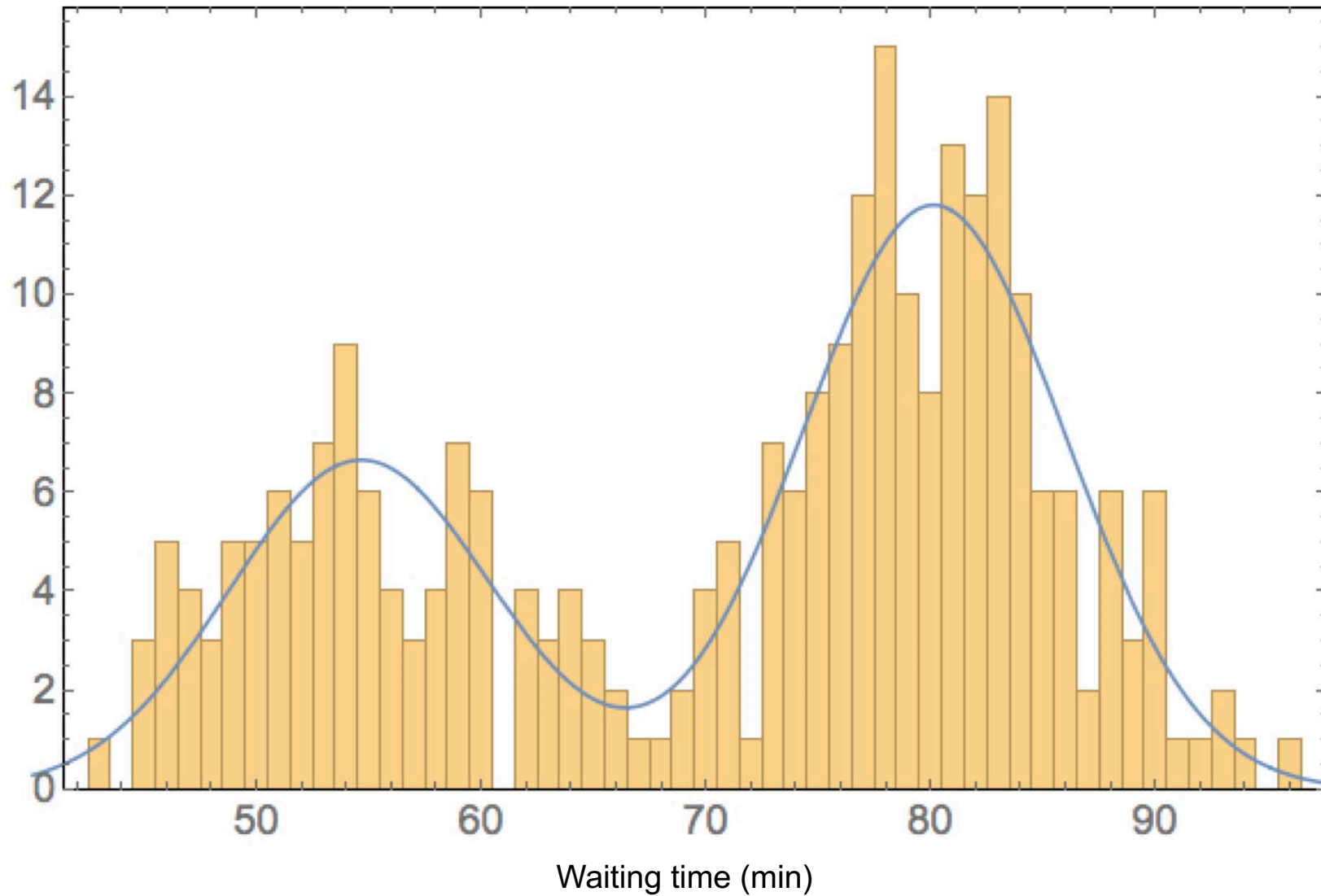
$$\mu_1^{(k+1)} = \frac{\sum_i p_i^{(k)} w_i}{\sum_i p_i^{(k)}}$$

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

$$\mu_2^{(k+1)} = \frac{\sum_i (1 - p_i^{(k)}) w_i}{\sum_i (1 - p_i^{(k)})}$$



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



2. Neutron star mass range

(Finn, PRL 73 (1994) 1878)

Neutron stars:

- The densest objects this side of an event horizon, with a mean density $\approx 10^{15} \text{ g cm}^{-3}$. Four teaspoons contain as much mass as the Moon.
- The largest surface gravity, about $10^{14} \text{ cm s}^{-2}$, 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 largest 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 neutrons 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 Telegraph*.

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].

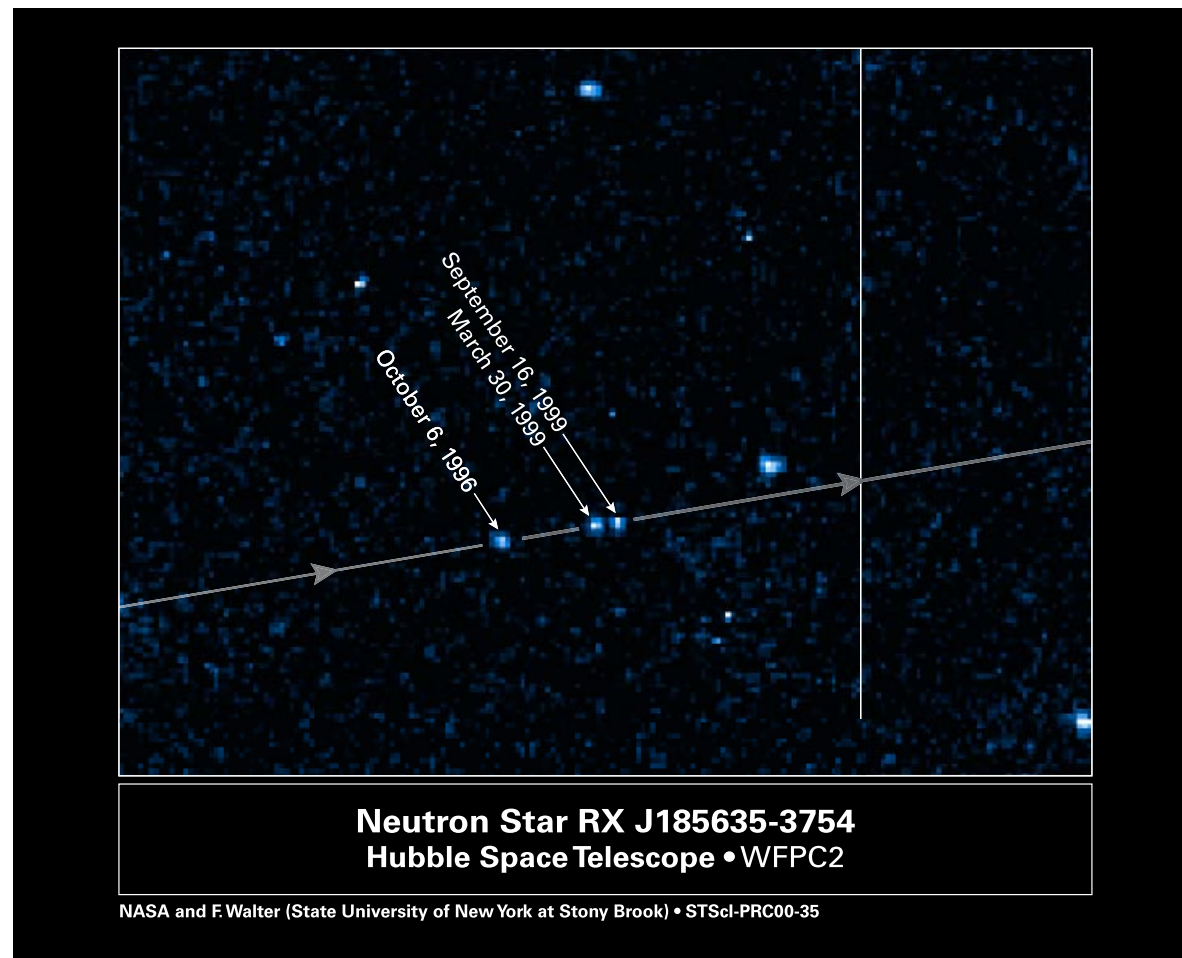
1974 The first binary pulsar, PSR 1913+16, discovered by Hulse and Taylor [15] (Nobel Prize 1993). It's orbital decay is the first observation [16] proving existence of gravitational radiation. Lattimer and Schramm [17] suggest decompressing neutron star matter from merging compact binaries leads to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



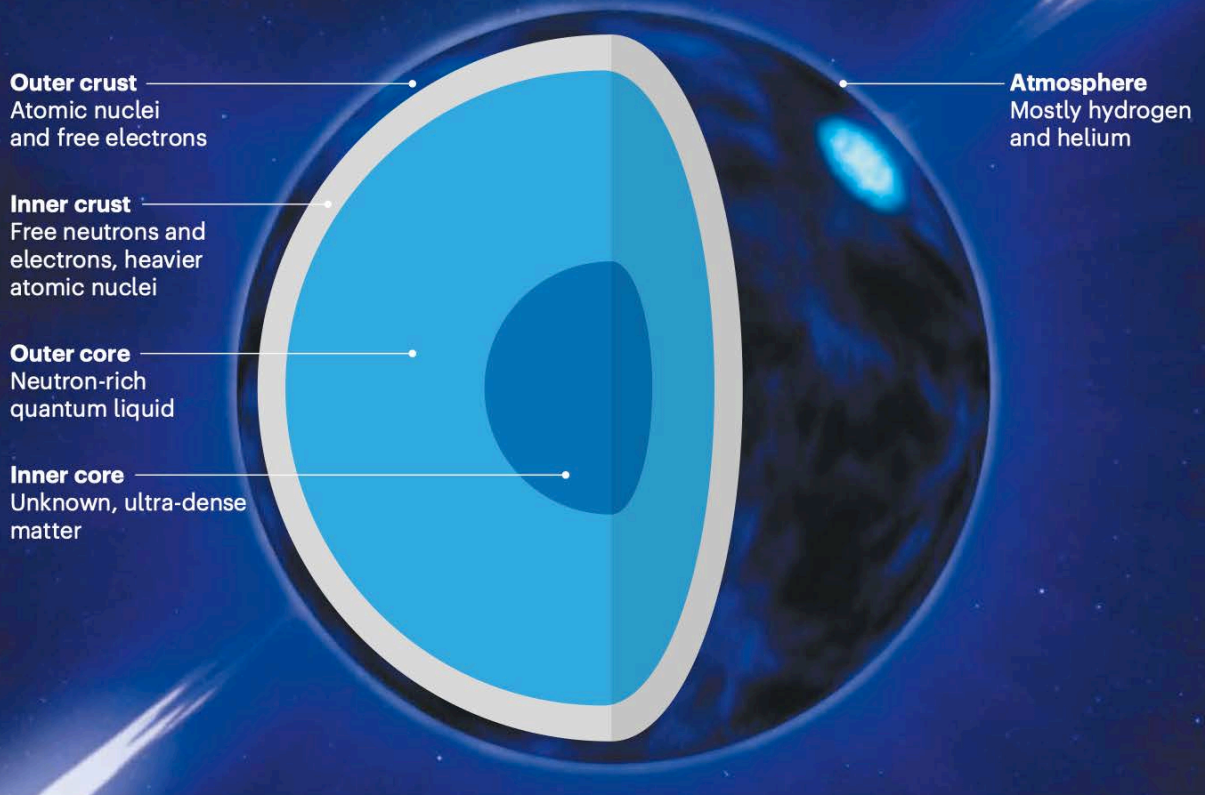
The Motion of RX J185635-3754 - The Nearest Neutron Star to Earth

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.

DENSE MATTER

Neutron stars get denser with depth. Although researchers have a good sense of the composition of the outer layers, the ultra-dense inner core remains a mystery.



Core scenarios

A number of possibilities have been suggested for the inner core, including these three options.

- u Up quark
- d Down quark
- s Strange quark
- d̄ Anti-down quark



Quarks

The constituents of protons and neutrons — up and down quarks — roam freely.



Bose-Einstein condensate

Particles such as pions containing an up quark and an anti-down quark combine to form a single quantum-mechanical entity.



Hyperons

Particles called hyperons form. Like protons and neutrons, they contain three quarks but include 'strange' quarks.

SOURCE: ADAPTED FROM NASA GODDARD SVS

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

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

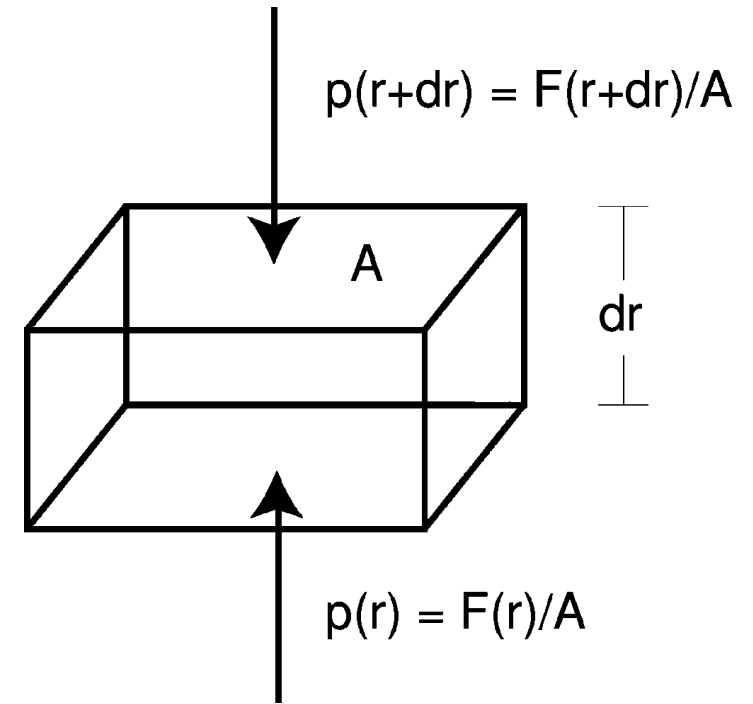
energy density

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



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

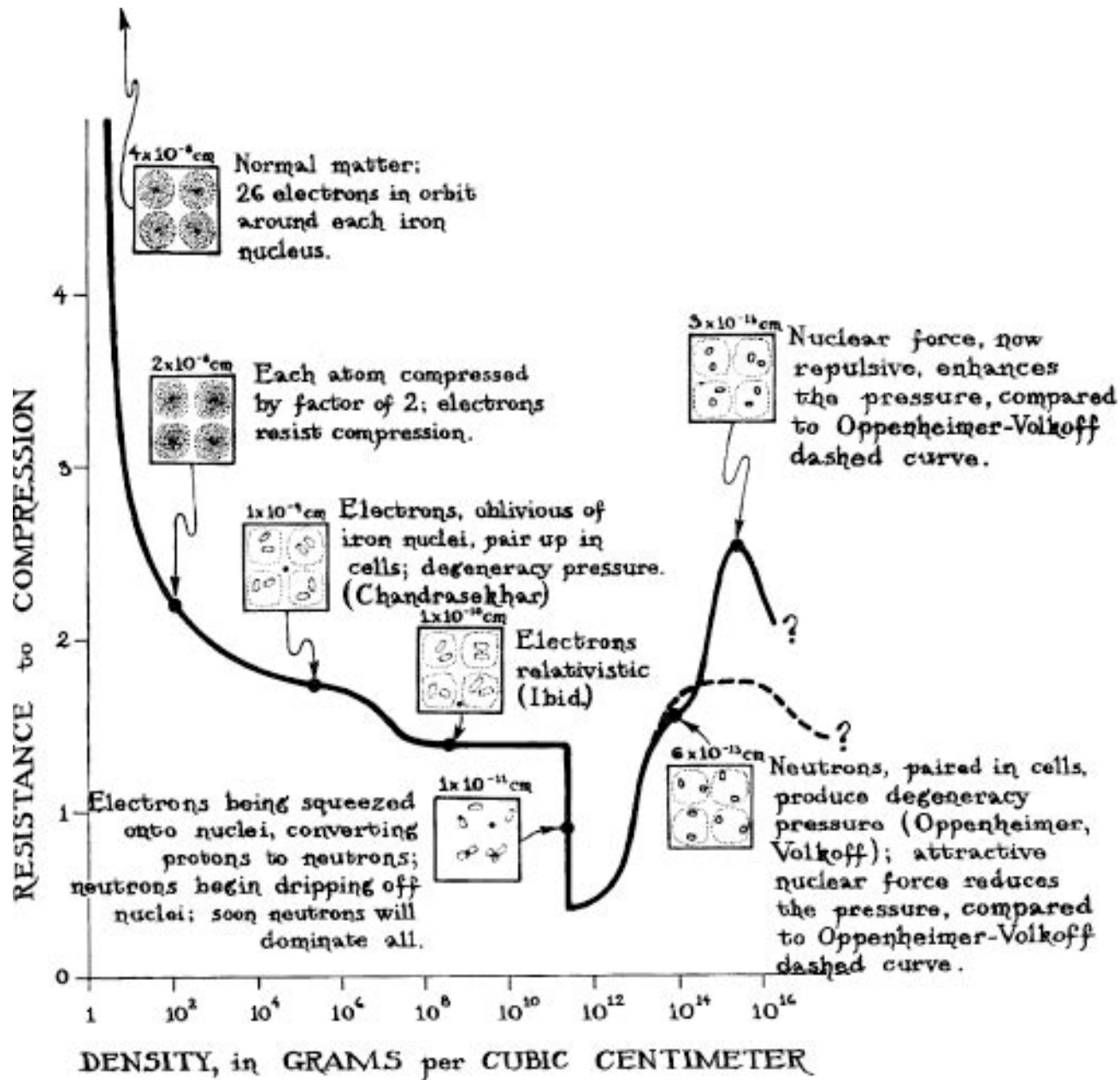
$$\mathcal{M}(r) = 4\pi \int_0^r \rho(r')r'^2 dr' = \frac{4\pi}{c^2} \int_0^r \epsilon(r')r'^2 dr'$$



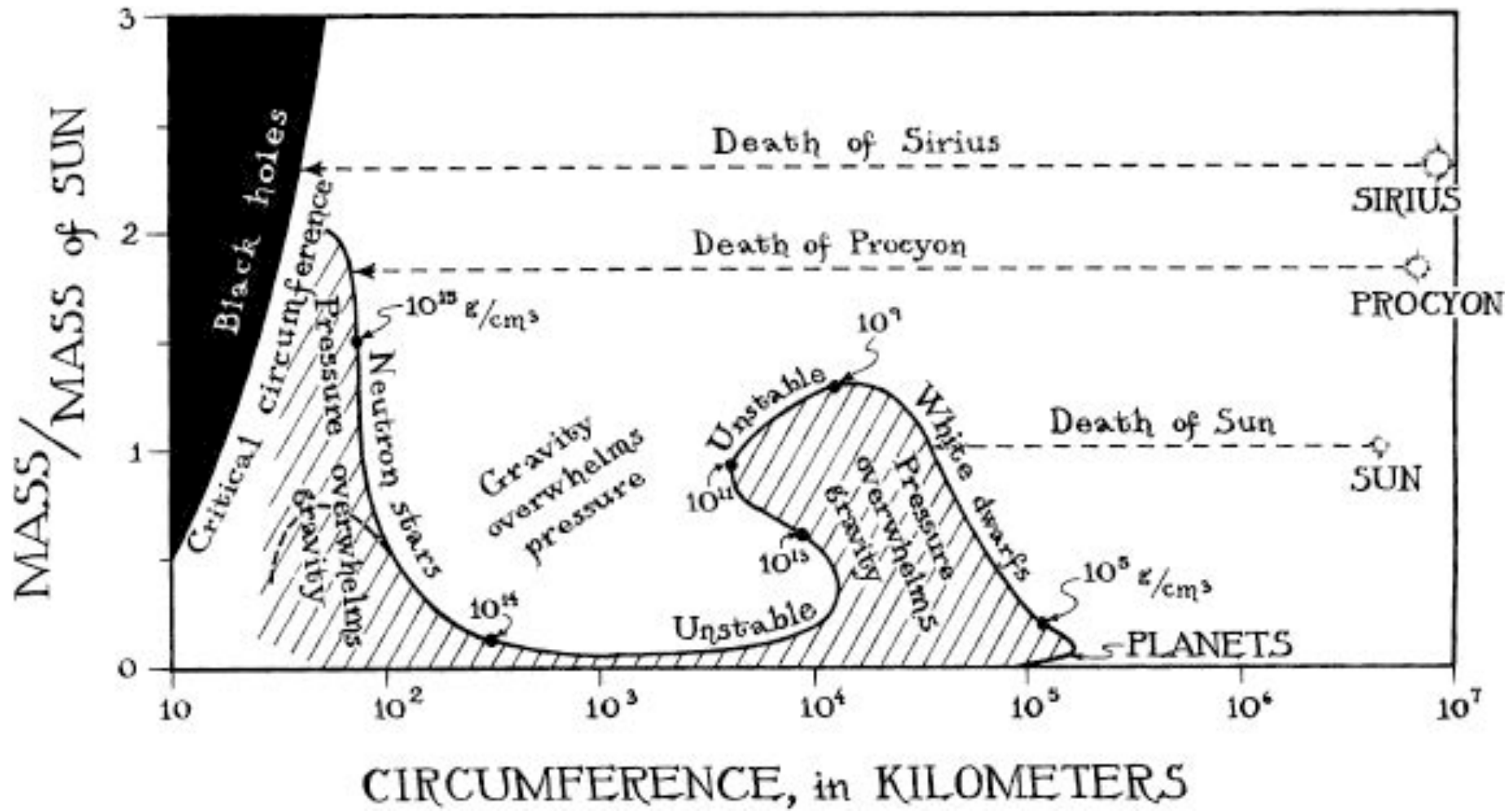
The complete, relativistic equation, contains corrections that involve the mass-energy density (the energy density and pressure are connected by the Equation Of State, EOS).

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

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

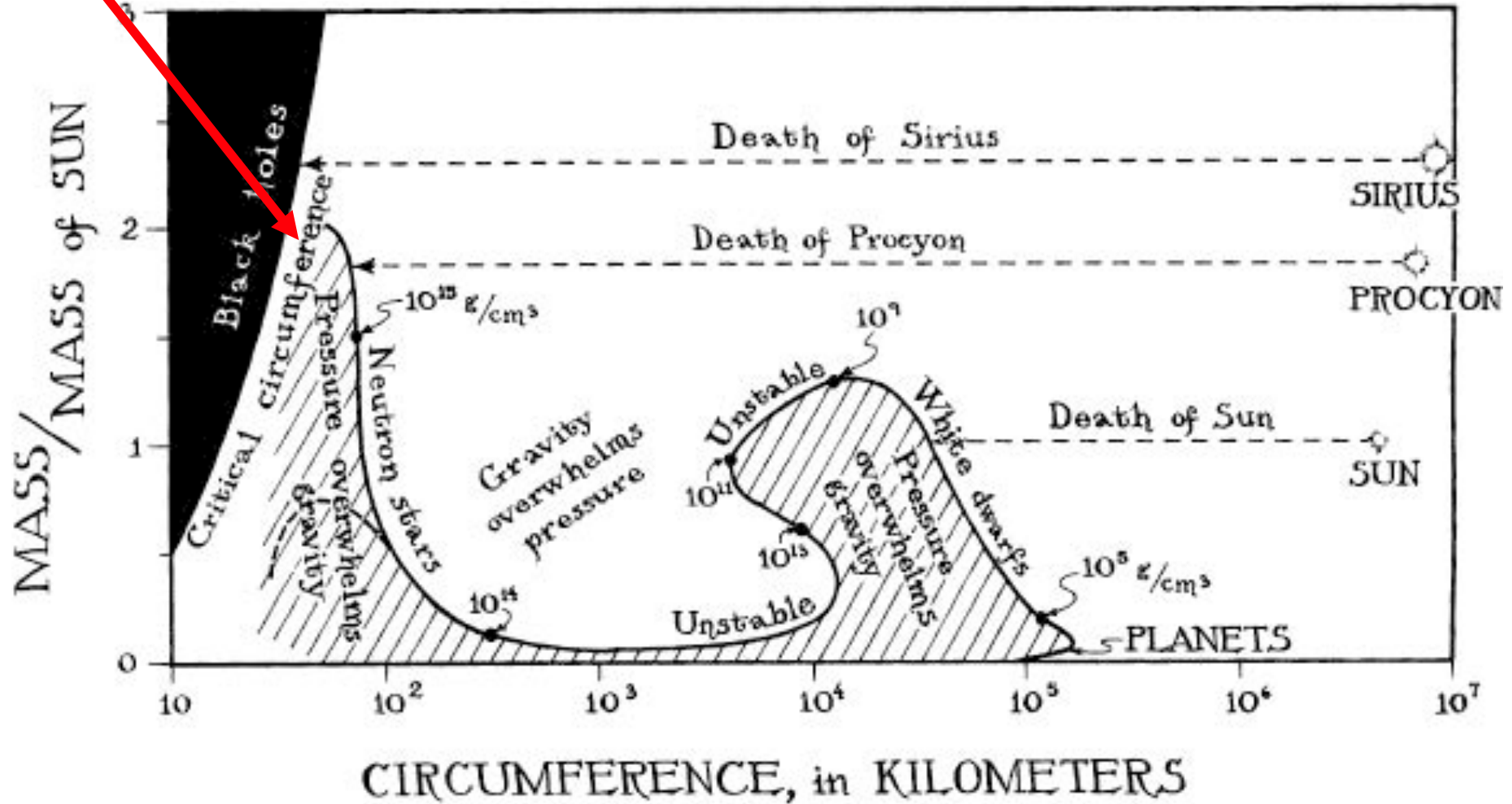


from K. S. Thorne: "Black Holes and Time Warps", Norton (1994)



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)

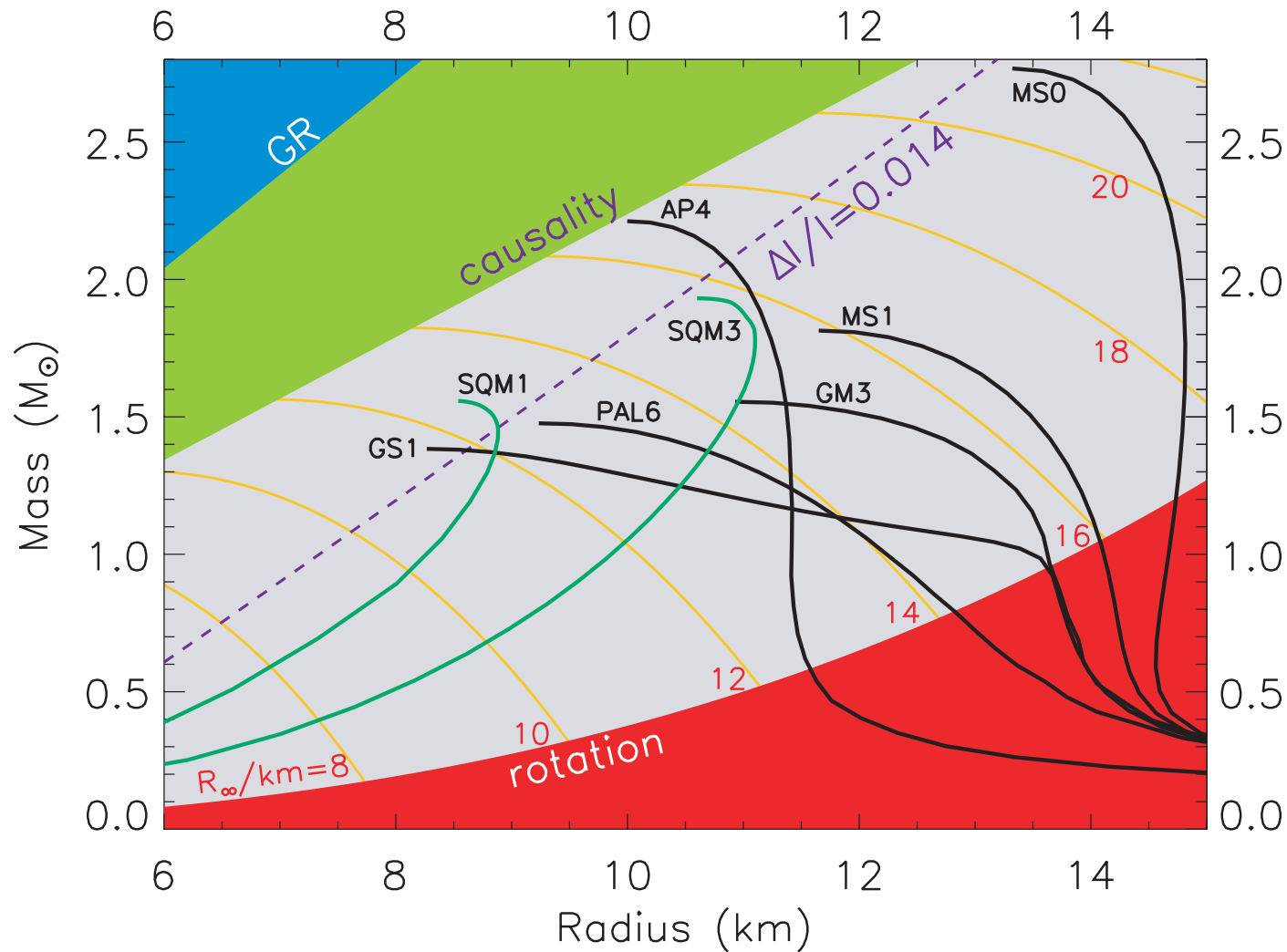
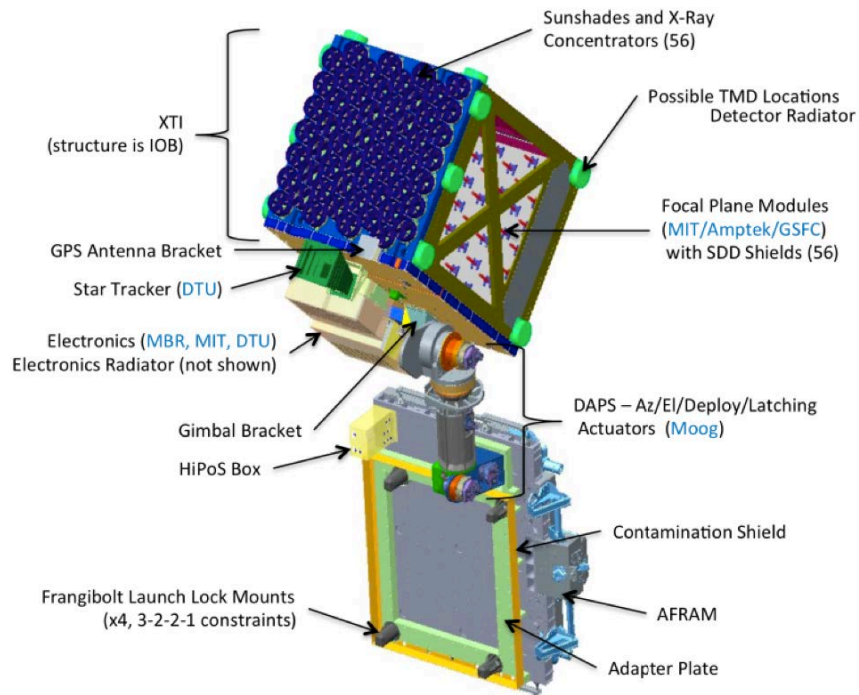


Fig. 2. Mass-radius diagram for neutron stars. Black (green) curves are for normal matter (SQM) equations of state [for definitions of the labels, see (27)]. Regions excluded by general relativity (GR), causality, and rotation constraints are indicated. Contours of radiation radii R_{∞} are given by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

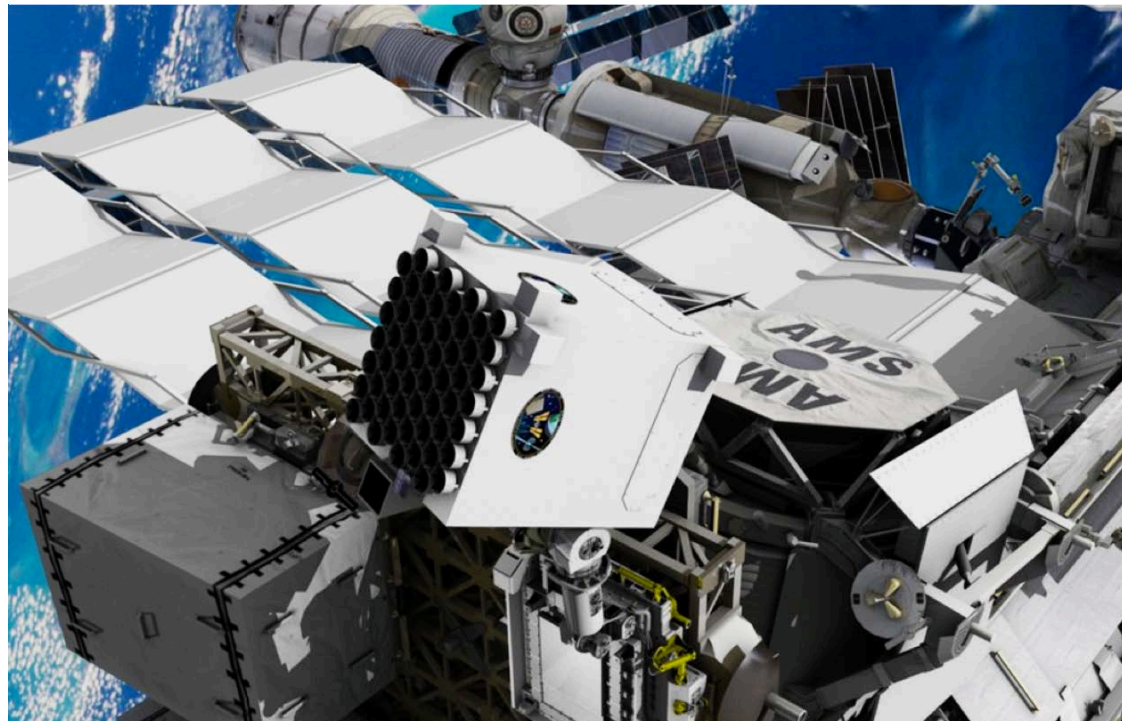
The mass-radius relation is important to pin down the EOS, and hence the composition of the neutron star nucleus.

The NICER (Neutron star Interior Composition ExploreR) mission is expected to provide crucial information in the very near future (resolution obtained on radius ~ 0.5 km)



The NICER instrument onboard the ISS

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. $M \approx 1.3 - 1.4 M_{\odot}$; 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

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.

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	σ_M/M_\odot	m_c/M_\odot	σ_{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	σ_f/M_\odot	M/M_\odot	σ_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

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

$$f, M, m_c \rightarrow \hat{f}, \hat{M}, \hat{m}_c$$

total mass

mass of companion

measured values

Gaussian likelihoods:

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

anyone of the variables listed above

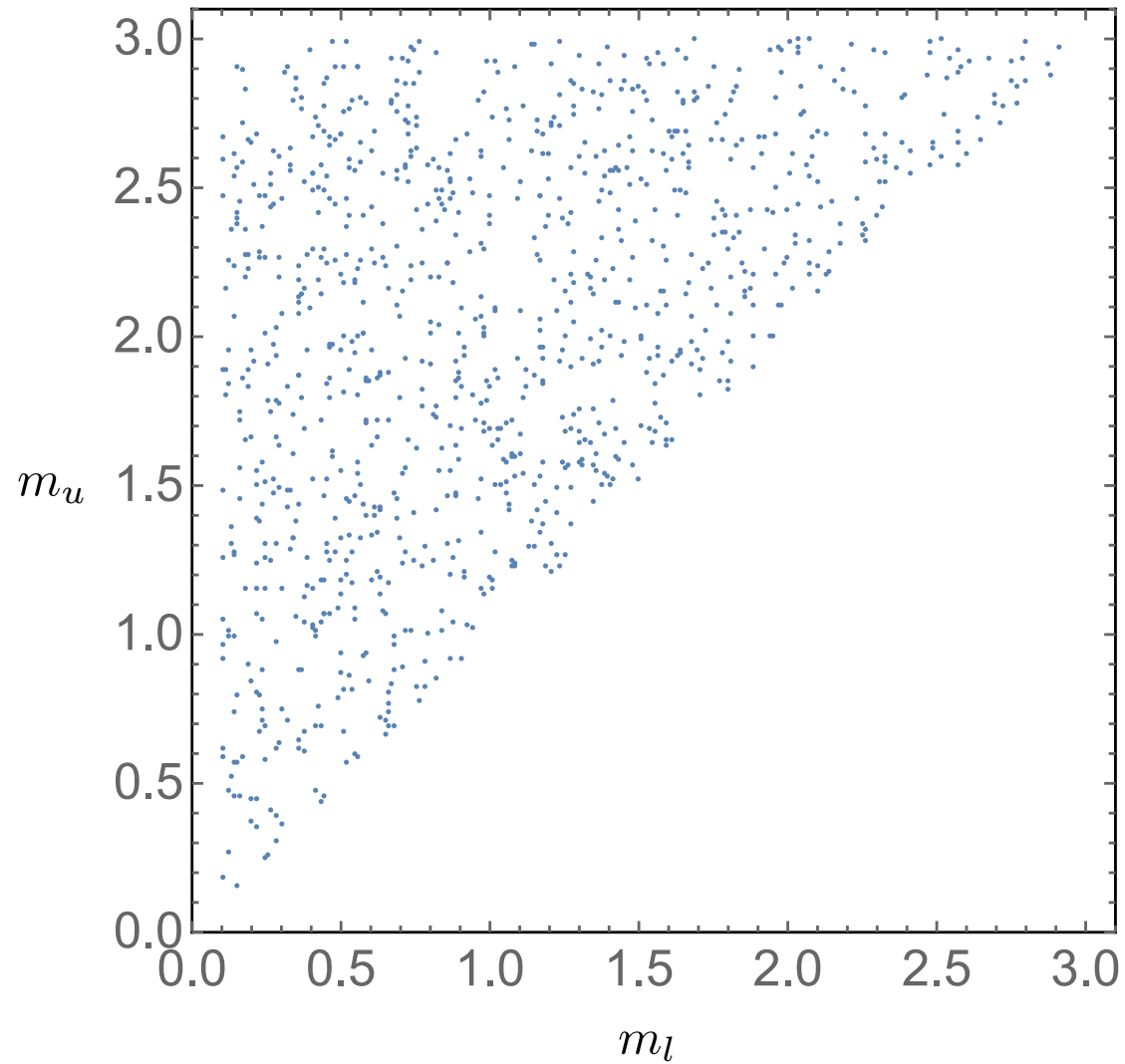
Our target: determine the upper and lower bound for the NS mass distribution.

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

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

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

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

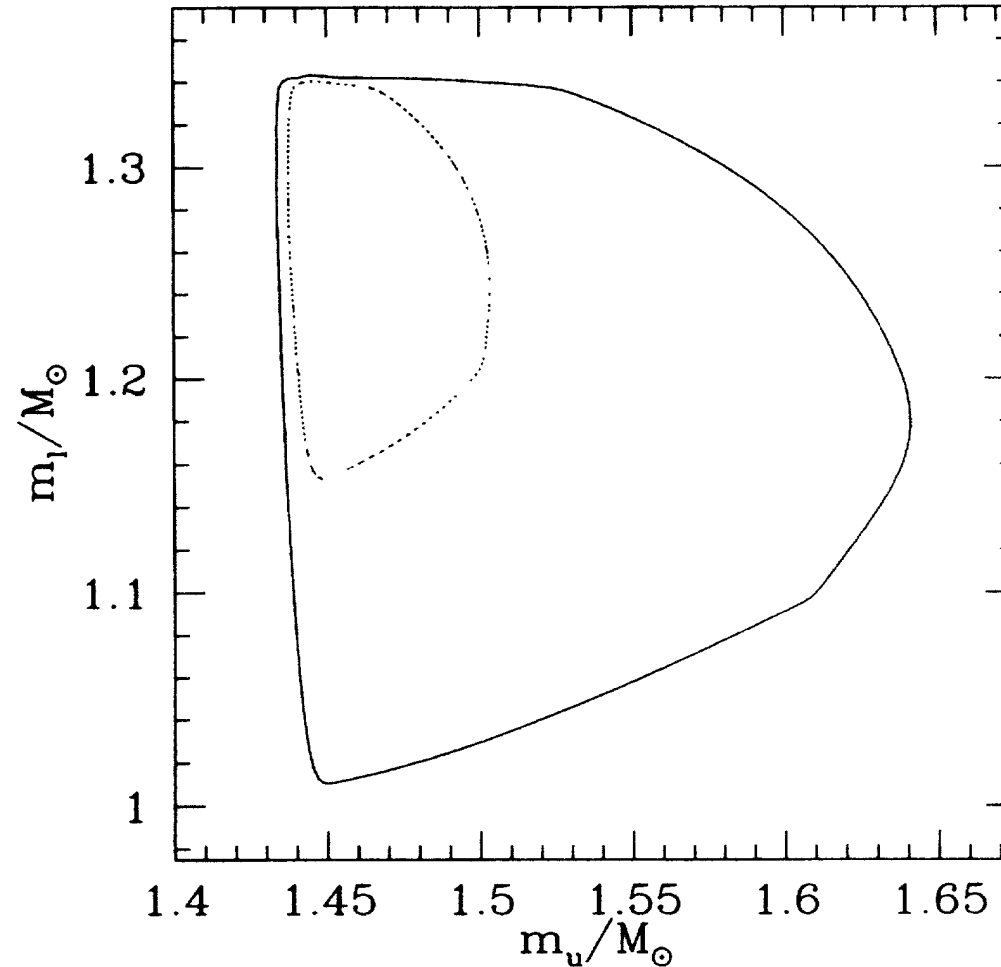


FIG. 1. Assuming ns masses are uniformly distributed between m_l and m_u , observations of PSRs 1534+12, 1913+16, 2127+11C, and 2303+46 determine the joint probability distribution for m_l and m_u . Shown here are contours enclosing regions of 68% (dotted) and 95% (solid) of this distribution.

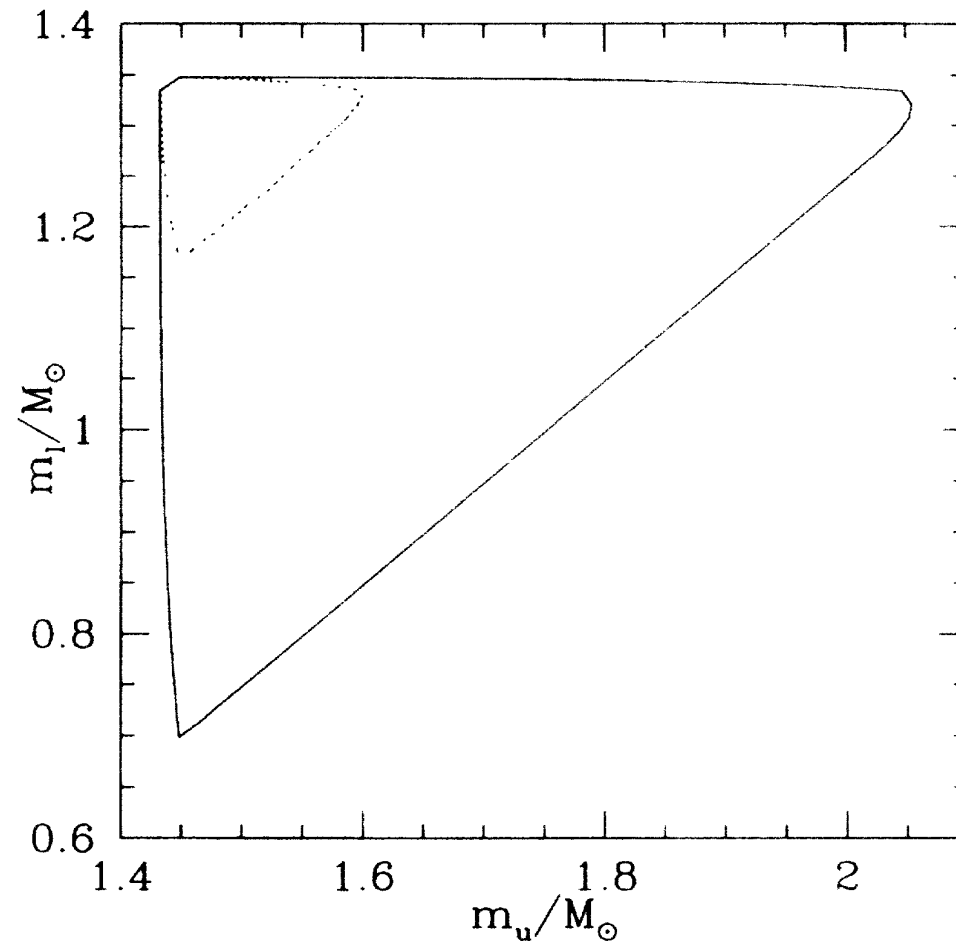


FIG. 2. As in Fig. 1, except that the contours are based on the constraints provided by observations of PSRs 1534+12 and 1913+16.

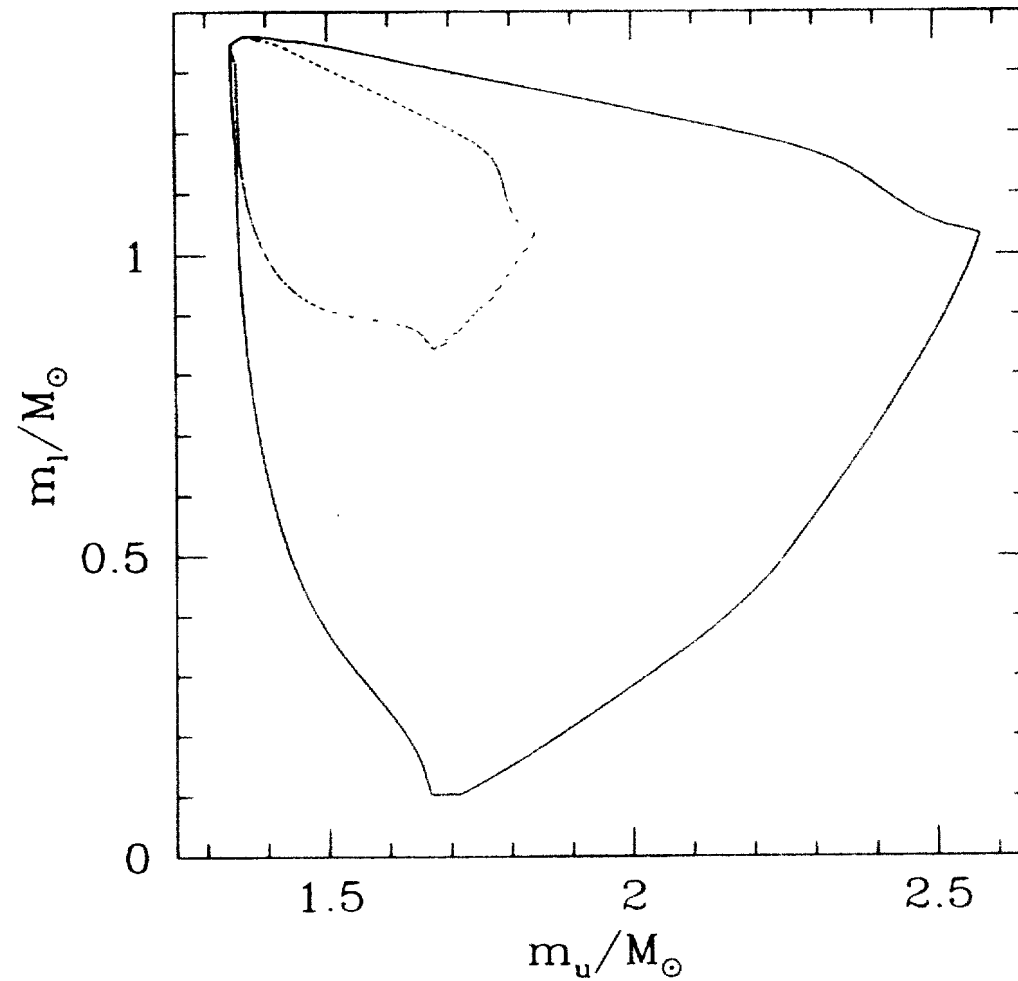


FIG. 3. As in Fig. 1, except that the contours are based on the constraints provided by observations of PSRs 2127+11C and 2303+46.

Additional references:

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.