Introduction to Bayesian Statistics - 5

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Two important computational techniques with a Bayesian basis

1. The EM algorithm
2. Image processing techniques (MLM, MEM, etc.)

1. 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

evidence

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

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

$$\mathcal{L}\left(\mathbf{d},\boldsymbol{\theta}\right)=\prod_{i}p\left(d_{i}|\boldsymbol{\theta}\right)$$

and we estimate the parameters by maximizing the likelihood function

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

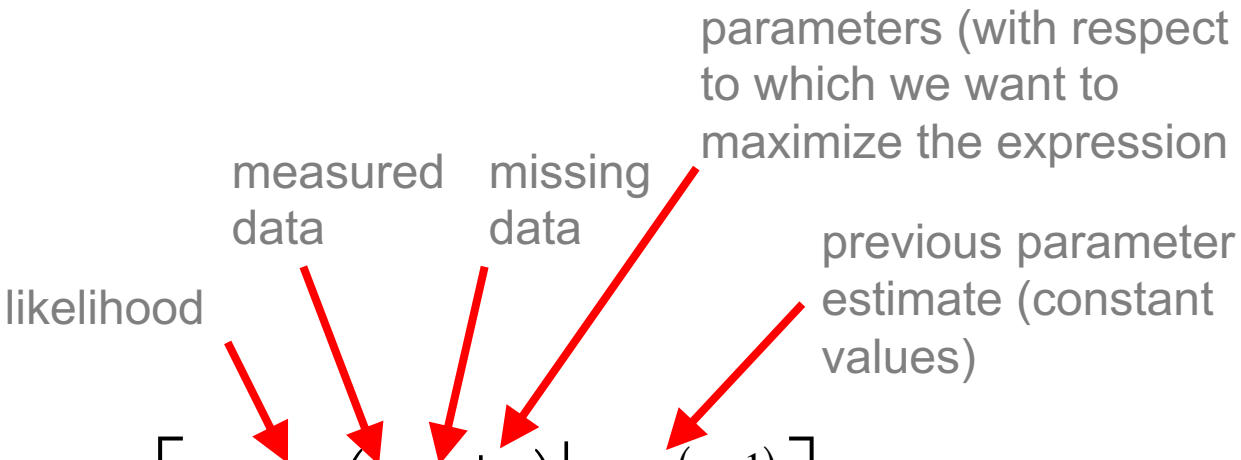
or, equivalently, its logarithm

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

(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:



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

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

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

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

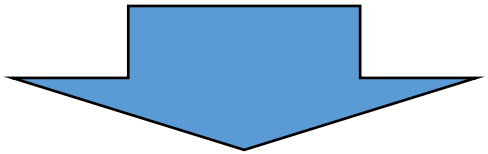
$$p(t') \propto \frac{1}{\theta} e^{-t'/\theta} \qquad 0 \le t' \le t$$

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

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

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

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



average log-likelihood

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

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

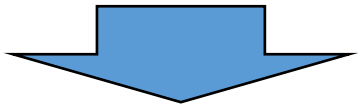
this ends the expectation step

the max of the mean likelihood

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

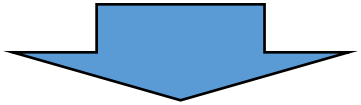
can be found by maximizing the approximate expression

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



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

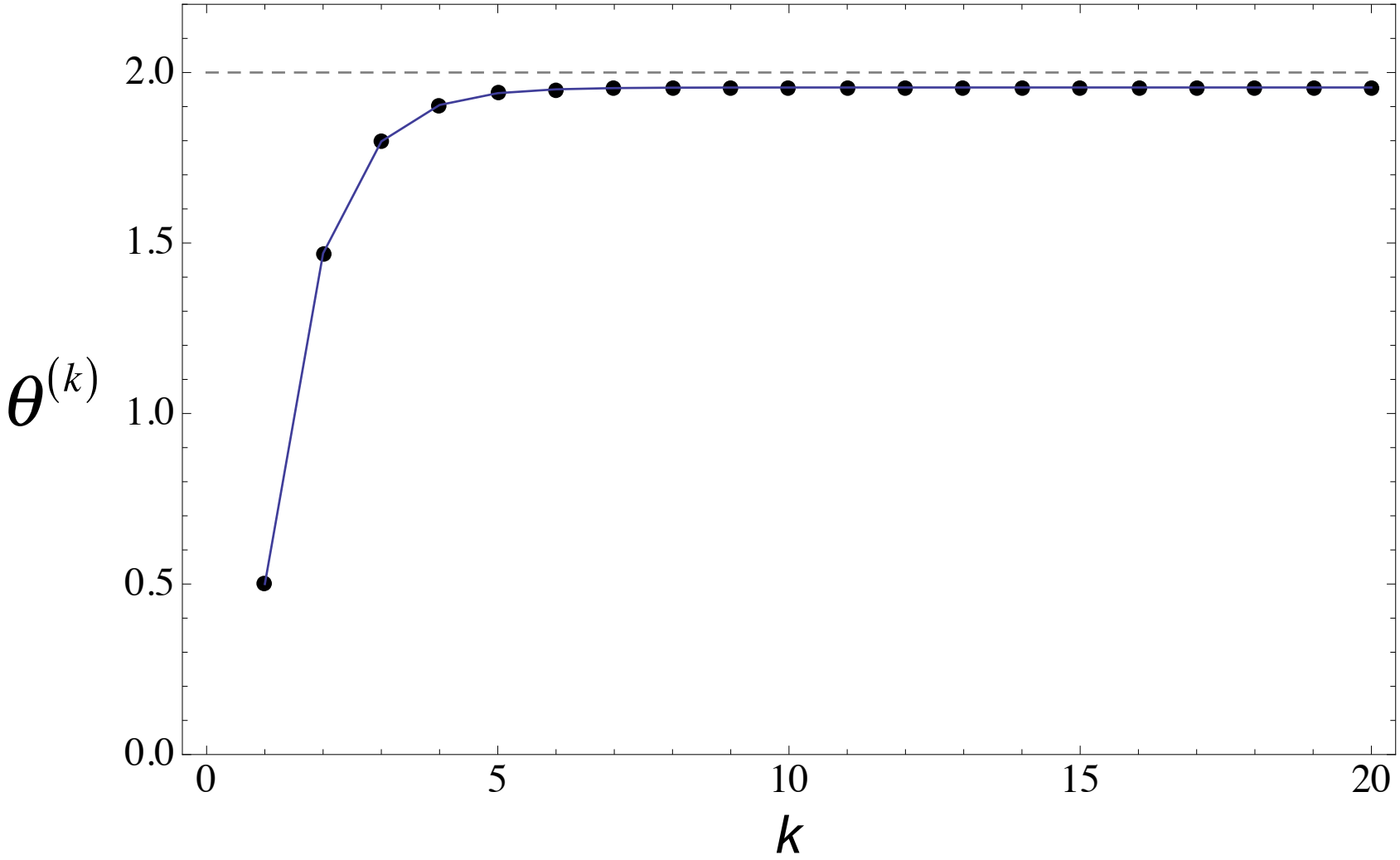
$$\frac{dQ}{d\theta} \approx -(n+m)\frac{1}{\theta} + \frac{1}{\theta^2}\left[n\langle u \rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)} \right) + (m-r)\left(\theta^{(k)} + t\right) \right] = 0$$



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

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.

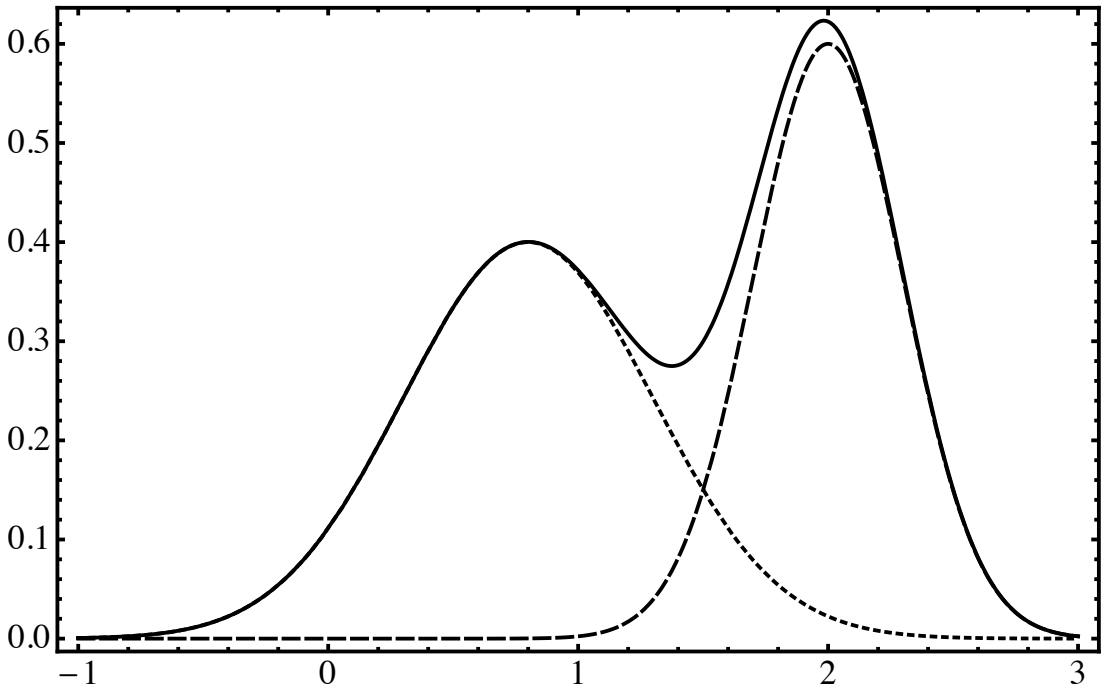


Important application of the EM method: parameters of “mixture models”.

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

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

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



Example: a Gaussian mixture model (M=2)

direct maximization of log likelihood

$$\log \mathcal{L}(\mathbf{x},\boldsymbol{\theta}) = \log \prod_n p(x_n|\boldsymbol{\theta}) = \sum_n \log p(x_n|\boldsymbol{\theta})$$

$$= \sum_n \log \left[\sum_{i=1}^{M} \alpha_i p_i(x_n|\boldsymbol{\theta}_i) \right]$$

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

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

($\boldsymbol{\theta}_i$ 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

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

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

$$\rightarrow \sum_{\mathbf{y}}\left[\log p\left(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}\right)\right] p\left(\mathbf{y}\middle|\mathbf{x},\boldsymbol{\theta}^{(i-1)}\right)$$

sum instead of integral, because the *y* variate is discrete

prior probabilities in the expression of the averaged log-likelihood

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

and now we use Bayes:

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

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

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

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

we find

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

$$= \sum_{\mathbf{y}}\sum_{k=1}^{N} \log\left[\alpha_{y_k} p_{y_k}\left(x_k \middle| \boldsymbol{\theta}_{y_k}\right)\right] \prod_{j=1}^{N} p\left(y_j \middle| x_j,\boldsymbol{\theta}^{(i-1)}\right)$$

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

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

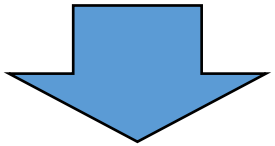
$$=\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{k=1}^{N}\sum_{\ell=1}^{M}\delta_{\ell,y_k}\log\left[\alpha_{\ell}p_{\ell}\left(x_k\middle|\boldsymbol{\theta}_{\ell}\right)\right]\prod_{j=1}^{N}p\left(y_j\middle|x_j,\boldsymbol{\theta}^{(i-1)}\right)$$

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

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

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

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



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

$$=\sum_{\ell=1}^{M}\sum_{k=1}^{N}\log\left[\alpha_\ell p_\ell\left(x_k\middle|\boldsymbol{\theta}_\ell\right)\right]\left\{\sum_{y_1=1}^{M}\cdots\sum_{y_{k-1}=1}^{M}\sum_{y_{k+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{\substack{j=1\\j\neq k}}^{N}p\left(y_j\middle|x_j,\boldsymbol{\theta}^{(i-1)}\right)\right\}p\left(\ell\middle|x_k,\boldsymbol{\theta}^{(i-1)}\right)$$

$$=\sum_{\ell=1}^{M}\sum_{k=1}^{N}\log\left[\alpha_\ell p_\ell\left(x_k\middle|\boldsymbol{\theta}_\ell\right)\right]\left\{\prod_{\substack{j=1\\j\neq k}}^{N}\sum_{y_j=1}^{M}p\left(y_j\middle|x_j,\boldsymbol{\theta}^{(i-1)}\right)\right\}p\left(\ell\middle|x_k,\boldsymbol{\theta}^{(i-1)}\right)$$

$$=\sum_{\ell=1}^{M}\sum_{k=1}^{N}\log\left[\alpha_\ell p_\ell\left(x_k\middle|\boldsymbol{\theta}_\ell\right)\right]p\left(\ell\middle|x_k,\boldsymbol{\theta}^{(i-1)}\right)$$

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

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

$$= \sum_{\ell=1}^{M}\sum_{k=1}^{N} \ln \alpha_\ell \; p_\ell(x_k, \boldsymbol{\theta}^{(i-1)}) + \sum_{\ell=1}^{M}\sum_{k=1}^{N} \ln p(\ell|x_k, \boldsymbol{\theta}) \; p_\ell(x_k, \boldsymbol{\theta}^{(i-1)})$$

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.

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

$$= \sum_{k=1}^{N} \left[\frac{1}{2c_m} - \frac{1}{2}(x_k - \mu_m)^2 \right]\; p(m | x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)})$$

$$= \sum_{k=1}^{N} \left[\frac{\sigma_m^2}{2} - \frac{1}{2}(x_k - \mu_m)^2 \right]\; p(m | x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = 0$$

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

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

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

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

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

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

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

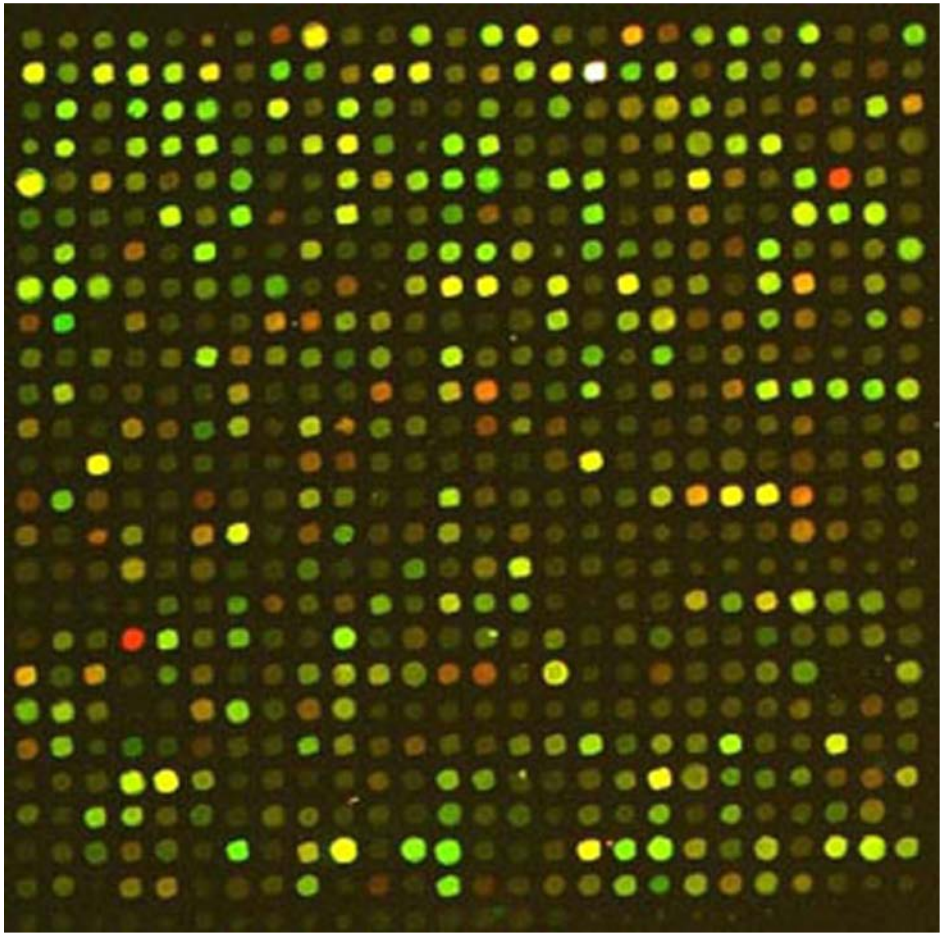
$$\left(\sigma_m^{(i)}\right)^2=\frac{\sum_{k=1}^{N}\left(x_k-\mu_m^{(i)}\right)^2 p\left(m\middle|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)}{\sum_{k=1}^{N}p\left(m\middle|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)}$$

We remark that the probabilities

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

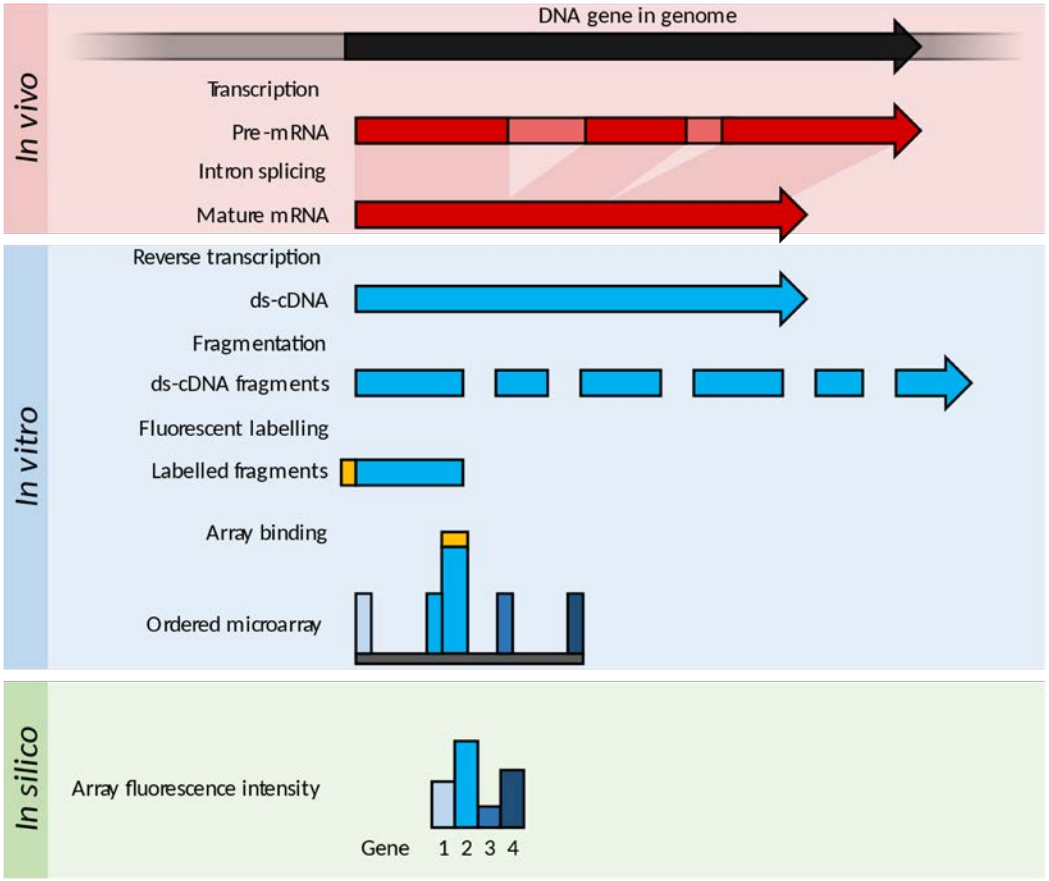
are an estimate of the frequencies of the y_n using the observed data x_n, and this amounts to a classification (selection of one of the component distributions).

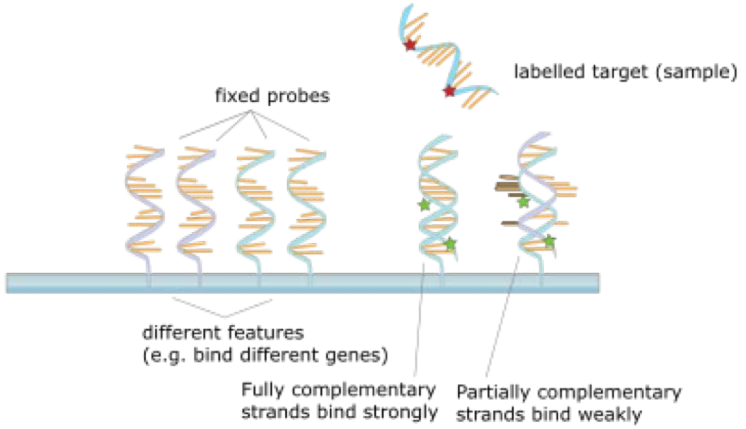
Example: classification of response of DNA microarrays.



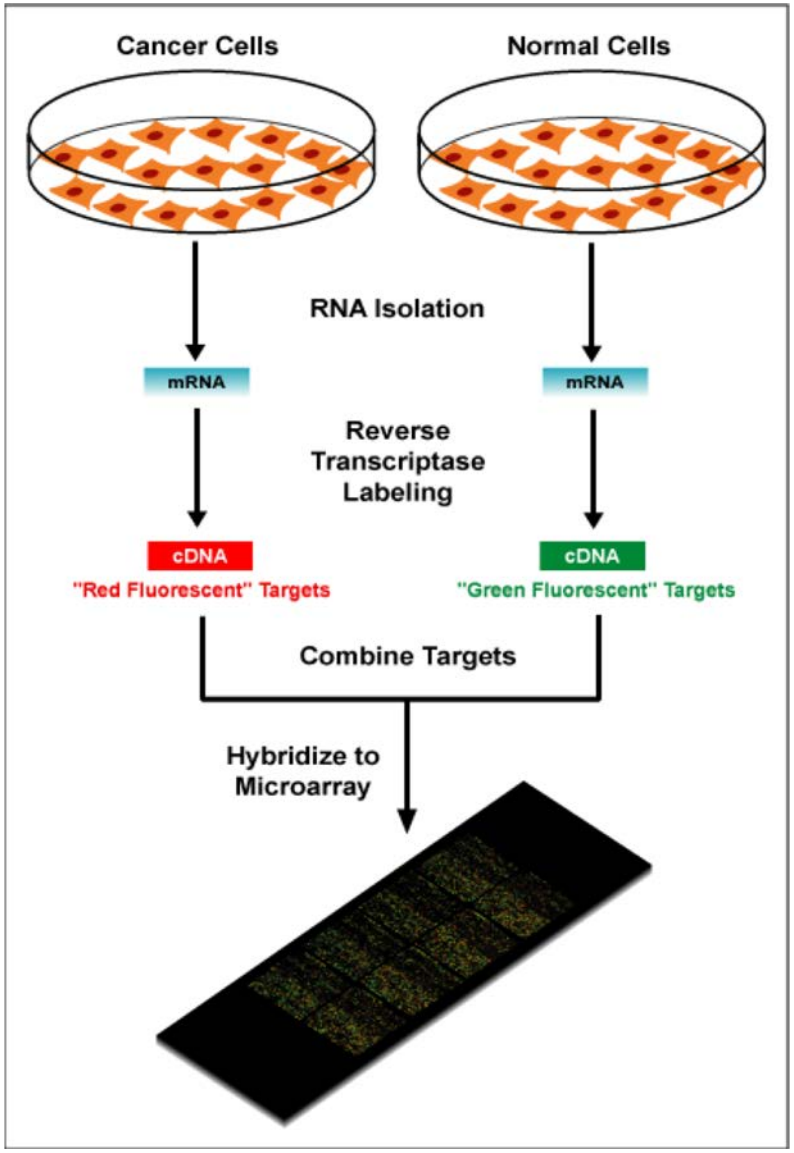
Microarray image
from:http://www.wormbook.org/chapters/www_germlinegenomics/germlinegenomics.html

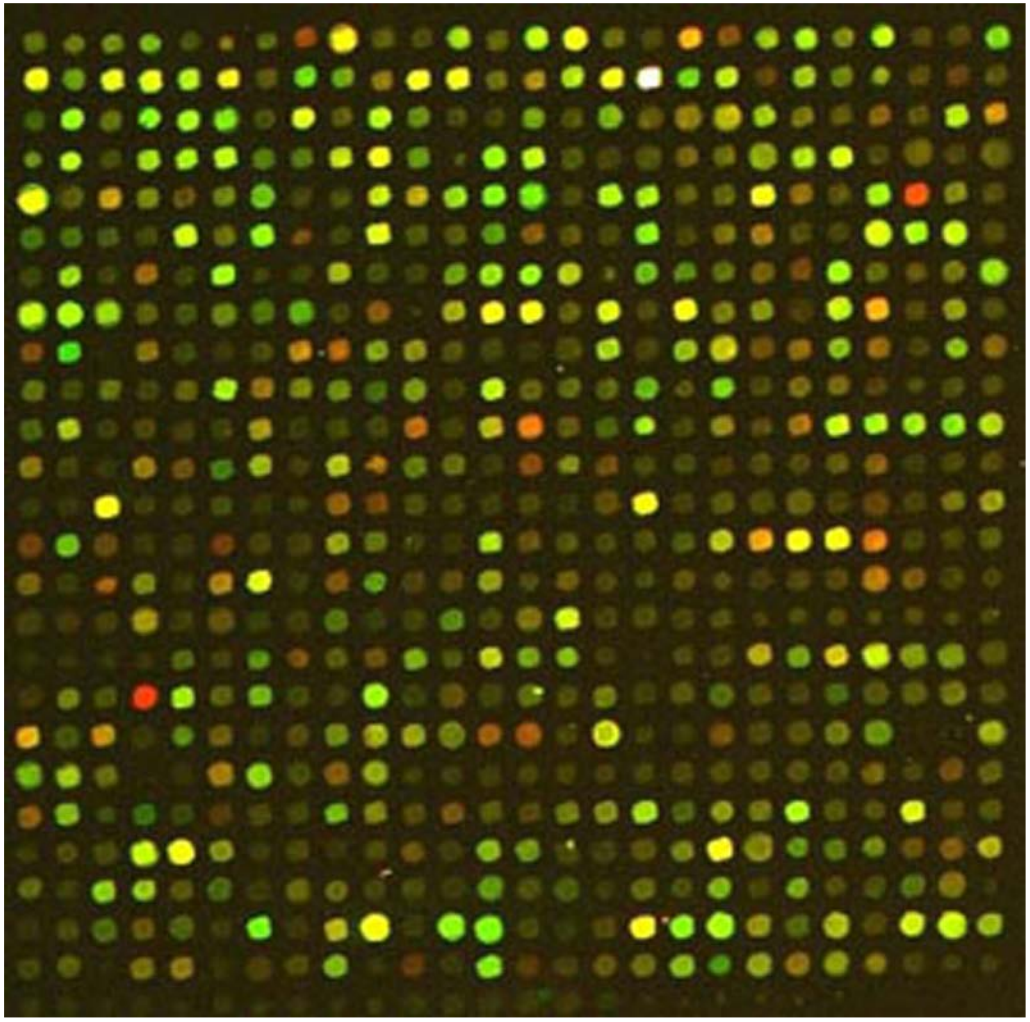
Microarray: lab on a chip





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)

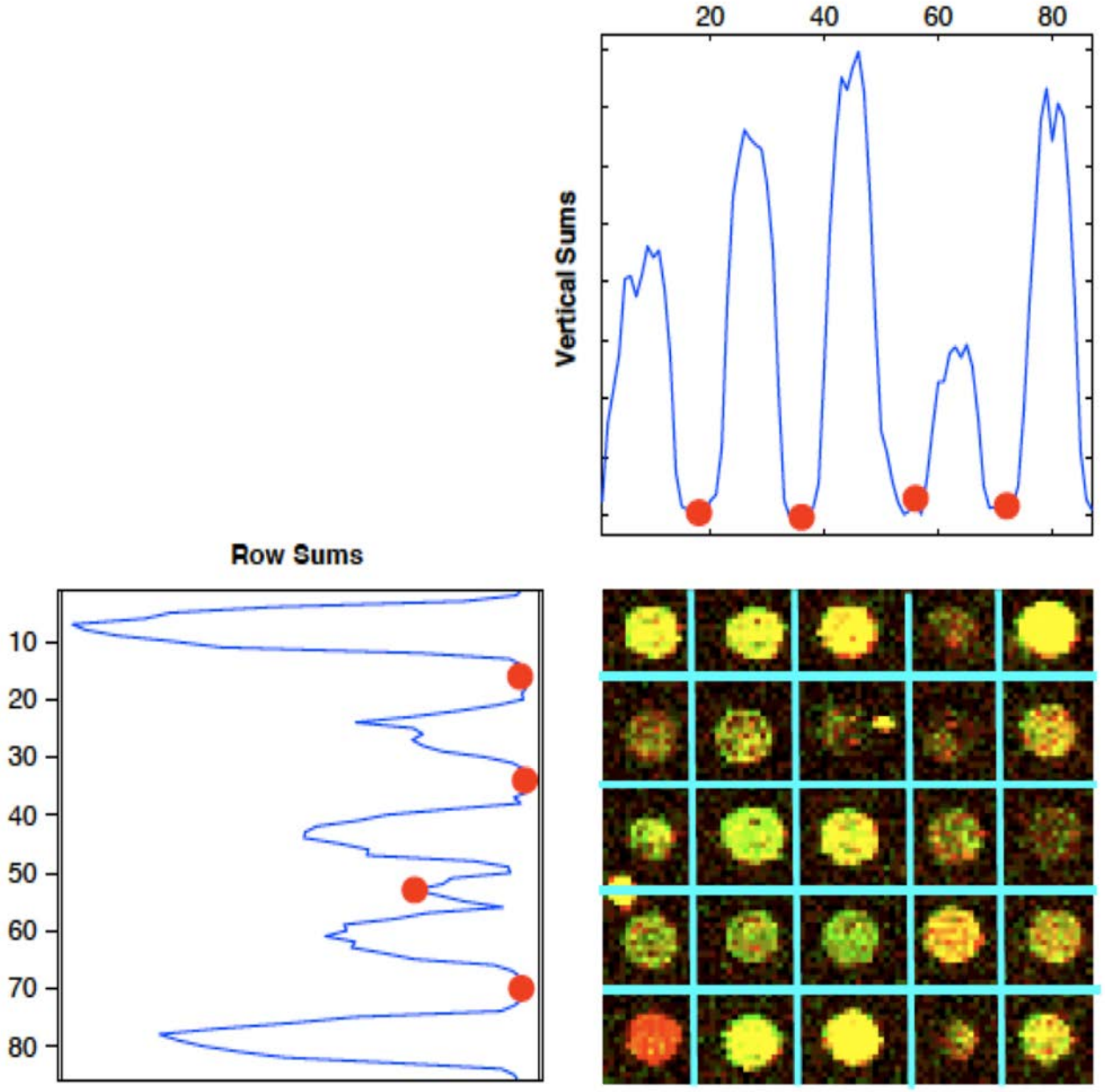




Microarray image from:http://www.wormbook.org/chapters/www_germlinegenomics/germlinegenomics.html

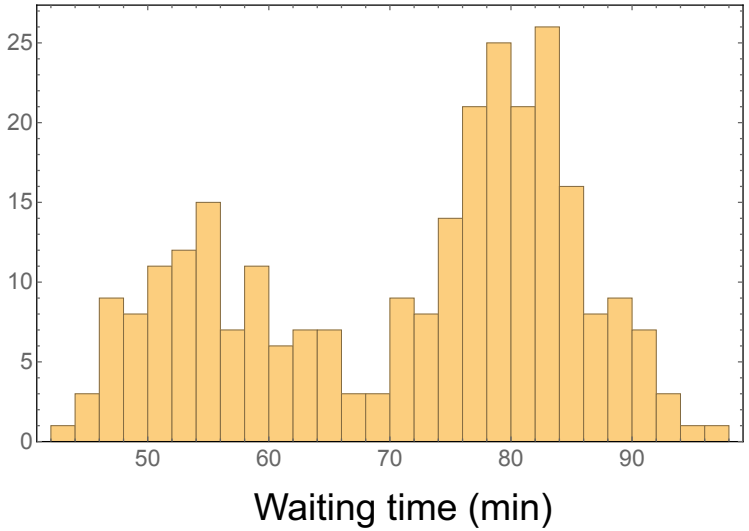
Further informations on DNA microarrays:http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html

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

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

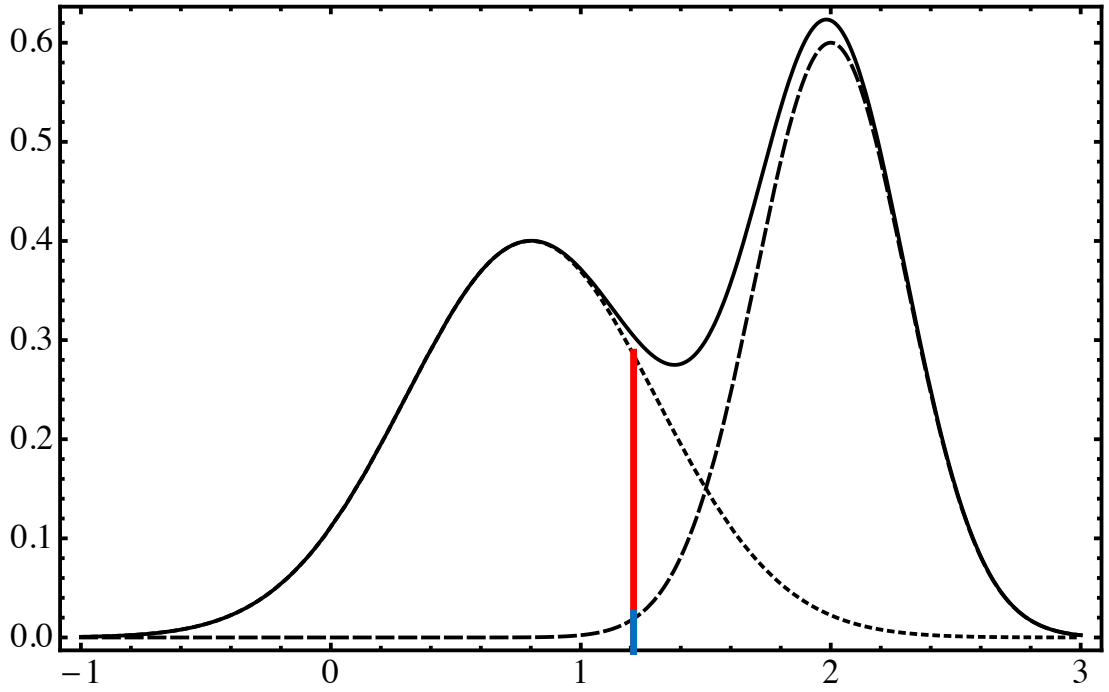
Again, we substitute the likelihood with the new one

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

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

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



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 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:

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

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

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

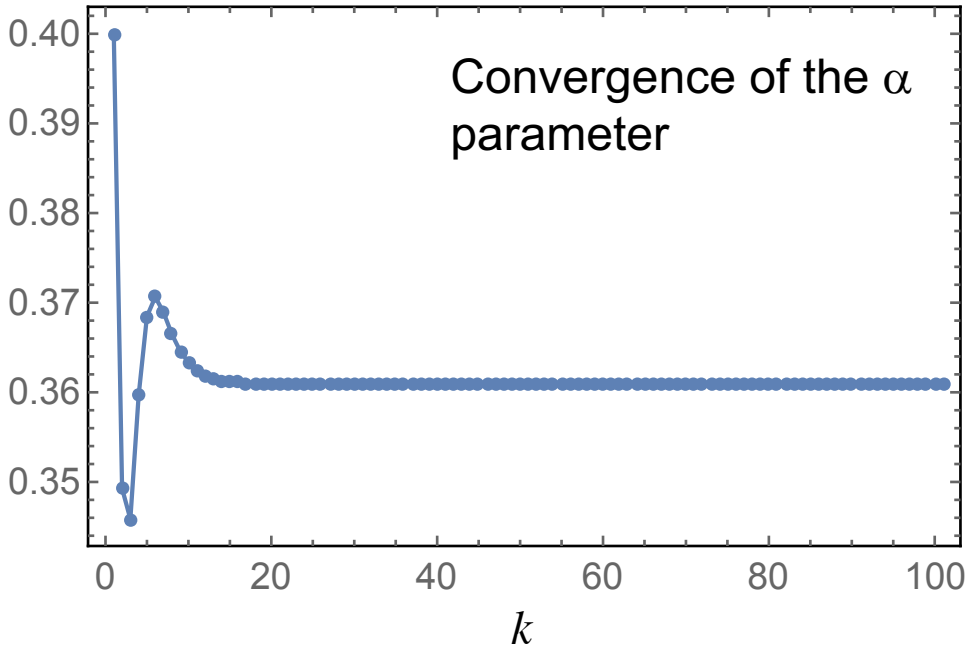
Finally we have the following set of equations:

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

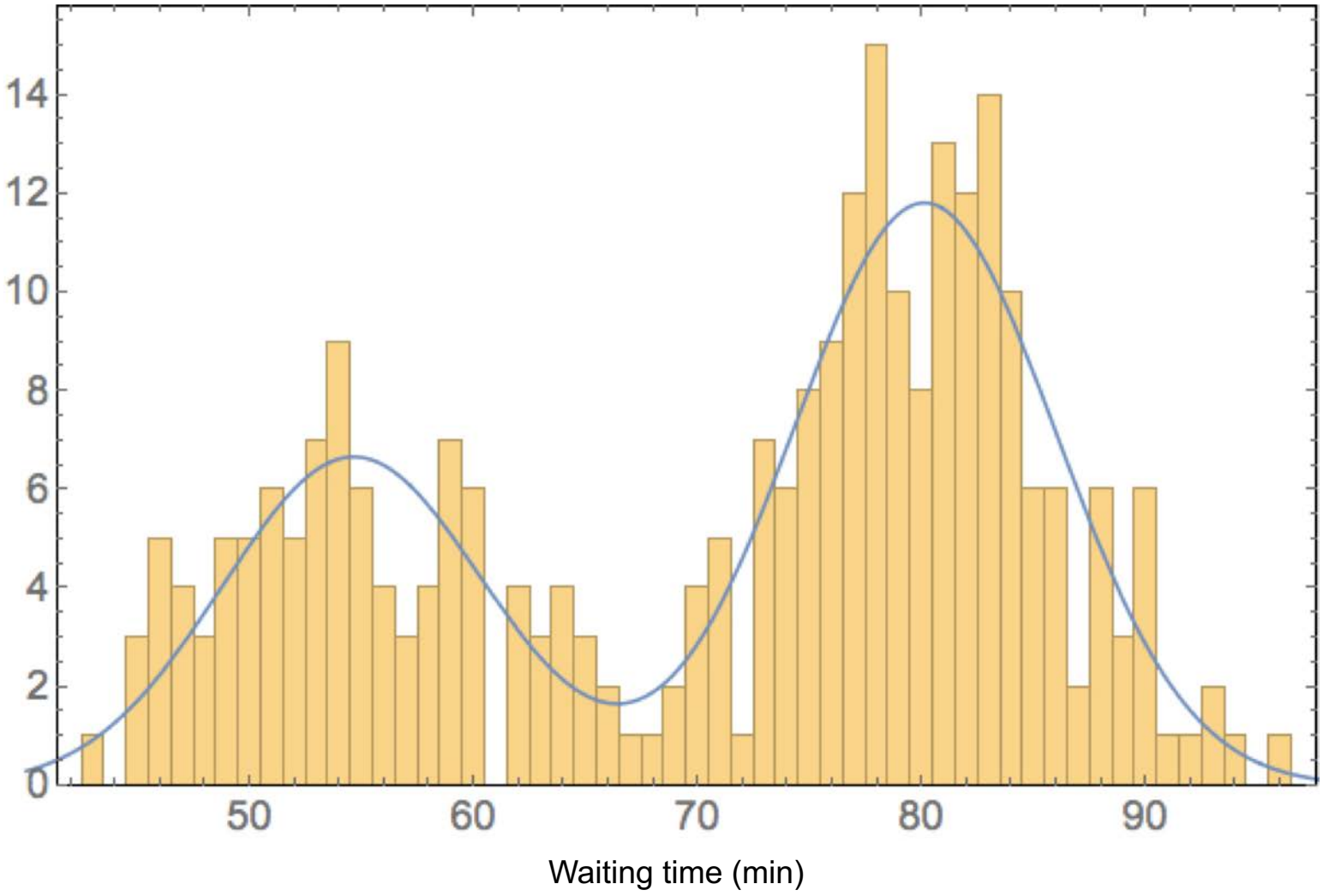
$$\alpha^{(k+1)} = \frac{1}{N}\sum_i p_i^{(k)}$$

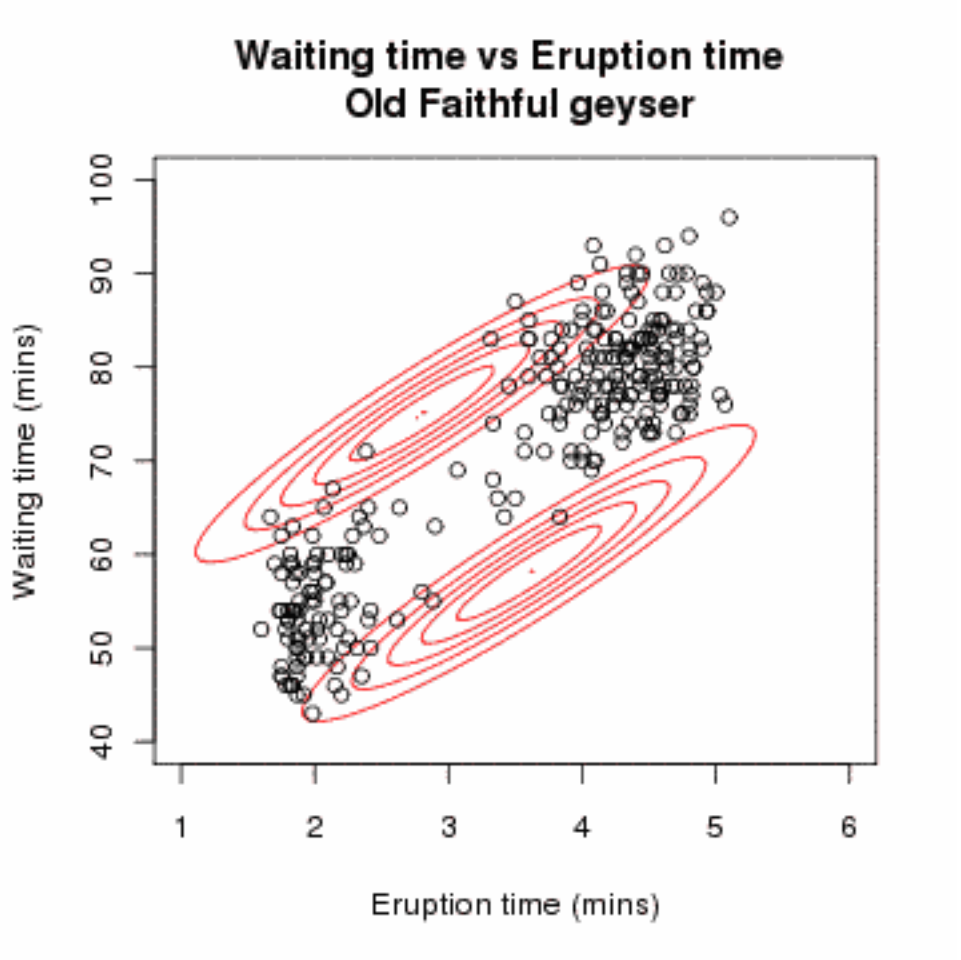
$$\left(\sigma_1^{(k+1)}\right)^2 = \frac{\sum_i p_i^{(k)}(w_i - \mu_1^{(k)})^2}{\sum_i p_i^{(k)}}; \qquad \mu_1^{(k+1)} = \frac{\sum_i p_i^{(k)} w_i}{\sum_i p_i^{(k)}}$$

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


Convergence of the α parameter
0.40
0.39
0.38
0.37
0.36
0.35
0
20
40
60
80
100
k

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




Waiting time vs Eruption time
Old Faithful geyser
Waiting time (mins)
40
50
60
70
80
90
100
1
2
3
4
5
6
Eruption time (mins)

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Help With R

Getting Help

Documentation

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 3.4.0 (You Stupid Darkness)** has been released on Friday 2017-04-21.
- **R version 3.3.3 (Another Canoe)** has been released on Monday 2017-03-06.
- **useR! 2017** (July 4 - 7 in Brussels) has opened registration and more at http://user2017.brussels/
- Tomas Kalibera has joined the R core team.
- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kalibera, and Balasubramanian Narasimhan.
- **The R Journal Volume 8/1** is available.
- The **useR! 2017** conference will take place in Brussels, July 4 - 7, 2017.
- **R version 3.2.5 (Very, Very Secure Dishes)** has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.
- **Notice XQuartz users (Mac OS X)** A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating over insecure channels.

faithful {datasets} R Documentation

Old Faithful Geyser Data

Description

Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

Usage

```
faithful
```

Format

A data frame with 272 observations on 2 variables.

[,1] eruptions numeric Eruption time in mins
[,2] waiting numeric Waiting time to next eruption (in mins)

Details

A closer look at `faithful$eruptions` reveals that these are heavily rounded times originally in seconds, where multiples of 5 are more frequent than expected under non-human measurement. For a better version of the eruption times, see the example below.

There are many versions of this dataset around: Azzalini and Bowman (1990) use a more complete version.

Source

W. Härdle.

References

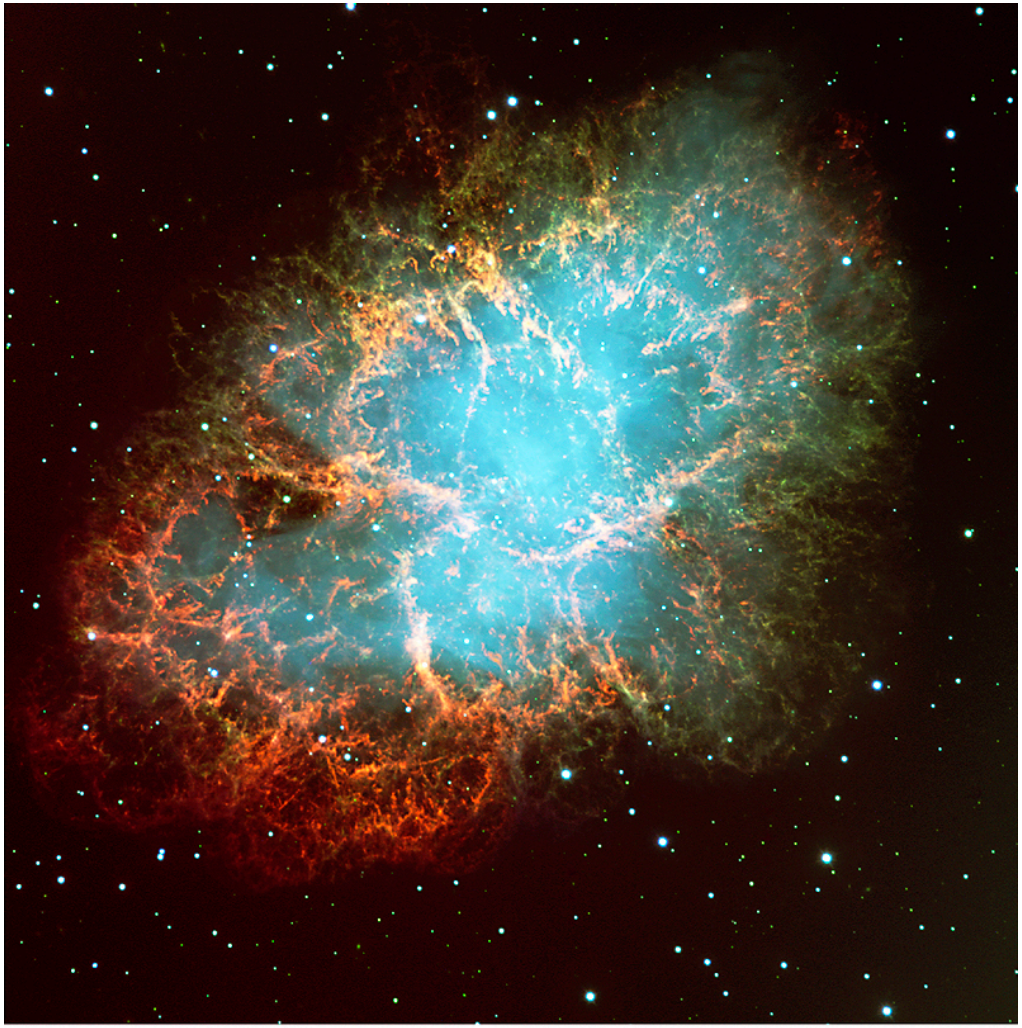
Härdle, W. (1991) *Smoothing Techniques with Implementation in S*. New York: Springer.

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics* **39**, 357–365.

See Also

`geyser` in package **MASS** for the Azzalini–Bowman version.

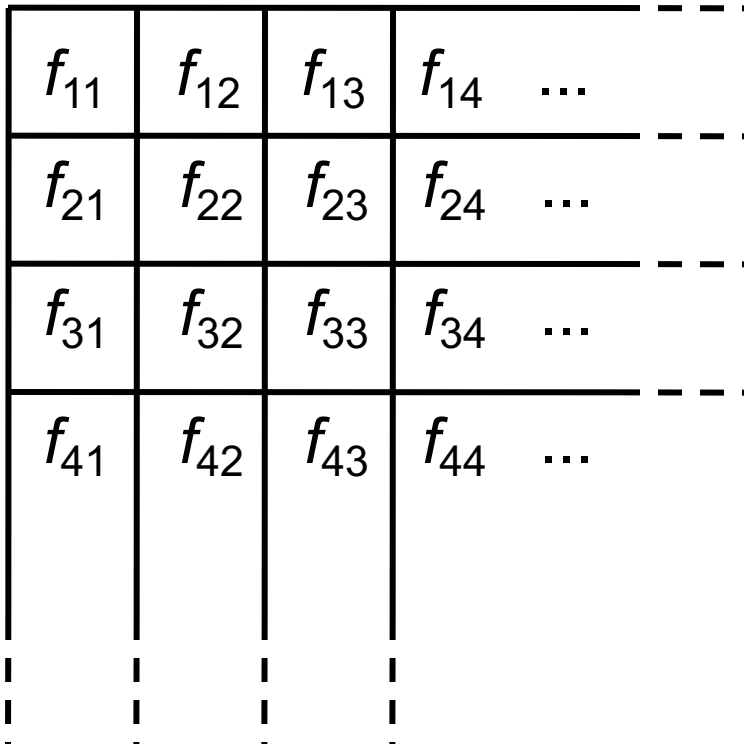
2. Image processing techniques (MLM, MEM)



The Crab Nebula in Taurus (VLT KUEYEN + FORS2)

ESO PR Photo 40f/99 (17 November 1999)

© European Southern Observatory



pixel map

true pixel vector **f**

posterior pixel distribution

likelihood

a priori pixel distribution

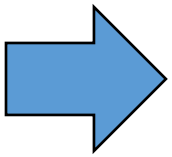
$$P(\mathbf{f}|\mathbf{g}) = \frac{P(\mathbf{g}|\mathbf{f})}{P(\mathbf{g})} P(\mathbf{f}) \propto P(\mathbf{g}|\mathbf{f}) P(\mathbf{f})$$

Bayesian estimate of true pixel vector from observed pixel vector

We estimate the true pixel distribution taking the pixel vector that maximizes the posterior distribution (MAP estimate: Maximum A Posteriori estimate).

This depends on the prior distribution

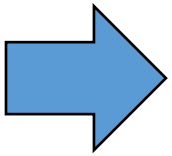
$P(\mathbf{f})$ flat prior



Maximum Likelihood Estimate (MLE)

$$P(\mathbf{f}|\mathbf{g}) \propto P(\mathbf{g}|\mathbf{f})P(\mathbf{f}) \propto P(\mathbf{g}|\mathbf{f})$$

$P(\mathbf{f})$ entropic prior



Maximum Entropy Method (MEM)

Notice that

$$\log P(\mathbf{f}|\mathbf{g}) \approx \log P(\mathbf{g}|\mathbf{f}) - \left[-\log P(\mathbf{f})\right]$$

therefore we obtain the estimate $\hat{\mathbf{f}}$ by maximizing the likelihood with the *penalty function*

$$\left[-\log P(\mathbf{f})\right]$$

Experiments have been tried with many different penalties, many of them barely justified on probabilistic grounds (or not at all!)

Let ***f*** be the vector of "true values" (uncorrupted intensities of an image, a spectrum, etc. ...), and translate these values into counts

$$n_i = \lfloor \alpha f_i \rfloor$$

(i = 1, ..., M). The least informative prior is that for a structureless image is uniform, and the probability of one count at the i-th position is just 1/M.

Likewise, the probability of a given vector of values where the total number of counts is N, is given by the multinomial probability

$$P(\mathbf{n}) = \frac{N!}{n_1!n_2!\ldots n_M!}\left(\frac{1}{M}\right)^N; \qquad \sum_k n_k = N$$

Using Stirling's approximation

$$n! \approx n^n e^{-n} \qquad \ln n! \approx n \ln n - n$$

we find, with the definition $p_i = f_i \Big/ \sum_{k=1}^{M} f_k$

$$\ln P(\mathbf{n}) \approx \left(N \ln N - N\right) - \sum_{i=1}^{M} \left(n_i \ln n_i - n_i\right)$$

$$= N \ln N - \sum_{i=1}^{M} n_i \ln n_i$$

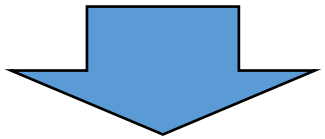
$$\approx -\alpha \sum_{i=1}^{M} f_i \ln f_i + \text{cost}.$$

entropic prior

$$P(\mathbf{n}) \propto \exp\left[-\alpha \sum_{i=1}^{M} f_i \ln f_i\right] \propto \exp\left[-\alpha \sum_{i=1}^{M} p_i \ln p_i\right] = \exp\left[\alpha S(\mathbf{f})\right]$$

Using the entropic prior and Bayes' theorem we find

$$P(\mathbf{f}) \propto \exp\left[\alpha S(\mathbf{f})\right]$$

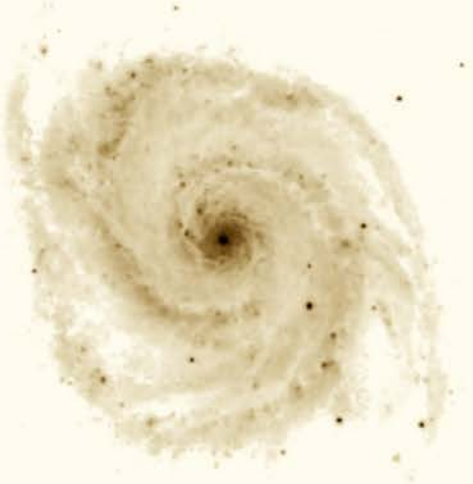


$$P(\mathbf{f}|\mathbf{g}) \propto P(\mathbf{g}|\mathbf{f})P(\mathbf{f}) \propto P(\mathbf{g}|\mathbf{f})\exp\left[\alpha S(\mathbf{f})\right]$$

$$\log P(\mathbf{f}|\mathbf{g}) \approx \log P(\mathbf{g}|\mathbf{f}) + \alpha S(\mathbf{f})$$

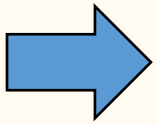
therefore we find the combination of pixels (i.e., the **f** vector) that maximizes the posterior distribution by maximizing a linear combination of likelihood and Shannon's entropy.

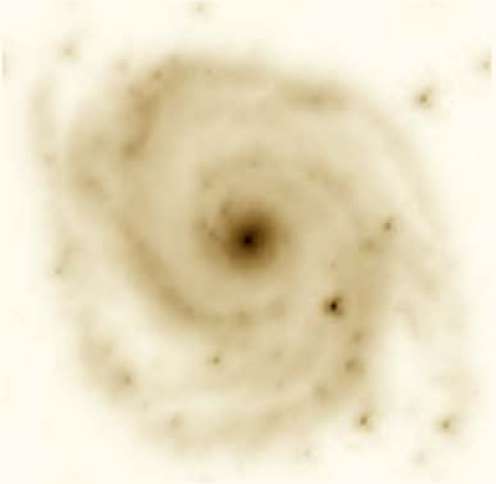
Image likelihood: 1. the observation model



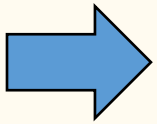
true image
of a galaxy

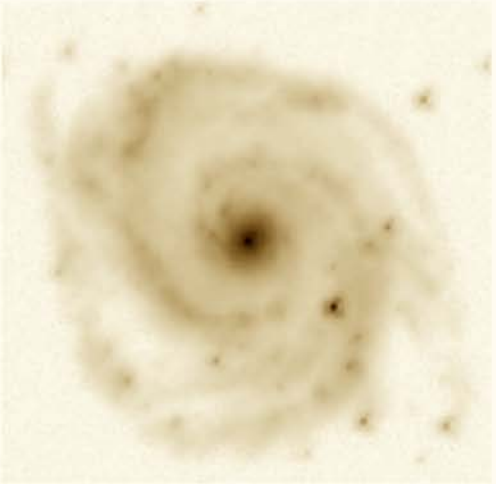
PSF
(Point Spread
Function





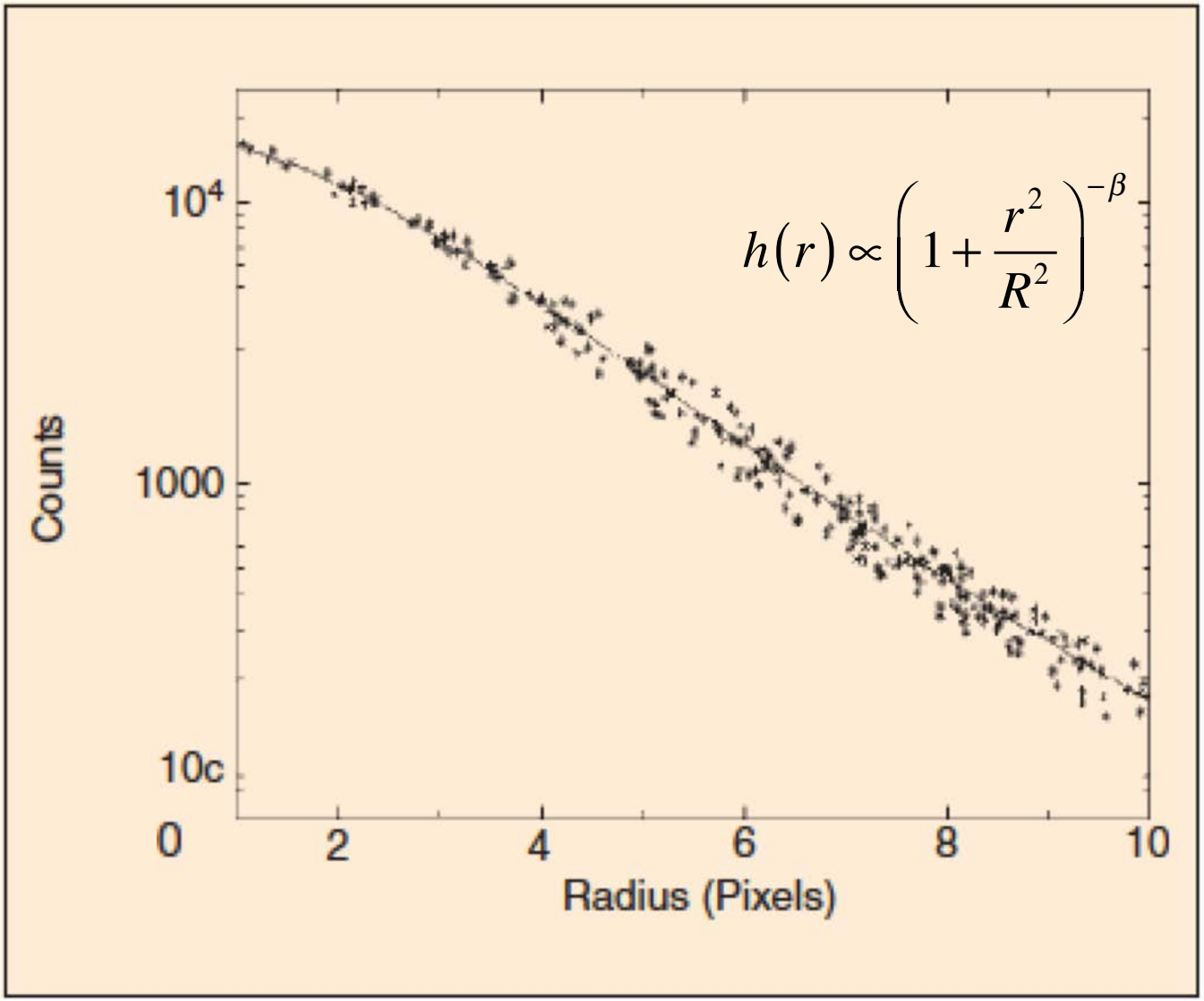
Noise





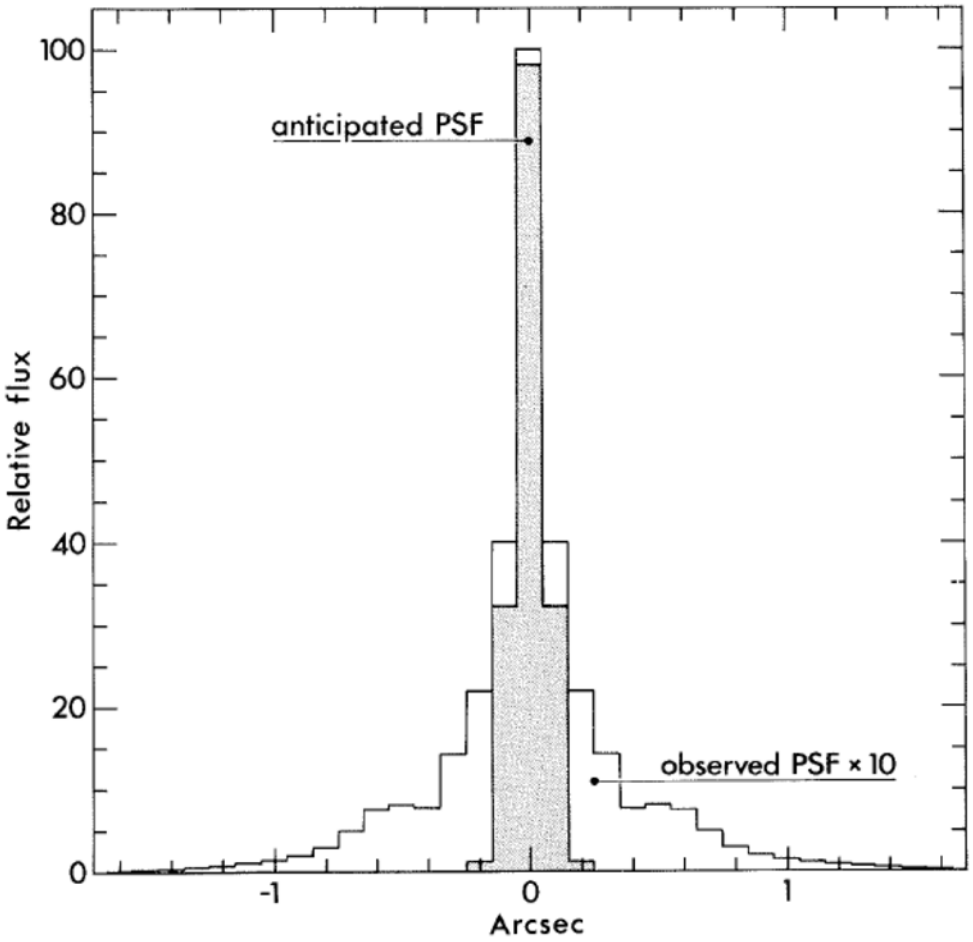
(example from Eric Thiebaut)

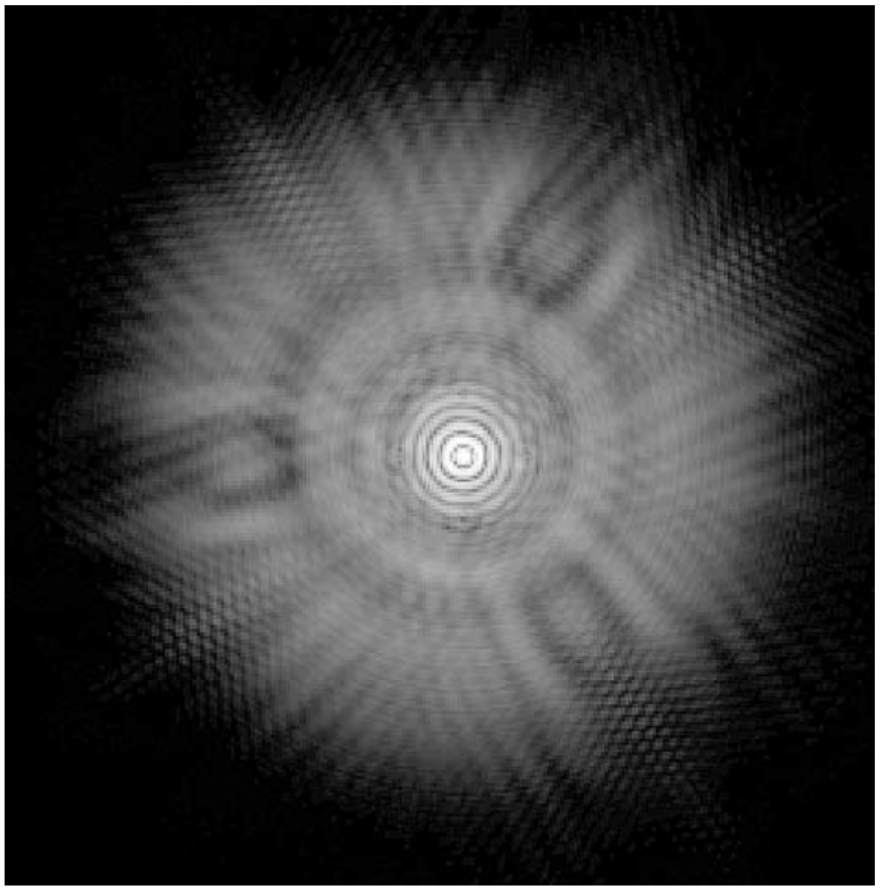
“dirty image”



PSF from atmospheric turbolence

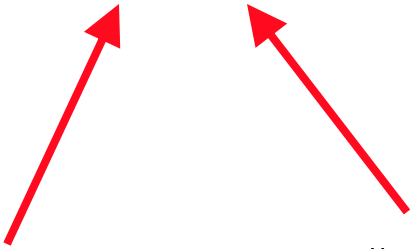
The Hubble PSF before the first servicing mission





In general the effect of the PSF is modeled by a linear operator

$$\mathbf{f} \rightarrow \mathbf{Hf}$$



action of optical system on true image is modeled by matrix **H**

"true" pixel vector

Image likelihood: 2. the noise model (degradation model)

Gaussian noise model $$P\left(\mathbf{g}|\mathbf{f}\right) \propto \exp\left[-\frac{\left(\mathbf{g}-\mathbf{Hf}\right)^2}{\sigma^2}\right]$$

Poisson noise model $$P\left(\mathbf{g}|\mathbf{f}\right) \propto \prod_n \frac{\left(\mathbf{Hf}\right)_n^{g_n}}{g_n!} \exp\left[-\left(\mathbf{Hf}\right)_n\right]$$

(Poisson noise mostly from detection process, Gaussian noise mostly from electronics or from approximation of Poisson noise)

sometimes we can use the Gaussian approximation of Poisson noise

$$P(\mathbf{g}|\mathbf{f}) \propto \prod_n \frac{(\mathbf{Hf})_n^{g_n}}{g_n!} \exp\left[-(\mathbf{Hf})_n\right]$$

$$\approx \prod_n \exp\left[-\frac{\left(g_n - (\mathbf{Hf})_n\right)^2}{2(\mathbf{Hf})_n}\right]$$

$$= \exp\left[-\sum_n \frac{\left(g_n - (\mathbf{Hf})_n\right)^2}{2(\mathbf{Hf})_n}\right]$$

Gaussian noise only:

maximize linear combination of entropy and chi-square

$$\log P(\mathbf{f}|\mathbf{g}) \approx \alpha S(\mathbf{f}) - \frac{(\mathbf{g}-\mathbf{H}\mathbf{f})^2}{\sigma^2}$$

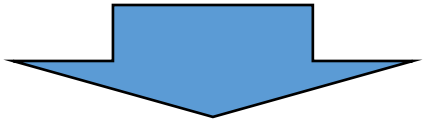
$$= \alpha S(\mathbf{f}) - \sum_n \frac{\left(g_n - (\mathbf{H}\mathbf{f})_n\right)^2}{\sigma^2}$$

$$= \alpha S(\mathbf{f}) - \chi^2(\mathbf{f})$$

Combined noise model

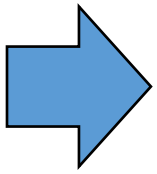
detector noise: Poisson noise
electronic noise: Gaussian noise

$$P(\mathbf{g}|\mathbf{f}) = \prod_n \sum_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(g_n - k)^2}{\sigma^2}\right] \frac{(\mathbf{Hf})_n^k}{k!} \exp\left[-(\mathbf{Hf})_n\right]$$



maximize

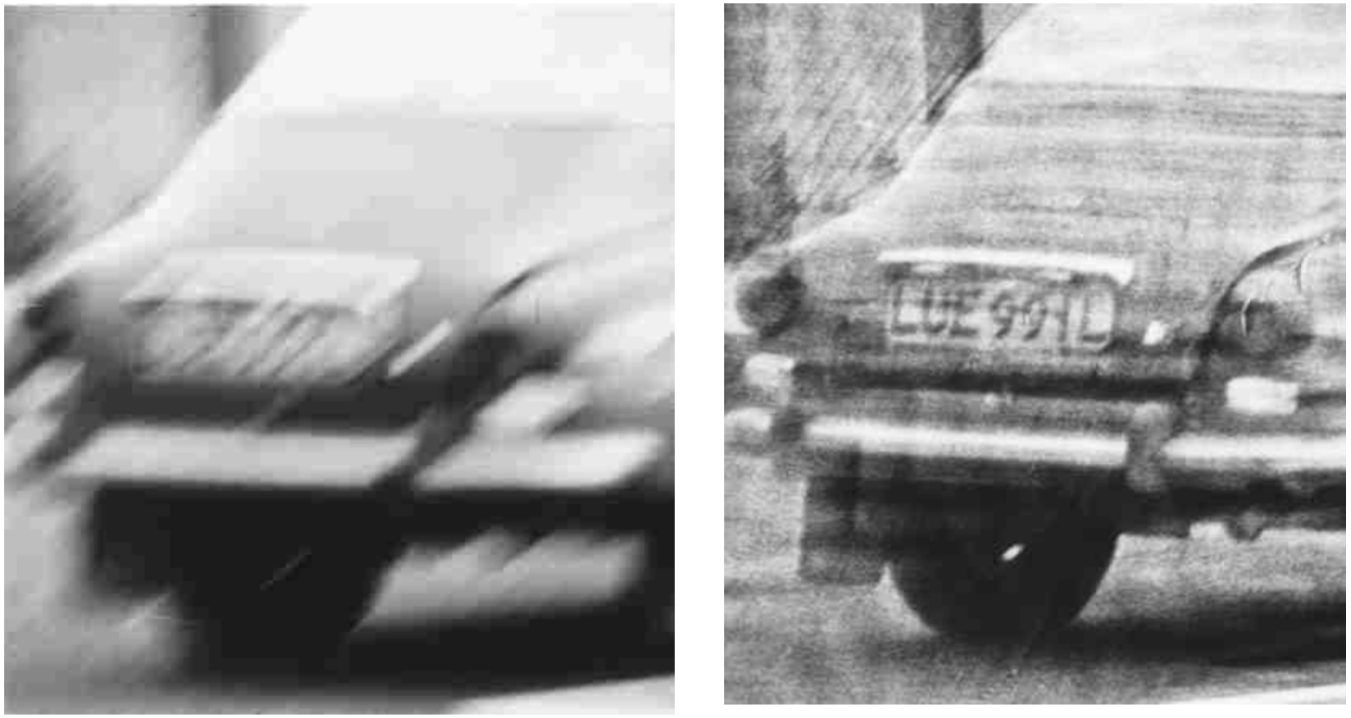
$$\log P(\mathbf{f}|\mathbf{g}) = \alpha S(\mathbf{f}) + \sum_n \log\left\{\sum_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(g_n - k)^2}{\sigma^2}\right] \frac{(\mathbf{Hf})_n^k}{k!} \exp\left[-(\mathbf{Hf})_n\right]\right\}$$



numerical maximization procedure

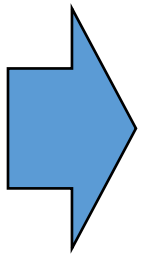
Applications of Max.Ent. to image processing

(J. Skilling , Nature **309** (1984) 748)



Car movement introduces linear correlations among pixels. The model of linear corrections does not allow direct inversion to find the corrected image because the number of variables is larger than the number of equations. The MaxEnt methods regularizes the problem and finds a reasonable solution.

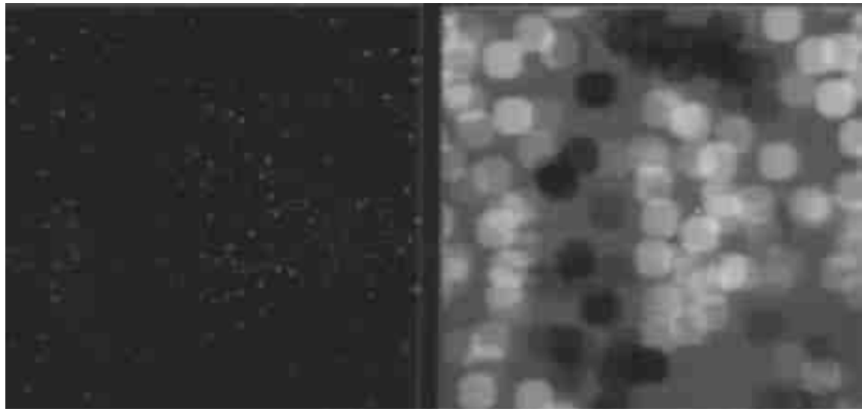
Reconstruction of missing data
(from http://www.maxent.co.uk)

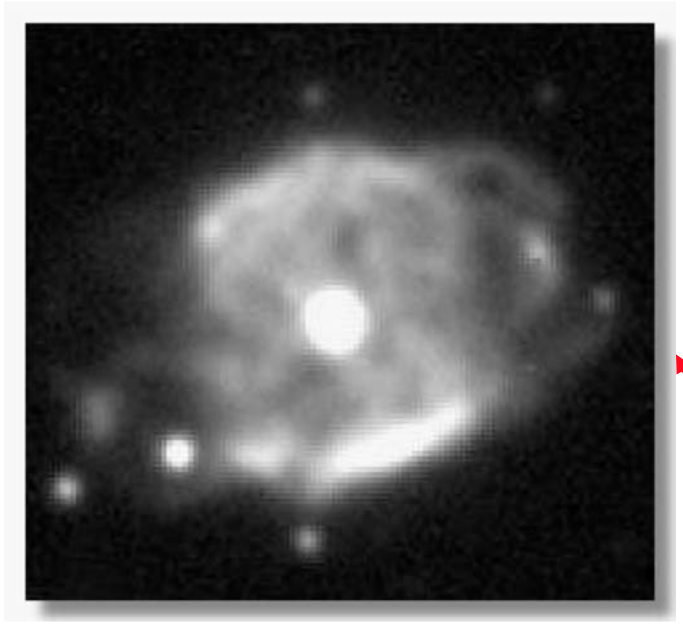


50%

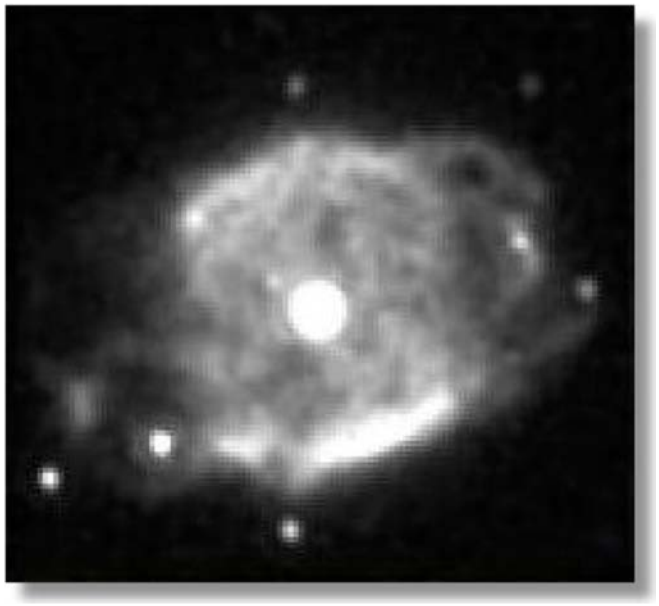
95%

99%

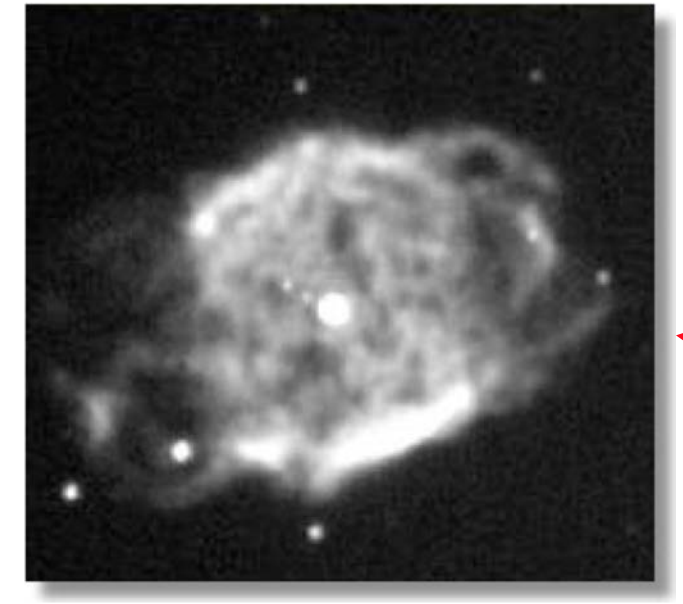




NGC 40



low resolution (MEM enhanced)



low resolution

high resolution

(from http://www.mirametrics.com)

MaxEnt Maximum Entropy Data Consultants Ltd.

Home
About MEDC
Applications
Examples
Products
Prices
Documents
Contact us
Search MEDC

Quick Search:

Search

John Skilling: Biographical information

John is Scientific Director of MEDC. He did his Ph.D. (on cosmic rays) in the Department of Physics at Cambridge University, and went on to become a Lecturer in the Department of Applied Mathematics and Theoretical Physics, and a Fellow of St Johns College.

In the late 1970s, another radio astronomer, Steve Gull, introduced him to the power of the Maximum Entropy Method. John wrote what was to become the first MemSys kernel system, and helped lay the Bayesian foundations for MEM. In 1981 he and Steve founded MEDC to exploit opportunities to apply MEM in other fields.

John resigned his Lectureship in 1990 in order to go fulltime with MSL and MEDC. Thanks to the wonders of modern technology John is able to telecommute from his new home in the West of Ireland, and he makes regular visits to clients both in the UK and further afield.

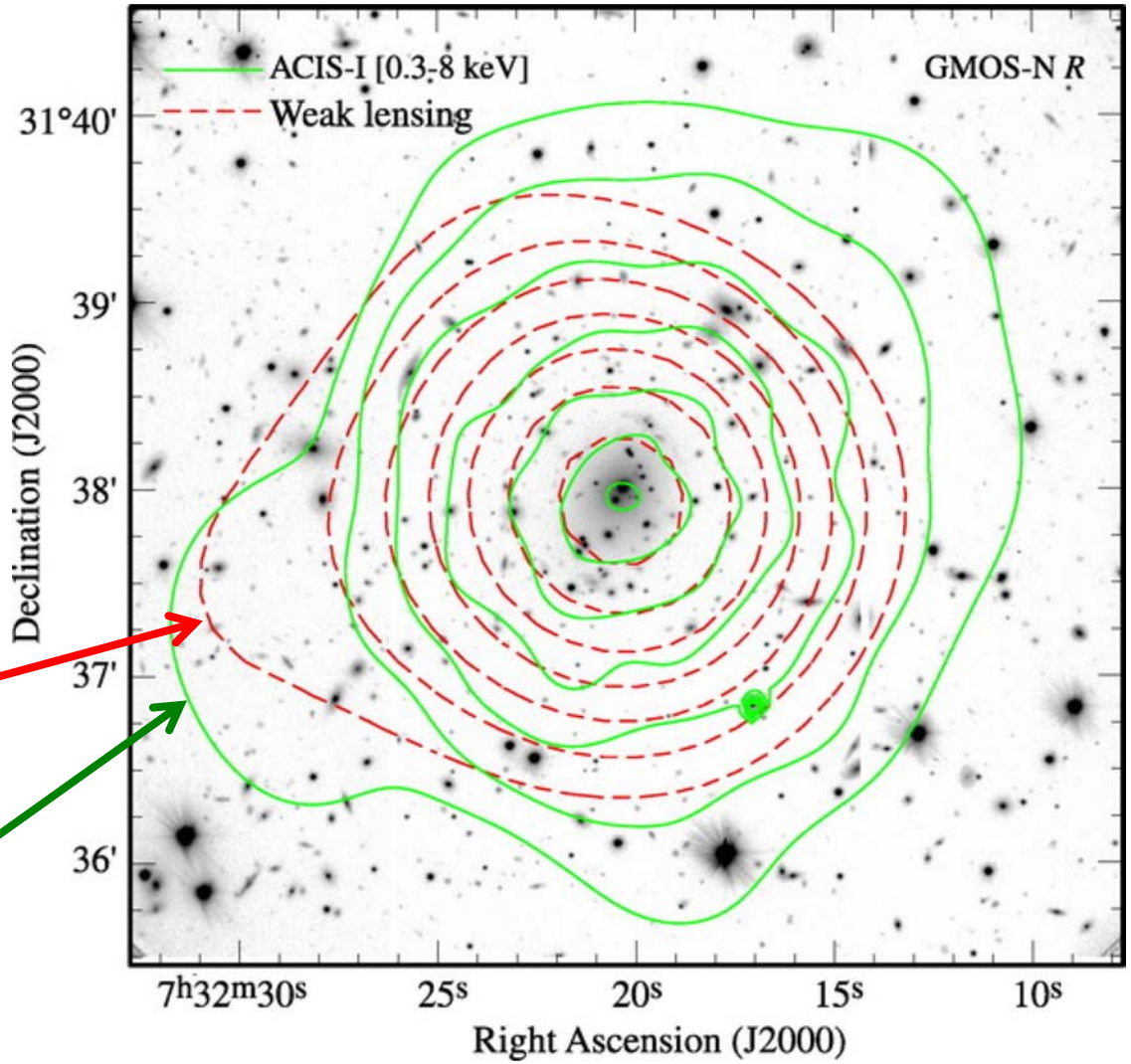
Home | Applications | Products | Prices | Documents | About MEDC | Contact Us | Full search

©MEDC Ltd. Last revised Wed Sep 19 22:19:39 2007

http://www.maxent.co.uk/ (currently not working - May 2017)

Example of LensEnt usage (Bridle et al, 1998):

reconstruction of mass density from lensing data, using Max Ent



reconstructed mass density

X-ray emission data

GMOS image of the central region of Abell 586 with **logarithmically spaced X-ray isophotes (*solid lines) and weak-lensing reconstructed mass density (dashed lines) superposed*.** *The X-ray point source near the southwest corner is the Seyfert 1 galaxy C171_3650.*
(from Cypriano et al., ApJ, **630** (2005) 38-49)

Many related methods: e.g. the Richardson-Lucy (RL) algorithm

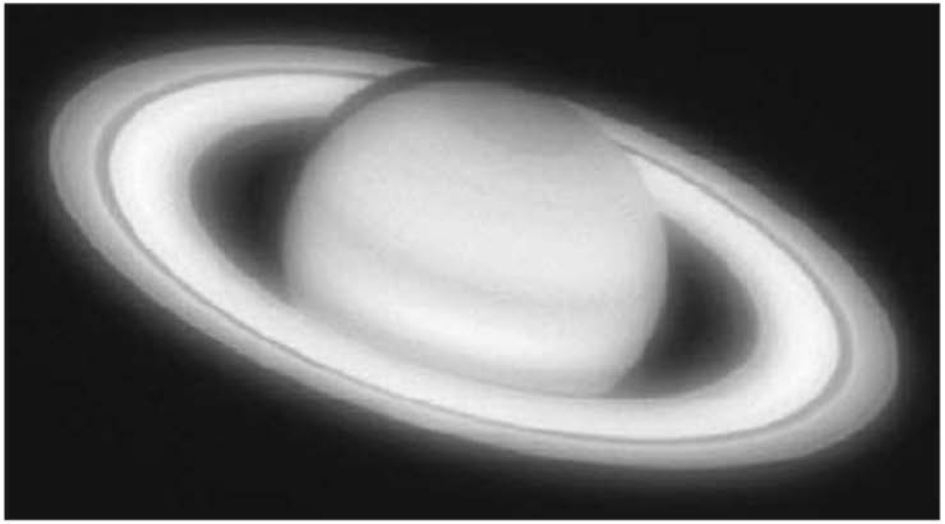
noise model: Poisson noise
prior: flat prior

$$P(\mathbf{f}|\mathbf{g}) \propto \prod_n \frac{(\mathbf{Hf})_n^{g_n}}{g_n!} \exp\left[-(\mathbf{Hf})_n\right] P(\mathbf{f})$$

$$\log P(\mathbf{f}|\mathbf{g}) \approx \sum_n \left[-(\mathbf{Hf})_n + g_n \log(\mathbf{Hf})_n\right] + const.$$

maximize this posterior distribution

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \sum_n \left[-(\mathbf{Hf})_n + g_n \log(\mathbf{Hf})_n\right]$$



▲ *8. Raw image of planet Saturn obtained with the WF/PC camera of the HST.*

▲ *9. Reconstruction of the image of Saturn using the R-L algorithm.*

NGC 604 in Spiral Galaxy M33

Hubble Heritage

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.

MaxEnt and image processing

• J. Skilling et al., Mon. Not. R. astr. Soc. **187** (1979) 145

• J. Skilling , Nature **309** (1984) 748

• R. Narayan and R. Nityananda, Ann. Rev. Astron. Astrophys. **24** (1986) 127

• R. Molina et al., IEEE Signal Proc. Magazine (marzo 2001) 13

• J. Skilling, A. W. Strong and K. Bennett, Mon. Not. R. astr. Soc. **187** (1979) 145

• J. Skilling and R. K. Bryan, Mon. Not. R. astr. Soc. **211** (1984) 11

• S. L. Bridle et al, Mon. Not. R. astr. Soc. **299** (1998) 895