# Using the Fisher information matrix in ML reconstruction

Johan Nuyts
K.U.Leuven, Belgium

# Outline

- The ML solution
- FIM for Gaussian likelihood
- FIM for Poisson likelihood
- FIM for Penalized likelihood
- Applications

# The ML solution

- Computers are faster
- Iterating ML and MAP to near convergence
- analyzing the ML and MAP estimates

ML solution: image that **maximizes** L



The term “ML image” is often used to denote an image obtained after a finite number of iterations (not too many to avoid the “noise deterioration”). However, this image does not have maximum likelihood. The actual ML image (the solution of the ML-problem) is only obtained after an infinite number of iterations.

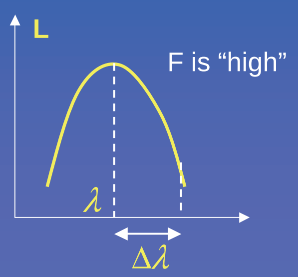
# The ML solution

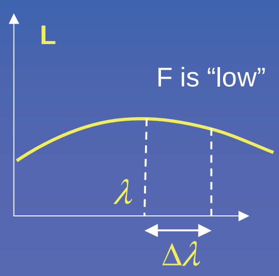
$$L = \sum_i y_i \ln\left(\sum_j a_{ij}\lambda_j\right) - \sum_j a_{ij}\lambda_j$$

$$\nabla L = 0$$

$$L(\lambda + \Delta\lambda) = L(\lambda) + \Delta\lambda' \; \nabla^2 L \;\; \Delta\lambda + \ldots$$

Expectation of $-\nabla^2 L$ is the **Fisher information matrix**





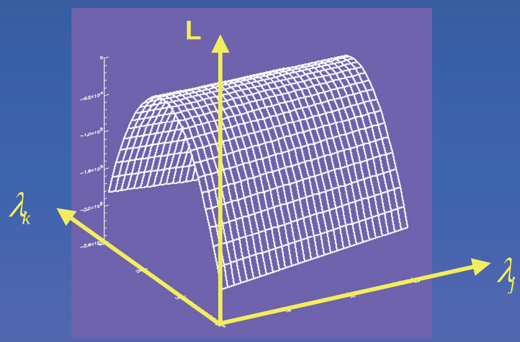

Because the first derivatives are zero at the maximum, the shape of the top is dominated by the second derivatives. The expected value of the second derivatives of the likelihood is called the Fisher information matrix [93,94,95]. See also [100] for more documentation on statistical reconstruction.

With high second derivatives, the maximum is well defined. Noise on the measurement y will propagate into L, but the effect on the position of the maximum will be limited.

If the second derivatives are small, the top is flat, and a bit of noise on *y* may move the position of the maximum considerably. In that case, the ML image will be very sensitive to noise on the projection data. Thus, the Fisher information matrix describes how much information about the reconstruction image is provided by the data.

The Fisher information matrix is large:its size is $N \times N$, with N the number of pixels in the reconstruction image (typically N is about $10^4$ in 2D emission tomography).

# The ML solution



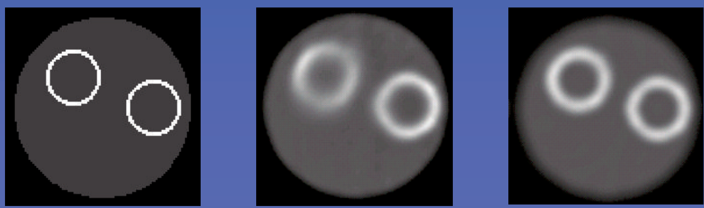

Fisher information is directional



The curvature of the likelihood function can be very different in different directions. In that case, the data provide more information about some pixel values (or combination of pixel values) than about others.

## Why computing F?

- analytical epressions for
  - impulse response
  - (co)variance
- helps in development of
  - fast approximate computation of variance, SNR...
  - penalties (priors), e.g. for uniform resolution
  - approximate prewhitening (noise decorrelation)



# Gaussian likelihood

A: system matrix (projection operator)
λ: reconstruction image
y: measured projection

$$A\lambda = y$$

Covariance of y:

$$C_y = \text{expectation}(y - \bar{y})(y - \bar{y})'$$

Uncorrelated measurement noise in emission tomography

$$= Diag[\sigma_i]$$

log-likelihood,

$$L = -\sum_i \frac{(\sum_j a_{ij}\lambda_j - y_i)^2}{\sigma_i^2}$$

$$= -(A\lambda - y)' C_y^{-1} (A\lambda - y)$$



The logarithm of a Gaussian is a parabola, so the quadratic approximation is exact here.

## ML for Gaussian

$$L = -(A\lambda - y)' C_y^{-1} (A\lambda - y)$$ the log-likelihood

$$\frac{\partial L}{\partial \lambda} = -A' C_y^{-1} A \, \lambda + A' C_y^{-1} y = 0$$ finding the maximum

$$\Downarrow$$

$$\lambda = [A' C_y^{-1} A]^{-1} A' C_y^{-1} y$$ the ML estimate

Here, the ML estimate is “efficient”: unbiased and optimal

$$\text{expectation}(\lambda) = \overline{\lambda} = [A' C_y^{-1} A]^{-1} A' C_y^{-1} \overline{y}$$



An estimator is unbiased if its expecation equals the true value. The best possible unbiased estimator is called an efficient estimator [93,94].

# FIM and covariance

**FIM**: expectation of $-\nabla^2 L$

$$L = -(A\lambda - y)' C_y^{-1} (A\lambda - y)$$

$$\nabla L = -A' C_y^{-1} A\, \lambda + A' C_y^{-1} y$$

$$-\nabla^2 L = \boxed{F = A' C_y^{-1} A}$$

**covariance (λ)**:

$$E(\lambda - \bar{\lambda})(\lambda - \bar{\lambda})'$$

$$= E\ [A' C_y^{-1} A]^{-1} A' C_y^{-1}\ (y - \bar{y})\ (y - \bar{y})'\ A C_y^{-1} [A' C_y^{-1} A]^{-1}$$

$$= [A' C_y^{-1} A]^{-1} A' C_y^{-1}\ C_y\ A C_y^{-1} [A' C_y^{-1} A]^{-1}$$

$$= \boxed{[A' C_y^{-1} A]^{-1}} = F^{-1}$$



For the efficient estimator, the covariance matrix is the inverse of the Fisher information matrix. The covariance is a measure of the uncertainty, the Fisher information is a measure of the certainty.

# Poisson likelihood: FIM

**FIM**: expectation of $-\nabla^2 L$

$$L = \sum_i y_i \ln(\sum_j a_{ij}\lambda_j) - \sum_j a_{ij}\lambda_j$$

$$\frac{\partial L}{\partial \lambda_j} = \sum_i a_{ij} \frac{y_i}{\sum_k a_{ik}\lambda_k} - a_{ij}$$

$$\frac{\partial^2 L}{\partial \lambda_j \partial \lambda_k} = -\sum_i a_{ij} a_{ik} \frac{y_i}{(\sum_i a_{ik}\lambda_k)^2}$$

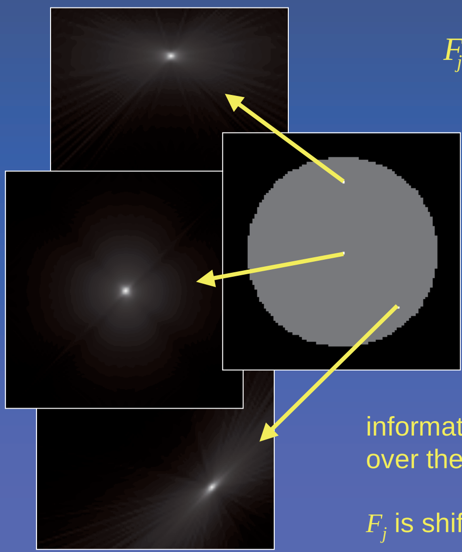
expectation of $-\frac{\partial^2 L}{\partial \lambda_j \partial \lambda_k}$ $= \sum_i a_{ij} a_{ik} \frac{\bar{y}_i}{\bar{y}_i^2} = \sum_i \frac{a_{ij} a_{ik}}{\bar{y}_i} = [A' C_y^{-1} A]_{jk}$

$\boxed{F = A' C_y^{-1} A}$ **same as Gaussian, with** $C_y^{-1} = Diag[\frac{1}{\bar{y}_i}]$



The Fisher information for the Poisson likelihood is the same as for the Gaussian likelihood [93,94,95]. This is not too suprising, as the Poisson distribution is well approximated by a Gaussian distribution y with var(y) = mean(y).

# Fisher information row j



$$F_{jk} = F_{kj} = \sum_i \frac{a_{ij} a_{ik}}{\bar{y}_i}$$

- make emtpy image
- set pixel j to 1
- projection: $a_{ij}$
- divide: $a_{ij} / \bar{y}_i$
- backproject: $\sum_i a_{ik} a_{ij} / \bar{y}_i$

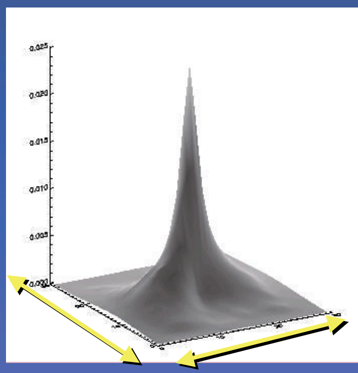
information about pixel j is spread over the image

$F_j$ is shift variant low pass filter



The diagonal elements of the Fisher information are higher, but the off-diagonal elements are definitely not zero. The pixel values in the reconstructed image are not independent.

# Fisher information row j



30 pixels

Fj: low pass filter
(approx.) local support



The Shepp-Logan software phantom (100x100 pixels) was used as the true activity distribution in a PET simulation (taking into account photon attenuation and shift-invariant detector blurring).

The surface plot shows a convolution kernel (30x30 pixels) extracted from row j (or column j) of the corresponding Fisher information matrix. Pixel j is indicated by the arrow. The asymmetrical attenuation and asymmetrical activity distribution yields an asymmetrical “Fisher information convolution kernel”. The inclusion of detector blurring widens that convolution kernel.

The correponding row of the covariances of pixel j is well approximated as the inverse of this kernel. Because the Fisher information acts as a low pass filter, the covariance matrix acts as a high pass filter.

# Poisson likelihood: covariance

**covariance (λ)**: $E(\lambda-\bar{\lambda})(\lambda-\bar{\lambda})' =$ **?**

$$\neq F^{-1}$$

MLEM is **not** an efficient estimator (it is biased)

- Do we need to know cov(λ) ?
  = is uncertainty more interesting than certainty?
- If we need cov(λ),
  can we approximate it as $F^{-1}$ ?



The Fisher information comes as a "well-behaved" blob. In contrast, the covariance matrix is a wildly oscillating function (similar to the impulse response of the ramp filter). For that reason, the diagonal of the Fisher information matrix can be used as a reasonable measure of the certainty about a pixel value, while variance (the diagonal of the covariance matrix) may actually be a poor estimate of the uncertainty about that value [83].

## Penalized likelihood (= MAP)

$$-\beta \sum_j \sum_k w_{jk}(\lambda_j - \lambda_k)^2 \;=\; -\lambda' R \lambda$$ log-prior

$$L = -(A\lambda - y)' C_y^{-1} (A\lambda - y) - \beta\, \lambda' R \lambda$$ log-posterior

$$\frac{\partial L}{\partial \lambda} = -A' C_y^{-1} A\, \lambda + A' C_y^{-1} y - \beta\, R\lambda = 0$$

$$\Downarrow$$

$$\lambda = [A' C_y^{-1} A + \beta R\,]^{-1}\, A' C_y^{-1}\, y$$ MAP estimate

$$C_\lambda = [F + \beta R]^{-1} F [F + \beta R]^{-1}$$ MAP covariance:



Note that λ'Rλ is a scalar, so it is equal to its transpose. Consequently, we have λ'Rλ = λ'R' λ = 0.5 λ'(R+R') λ.

It follows that only the effect of the prior is only determined by the symmetrical component of R.

# MAP impulse response

the MAP estimate

$$\lambda = [A'C^{-1}A + \beta R]^{-1} A'C^{-1} y$$

$$\bar{\lambda} \Rightarrow \bar{\lambda} + \delta_j$$

$$\bar{y} \Rightarrow \bar{y} + A\delta_j$$

$\bar{\lambda}$ *= true image*

$\lambda$ *= reconstruction*

$$\lambda + [A'C^{-1}A + \beta R]^{-1} A'C^{-1} A\delta_j$$

$$= \lambda + [F + \beta R]^{-1} F\delta_j$$

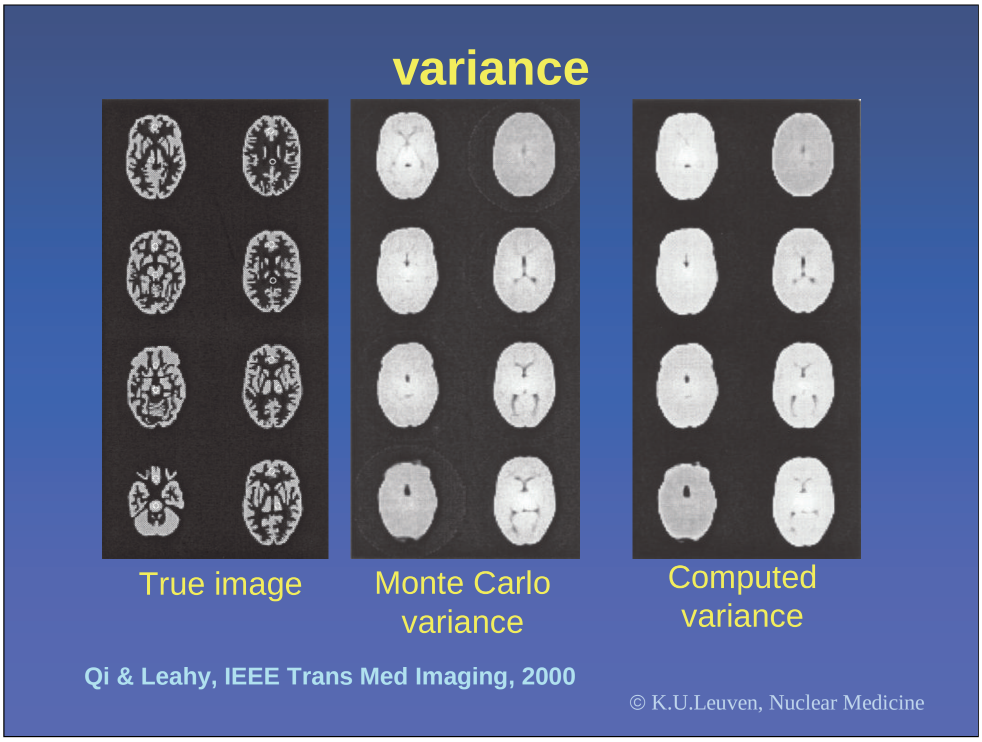
- If R is high pass, then $[ F + \beta R ]^{-1}F$ is low pass
- Compute with FFT, assuming local smoothness, local support



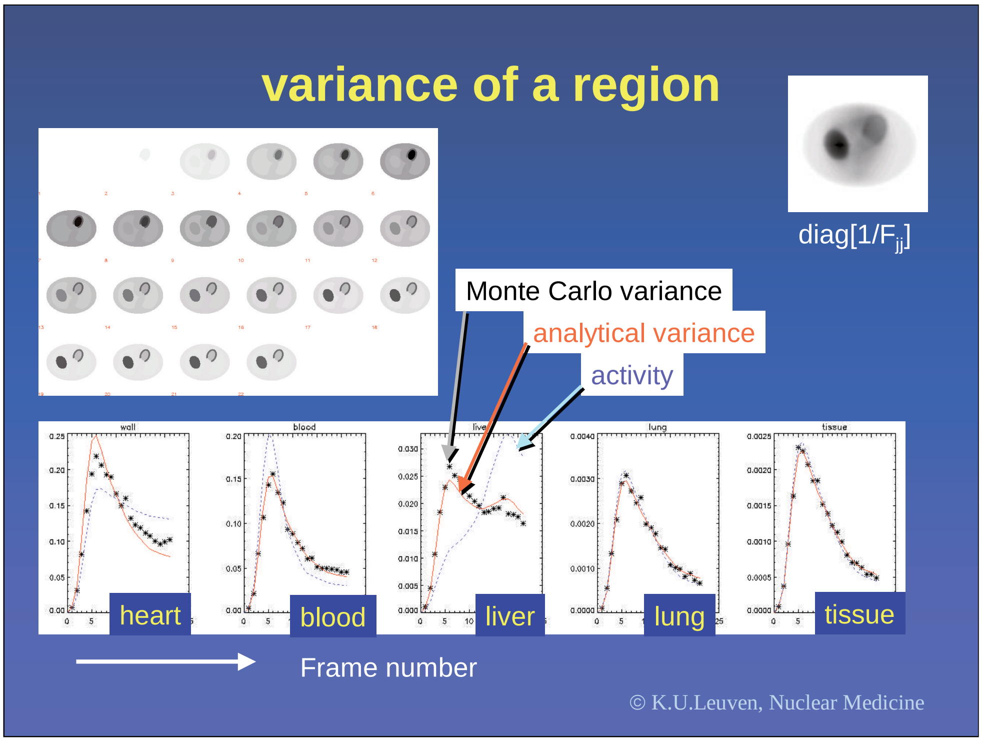
$\delta_j$ is an image with zeros everywhere, except in pixel j. This is the impulse. Due to the impulse, the expectation of the measurement changes with $A\delta_j$. The impulse response is the resulting change in the reconstruction $\lambda$. The impulse response is a blob, obtained by convolving the impulse with the shift-variant low pass filter $[F + \beta F]^{-1}F$.

# Applications

- predict variance of ML or MAP images
- predict and optimize MAP local impulse response
- prewhiten ML or MAP images




Qi and Leahy [90] derived analytical expressions for the contrast recovery coefficient (the central value of the impulse response) and variance in every pixel, for a simulated microPET acquisition. This allows them to design a prior that yields (nearly) shift invariant contrast recovery.




These plots illustrate the performance of a simple measure of uncertainty (based on the diagonal of the Fisher information matrix) about region values [83]. It could be used in weighted least squares analysis of region-of-interest values. In practice, it is often assumed that the variance is proportional to the mean, but in some regions, this is clearly not a valid assumption.

## variance
## for computer observers

log of likelihood ratio (SKE,BKE):

$$\ln \frac{p(\lambda \mid \lambda_0)}{p(\lambda \mid \lambda_1)} = (\lambda - \lambda_0)' C_\lambda^{-1} (\lambda - \lambda_0) - (\lambda - \lambda_1)' C_\lambda^{-1} (\lambda - \lambda_1)$$

drop constant terms

$$q = (\lambda_1 - \lambda_0)' C_\lambda^{-1} \lambda$$

approximation

$$q = (\lambda_1 - \lambda_0)' (F + \beta R)^{-1} F (F + \beta R)^{-1} \lambda$$

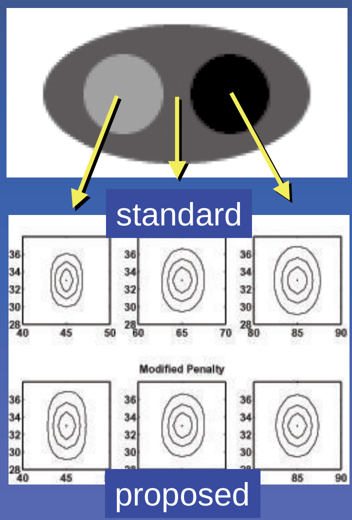
**Bonetto, Qi, Leahy, IEEE Trans Nucl Sci, 2000**



The channelized Hotelling observer [97-99] contains a prewhitening step (= decorrelating noise, by multiplying with the inverse of the covariance matrix). Gifford et al [97] estimate the covariance at the output of the channels, where the covariance matrix is much smaller. This reduces the amount of cpu needed to estimate it with Monte Carlo simulations.

Bonetto et al [87] use the Fisher information, which eliminates the need for Monte Carlo alltogether.

## uniform local impulse response




estimate “strength” of likelihood in j as $F_{jj}$

set “strength” of penalty to

$$w_{jk} = \sqrt{F_{jj} F_{jk}}$$

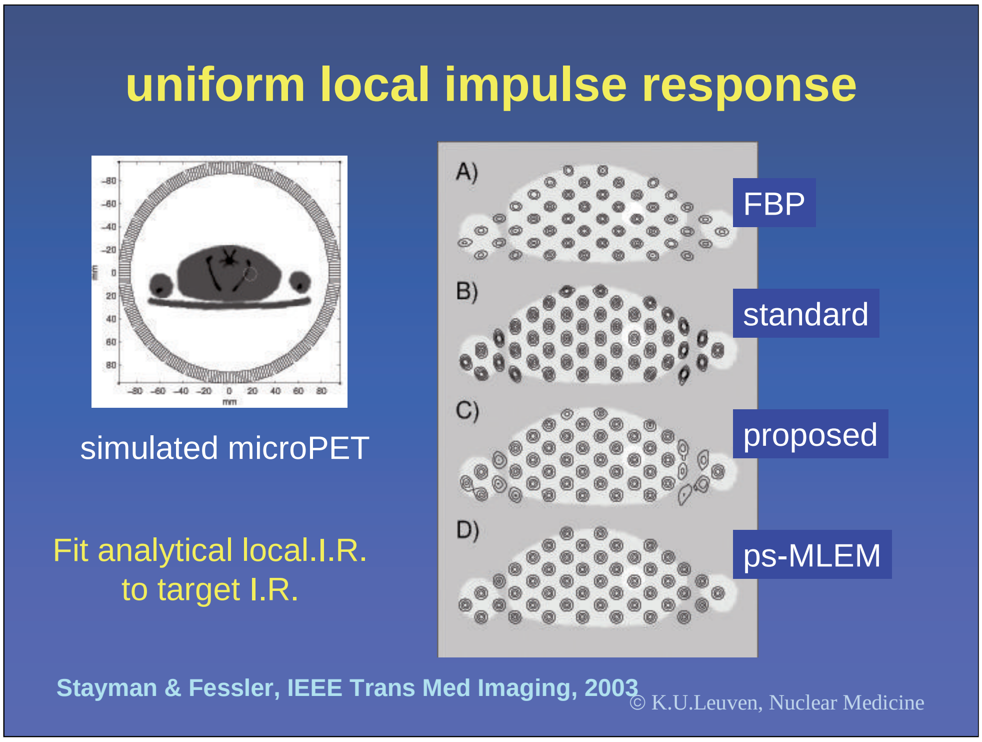
in

$$-\beta \sum_j \sum_k w_{jk} (\lambda_j - \lambda_k)^2$$

Fessler & Rogers, IEEE Trans Image Proc, 1996



Fessler and Rogers [59]. This method adapts the amplitude of the penalty to the strength of the likelihood. The resulting mean FWHM of the local impulse response is nearly shift invariant. However, the local impulse response still has position dependent asymmetry. This happens because the local penalty still had the same strength in every direction, while the amount of information present in the data is orientation dependent.




Stayman and Fessler [88-89]. The analytical expression for the local impulse response takes into account the orientation dependence of the certainty. The parameters of the penalty are determined by fitting the local impulse response to a symmetrical target function. This makes the local strength of the penalty different in different directions, and yields a virtually perfect  impulse reponse.

“ps-MLEM” denotes post-smoothed MLEM.

“standard” denotes shift-invariant quadratic penalty.

# Fisher information about pixel differences

penalty penalizes pixel value **differences**

$\Downarrow$

FIM for estimating pixel value differences:

$$\text{FIM-diagonal}(\lambda_j - \lambda_k) = -E\left(\frac{\partial^2 \text{Likelihood}}{\partial(\lambda_j - \lambda_k)^2}\right)$$

$$= \sum_i \frac{(c_{ij} - c_{ik})^2}{\bar{y}_i}$$



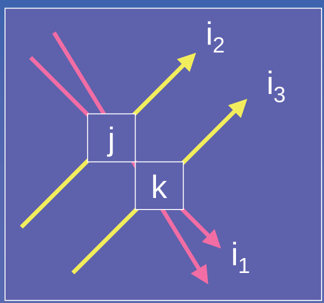
The parameters $\lambda_j$ and $\lambda_k$ are eliminated by rewriting them as a linear combination of ($\lambda_j+\lambda_k$) and ($\lambda_j-\lambda_k$). That allows to compute the Fisher information for estimating pixel differences [96].

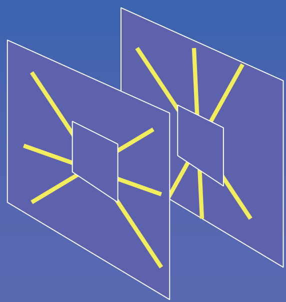
# uniform local impulse response

$$\sum_i \frac{(c_{ij} - c_{ik})^2}{\bar{y}_i} \approx \sum_{i \in S_{j-k}} \frac{c_{ij}^2}{\bar{y}_i}$$

$S_{j\text{-}k}$ = set of projections perpendicular to j-k

- info in projection lines that hit 1 pixel but not both
- 3D penalty does not smooth well in axial direction





Only projections intersecting one of the pixels and not both pixels, provide information about the difference between the pixel values.

Most tomographical systems have a cylindrical symmetry. From the expressions, it follows that these systems acquire more information about axial pixel differences than about transaxial pixel differences.

- If the neighboring pixels j and k are located on the axis, then all projection lines intersecting j or k provide useful information.
- If the neighboring pixels j and k are located in a transaxial plane, then only part of the projection lines intersecting j or k provide information. The ones that intersect both pixels provide little or nothing.

It follows that for uniform resolution, axial differences must be penalized more than transaxial differences.

# comparison with post-smoothed MLEM

MAP local impulse response: $[F+\beta R]^{-1}F\delta_j$

MAP covariance: $[F+\beta R]^{-1}F[F+\beta R]^{-1}$

*uniform resolution*

MAP impulse response: $[F+\beta R]^{-1}F\delta_j = P\delta_j$

MAP covariance: $\boxed{PF^{-1}P}$

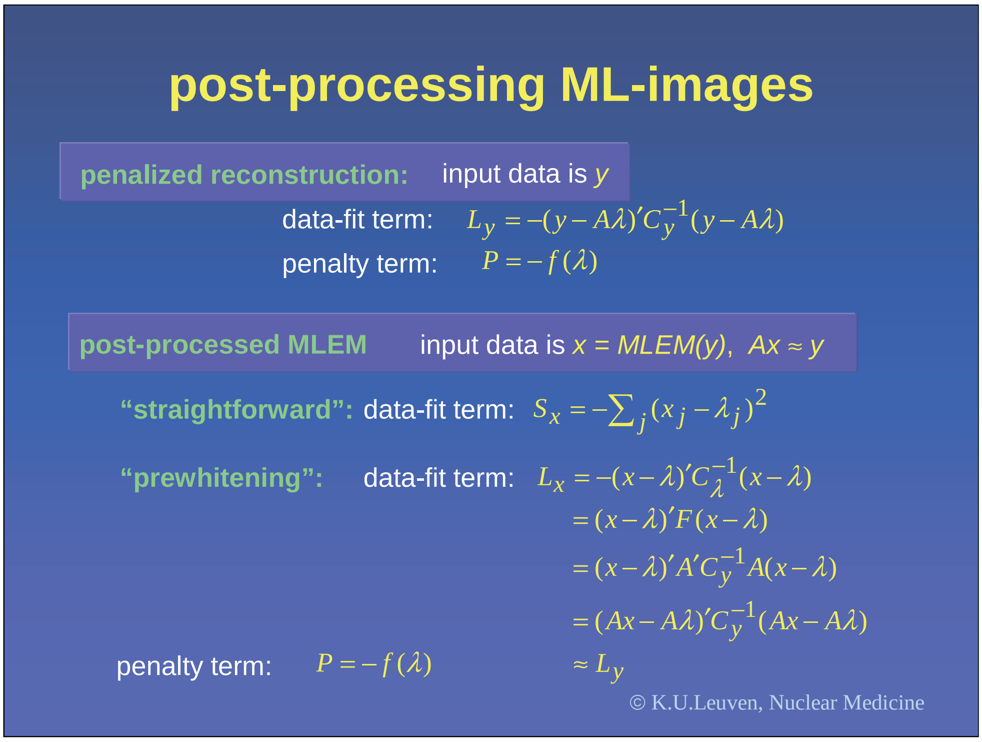
**ML covariance:** $F^{-1}$

**ps-ML covariance:** $\boxed{PF^{-1}P}$

**Stayman & Fessler, IEEE Trans Med Imaging, 2003**

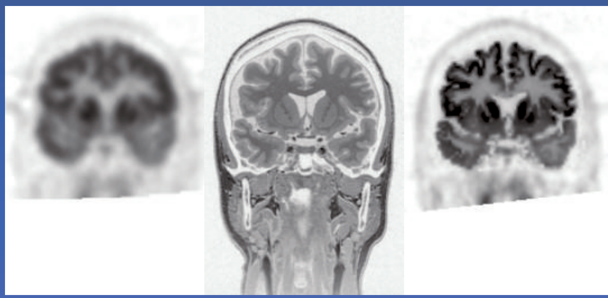


It was found that penalized likelihood for uniform resolution has similar noise characteristics as post-smoothed MLEM (at carefully matched resolution) [89,96]. Stayman and Fessler derived this explanation [89].




$L_y$ measures the agreement with the sinogram data y.
To compare MAP with post-processed ML, we need a similar smoothing effect in the post-processing. For that purpose, we use the same penalty, and we invent some similarity measure to penalize deviations from the unconstrained ML image x.
$S_x$ measures the agreement with the unconstrained ML reconstruction x, as the sum of squared differences.
$L_x$ measures the agreement with the unconstrained ML reconstruction x, taking into account the covariances between the pixels.
The derivation indicates that $L_x$ and $L_y$ are nearly equivalent. Consequently, post processing unconstrained MLEM images should be equivalent to penalized likelihood reconstruction, if the pixel covariances are respected (achieved by using the Fisher information matrix F).

## example: MAP with anatomical priors



- Gibbs prior in gray matter (relative difference)
- Intensity prior in white (with estimated mean)
- Intensity prior in CSF (mean = 0)

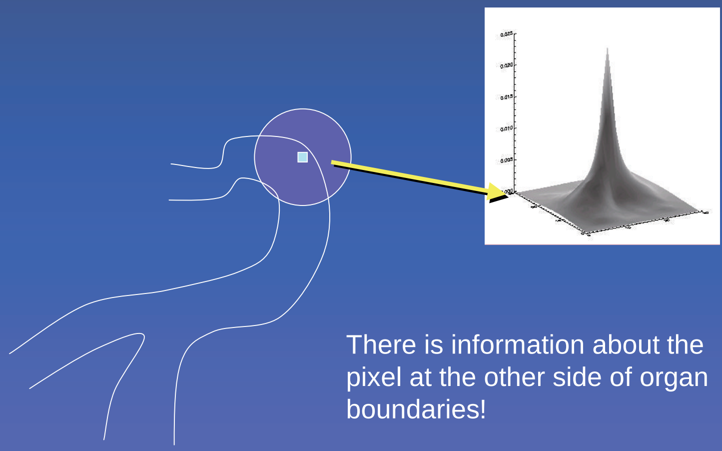
Baete et al, M5-4, MIC 2003



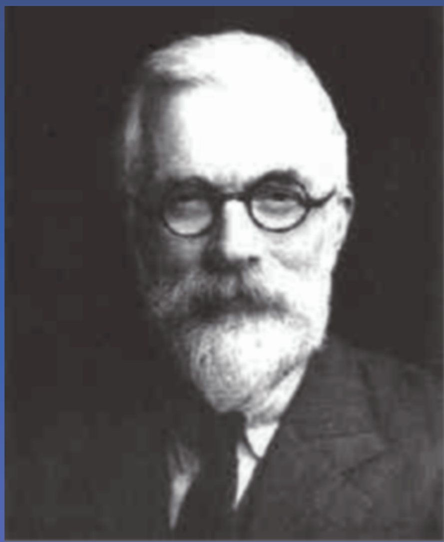
This is an example of using anatomical priors for MAP-reconstruction. This approach is taken as a case to evaluate the use of $L_x$ as a substitute of $L_y$ in a post-processing method, using a simple simulation.

In that simulation, we used the same prior as in MAP, and combined it with $S_x$ and $L_x$. Our experiment indicates that ignoring the covariances causes measureable loss of information [Nuyts et al, M5-2, MIC2003].

# post-processing ML-images



There is information about the pixel at the other side of organ boundaries!



**Ronald Aylmer Fisher**
**1890-1962**